Korean Journal of English Language and Linguistics, **Vol 25, September 2025, pp. 1299-1314**
DOI: 10.15738/kjell.25..202509.1299

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Enhancing L2 Learners' Affective Outcomes and Oral Proficiency Through AI-Chatbot Interaction*

Eun Young Kang (Kongju National University)

Received: July 6, 2025
Revised: September 11, 2025
Accepted: September 16, 2025

Eun Young Kang
Professor, Division of Liberal Arts
Kongju National University
Tel: 041) 521-9735
Email: ekang@kongju.ac.kr

* The manuscript of this research was completed while the author was on research leave abroad as a visiting scholar at Simon Fraser University during the 2024 academic year.

ABSTRACT

Kang, Eun Young. 2025. Enhancing L2 learners' outcomes and oral proficiency through AI Chatbot interaction. *Korean Journal of English Language and Linguistics* 25, 1299-1314.

This study investigated the effects of AI chatbot interaction on EFL learners' foreign language speaking anxiety (FLSA), willingness to communicate (WTC), and speaking proficiency. Forty-eight Korean EFL learners were assigned to either an experimental (n = 24) or a comparison group (n = 24). The experimental group engaged in conversations to complete speaking tasks using the mobile-based app ChatGPT-4o, while the comparison group conducted face-to-face conversations with peers. Data collection included pre- and post-task assessments and questionnaires measuring FLSA and WTC. Speaking proficiency was evaluated across five dimensions: grammar, vocabulary, fluency, pronunciation, and task fulfillment. The findings indicated that the use of an AI chatbot during speaking tasks was significantly more effective than peer interaction in reducing FLSA and enhancing WTC and speaking proficiency. Notably, participants who interacted with ChatGPT showed significant gains in grammar, vocabulary, and task fulfillment, while no significant differences were found between groups for fluency and pronunciation. These results highlight the potential of integrating AI chatbots into EFL speaking instruction to help learners overcome affective barriers and enhance oral proficiency.

KEYWORDS

AI chatbot, ChatGPT, speaking proficiency, foreign language speaking anxiety, willingness to communicate

1. Introduction

Speaking is a fundamental skill in second language (L2), and it is especially essential for learners to articulate their ideas fluently and engage in effective communication. However, compared to other language skills, speaking is often perceived as the most challenging and anxiety-inducing for English as a Foreign Language (EFL) learners (Tai and Chen 2024). This difficulty is partly attributable to affective barriers such as fear of making mistakes, lack of self-confidence, and peer pressure, all of which can detrimentally affect learners' performance (Dizon, 2020). Numerous studies have emphasized the critical role of affective variables—such as anxiety and willingness to communicate (WTC)—in L2 speaking development (Dörnyei 2005). Notably, foreign language speaking anxiety (FLSA), defined as the apprehension experienced when communicating in the target language, poses a major obstacle to the development of speaking skills, impeding learners' ability to speak fluently and engage effectively in learning interactions (Hanafiah et al. 2022). Similarly, WTC, described as "a readiness to enter into discourse at a particular time with a specific person or persons, using an L2" (MacIntyre et al. 1998, p. 547), plays a crucial role in shaping the language learning process. Learners with low WTC often demonstrate reduced capacity to enhance their communication skills (MacIntyre 2017) and a reluctance to take risks when using the L2 during communication (Peng 2019). Thus, fostering a supportive and low-stress learning environment is essential for mitigating these affective barriers and promoting more effective language learning.

The rapid advancement of technologies such as artificial intelligence (AI) and natural language processing (NLP) has introduced dialogue-based systems, particularly conversational chatbots, as promising tools for language learning. These systems can be integrated into educational contexts in either ready-made or custom-designed forms (Lee et al. 2024). General-purpose chatbots are widely accessible online (e.g. Kuki, Cleverbot, and Mondly) or through speech-recognition devices like Amazon Alexa, Google Home, and Apple Siri. In addition, educators can create chatbots tailored to instructional objectives using user-friendly application programming interfaces (APIs), such as Dialogflow or Clova (Jeon 2024 Yang et al. 2022). Previous research has shown that chatbots can enhance communication opportunities for L2 learners (e.g., Ji et al. 2023), increase motivation and reduce speaking anxiety (e.g., Chien et al. 2022, Jeon 2024), provide language-related feedback (e.g., Hew et al. 2022), and support language assessment (e.g., Jeon 2023).

Recent developments in generative AI have built on this line of research and dramatically expanded the potential of chatbot technology for language education. Since late 2022, generative AI platforms such as ChatGPT, Bing Chat, Google Bard, and Gemini have offered L2 learners new opportunities for creative and flexible language learning. Unlike earlier chatbots (e.g., Eliza, Siri, and Alexa), these models provide advanced language generation, multimodal integration, and user customization, and they demonstrate strong linguistic accuracy (Liu et al. 2024). Among them, ChatGPT has emerged as the leading platform, capable of producing human-like dialogue.

