The JALT CALL Journal
ISSN 1832-4215, 22, 103853 (2026)
https://doi.org/10.29140/jaltcall.2026.103853

Castledown

Interacting with AI voice-based chatbots: A study on EFL students' satisfaction and speaking performance

Ho Thi Hien
University of Foreign Languages and International Studies, Hue University; Industrial University of Ho Chi Minh City, Vietnam
hthien.dhnn24@hueuni.edu.vn

Nguyen Quang Ngoan
Quy Nhon University, Vietnam
nguyenquangngoan@qnu.edu.vn

Nguyen Van Huy
University of Foreign Languages and International Studies, Hue University, Vietnam
nguyenvanhuy@hueuni.edu.vn

AI voice-based chatbots (AVCs) are increasingly integrated into EFL speaking instruction as potential conversational partners for speaking practice. However, studies on students' satisfaction with the chatbot-human interaction and its association with speaking gains remain limited. This study examines the learners' satisfaction with ChatGPT's spoken language during conversational interaction, the effects of AVC-supported speaking practice on students' oral performance and the association between satisfaction and speaking performance. A quantitative pretest-posttest design was employed with 116 Vietnamese EFL students over an 8-week intervention. Data were collected through pre- and post-speaking tests and a satisfaction questionnaire. The test scores were analyzed using paired-samples t-tests and effect sizes, descriptive statistics and correlation analysis were used to examine the learners' satisfaction levels and their relationship with speaking improvement. Overall, learners reported moderate to high satisfaction with ChatGPT-generated language, with mean scores ranging from 3.48 to 4.34 on the 5-point Likert scale. In addition, learners' speaking scores were observed to increase from pre-test to post-test ($p < 0.05$), with a large effect size (Cohen's $d = 1.34$). However, the correlation analysis revealed no statistically significant association between satisfaction levels and speaking gains ($r = 0.087$, $p > 0.05$). Theoretically, the study contributes empirical evidence

to chatbot-human interaction studies by suggesting that AVCs can support learners' speaking improvement. Importantly, it also shows that learners' satisfaction with the learning experience is not necessarily associated with their speaking gains. Pedagogically, the study highlights the need for more instructional scaffolding in AI-supported speaking practice.

Keywords: AI voice-based chatbots, learner satisfaction, CHISM, satisfaction and speaking performance, chatbot-human interaction

Introduction

The advent of Generative Artificial Intelligence (GenAI) has significantly transformed language instruction in recent years, primarily through the integration of AI text-based and voice-based chatbots as adaptive virtual partners or tutors. These technologies have shifted the way language learners learn from human-human interaction to chatbot-human interaction (Belda-Medina & Calvo-Ferrer, 2022; Belda-Medina & Kokošková, 2023; Çakmak, 2022). An AI voice-based chatbot (hereafter, AVC) or an AI conversational agent is defined as a computer system that mimics natural conversations with humans in spoken language using natural language processing and speech technologies (Allouch et al., 2021; Belda-Medina & Kokošková, 2023; Yusuf et al., 2025). AVCs have been evidenced to be a potential solution for the long-standing challenges in conventional teaching and learning speaking. EFL learners in traditional speaking classes face a lot of challenges, such as limited opportunities to have conversations with fluent speakers or native speakers (Nguyen & Burke, 2024), unwillingness to communicate because of the fear of being judged when making mistakes, limited vocabulary, poor grammar, unintelligible pronunciation, and the fear of being judged on their level of English (Chulee et al., 2023), and the potential of receiving unfavorable feedback from peers or teachers (Tai & Chen, 2020).

Research has shown that AVCs can function as virtual speaking partners that can help EFL learners improve speaking skills (Azizimajd, 2023; Duong & Suppasetseree, 2024; Lim, 2024; Yusuf et al., 2025). In addition to linguistic gains, previous studies have found that practicing speaking with AVCs is beneficial for affective-related aspects, such as reducing anxiety levels, increasing learning motivation, willingness to communicate, and confidence (Azizimajd, 2023; Duong & Suppasetseree, 2024; Fathi et al., 2024). Despite the growing number of studies on the application of AVCs in language learning, empirical evidence on learners' satisfaction with chatbot-human interaction and its relationship with speaking performance remains limited. To date, only two studies conducted in Western countries (Belda-Medina & Calvo-Ferrer, 2022; Belda-Medina & Kokošková, 2023) have examined how students are satisfied with the experience of interacting with the selected chatbots. The results indicated relatively low levels of satisfaction with these novel ways of practicing speaking, due to students' high proficiency in speaking and the inappropriate selection of chatbots for instructional purposes. Notably, these studies did not

investigate whether there is alignment or mismatch between learners' satisfaction with the chatbot-generated language and their speaking performance, leaving an important gap in the literature.

To address this gap, the present study investigates EFL students' satisfaction with the language experience produced by an AVC during an 8-week speaking practice intervention. It also examines changes in students' speaking performance over time and explores the correlational relationship between students' satisfaction levels and their speaking gains. The study was conducted at a Vietnamese university, where the students' English proficiency level is relatively low, yet the need to speak English as a second language has recently been emphasized nationwide (Decision 2371/QĐ-TTg). Additionally, Vietnamese EFL students have been experiencing substantial difficulties in developing speaking skills, learning to speak in class, resulting in low speaking proficiency despite a long period of English learning (McAleavy et al., 2018; Nguyen & Burke, 2024; Pham, 2021).

Literature review

AI voice-based chatbots

The development of AVCs has evolved from simple rule-based programs to generative artificial intelligence since the 1960s. The early versions of chatbots (e.g., Eliza, Parry, Alice) marked the introduction of human-machine dialogues despite certain limitations such as chatbots' use of pre-programmed dialogue structures, limited responsiveness, and inability to accommodate diverse learner proficiency levels (Černý, 2022). In the 2010s, more voice-activated personal assistants were developed, such as Siri in 2010, Duolingo in 2012, Mondly in 2013, Elsa Speak in 2015, and Google Assistant in 2016. These chatbots were driven by the advancements in AI and Natural Language Processing technologies that enable synchronous, task-oriented spoken interaction within contextualized simulated scenarios (Černý, 2022). However, many of these systems were acknowledged to provide insufficient feedback, focusing primarily on surface-level pronunciation accuracy rather than discourse-level interaction or communicative competence (Kuhail et al., 2022). Since 2020, the incorporation of AI and Large Language Model revolution has opened a new era of chatbots in education, particularly ChatGPT, developed by OpenAI in 2022. This generation of chatbots includes text-based and voice-based functions to facilitate language learners based on their needs.

Impacts of AVCs on speaking performance

Although research on using AVCs for speaking instruction is in its early stages, the existing studies showed that AI-mediated speaking activities benefit EFL learners in many ways. AVCs were found to help students speak more fluently with fewer hesitations and pauses, and higher speech rates (Azizimajd, 2023; Duong & Suppasetseree, 2024; Kang, 2022; Lim, 2024). Students were able to

produce more precise and accurate pronunciation after speaking with AVCs (Çakmak, 2022; Duong & Suppasetseree, 2024); expand lexical ranges on related topics (Fathi et al., 2024; Shikun et al., 2024; Zou et al., 2024). Eragamreddy (2025) highlights additional benefits of conversing with AVCs, including reducing hesitation, improving word choice, producing more accurate sentence structures via real-time corrective feedback, articulating more natural intonation, stress, and improving turn-taking, topic maintenance, and conversational repair strategies.

To explain why students performed better when they conversed with virtual speaking partners, past studies found that AVCs helped students reduce anxiety when speaking (Azizimajd, 2023; L. T. Yu et al., 2020), provided learners with instant feedback (Belda-Medina & Kokošková, 2023; Kim et al., 2021; Shikun et al., 2024), a friendly-speaking environment with 24/7 availability (Fathi et al., 2024; Yusuf et al., 2025), and endless repetitions, and multimodal functions (Koç & Savaş, 2025).

In the Vietnamese educational context, the studies on AI-mediated speaking instructions are less extensive but increasingly visible. Duong and Suppasetseree (2024) conducted a quasi-experimental study with 30 non-English-major Vietnamese undergraduate students who had 6 weeks to interact with the Andy English Bot during online classes. The results indicated significant improvement in learners' speaking skills, especially in using suitable hedging words and grammar structures. Accuracy was improved because the chatbot used correct grammar and spoke like a native speaker, and fluency improved because the chatbot waited patiently and required clearer speech for recognition. H. H. Nguyen (2025) conducted a larger-scale research with 240 university students (both English majors and non-English majors of different English proficiency levels. The participants had 16 weeks to use SpeakAI daily after class. The results showed significant improvement in fluency, coherence, and pronunciation, followed by vocabulary range, grammatical range, and accuracy. In the meantime, T. H. Nguyen (2025) conducted research with 38 university students at a private university in Vietnam. The participants had eight weeks to practice speaking with ChatGPT-4 three times a week in various speaking tasks such as role-plays, problem-solving, and opinion-sharing. The findings showed significant improvement in all aspects of speaking skills (i.e., pronunciation, fluency and coherence, grammatical accuracy, lexical resource).

These are very early studies on the use of AVCs to transform traditional speaking instruction. Students were offered opportunities to converse with chatbots rather than with a human peer. Overall, the studies showed that chatbot-human interaction improved students' speaking skills. However, they mostly employed non-English majors whose learning purposes, language proficiency levels, and communicative demands differ substantially from those of English majors. In addition, previous studies implemented chatbot-based speaking practice as out-of-class or self-directed activities (Duong & Suppasetseree, 2024; H. H. Nguyen, 2025). It might be difficult for researchers and teachers to monitor learners' actual interactional patterns and ensure the quality of speaking practice. This study provides a different perspective to observe how

English-major students in a Vietnamese university interact with an AVC in a classroom setting. The students were offered a ChatGPT Plus account for practicing speaking with a chatbot as a conversational speaking partner during structured speaking activities. This design allowed us to observe students' actual interaction behaviors, provide consistent practice opportunities, and systematically examine speaking outcomes.

Relationship between learners' satisfaction and learning outcomes

Learners' satisfaction is defined as a feeling or attitude of fulfillment when learners get what they expected during the learning activities or processes (Shee & Wang, 2008; Q. Yu, 2022). This is a psychological state formed from learners' subjective evaluation after comparing their learning expectations and the actual perceived learning conditions, such as course content, learning environment, and instructional methods (Fernandes et al., 2013; Gameel, 2017; Q. Yu, 2022). Learning satisfaction is an important factor for both language learners and educators. Firstly, when learners are satisfied with their learning experience, they can put more effort into their self-language learning, feel more motivated to attend classes, and reduce the likelihood of dropping out of school than unsatisfied students (Elliott & Shin, 2002; Gao, 2010; Özgüngör, 2010). Moreover, understanding learners' satisfaction can help language educators make necessary adjustments to instructional design in response to the market needs of courses and programs, thereby enhancing the learning experience (Asakereh & Dehghannezhad, 2015; Elliott & Shin, 2002; Özgüngör, 2010).

Prior studies have evidenced controversial findings on the relationship between learners' satisfaction and their language performance. On the one hand, learners' satisfaction strongly correlated with their learning outcomes. Students with higher satisfaction were likely to be higher achievers in language learning. For example, Asakereh and Dehghannezhad (2015) found that students' satisfaction with speaking classes had a significant positive correlation with speaking skills achievement through the results of Pearson correlation analyses. Similarly, Qutob (2018) and Khiat (2014) reveal a positive association between students' satisfaction with learning materials, the language teacher, and acquired speaking skills. For online education, prior studies provided consistent findings that learners' satisfactory ratings for learning management systems, online courses, course structure, learner interaction, and instructor presence correlated with their language learning outcomes (Gray & DiLoreto, 2016; Li & Ni, 2024; Shin & Sok, 2023). On the other hand, studies found no statistical difference between learners' satisfaction with learning experience and language learning outcomes in different learning contexts such as distance programs (Khan & Iqbal, 2016), e-learning (Yılmaz, 2023), online learning (Anthonysamy & Singh, 2023; Murtiningsih et al., 2023; Shin & Sok, 2023). Our study delved into this association in the context of chatbot-human interaction in speaking classes to investigate whether learners' satisfaction is associated with their speaking gains.

Previous studies on learners' satisfaction with chatbot-human interaction

To date, a few studies have examined learners' satisfaction with chatbot-human interaction in practicing speaking skills. The first of which was conducted by Belda-Medina and Calvo-Ferrer (2022) with undergraduate teacher-students in two universities in Spain and Poland, whose English proficiency levels were at upper-intermediate to advanced (approximately B2–C1), some of them were native speakers of English. It measured ESL learners' satisfaction with language experience produced by three conversational agents (Replika, Kuki, Wysa) through a questionnaire grounded in the CHISM. The findings showed relatively high satisfaction with Replika and moderate satisfaction with Kuki and Wysa. While chatbots were highly praised for their wide lexical range, the quick response interval, and natural conversational flow, the learners showed dissatisfaction with the AVCs' ability to detect and correct learners' linguistic mistakes. This study was among the earliest to record learners' voices regarding their satisfaction after interacting orally with chatbots rather than human partners. However, the research design was quite complicated and insufficient to support such conclusions. Accordingly, the participants were required to select any of the three available AVCs to practice, which led to an uneven exposure across the AVCs. If a chatbot matched their preference, they were likely to interact more and gain a greater understanding, while other AVCs were largely neglected. Another study evaluated satisfaction levels of college students in Spain and Czech Republic with the interaction with four AVCs, namely Mondly, Andy, John Bot, and Buddy.ai (Belda-Medina & Kokošková, 2023). The participants' English levels were also from B2 to C2. They practiced speaking with these AVCs after class and completed assessment reports after each practice session, submitted transcripts, and screenshots of their interaction. The satisfaction was measured based on an adapted version of the CHISM (Belda-Medina & Calvo-Ferrer, 2022). The results indicated a low satisfaction level with the language-related aspects produced by the AVCs. In line with the aforementioned research, the learners in this study reported dissatisfaction with the chatbots' error detection and correction. Moreover, the participants were also dissatisfied with the AVCs' ability to vary sentence complexity and adaptability to context. The study provided a more refined version of the adapted version of the CHISM, yet its research design posed some limitations, such as the inappropriate selection of the chatbot for the participants' level. For instance, Buddy.ai was originally designed for kids whose proficiency levels were not as high as those of university students at levels B2 to C2. In addition, the learners were offered four AVCs with different features and designs; they had freedom in selecting the chatbot to practice, resulting in similar preference bias as mentioned above.