However, despite its widespread adoption, empirical research on ChatGPT's effects on L2 learning in Korean EFL contexts remains scarce, with limited attention to speaking interaction. In particular, little is known about how ChatGPT influences affective outcomes such as foreign language anxiety and willingness to communicate, as well as learners' speaking performance. To address this gap, the present study investigates the effects of ChatGPT-mediated interaction on both affective outcomes and oral proficiency. Unlike earlier chatbots that relied on retrieval-based methods and fixed templates, ChatGPT employs a generative model capable of producing human-like, contextually appropriate responses across diverse topics. With integrated speech recognition and text-to-speech (TTS) functions, it also enables voice-based interaction on devices such as smartphones and computers. Building on these capabilities, the present study explores the educational potential of ChatGPT by examining learners' speaking proficiency, anxiety levels, and willingness to engage in English conversation. In doing so, it

seeks to provide insights into the role of AI-mediated interaction in enhancing L2 speaking development and fostering more effective language learning experiences.

2. Literature Review

2.1 Foreign Language Speaking Anxiety (FLSA)

Speaking in a foreign language is often associated with the risk of embarrassment or humiliation. Learners commonly fear making errors due to limited proficiency and may avoid situations that could result in public embarrassment. This emotional difficulty is often referred to as FLSA, which encompasses the negative emotional responses triggered during foreign language speaking tasks (Bashori et al. 2021). To assess such anxiety, researchers have developed various instruments, among which the Foreign Language Classroom Anxiety Scale (FLCAS) by Horwitz et al. (1986) remains one of the most widely used. The FLCAS has been adapted to measure anxiety associated with various language skills, including listening, speaking, reading, and writing (Bashori et al. 2022). Building on this foundation, Öztürk and Gürbüz (2014) specifically developed an FLSAS to assess speaking-related anxiety. The importance of these tools is evident, as research consistently demonstrates that FLSA significantly impairs L2 learners' speaking proficiency, thus hindering oral performance (Mukminin et al. 2015).

Recent studies have explored the role of AI chatbots in reducing EFL learners' speaking anxiety, though findings have been inconsistent. For instance, Shazly (2021) conducted a study with Egyptian college EFL learners and found that despite providing sufficient opportunities for speaking practice, AI chatbots (e.g., Mondly and Web-based chatbots) failed to alleviate learners' speaking-related anxieties. In contrast, Jeon (2024) reported positive effects of AI chatbot use on FLSA. Drawing on in-depth interviews with beginning-level primary school EFL students, the study showed that interaction with a customized AI chatbot, built with Google's Dialogflow, reduced learners' fear of making mistakes. It also helped lower their anxiety about being judged by peers or teachers.

Such discrepancies may be partially explained by differences in the types of AI chatbots employed. In Shazly's study, general-purpose, non-educational chatbots were used, lacking goal-oriented dialogues tailored for language learning. It was suggested that these chatbots might have heightened learners' apprehension about proactively initiating conversations. Conversely, Jeon's study implemented a customized, education-focused chatbot designed to meet learners' specific needs, likely contributing to more positive outcomes. These findings imply that well-structured speaking tasks and teacher support may be necessary to help learners feel less overwhelmed when interacting with non-educational AI chatbots. Nonetheless, further research is needed to systematically examine the affective effects of AI-chatbot use, particularly regarding FLSA.

2.2 Willingness to Communicate (WTC)

WTC is defined as a learners' readiness to engage in L2 communication in specific situations with particular individuals or groups (MacIntyre et al. 1998). It encompasses both stable, trait-like dispositions and fluctuating, situationally dynamic states. Trait-like WTC refers to relatively stable personality characteristics—such as L2 anxiety, motivation, gender, age—that remain consistent across time and contexts. In contrast, situated-dynamic WTC represents a temporary, changeable state of readiness influenced by factors such as topics, classroom environment, and familiarity with interlocutors (Peng 2022).

In recent years, a growing body of research has examined whether emerging technologies can provide low-anxiety environments for L2 communication and positively impact learners' WTC. Technologies investigated include virtual reality (VR) tools such digital games (Reinders and Wattana 2014) and mobile instance messaging applications (Shadiev et al. 2022). These studies generally report that technological tools play a positive role in enhancing learners' WCF. However, investigations specifically exploring the relationship between AI chatbots and WTC remain limited.

A recent study by Tai and Chen (2024) examined the impact of AI chatbots on learners' WTC, focusing on how interaction with Google Assistant influenced EFL learners' communication behaviors and perceptions. Involving 112 junior high school EFL students, the study implemented a 15-day language learning program using Google Assistant. By comparing pre- and post-program WTC survey results, the researchers found that interaction with Google Assistant significantly enhanced learners' WTC, increased communicative confidence, and reduced language anxiety.

Similarly, Kim and Su (2024) conducted a study assessing the effects of an AI chatbot-based speaking activity course on 20 elementary-level Korean language learners' WTC. The researchers developed a chatbot program using Danbee AI, which supports both text-based and voice-enabled interactions. After participating in eight sessions, the learners reported reduced anxiety levels, which in turn led to enhanced willingness to communicate in Korean.

While these findings suggest that AI chatbot integration can positively influence learners' WTC, additional research remains necessary to substantiate and expand upon these results. The present study seeks to contribute to this growing body of literature by examining the impact of a different AI chatbot, ChatGPT, on L2 learners' WTC. By doing so, it aims to provide further insights into the potential role of AI-mediated interaction in fostering greater communicative willingness in L2 learning contexts.

2.3 AI Chatbot as Communication Tools for L2 Speaking

A chatbot is a virtual agent that interacts with users and process their inputs through a computer application (Chiu et al. 2024). Recent advancements in AI technologies—such as NLP, automatic speech recognition, and large language models—have significantly enhanced chatbots' capabilities, enabling them to deliver more human-like interactions and improved conversational abilities (Chiu et al. 2024). AI chatbots like ChatGPT are now capable of conducting coherent, back-and-forth conversations across a wide range of topics and scenarios. They can adopt specific roles, such as a job interviewer, customer, or friend, and generate responses appropriate to the assigned role.

However, despite their ability to provide immediate and personalized feedback, L2 researchers have argued that general-purpose chatbots, which are not specifically designed for language learning, may not serve as ideal conversational partners for L2 learners. Teachers, for example, have limited control over the scope of conversations, and chatbot responses may occasionally deviate from the intended lesson objectives (Kim et al. 2022). Furthermore, many English learners may struggle to sustain extended conversation with chatbots without additional instructional support.

To address these challenges and optimize the educational use of generative AI chatbots, researchers recommend combining chatbot-based L2 speaking tasks with instructor-led guidance (Lee et al. 2020). Examples of such L2 speaking tasks include role-plays simulating real-life communicative scenarios, such as making a restaurant reservation, participating in job interviews, or seeking and giving advice on personal or health-related issues. In addition to integrating well-structured tasks, instructors can guide learners in tailoring chatbot interactions by

modifying prompts to match their L2 proficiency. For instance, ChatGPT allows users to adapt its responses by elaborating, paraphrase, adjusting response speed, or offering corrective feedback when prompted. Through the strategic combination of task-based activities and scaffolded prompting strategies, teachers can leverage general-purpose chatbots, such as ChatGPT, for language learning purposes.

While a variety of general-purpose AI chatbots exist, much of the recent L2 research has focused on intelligent personal assistants (IPAs) such as Apple's Siri, Amazon's Alexa, Google Assistant. For example, Dizon (2020) examined the impact of Alexa on EFL learners' speaking skills. In this study, Japanese college students engaged in 12-minute interaction sessions with Alexa during class over a 10-week period. The findings revealed that Alexa provided learners with ample speaking practice opportunities and facilitated diverse forms of interaction. Similarly, Tai and Chen (2024) explored the use of IPAs to develop EFL learners' oral proficiency. In their study, Chinese-speaking ninth graders interacted with Google Assistant to perform tasks such as asking questions, narrating, describing, and expressing opinions. The results indicated that using IPAs significantly enhanced learners' oral proficiency compared to conventional classroom instruction.

While these studies by Tai and Chen (2024) and Dizon (2020) suggest that IPAs have the potential to enhance speaking skills, IPAs are limited in their ability to sustain extended or complex dialogues compared to more advanced generative AI-based chatbots like ChatGPT. IPAs typically support task-oriented queries through pre-programmed commands or integrated databases, whereas generative AI-powered chatbots are capable of conducting open-ended, free-flowing conversations that dynamically adapt to a wide range of topics. Consequently, generative AI systems such as ChatGPT can predict, comprehend, and produce human-like text, offering transformative potential for L2 language learning.

Despite the promising developments, the integration of generative Al-based chatbots in language education remains an emerging area of inquiry. Liu and Ma (2024) emphasize the need for more comprehensive and rigorous studies to evaluate their pedagogical effectiveness and feasibility for L2 learning contexts. In response to these gaps, the present study seeks to examine the feasibility of using ChatGPT for L2 learning and its impact on key affective factors, namely FLSA and WTC. Specifically, the study addresses the following questions:

1) Does engaging in conversations with ChatGPT positively impact on EFL learners' FLSA?
2) Does engaging in conversations with ChatGPT positively impact on EFL learners' WTC?
3) Does conversing with ChatGPT significantly improve EFL learners' speaking proficiency?

3. Method

3.1 Participants

Forty-eight first-year university students participated in the study. They were aged between 18-20 years (M = 19.23) and were primarily enrolled in science and engineering majors, taking a required General English course. The course was conducted for two hours each week over a 15-week semester. Participants took a practice TOEIC test for level placement prior to the study, and their scores ranged from 550 to 780, corresponding to the low-intermediate level (B1) according to the Common European Framework of Reference for Languages (Council of Europe 2001). Based on these scores, students were assigned to two classes, which were randomly designated as the experimental group and the comparison group. No significant differences were found between the groups' TOEIC scores at baseline, $t(46) = -0.02$, $p = .98$.