Despite providing valuable insights into learners' perception of chatbot-human interaction, these studies exposed some limitations. First, there were conducted in Western universities, where students' speaking proficiency ranged from upper-intermediate to advanced levels (B2–C2), even native speakers (Belda-Medina & Calvo-Ferrer, 2022; Belda-Medina & Kokošková, 2023). These proficient users of English tend to have higher expectations for the chatbot's

language use, which might not reflect the needs of EFL learners in EFL contexts. Second, participants in those studies reported inconsistent satisfaction levels with different AVCs. As each AVC differs in features, interaction design, feedback mechanisms, and target users, the inclusion of multiple AVCs in the same intervention might result in learners' confusion and preference bias.

To address those limitations, our study adapted the LEX dimension of the CHISM to investigate Vietnamese university students' satisfaction with language-related aspects produced by ChatGPT. Moreover, we aimed to examine the association between learners' satisfaction and speaking gains. This research design enhances the study's novelty in several ways. First, the present study focused on EFL learners with lower levels of English proficiency (A2–B1) who had a strong desire to improve speaking skills for future academic, professional, and communicative purposes. By providing these students with opportunities to engage in real-time spoken interaction through ChatGPT's advanced voice mode, we aimed to create more accessible speaking practice without reliance on native-speaker interlocutors. Second, by employing a single AVC, the study controls for design-related variability across chatbots. Thus, reducing confounding factors related to system features. Finally, the study examines students' satisfaction levels in relation to their speaking performance to reveal if learners' perceived experience aligns with or diverges from their subjectively measured speaking outcomes. Consequently, we seek to answer three research questions:

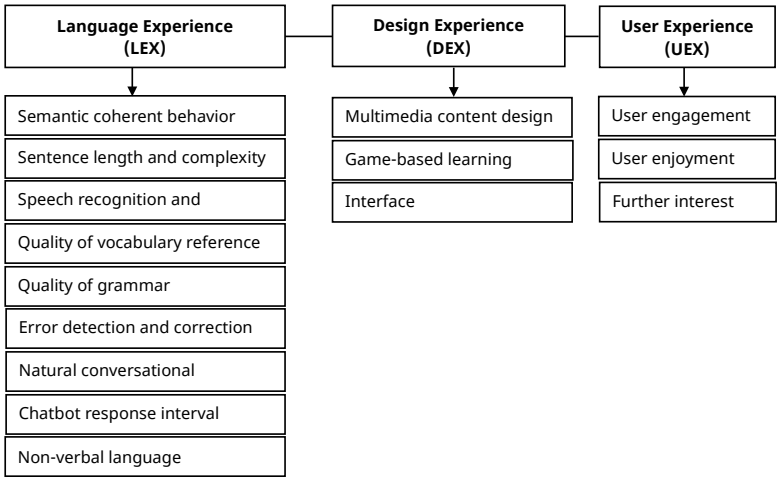
RQ1. To what extent are students satisfied with the language experience produced by the AI voice-based chatbot?

RQ2. To what extent does learners' speaking performance improve after engaging in AI-based chatbot speaking practice?

RQ3. Is learners' satisfaction with the language experience associated with their speaking outcomes?

Theoretical framework

This study was grounded in the Chatbot–Human Interaction Satisfaction Model (CHISM) (Belda-Medina & Kokošková, 2023), which was developed to conceptualize and measure learners' satisfaction when interacting with educational chatbots. CHISM was first created in 2022 (Belda-Medina & Calvo-Ferrer, 2022) and then revised in 2023 with some slight modifications of wording and features of the selected chatbots used. Both versions measure students' satisfaction with chatbot-human interaction in three macro dimensions, including Language Experience (LEX), Design Experience (DEX), and User Experience (UEX) (see Figure 1). LEX includes language-related aspects produced by the chatbot as a speaking partner, DEX items focus on the features and designs of the AVCs, and UEX comprises items on learners' overall perception of the experience.



Note. This diagram shows the structure of CHISM (Belda-Medina & Kokošková, 2023).

Figure 1. Chatbot-Human Interaction Satisfaction Model (CHISM)

Language experience dimension

Our study adapted the LEX dimension of the CHISM framework (Belda-Medina & Kokošková, 2023) by excluding the *non-verbal language* because this function (i.e. Emojis, memes, stickers, GIFs, and other visual representations) is often seen in the text-based mode of ChatGPT, but not in the voice-based version. Moreover, ChatGPT does not have an embodied avatar with visual cues, gestures, and symbols to convey meaning and emotions in the absence of spoken words. Therefore, the LEX was measured through eight constructs with the specific operational definitions as in the original version. Firstly, *semantic coherent behavior* refers to ChatGPT's ability to maintain a contextually relevant and meaningful dialogue with the user. Secondly, *sentence length and complexity* focuses on the structure and variety of the sentences that ChatGPT used in its responses. Thirdly, *quality of vocabulary reference* refers to ChatGPT's use of diverse and topic-related vocabulary in context. Fourthly, *quality of grammar explanations* means ChatGPT's ability to explain grammar rules accurately and clearly. Fifthly, *error detection and correction* refers to ChatGPT's ability to detect the users' linguistic errors and correct them. Sixthly, *speech recognition and synthesis* refers to the AVC's ability to recognize and understand users' language input that converts spoken language into written text. Seventhly, *natural conversational interaction* pertains to ChatGPT's ability to resemble the natural flow and dynamics of human conversation, such as understanding and responding appropriately to user inputs, demonstrating empathy, and adapting the language style and tone to suit the learner's preferences. Finally, *chatbot response interval* relates to the time it takes for a chatbot to generate a response following text or voice-based input from the learner.

Our study only focused on examining learners' satisfaction with LEX dimension because of the following reasons. First, it aimed to explore the impacts

of a new teaching technique on students' speaking skills, so LEX ensured a closer conceptual alignment between how students felt about the language experience and their speaking gains. Moreover, restricting the analysis to LEX helped reduce construct-irrelevant variance arising from non-linguistic aspects of chatbot use, such as the technical features (DEX) or usability issues (UEX). These aspects were undeniably important to explore the comprehensive satisfaction with chatbot-human interaction, but they might not be central to explaining changes in speaking outcomes. Third, the study represented an initial application of CHISM in the context of chatbot-assisted speaking practice with Vietnamese university students. Beginning with a single, theoretically central dimension would make the measurement more focused and manageable. Therefore, it allowed the researchers to provide a clearer answer to the question of whether satisfaction with language experience is associated with speaking outcomes.

Methodology

Research design and context

To measure learners' satisfaction levels with the chatbot-human interaction experience and how their speaking skills change, we adopted a one-group pretest-posttest research design. In the current research setting, the decision to exclude a control group was made for practical and ethical concerns. First, the intervention was conducted in a normal classroom setting where random assignment of individuals was unfeasible. As institutional regulations, all classes must be pre-registered before the semester starts. Second, all students enrolling in the same compulsory speaking course were required to receive equal learning conditions. Although this approach is weaker than design with control groups for establishing a causal relationship, it is often used in field research because the pretest provides a baseline measure for change (Shadish et al., 2002). Therefore, this research design is often used to explore the participants' observed change or growth under real instructional conditions over time (Shadish et al., 2002).