All participants had studied English as a required subject for nine years before entering college; however, their English use had been largely limited to receptive skills, reflecting a reading- and grammar-oriented instructional approach commonly observed in Korean secondary education. Consequently, participants generally exhibited reluctance to communicate in English, primarily due to limited speaking proficiency and low self-confidence. In an effort to address this issue, learning activities incorporating AI chatbot-based interaction were designed for the experimental group.

3.2 Speaking Tasks

The participants engaged in total of nine communicative speaking tasks, completing one task per week over the duration of the study (see Table 1). The experimental group performed the tasks by interacting with ChatGPT-4o installed on their smartphones, while the comparison group completed the tasks through face-to-face peer interaction in pairs. The speaking tasks were categorized into three types: information gap, reasoning gap, and opinion gap, following Ellis's (2018) framework.

Each task focused on practical everyday topics relevant to the participants' lives, such as friendship, travel, and television celebrities. Speaking tasks were implemented according to the principles of Task-based Language Teaching (TBLT, Willis 1996), consisting of three stages–pre-task, during-task, and post-task. In the pre-task stage, participants were introduced to the general topic and encouraged to reflect on relevant linguistic elements necessary for completing the task. This stage lasted approximately 20 minutes and included a collective review of useful expressions, vocabulary items, or grammatical structures. In the during-task stage, participants used English to complete the main speaking activity, which lasted approximately 10 to 15 minutes. Both groups received a handout outlining task objectives and instructions. In the post-task stage, activities were flexibly structured to provide additional opportunities for language practice and experimentation with target forms, lasting about 10 minutes.

Table1. Overview of Tasks Utilized in the Study

Session	Task type	Task Description
1	Information gap	Describing and guessing three different members of a chosen K-pop group
2	Reasoning gap	Sharing information about tourist destinations and deciding on a holiday location
3	Opinion gap	Sharing and discussing opinions on how to build a strong friendship
4	Information gap	Describing and guessing three famous Korean actors
5	Reasoning gap	Selecting four items from a list to take on a journey and explaining the selection
6	Opinion gap	Discussing favorite and least favorite animals
7	Information gap	Describing and guessing three favorite TV shows or movie characters
8	Reasoning gap	Selecting five items for survival on a deserted island and justifying choices
9	Opinion gap	Sharing and discussing qualities that make a good parent

3.3 Instruments

3.3.1 ChatGPT

In this study, ChatGPT-4o was employed as an instructional tool to facilitate students' practice of English conversational skills. The model was selected due to its free accessibility and its capability to engage users in voice-based dialogues through the "Talk to ChatGPT" interface. Dialogues were initiated using instructor-designed prompts. Participants in the experimental group completed speaking tasks by interacting with ChatGPT-4o on their smartphones using headphones. The interactions between individual students and ChatGPT-4o were saved as text files within the Talk-to-ChatGPT application and later submitted to the instructor as conversation scripts. To establish a parallel condition, the comparison group's face-to-face interactions were also audio-recorded.

3.3.2 Speaking assessment

Participants' English-speaking proficiency was assessed both before and after the intervention using pretest and posttest measures. Each assessment was conducted individually, and audio recordings were collected for subsequent analysis. The assessments were modeled after the TOEIC speaking test, focusing on two task types: describing a picture and expressing an opinion. These tasks required participants to describe everyday events and answer simple questions on familiar topics, evaluating their ability to describe people and activities, narrate simple events, and express opinions. Participants were allotted 45 seconds to prepare and 45 seconds to respond to each task. One item was used to assess each task type, and the same test was administered for both the pretest and posttest (see Appendix A). A 10-week interval between the pretest and posttest minimized the potential practice effects.

Brown's (2001) oral proficiency scoring categories were slightly adapted to evaluate 5 specific components of speaking performance: grammar, vocabulary, fluency, pronunciation, and task fulfillment. These aspects measured participants' ability to use accurate linguistic structures and vocabulary, produce smooth speech with appropriate pronunciation and intonation, and fulfill the communicative demands of each task with relevant and organized content. Each participant completed two speaking tasks—one picture description and one opinion expression—and for each task, performance in the five categories was rated on a 1 to 5 scale. Thus, for each category, the total possible score ranged from 2 to 10, yielding an overall speaking score range of 10 to 50. This scoring method allowed for a comprehensive assessment of each participant's speaking proficiency across both task types. The results were evaluated by two trained raters, both native English speakers teaching university-level EFL courses in Korea. Interrater reliability, assessed using Cohen's Kappa coefficient, high, with values of 0.86 for the pretest and 0.94 for the posttest, indicating strong consistency between the two raters.

3.3.3 Foreign Language Speaking Anxiety Scale

To assess participants' levels of FLSA, this study employed the FLSAS created by Öztürk and Gürbüz (2014). The questionnaire included 18 items selected from the original 33 items of the FLCAS developed by Horwits et al. (1986). Each item presented a hypothetical classroom speaking situation (e.g., "I tremble when I know that I am going to be called on in English classes") and asked participants to rate their level of anxiety on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The scores ranged from 18 to 90. According to Öztürk and Gürbüz (2014), scores exceeding 72 indicates high speaking anxiety, scores between 54 and 72 indicate

moderate anxiety, and scores below 54 reflect low speaking anxiety. The FLSA demonstrated high internal consistency in the current study, with Cronbach's alpha, with reliability coefficients of 0.89 for the pretest and 0.91 for the posttest.

3.3.4 WTC scale

Participants' WTC was assessed using a scale developed by Peng and Woodrow (2010), consisting of 10 self-report items. The scale included items representing typical classroom communication tasks, such as asking classmates or teachers clarification questions, giving short speeches, or participating in role-plays in English. Participants rated each item on a 5-point Likert scale: 1 (*Definitely not willing*), 2 (*Probably not willing*), 3 (*not sure*), 4 (*Probably willing*), and 5 (*Definitely willing*). Total scores on the WTC scale ranged from 10 to 50. Internal consistency was acceptable, with Cronbach's alpha coefficients of 0.71 for the pretest and 0.79 for the posttest.

3.4 Data Collection and Analysis

This study was conducted over a period of 13 weeks during regular class hours, with one of the two weekly 50-minute sessions allocated to the research activities. Prior to the commencement of the study, the researcher explained the general objectives to the students, and participation was limited to those who voluntarily provided informed written consent. Data were collected through pre- and post-questionnaires assessing FLSA and WTC, as well as pre- and posttests of speaking proficiency.

All statistical analyses were conducted using SPSS version 25.0. First, independent samples *t*-tests were performed to determine whether any significant differences existed between the experimental and comparison groups in participants' initial levels of anxiety, WTC, and speaking proficiency. Subsequently, a one-way analysis of covariance (ANCOVA) was conducted to examine differences in posttest scores between the two groups, while controlling for pretest scores. In the analysis, the pretest served as the covariate, the posttest scores as the dependent variable, and group membership (experimental vs. comparison) as the independent variable.

To further investigate whether there were statistically significant differences between the experimental and comparison groups across multiple dimensions of speaking performance, a one-way multivariate analysis of variance (MANOVA) was conducted on the posttest speaking scores. The five dependent variables were grammar, vocabulary, fluency, pronunciation, and task fulfillment, with group membership serving as the independent variable.

4. Results

4.1 Participants' FLSA

To ensure group equivalence at the outset of the study, an independent samples *t*-test was conducted to compare participants' pre-questionnaire scores. The assumptions for the *t*-test were met: both groups demonstrated normality (Shapiro-Wilk: $p = .297$ for experimental; $p = .659$ for comparison group), and Levene's test confirmed homogeneity of variances, $F(1, 46) = 0.34$, $p = .564$. No assumption violations were observed. The *t*-test revealed no statistically significant difference between the experimental group ($M = 71.79$, $SD = 7.35$) and the comparison

group (M = 72.63, SD = 8.20), $t(46) = -0.37$, $p = .713$, partial $\eta^2 = .003$, 95% CI [0.00, 0.11], indicating that the two groups were comparable prior to the intervention.

According to the FLSA scoring criteria (Öztürk and Gürbüz 2014), the mean scores of both groups reflected high levels of speaking anxiety at pretest. However, as displayed in Table 1, FSLA levels decreased at the posttest, with the experimental group exhibiting a greater reduction in mean posttest scores (M = 62.21, SD = 8.36) compared to the comparison group (M = 69.67, SD = 8.38). To determine whether these differences were statistically significant, a one-way ANCOVA was conducted.

Prior to conducting the ANCOVA, key assumptions were tested and confirmed. The assumption of homogeneity of regression slopes was met, as the interaction between pretest scores and group was not significant, $F(1, 44) = 0.28$, $p = .596$. The residuals were normally distributed ($p = .960$), no outliers were detected based on standardized residuals, and the scatterplot of residuals versus predicted values indicated no violation of linearity.

A one-way ANCOVA was then conducted to compare FLSA scores between the experimental and comparison groups, controlling for pretest scores. As shown in Table 2, after adjusting for pretest differences, there was a significant effect of group on posttest scores, $F(1, 45) = 56.59$, $p < .001$, partial $\eta^2 = .56$, 95% CI [0.39, 0.66]. Pretest scores were also a significant covariate, $F(1, 45) = 213.88$, $p < .001$, partial $\eta^2 = .83$, 95% CI [0.73, 0.88]. These results suggest that, after accounting for initial anxiety levels, participants who interacted with ChatGPT experienced significantly greater reductions in speaking anxiety compared to those who engaged in peer conversations. This indicates that ChatGPT-based interaction may be more effective in lowering EFL speaking anxiety than peer-based interaction.

Table 2. Descriptive Statistics for FLSA and WTC Scores by Group and Time Point

Group	FLSA		WTC	
	Pretest	Posttest	Pretest	Posttest
	M (SD)	M (SD)	M (SD)	M (SD)
Experimental	71.79 (7.35)	62.21 (8.36)	18.92 (5.32)	30.04 (6.89)
Comparison	72.63 (8.20)	69.67 (8.38)	19.38 (5.14)	23.62 (7.27)

Table 3. ANCOVA Result for FLSA

Source	SS	df	MS	F	p
Group	672.32	1	672.32	56.59	< .001
Pretest	2541.29	1	2541.29	213.88	< .001
Error	534.67	45	11.88		

4.2 Participants' WTC

Table 1 shows that WTC scores increased in both groups; however, the experimental group exhibited a markedly greater gain, suggesting a stronger effect of the intervention. Statistical analyses were conducted to examine group differences in WTC.