It included a group of participants who participated in eight sessions interacting with ChatGPT as a speaking partner. The study was conducted in a state university in the South of Vietnam from August to October 2025. The students participating in the study were English-major sophomores undertaking a Speaking 2 course. The expected learning outcomes of this course are that students will achieve a proficiency level of B1 in speaking skills. The course book was New Interaction 2 (Tanka & Most, 2020). The students have a 100-minute speaking lesson weekly. The normal class size of the current research site is about 40 students. Before collecting data, a gatekeeper letter was obtained to document institutional approval from the university, faculty, and relevant academic divisions. For integrating AI-assisted speaking activities, the three speaking classes were conducted in a Lab room with the availability of digital conditions such as internet connection, a computer with a microphone and

headphones for each student, a big screen for teachers to show to the whole class, and a network support system to connect the teacher and the students' screens.

Participants

The study used a convenience sample of English-major students at a Vietnamese state university. The students were from three intact classes who were currently undergoing the English Speaking 2 course. Employing participants from intact classes enhanced ecological validity and natural dynamics of classroom learning, which is widely supported in educational research (Creswell & Creswell, 2017; Dörnyei, 2007). There were originally 122 students in the classes, but some of them did not participate in the data collection procedure. Therefore, the final number of valid participants was 116, with 36 males (31%) and 80 females (69%). Following the rule-of-thumb recommendations for classroom-based intervention studies, which suggest that a minimum of 30 participants is adequate for a pre-post design and 50 for a correlational test (VanVoorhis & Morgan, 2007), the sample size is large enough to support statistical inferences. The mean age of the group was 19.22 years (SD = .789). All participants have been studying English for more than 5 years, and their self-reported speaking proficiency ranged from A2 to B1 at the beginning of the course. All of them were familiar with technological skills to conduct a conversation with ChatGPT. At the beginning of the course, the students were informed of the chatbot-assisted speaking activities, the purposes of the research and their voluntary participation, and their right to withdraw without penalty.

Instruments

In order to investigate the satisfaction levels, speaking performance changes over time, and their relationship, an 8-week project entitled "GPT Talk Time" was run for the classes. The class procedure and topics of the conversations followed the curriculum and the textbook. Over the eight weeks, the students engaged in weekly AI-mediated speaking activities for about 30 minutes, during which they spoke with ChatGPT as a speaking partner rather than a peer. Before the intervention, all students were trained on how to use the guiding prompts (e.g., setting roles, introducing opinions, giving reasons, asking follow-up questions). Although the prompts were scripted in weekly handouts, learners were encouraged to extend the conversation freely. This design aimed to balance consistency across students with opportunities for natural, open-ended interaction. Moreover, the chatbot was prompted to act as a conversational partner, provide corrective feedback when major errors affected comprehension, detect frequent errors, and offer brief suggestions on vocabulary, grammar, or fluency at the end of the conversation. Another important point is to deal with potential communication breakdowns. In the training sessions, learners were informed of possible issues and suggested solutions. For example, they were provided with and guided strategies to negotiate for meaning

in case the conversation went wrong. Interaction breakdowns due to technical issues, such as internet connection and microphones, were also monitored through classroom observation and weekly review of selected chat logs. In such cases, the teacher provided brief scaffolding or clarification to help students resume the task.

Data were collected through a satisfaction questionnaire and a pre-test and post-test. The questionnaire was designed in the light of the CHISM framework (Belda-Medina & Kokošková, 2023) with 14 items in the format of a 5-point Likert scale (1= very dissatisfied to 5 = very satisfied). The questionnaire was reviewed by three colleagues who hold PhD degrees and are currently teaching English for academic purposes to ensure content validity and appropriateness. The feedback from those experts was synthesized and reviewed by the researchers. Then, a pilot test with a small group of 15 students was conducted to check item clarity and time to respond. After this phase, the final version of the questionnaire was polished regarding the wording, structure, and clarity. Then it was translated into Vietnamese for the official participants to understand easily. Cronbach's alpha value for the sections of the questionnaire was .876 (LEX items), indicating its good internal consistency (Pallant, 2020). Most items showed acceptable corrected item-total correlations above the .30 threshold, suggesting strong alignment with the overall scale. Although LEX3 and LEX4 had lower correlations (.227 and .212, respectively), we decided to keep them as two important items of the questionnaire for two reasons. First, if they were removed, Cronbach's Alpha would only increase slightly, which indicates their minimal impact on the overall reliability. Secondly, these items were included in the theoretical framework, so retaining them helps preserve the content validity of the scale.

The speaking pre- and post-tests were designed to measure the changes in students' speaking performance at the beginning and end of the intervention. They include identical questions to avoid the task difficulty effect (Brown & Abeywickrama, 2019). Moreover, to mitigate the risk of memorization or rehearsed responses, the post-test was conducted more than two months after the pre-test. In addition to the sufficiently long interval between the pre- and post-tests, students were not informed in advance that the same questions would be reused, and the scoring focused on general speaking performance (e.g., fluency, vocabulary, grammar, pronunciation, and discourse management) rather than content recall. The test structure adopted the Vietnamese Standardized Test of English Proficiency (VSTEP) speaking test to ensure the test validity and reliability (Weir, 2005). The VSTEP speaking test consists of three tasks: *Social interaction, Problem solution,* and *Topic development.* The time to complete the test is approximately 12 minutes. It assesses test-takers' speaking performance based on five criteria, including fluency, vocabulary, grammar, pronunciation, and discourse management. Its analytic rubric describes specific indicators for scoring test-takers' speaking performance, ranging from 0 to 10 for each criterion. Students participate in the computer-based test with a screen and audio recording. Students' recordings were coded with the class name plus a number (e.g., A1). The tests were rated by the researcher and another invited rater who

holds a Master of Arts degree; both raters were trained and had more than 4 years of experience evaluating the VSTEP speaking test. To ensure consistency between the raters, 20 random recordings were selected and rated separately by the raters. The inter-rater agreement was calculated based on an acceptable difference of ±1 point between raters. Seventeen out of twenty selected recordings fell within the acceptable range, accounting for an agreement rate of 85%, which was considered highly acceptable (McHugh, 2012). Then the raters discussed and resolved the rating differences. Finally, the researcher continued to grade the rest recordings.

Data collection procedure

The procedure to collect data for the study, designed to span 11 weeks, was summarized in Table 1.

Table 1. Data collection timeline

Week 1	Preparation and training	– Introduce the project "GPT Talk Time" – Inform and reach a consent – Provide a trial activity for students to have a conversation with ChatGPT – Introduce VSTEP speaking test format and scoring criteria – Provide training on how to use OBSscreen-recording software
Week 2	Pre-test and training on dealing with communication breakdowns	– Distribute pre-test – Provide a ChatGPT Plus account for the students – Provide training on chatbot settings – Deliver a handout with techniques to resolve communication breakdowns when conversing with ChatGPT
Weeks 3–10	GPT Talk Time project	– Introduce the topic – Review related vocabulary and grammatical structure – Deliver the handout and give instructions – Ask students to prepare ideas for 5 minutes – Ask students to engage in a conversation with ChatGPT to complete the handout in 20 minutes – Ask students to share their opinions with the whole class – Provide feedback and review the lesson
Week 11	Post-test and questionnaire	– Distribute post-test – Distribute questionnaire

Data analysis

To measure students' satisfaction with linguistic aspects produced by ChatGPT, the item responses on LEX dimension were first checked for internal consistency (Cronbach's α). Satisfaction levels were calculated and described descriptively using means and standard deviations with SPSS 26.