First, assumptions for the independent samples t-test were tested and satisfied: Both groups demonstrated normality (Shapiro-Wilk: $p = .407$ for experimental; $p = .571$ for comparison), and Levene's test confirmed homogeneity of variances, $F(1, 46) = 0.60$, $p = .442$. No significant pre-intervention difference was observed between the experimental (M = 18.92, SD = 5.32) and comparison (M = 19.38, SD = 5.14) groups, $t(46) = -1.08$, $p = .284$, partial $\eta^2 = .03$, 95% CI [0.00, 0.15].

To assess whether post-intervention differences were statistically significant, a one-way ANCOVA was conducted using pretest scores as a covariate. Prior to the analysis, key assumptions were examined and met. The assumption of homogeneity of regression slopes was satisfied, as the interaction between pretest scores and group was not significant, $F(1, 44) = 0.10$, $p = .748$. Residuals were normally distributed ($p = .960$), no outliers were identified based on standardized residuals, and a scatterplot of residuals versus predicted values indicated linearity.

As shown in Table 3, the ANCOVA revealed a statistically significant effect of group on posttest WTC scores after controlling for pretest scores, $F(1, 45) = 8.27$, $p = .006$, partial $\eta^2 = .17$, 95% CI [0.01, 0.37]. Participants in the experimental group demonstrated significantly greater willingness to communicate than those in the comparison group. This suggests that interaction through ChatGPT was more effective in enhancing learners' communicative confidence and engagement in English than peer interaction.

Table 4. ANCOVA Result for WTC

Source	SS	df	MS	F	*p*
Group	428.18	1	428.18	8.27	0.006
Pretest	23.23	1	23.23	0.45	0.506
Error	2329.73	45	51.77		

4.3 Participants' Speaking Proficiency

An independent samples *t*-test was conducted to compare pretest speaking total scores between the experimental and comparison groups. Assumption checks confirmed that the data were normally distributed for both groups (Shapiro-Wilk: $p = .122$ for the experimental group, $p = .453$ for the comparison group), and that variances were equal (Levene's test: $p = .704$). The results revealed no significant difference between the experimental ($M = 22.64$, $SD = 4.87$) and comparison groups ($M = 23.04$, $SD = 5.45$), $t(46) = -0.27$, $p = .786$, partial $\eta^2 = .00$, 95% CI [0.00, 0.10], indicating comparability at baseline.

Following the intervention, the total speaking test scores of the two groups diverged at posttest, as shown in Table 4. While both groups demonstrated improvement, the experimental group achieved significantly higher scores ($M = 32.04$, $SD = 3.95$) compared to the comparison group ($M = 27.08$, $SD = 5.54$). To confirm whether these differences were statistically meaningful after accounting for initial proficiency, a one-way ANCOVA was conducted with posttest scores as the dependent variable and pretest scores as the covariate. Assumptions of normality, homogeneity of variances, linearity, and homogeneity of regression slopes were checked and satisfied. The ANCOVA revealed a significant group effect, $F(1, 45) = 34.05$, $p < .001$, partial $\eta^2 = .43$, 95% CI [0.24, 0.58], indicating that the experimental group significantly outperformed the comparison group on the posttest after controlling for pretest scores. Pretest scores were also a significant covariate, $F(1, 45) = 76.02$, $p < .001$, partial $\eta^2 = .63$, 95% CI [0.46, 0.73].

To further examine which specific areas of speaking proficiency showed greater improvement, a one-way MANOVA was conducted on the posttest speaking scores across five skill categories: grammar, vocabulary, fluency, pronunciation, and task fulfillment. Prior to conducting the MANOVA, the assumptions of multivariate normality, linearity, homogeneity of variance-covariance matrices, and absence of multicollinearity were checked and found to be satisfied. There was a statistically significant multivariate effect of group, *Wilks'* $\Lambda = .298$, $F(5, 42) = 9.79$, $p < .001$, partial $\eta^2 = .54$, 95% CI [0.33, 0.68]. Follow-up univariate ANOVAs revealed that the experimental group performed significantly better than the comparison group in grammar, $F(1, 46) = 20.83$, $p < .001$, partial $\eta^2 = .31$, 95% CI [0.13, 0.49]; vocabulary, $F(1, 46) = 12.50$, $p < .001$, partial $\eta^2 = .21$, 95% CI [0.05,

0.38]; and task fulfillment, $F(1, 46) = 9.36$, $p = .004$, partial $\eta^2 = .17$, 95% CI [0.02, 0.35]. No significant group differences were found for fluency or pronunciation at posttest, $t(46) = 0.86$, $p = .392$, partial $\eta^2 = .02$, 95% CI [0.03, 0.16]; and $t(46) = 0.42$, $p = .676$, partial $\eta^2 = .00$, 95% CI [0.05, 0.12], respectively.