Students' speaking performance was measured at the beginning and end of the project. To quantify improvement, we computed a gain score (posttest

scores minus pretest scores). To have an appropriate selection of the statistical tests, we first checked if the test scores were normally distributed or not by using the Shapiro–Wilk test, skewness and kurtosis values, and visual inspection of the histogram (Ghasemi & Zahediasl, 2012). The Shapiro–Wilk test was not significant, $W = .988$, $p = .367$ ($p > 0.05$), indicating that the gain scores were normally distributed. Skewness (−0.13) and kurtosis (−0.24) were also within acceptable ranges. To identify the statistical significance of students' speaking performance improvement, the speaking test scores were analyzed using t-tests for paired samples (Field, 2024). Cohen's d was calculated using the standard formula for paired samples, in which the effect size equals the t-value divided by the square root of the sample size ($d = t / \sqrt{n}$). This method is recommended for within-subject designs because it standardizes the mean difference using the variability of the difference scores (Fritz et al., 2012).

Finally, to examine the relationship between students' satisfaction and their speaking improvement, we checked for normality of the data distribution of the satisfaction scores by using a similar test as above. The results indicated that satisfaction scores were normally distributed ($W = .984$, $p = .177$ [$p > 0.05$]), then a Pearson's correlation was used to calculate the association between satisfaction levels and speaking improvements of the participants (Field, 2024).

Results

Firstly, as shown in Table 2, the study unveils how satisfied the students felt with the language experience produced by ChatGPT over the chatbot-human interaction sessions in the speaking class. Overall, the descriptive statistics show moderate to high satisfaction levels with ChatGPT's linguistic and interactional performance. All the items had the mean scores ranging from 3.48 to 4.34 on the 5-point Likert scale. Students expressed the highest level of satisfaction with ChatGPT's response speed (M = 4.34, SD = .685), followed by error detection (M = 4.21, SD = .829), vocabulary range (M=4.19, SD = .745), and topic-relevant vocabulary (M = 4.18, SD = .705). In contrast, the participants showed relatively low satisfaction with ChatGPT's use of complex grammatical structures (M = 3.48, SD = .964) and sentence length (M = 3.66, SD = .835), indicating comparatively lower satisfaction with these aspects of ChatGPT's linguistic output. Standard deviations ranged from 0.622 to 0.964, indicating moderate variability in learners' perceptions across items.

Table 2. Descriptive results for students' satisfaction levels with LEX (N=116)

	Mean	Std deviation
LEX 1. ChatGPT's on-topic responses	4.16	0.654
LEX 2. ChatGPT's relevant responses	4.17	0.622
LEX 3. ChatGPT's use of sentence length	3.66	0.835
LEX 4. ChatGPT's use of grammar complexity	3.48	0.964
LEX 5. ChatGPT's use of vocabulary range	4.19	0.745
LEX 6. ChatGPT's use of topic-appropriate vocabulary	4.18	0.705
LEX 7. ChatGPT's grammar explanation accuracy	4.01	0.786
LEX 8. ChatGPT's grammar explanation clarity	3.79	0.909
LEX 9. ChatGPT's ability to detect errors	4.21	0.829
LEX 10. ChatGPT's corrective feedback	4.10	0.869
LEX 11. ChatGPT's voice recognition	3.88	0.876
LEX 12. ChatGPT's voice synthesis	3.91	0.894
LEX 13. ChatGPT's ability to maintain natural conversation flow	3.98	0.791
LEX 14. ChatGPT's response speed	4.34	0.685

Regarding students' speaking improvement after engaging in the 8-week project (excluding preparation and data collection time), interacting with ChatGPT as a speaking partner, Table 3 shows the results of a paired t-test for the overall speaking gains by comparing the post-test and pre-test scores. The results showed that students' speaking scores improved significantly from pre-test to post-test, $t(115) = 14.43$, $p < 0.01$, with a mean gain of 1.59 points (SD = 1.19). In addition, the result of the Cohen's d test indicated a large effect size (Cohen's $d = 1.34$), which reflected a substantially observed improvements in students' pretest and posttest scores following the intervention.

Table 3. Results of a paired t-test for students' overall speaking improvement

	Paired differences		t	df	Sig. (2-tailed)
	Mean	Std deviation			
Overall_posttest – Overall_pretest	1.59	1.19	14.43	115	0.00

Looking at more details of speaking skills improvement, a series of paired-samples t-tests was conducted to examine students' changes across the five analytic speaking subskills. These criteria were grounded in the rubric for evaluating speaking skills of the VSTEP test. Table 4 showed the results of the paired t-tests for speaking subskills changes. Results indicated statistically significant gains in all areas ($p < 0.01$). Among the five criteria, the most substantial increase was observed in *discourse management* (M = 1.89, SD = 1.24), $t(115) = 16.37$, followed by *vocabulary use* (M = 1.65, SD = 1.22), $t(115) = 14.56$. Significant improvements were also found in *grammar* (M = 1.52, SD = 1.30), $t(115) = 12.62$,

fluency (M = 1.47, SD = 1.41), *t*(115) = 11.19, and *pronunciation* (M = 1.42, SD = 1.54), *t*(115) = 9.96.

Table 4. Results of a paired samples t-test for speaking performance by criteria

		Paired differences				
		Mean	Std deviation	t	df	Sig. (2-tailed)
Pair 1	Vocab_posttest – Vocab_pretest	1.65	1.21	14.56	115	0.000
Pair 2	Gram_ posttest – Gram_ pretest	1.51	1.30	12.62	115	0.000
Pair 3	Pronun_ posttest – Pronun_ pretest	1.42	1.53	9.96	115	0.000
Pair 4	Fluency_ posttest – Fluency_ pretest	1.47	1.41	11.19	115	0.000
Pair 5	Discourse_ posttest – Discourse_ pretest	1.89	1.24	16.37	115	0.000

The association between students' satisfaction and speaking performance was shown in Table 5. We conducted Pearson's correlation to explore the relationship between students' feeling with the learning experience and their speaking gains over eight weeks. Surprisingly, the results indicate that the correlation between learners' satisfaction levels and their speaking gain scores was positive but very small and not statistically significant, $r(114) = .087$, $p > 0.05$. This result suggests that students' satisfaction levels were not associated with the observed improvement in their speaking performance.

Table 5. Pearson's Correlation between students' satisfaction with LEX and speaking gains (N = 116)

		Speaking gain	Language experience
Speaking gain	Pearson Correlation	1	0.087
	Sig. (2-tailed)		0.350
Language experience	Pearson Correlation	0.087	1
	Sig. (2-tailed)	0.350	

Discussion

The present study aimed to provide an objective analysis of Vietnamese university students' satisfaction with language experience as interacting with an AVC, and the observed speaking improvement. It also examined the correlation between how the learners subjectively felt about the learning experience and their speaking gains. The following sections discuss the findings of the study in greater depth to address the three formulated research questions.