Table 5. Descriptive Statistics for Speaking Scores by Group and Time Point

Category	Group	Pretest *M*	Pretest *SD*	Posttest *M*	Posttest *SD*
Grammar	Experimental	4.84	1.37	7.60	1.53
	Comparison	4.36	1.25	5.48	1.53
Vocabulary	Experimental	4.32	1.38	6.80	1.19
	Comparison	4.64	1.22	5.52	1.16
Fluency	Experimental	4.20	1.15	5.40	0.82
	Comparison	4.60	1.32	5.12	1.36
Pronunciation	Experimental	4.40	1.22	5.72	1.24
	Comparison	4.80	1.53	5.56	1.39
Task fulfillment	Experimental	4.88	1.48	6.52	1.08
	Comparison	4.84	1.25	5.40	1.26
Total score	Experimental	22.64	4.87	32.04	3.95
	Comparison	23.04	5.45	27.08	5.54

Note. The maximum possible total score was 50, with a maximum of 10 points per category.

Table 6. ANCOVA Result for Speaking Posttest Scores

Source	SS	df	MS	F	*p*
Group	999.37	1	999.37	34.05	< .001
Pretest	2228.88	1	2228.88	76.02	< .001
Error	1321.84	45	29.37		

5. Discussion

The goal of this study was to determine whether conversing with a chatbot positively impacts EFL learners' affective outcomes and English oral proficiency. The findings yield important insights into the potential benefits of using ChatGPT as a pedagogical tool.

First, ChatGPT-based interaction was found to be effective in reducing learners' speaking anxiety. In the pre-questionnaire, participants exhibited extremely high levels of FLSA, consistent with previous observations of Korean EFL learners (Kim 2018). Language learners often experience anxiety when required to demonstrate their linguistic abilities publicly, and Korean EFL learners, in particular, report higher anxiety levels during speaking tasks compared to reading or writing activities (Kim 2018). The positive outcomes observed in this study align with findings from previous research (e.g., Dizon 2020, Tai and Chen 2024), suggesting that AI chatbot interactions can help mitigate learner anxiety. One possible explanation is the reduced social pressure in chatbot-mediated tasks: the artificial nature of chatbots provides a low-stakes environment that promotes emotional safety, allowing learners to make mistakes without fear of judgment or loss of face (Tai and Chen 2024). This supportive context fosters greater confidence and comfort in using the target language. Lower anxiety levels, in turn, may reduce

learners' affective filters (Krashen 1982), thereby facilitating more effective language acquisition. Consequently, interacting with ChatGPT may help learners feel more at ease during speaking practice, contributing to a reduction in FLSA.

In addition to reducing speaking anxiety, the study found that engaging in English speaking practice with ChatGPT significantly enhanced Korean EFL learners' WTC. This result is consistent with prior research (e.g., Kim and Su 2024, Tai and Chen 2024), which reported that AI chatbots foster greater communicative confidence and reduce anxiety, thereby promoting learners' readiness to use the L2. These findings can be interpreted within the framework of MacIntyre et al.'s (1998) heuristic model of WTC, which emphasizes the role of affective factors—such as anxiety and self-confidence—as proximal antecedents to L2 communication. The inverse relationship between anxiety and WTC observed in the present study supports this model: learners in the experimental group reported reduced anxiety and enhanced confidence, making them more willing to engage in English communication. The nonjudgmental and supportive environment provided by ChatGPT interactions appear to have contributed to this positive shift.

Moreover, AI chatbot-based interactions significantly enhanced speaking proficiency among Korean EFL learners. The experimental group demonstrated greater improvement in overall speaking skills compared to the comparison group, with particular gains in grammar, vocabulary, and task fulfillment. These findings are consistent with prior studies (e.g., Dizon 2020, Tai and Chen 2024), which highlighted the effectiveness of chatbot-mediated practice in developing L2 oral proficiency.

One plausible explanation for the observed gains lies in ChatGPT's ability to process non-native input and provide linguistically rich, adaptive, and comprehensive responses tailored to learners' proficiency levels. This dynamic interaction may have provided participants with meaningful input, supporting vocabulary acquisition, grammar development, and improved communicative competence. Learners were often able to immediately incorporate new vocabulary, expressions or grammatical structures encountered during chatbot interactions. In contrast, students practicing with peers of similar proficiency levels may have received more limited or less accurate linguistic input.

Additionally, the communicative tasks employed—such as information gap activities, opinion exchanges, and decision-making tasks—promoted goal-oriented interaction, requiring learners to engage in purposeful language use. The combination of structured task design and AI-mediated conversations may have created optimal conditions for meaningful output and learner engagement. These findings support the Lee et al.'s (2020) argument that task-specific chatbot interactions are particularly effective in fostering L2 speaking development.