Students' satisfaction with LEX dimension

For research question 1 (*To what extent are students satisfied with the language experience produced by the AI voice-based chatbot?*), the findings show learners' moderate to high levels of satisfaction with LEX when conversing with ChatGPT, with mean scores ranging from 3.48 to 4.34. This result suggests that in the context of chatbot-human interaction, language produced by ChatGPT was generally able to meet the learners' expectations. This finding was in line with previous studies on other AVCs, such as Replika (Belda-Medina & Calvo-Ferrer, 2022) and Translation Learning Chatbot (Aleedy et al., 2025). However, this finding suggests that learners' satisfaction with LEX might depend on the specific chatbot used. Earlier research that involved multiple chatbots showed that learners could be satisfied with one system but dissatisfied with another (Belda-Medina & Calvo-Ferrer, 2022; Belda-Medina & Kokošková, 2023). Another factor that might lead to different satisfaction levels could be learners' English proficiency levels. The present study included EFL learners of lower levels of English proficiency (A2-B1) whose demand for improving English for academic and professional purposes was substantial, yet they had limited chances to talk to proficient speakers and receive corrective feedback (Çakmak, 2022; Nguyen & Burke, 2024). Having a virtual speaking partner with a native-like voice (Duong & Suppasetseree, 2024; Tai & Chen, 2020; Zou et al., 2024) that is supportive and available 24/7 could potentially offer a satisfying experience in practicing speaking. These students' satisfaction levels differed from those of highly proficient learners (B2-C2), as in previous studies (Belda-Medina & Calvo-Ferrer, 2022; Belda-Medina & Kokošková, 2023), who were not satisfied with chatbot-generated language because of their high expectations and demanding communication purpose.

Our study also revealed that the highest satisfaction ratings were associated with ChatGPT's *response speed* (M = 4.34, SD = .685), *relevant and on-topic responses* (M = 4.16, SD = .654 and M = 4.17, SD = .622 respectively), *error detection and correction* (M = 4.21, SD = .829 and M = 4.10, SD = .669 respectively), and *vocabulary range and relevance* (M = 4.19, SD = .745 and M = 4.18, SD = .705 respectively). This means learners particularly valued ChatGPT's ability to maintain fluent, contextually appropriate conversations. In contrast, participants reported lower mean scores for *sentence length* (M = 3.66, SD = .835) and *grammar complexity* (M = 3.48, SD = .964). It means that some learners perceived the limitations in the depth or sophistication of ChatGPT's linguistic output. These findings partially contradict Belda-Medina and Calvo-Ferrer's (2022), who emphasized that learner dissatisfaction with chatbots' *error detection and correction* is believed to be important in second language learning (Lyster et al., 2013). Overall, the results suggested that learner satisfaction with the AVC as a speaking partner depended significantly on how smooth, responsive, and relevant the interaction was. At the same time, higher expectations for grammatical accuracy and instructional clarity were raised as a challenge. It somehow emphasized the role of teachers' guidance to support the optimum ability of chatbot-human interaction.

Observed impacts of chatbot-human interaction on speaking performance

For addressing research question 2 (To what extent does learners' speaking performance improve after engaging in AI-based chatbot speaking practice?), the results showed that students' speaking outcomes at the end of the project were higher than those at the beginning of the project (post-pre gain ($t(115) = 14.43$, $p < 0.01$). In other words, there was an observable improvement in students' speaking performance after conversing with ChatGPT for 8 sessions. The most notable gains were in discourse management and vocabulary use. It suggests that engaging in the interaction with ChatGPT somehow supported learners' ability to organize ideas and access lexical resources during spoken production. This finding aligns with prior studies, which emphasized the role of AVCs in facilitating learners' fluency, idea development, and lexical expansion through repeated, low-anxiety practice (Fryer & Carpenter, 2006; Kohnke et al., 2023; Tai & Chen, 2020). Improvements in grammar, fluency, and pronunciation further align with findings from AI-assisted speaking research, indicating that frequent exposure to accurate input and corrective feedback can promote measurable speaking development (Kohnke et al., 2023; Ouyang & Jiao, 2021). Taken together, the findings suggested that AVCs can be a potentially beneficial speaking partner that might support EFL learners in multiple subskills of speaking performance. With teachers' effective instructional design and guidance, EFL students can use ChatGPT as a conversational partner to practice speaking. By so doing, these tools may help address some of the ongoing challenges in speaking instruction, such as a lack of opportunities to speak with proficient or native speakers, and insufficient time for feedback.

Association between students' satisfaction with LEX and their speaking performance

For research question 3 (*Is learners' satisfaction with the language experience associated with their speaking outcomes?*), the findings surprisingly showed that learners' satisfaction with learning experience was not statistically correlated with their improved speaking skills ($r = .087$, $p > 0.05$). In other words, students who felt more satisfied with conversing with ChatGPT as a speaking partner did not necessarily achieve greater speaking gains than those who reported lower satisfaction. Possible reasons for this null correlation might include novelty effect, insufficient time-on-task, and topic misalignment. In the research context, EFL learners were conventionally more familiar with conversing with human interlocutors, in which they faced several challenges, such as fear of being judged by the teacher or other classmates, anxiety about making mistakes, and limited opportunities to communicate with proficient speakers (Chulee et al., 2023; Nguyen & Burke, 20252024; Tai & Chen, 2020). When they were given the chance to practice speaking with a chatbot, which is always available for interaction, and offers enthusiastic encouragement (Fathi et al., 2024; Yusuf et al., 2025), and can function as a more knowledgeable partner (Li et al, 2025), they may have evaluated this novel learning experience positively and reported high levels of satisfaction. Another explanation for the lack of a

significant correlation between learners' satisfaction and speaking improvement might be related to insufficient time-on-task. Although the intervention lasted eight weeks, such in-class practice might not provide enough time to rehearse and improve speaking skills. Learners might need more practice time both in class and out of class, as practice enables the gradual transformation of performance from controlled to automatic, leading to language automatization (DeKeyser, 2007). In-class tasks would provide them with guidance to replicate or repeat the task at home, so that their satisfaction level might be correlated with speaking improvement. One more reason that might account for the null correlation between satisfaction and speaking improvement is the misalignment between practice tasks and speaking tests. The intervention was conducted in a real classroom context where task topics were aligned with the units in the selected textbook, and task designs were communicative-oriented. However, the topics, task format, and rating scales used in the pretest and posttest were different from what learners had practiced during the intervention. Such task-test mismatch has been identified as a common issue in language assessment and can undermine the validity of inferences drawn from test scores (Fulcher, 2014).

The lack of correlation between learners' satisfaction with learning experience and their speaking gains found in the present study is somewhat consistent with prior research. For instance, Anthonysamy and Singh (2023), Murtiningsih et al. (2023), and Yilmaz (2023) revealed that satisfaction was not a significant predictor of academic performance in online and distance learning contexts. However, our results contradicted some studies that found a significant positive association between learners' satisfaction and learning. For instance, Qutob (2018) and Khiat (2014) highlighted the statistical correlation between learners' satisfaction with learning materials, language teacher, and learning experience, and speaking performance in traditional speaking classes. Gray and DiLoreto (2016), Li and Ni (2024) reported a slight association between learners' satisfaction with online and distant learning experience and their language learning outcomes. One possible explanation for this inconsistency lies in differences in learning context, instructional modes, and the way satisfaction was measured. While earlier research examined satisfaction in traditional or online courses where instruction, materials, and teacher support were closely aligned with learning outcomes (Gray & DiLoreto, 2016; Khiat, 2014; Qutob, 2018), satisfaction with chatbot-mediated speaking practice might reflect learners' affective comfort and convenience rather than direct gains in speaking performance.

Conclusion

This study aimed to examine Vietnamese university students' satisfaction with the language experience produced by ChatGPT as a speaking partner, their speaking gains after 8 speaking sessions in class with the AVC, as well as the relationship between satisfaction and speaking gains. The results indicate that the students were moderately to highly satisfied with the chatbot's linguistic

production during the conversations between the human and the chatbot. Another finding of this study is that results from the paired t-test showed significant improvement in EFL learners' oral performance after practicing speaking with ChatGPT, especially vocabulary range, fluency, and the use of cohesive devices. Although students enjoyed the learning experience and their speaking gains were significant, interestingly, there was no statistical correlation between students' enjoyment of the learning experience and their speaking performance.

Based on those findings, we propose several pedagogical implications for better AI-mediated speaking instructions. First, AVCs might function as a conversational partner to help learners improve speaking skills, particularly by increasing fluency, expanding vocabulary range, and promoting cohesive discourse. Despite their potential in enhancing learners' oral performance, AVCs should not be seen as a magic wand that could meet overly high expectations from both learners and teachers. Instead, they could be supportive assistants only when they are in the hands of teachers who are capable of organizing learning activities and know how to foster learners' positive emotions. For classroom practice, it would be necessary for teachers to design AI-mediated speaking tasks through semi-structured role-plays and task-based conversations, in which key guidance and prompts are scripted, yet learners are encouraged to extend their responses rather than engage in unguided free chatting. To ensure the quality of the AI-mediated speaking tasks, it would be important for teachers to provide initial training on dealing with technical issues and dialogue maintenance strategies; allow learners time to familiarize themselves with this novel learning experience; observe how the tasks work in practice, and make necessary adjustments to task design over time. Second, the absence of significant correlation between learners' satisfaction with the learning experience and their speaking gains suggests that teachers should design the tasks more specifically and easily for learners to follow. For example, each task should include clear task goals, guidance, and samples before the real chatbot-human sessions. With specific designs, learners might have a sense of accomplishment as well as satisfaction after the tasks. In addition, the same tasks should be repeated in the following sessions for learners to familiarize themselves, reflect, and improve the quality of the interaction. To promote more time-on-task, every in-class practice task should be well-organized to set a good model for learners to replicate at home, which might encourage learners to practice speaking with chatbots more after class. Teachers also redesign the format of tests to be matched with practice task design so that learners' test scores would not be affected by the test effect. Third, for chatbot designers, while the high level of satisfaction with the chatbot's linguistic output highlights the importance of maintaining natural, coherent, and contextually appropriate responses, it also points to the need for features that can support learners' noticing, elaboration, and form-focused feedback during interaction.

The present study made important theoretical and contextual contributions to the AI-mediated speaking instruction. Theoretically, it supports previous research that chatbot-human interaction is still a novel experience in language

education, which surprised and excited the learners, and to some extent, learners' speaking skills could be improved when they practice speaking with chatbots. It also adds an important finding that learners' satisfaction with their learning experience is not significantly correlated with improved speaking outcomes in chatbot-human interaction. Furthermore, in the Vietnamese educational context where EFL students tend to have limited chances to speak to a competent speaker both in and out of class, and insufficient time for feedback due to large class size, this study highlights the potential of using AVCs in general, ChatGPT in particular, as an accessible tool, a more knowledgeable speaking partner to support speaking practice. Although the present study was conducted in the Vietnamese context, these findings may also be relevant to other EFL contexts that share similar constraints in terms of input and speaking opportunities. However, caution is needed when generalizing the results, as differences in learners' characteristics and proficiency levels, and instructional settings may influence the effectiveness of chatbot-supported speaking practice.

This study has certain limitations that should be addressed in future research. First, we only focused on the linguistic-related aspects (LEX) of CHISM, suggesting that future studies should include both DEX and UEX to provide a more comprehensive understanding of chatbot-human interaction. In addition, this study used only quantitative data to explore satisfaction and changes in speaking skills, which highlights the need for mixed-methods or qualitative research design to provide deeper insights into learners' satisfaction and/or perceived speaking gains. Finally, future studies should include a control group to make more robust comparisons between human-human interaction and human-chatbot interaction in speaking improvement.

Declaration of conflicting interests

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

Acknowledgement

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

Use of generative AI

This manuscript made use of Grammarly to assist in paraphrasing and improving the clarity of text and ChatGPT for locating information, articles, and references. No AI tools were used for the generation of research hypotheses or the writing of conclusions. The final manuscript was authored and reviewed by the listed authors without additional AI input.

References

Aleedy, M., Atwell, E., & Meshoul, S. (2025). A multi-agent chatbot architecture for AI-driven language learning. *Applied Sciences, 15*(19), 10634. https://doi.org/10.3390/app151910634

Allouch, M., Azaria, A., & Azoulay, R. (2021). Conversational agents: goals, technologies, vision and challenges. *Sensors, 21*(24), 8448. https://doi.org/10.3390/s21248448

Anthonysamy, L., & Singh, P. (2023). The impact of satisfaction, and autonomous learning strategies use on scholastic achievement during Covid-19 confinement in Malaysia. *Heliyon, 9*(2), e12198. https://doi.org/10.1016/j.heliyon.2022.e12198

Asakereh, A., & Dehghannezhad, M. (2015). Student satisfaction with EFL speaking classes: Relating speaking self-efficacy and skills achievement. *Issues in Educational Research, 25*(4).

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

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI conversational partners in language learning. *Applied Science, 12*(17). https://doi.org/10.3390/app12178427

Belda-Medina, J., & Kokošková, V. (2023). Integrating chatbots in education: Insights from the Chatbot-Human Interaction Satisfaction Model (CHISM). *International Journal of Educational Technology in Higher Education, 20*(1), 62. https://doi.org/10.1186/s41239-023-00432-3

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment principles and class-room practices* (3rd Ed.). Pearson Education.

Çakmak, F. (2022). Chatbot-human interaction and its effects on EFL students' L2 speaking performance and anxiety. *Novitas-ROYAL, 16*(2), 113–136.

Černý, M. (2022). The history of chatbots: The journey from psychological experiment to educational object. *Journal of Applied Technical and Educational Sciences, 12*(3). https://doi.org/10.24368/jates322

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 Journal, 47*(4). https://doi.org/10.61871/mj.v47n4-5

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage.

DeKeyser, R. M. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology.* Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford University Press.

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

Elliott, K. M., & Shin, D. (2002). Student satisfaction: An alternative approach to assessing this important concept. *Journal of Higher Education Policy and Management, 24*(2), 197–209. https://doi.org/10.1080/1360080022000013518

Eragamreddy, N. (2025). Exploring the role of Google Assistant in enhancing EFL learners' speaking skills: A systematic review. *The JALT CALL Journal, 21*(3), 2223–2223. https://doi.org/10.29140/jaltcall.v21n3.2223

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

Fernandes, C., Ross, K., & Meraj, M. (2013). Understanding student satisfaction and loyalty in the UAE HE sector. *International Journal of Educational Management, 27*(6), 613–630. https://doi.org/10.1108/IJEM-07-2012-0082

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. https://doi.org/10.1037/a0024338

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning and Technology, 10*(3), 8–14. https://doi.org/10.64152/10125/44068

Fulcher, G. (2014). *Testing second language speaking.* Routledge.

Gameel, B. G. (2017). Learner satisfaction with Massive Open Online Courses. *American Journal of Distance Education, 31*(2), 98–111. https://doi.org/10.1080/08923647.2017.1300462

Gao, X. (2010). *Strategic language learning: The roles of agency and context.* Bristol: Multilingual Matters. https://doi.org/10.21832/9781847692450

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Gray, J. A., & DiLoreto, M. (2016). The effects of student engagement, student satisfaction, and perceived learning in online learning environments. *International Journal of Educational Leadership Preparation, 11*(1).

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). https://doi.org/10.15702/MALL.2022.25.2.09

Khan, J., & Iqbal, M. J. (2016). Relationship between student satisfaction and academic achievement in distance education: A case study of AIOU Islamabad. *FWU Journal of Social Sciences, 10*(2), 137–145.

Khiat, H. (2014). An analysis of the relationship between learning satisfaction and academic achievement of non-traditional learners in Singapore. *International Journal of Research Studies in Education, 3*. https://doi.org/10.5861/ijrse.2013.559

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

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

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Li, L., Zhang, X., Zou, B., & Yang, Q. (2025). AI partner or peer partner? Exploring AI-mediated interaction in EFL pronunciation from a socio-cultural perspective. *Learning, Culture and Social Interaction, 55*, 100958. https://doi.org/10.1016/j.lcsi.2025.100958

Li, H., & Ni, A. (2024). What contributes to student language learning satisfaction and achievement with learning management systems. *Behavioral Sciences, 14*(4), 271. https://doi.org/10.3390/bs1404027

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

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

McAleavy, T., Ha, T. T., & Fitzpatrick, R. (2018). *Promising practice: government schools in Vietnam*. ERIC.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Murtiningsih, S. R., Islamiyati, A. R., & Hidayatulloh, S. M. M. (2023). EFL pre-service teachers' satisfaction of online English learning and their academic achievements. *Indonesian Journal of English Language Teaching and Applied Linguistic, 8*(2), 169–183. https://doi.org/10.21093/ijeltal.v8i2.1548

Nguyen, H. H. (2025). AI for English speaking practice: A study of effectiveness and engagement among Vietnamese university learners. *Australasian Journal of Educational Technology, 41*(3), 45–61. https://doi.org/10.14742/ajet.10050

Nguyen, T. P. L., & Burke, R. (2024). Foreign language teaching in Vietnam: opportunities and challenges. In M. Hayden & T. L. Tran (Eds.). *Education in Vietnam: Making haste slowly* (pp. 145–158). Routledge. https://doi.org/10.4324/9781003298700

Nguyen, T. H. (2025). The impact of AI chatbot integration on Vietnamese EFL learners' speaking performance, fluency, and confidence. *Language, Technology, and Social Media, 3,* 2. https://doi.org/10.70211/ltsm.3026-7196.314

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence, 2,* 100020. https://doi.org/10.1016/j.caeai.2021.100020

Özgüngör, S. (2010). Identifying dimensions of students' ratings that best predict students' self-efficacy, course value and satisfaction. *Eurasian Journal of Educational Research, 10*(38), 146–163.

Pallant, J. (2020). *SPSS survival manual: A step-by-step guide to data analysis using IBM SPSS.* Routledge.

Pham, H. C. (2021). English language education in rural areas: Current issues, complexities and ways forward. *VNU Journal of Science: Education Research, 37*(4). https://doi.org/10.25073/2588-1159/vnuer.4538

Qutob, M. M. (2018). The relationship between EFL learners' satisfaction within the classroom environment and their speaking skills. *English Language Teaching, 11*(7), 116–124. http://doi.org/10.5539/elt.v11n7p116

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton, Mifflin and Company.

Shee, D. Y., & Wang, Y.-S. (2008). Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its applications. *Computers & Education, 50*(3), 894–905. https://doi.org/10.1016/j.compedu.2006.09.005

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

Shin, H. W., & Sok, S. (2023). Student satisfaction and perceived learning in an online second language learning environment: A replication of Gray and DiLoreto (2016). *ReCALL, 35*(2). https://doi.org/10.1017/S0958344023000034

Tai, T.-Y., & Chen, H. H.-J. (2020). The impact of Google Assistant on adolescent EFL learners' willingness to communicate. *Interactive Learning Environments, 31*(3), 1485–1502. https://doi.org/10.1080/10494820.2020.1841801

Tanka, J., & Most, P. (2020). *New Interactions 2.* Mc Graw Hill Press.

VanVoorhis, C. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43–50. https://doi.org/10.20982/tqmp.03.2.p043

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Hampshire: Palgrave McMillan.

Yılmaz, A. B. (2023). The relationship between satisfaction, interaction, e-learning readiness, and academic achievement in online learning. *Open Praxis, 15*(3), 199–213. https://doi.org/10.55982/openpraxis.15.3.578

Yu, L. T., Song, J. Y., & Chiu, F. Y. (2020). Using a three-dimension virtual world to reduce language anxiety and enhance English-speaking performance of EFL university learners: A collaborative project. *Taiwan Journal of TESOL, 17*(2), 65–89. https://doi.org/10.30397/TJTESOL.202010_17(2).0003

Yu, Q. (2022). Factors influencing online learning satisfaction. *Frontiers in Psychology, 13*, 852360. https://doi.org/10.3389/fpsyg.2022.852360

Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Pedagogical AI conversational agents in higher education: A conceptual framework and survey of the state of the art. *Educational Technology Research and Development, 73*(2), 815–874. https://doi.org/10.1007/s11423-025-10447-4

Zou, B., Wang, C., Yan, Y., Du, X., & Ji, Y. (2024). Exploring English as a foreign language learners' adoption and utilization 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

Satisfaction questionnaire

How satisfied are you with the following aspects of language experience when speaking with ChatGPT?
1 = Very dissatisfied
2 = dissatisfied
3 = Neutral
4 = Satisfied
5 = Very satisfied

Items	1	2	3	4	5
LEX 1. ChatGPT's on-topic responses					
LEX 2. ChatGPT's relevant responses					
LEX 3. ChatGPT's use of sentence length					
LEX 4. ChatGPT's use of grammar complexity					
LEX 5. ChatGPT's use of vocabulary range					
LEX 6. ChatGPT's use of topic-appropriate vocabulary					
LEX 7. ChatGPT's grammar explanation accuracy					
LEX 8. ChatGPT's grammar explanation clarity					
LEX 9. ChatGPT's ability to detect errors					
LEX 10. ChatGPT's corrective feedback					
LEX 11. ChatGPT's voice recognition					
LEX 12. ChatGPT's voice synthesis					
LEX 13. ChatGPT's ability to maintain natural conversation flow					
LEX 14. ChatGPT's response speed					