However, it is noteworthy that the study did not observe significant gains in fluency and pronunciation. While this may partially reflect the limited scope of the speaking assessment, intervention-specific factors could also help explain these findings. ChatGPT interactions are primarily driven by brief spoken input, which may not consistently encourage extended speech production necessary to develop fluency. Additionally, unlike human interlocutors or specialized pronunciation tools, ChatGPT does not automatically provide targeted corrective feedback on phonological accuracy or real-time pacing, which are essential for pronunciation and fluency improvement. As a result, learners may have had fewer opportunities to refine these aspects during the intervention period. Future studies could explore the integration of pronunciation-focused prompts or supplementary tools alongside chatbot interaction to address these areas more directly.

The findings of this study offer several pedagogical implications for language educators. First, integrating AI chatbots like ChatGPT into speaking tasks may provide learners with a psychologically safe space to practice oral communication without fear of negative evaluation. Such low-stress environments are especially beneficial in

high-anxiety EFL contexts, such as Korean, where speaking in front of peers often provokes considerable apprehension.

Second, regular chatbot-based interaction may help build learners' communicative confidence over time, encouraging greater willingness to participate in classroom discourse and real-world communication scenarios. Educators can leverage AI tools as supplemental, scalable, and personalized speaking practice, supporting differentiated instruction and promoting learner autonomy. This approach may help address individual differences in affective factors—such as anxiety and WTC—that influence language performance.

Third, integrating AI chatbots like ChatGPT offer linguistically rich and responsive input, essential for the development of speaking proficiency. In a nonjudgmental, low-pressure environment, learners may feel more comfortable experimenting with new vocabulary, grammatical structures, and complex ideas.

Finally, combining chatbot interaction with structured communicative tasks appears particularly beneficial. Tasks that require authentic interaction, problem-solving or collaborative decision-making—can be enhanced through chatbot mediation, ensuring consistent modeling of appropriate linguistic forms while maintaining learner engagement.

6. Conclusion

While this study provides valuable insights into the effects of chatbot-based interaction on EFL learners' speaking anxiety, WTC, and speaking proficiency, several limitations should be noted. First, the absence of qualitative data limited the ability to capture learners' subjective experiences with ChatGPT. Since the study examined attitudinal variables such as FLSA and WTC, qualitative interviews could have offered a more nuanced understanding. They might have revealed how learners engaged with the chatbot environment and how these experiences shaped their communicative behaviors. Second, the scope of the WTC measure was restricted. The ten-item questionnaire may have lacked the sensitivity to reflect participants' initial communicative tendencies. This limitation may explain why pretest WTC was not a significant covariate in the ANCOVA analysis. In contrast, FLSA showed predictive validity. These findings highlight the need for more comprehensive or alternative WTC instruments in future research. Third, although positive effects were observed in grammar, vocabulary, and task fulfillment, no significant gains emerged in fluency and pronunciation. This result should be interpreted with caution because the speaking test included only a limited number of items. Broader and more detailed proficiency assessments are needed to determine whether these trends can be replicated and generalized.

Beyond these measurement-related issues, the characteristics of the sample constrain the generalizability of the findings. Both groups were composed of science and engineering majors. Their technological background may have reduced resistance to chatbot use compared with learners from other fields. In addition, the participants' English proficiency was restricted to the CEFR B1 level (TOEIC 550-780). It remains uncertain whether the results extend to learners at higher or lower proficiency levels. Finally, the comparison group interacted only with peers of similar proficiency. This raises the question of whether chatbot-based interaction would remain more effective than peer interaction when learners engage with more advanced non-native or native speakers. Future studies should therefore diversify participant profiles, include a broader proficiency range, and examine interaction with interlocutors of varying linguistic expertise.

Despite these limitations, this study is among the few that highlight the promising potential of integrating AI chatbot-based interaction into EFL speaking instruction. The observed gains in WTC and reductions in speaking anxiety suggest that AI chatbots such as ChatGPT can serve as effective tools for enhancing learners'

communicative confidence and engagement. Particularly in high-anxiety EFL contexts like Korea, such tools may offer a valuable supplement to traditional instruction by providing learners with nonjudgmental, low-pressure opportunities to practice spontaneous speech. Overall, the findings underscore the potential of AI-mediated interaction to foster meaningful, low-anxiety speaking opportunities, ultimately promoting greater learner engagement and supporting the development of oral proficiency.

References

Bashori, M., R. van Hout, H. Strik and C. Cucchiarini. 2021. Effects of ASR-based websites on EFL learners' vocabulary, speaking anxiety, and language enjoyment. *System* 99, 102496.

Bashori, M., R. van Hout, H. Strik and C. Cucchiarini. 2022. Web-based language learning and speaking anxiety. *Computer Assisted Language Learning* 35(6), 1058-1089.

Brown, H. D. 2001. *Teaching by Principles: An Interactive Approach to Language Pedagogy*. Pearson Education.

Chien, Y., T. C. Wu, Lai and Y. Huang. 2022. Investigation of the influence of artificial intelligence markup language-based LINE ChatBot in contextual English learning. *Frontiers in Psychology* 13, Article 785752.

Chiu, T. K. F., B. L. Moorhouse, C. S. Chai and M. Ismailov. 2024. Teacher support and student motivation to learn with artificial intelligence (AI) chatbot. Interactive *Learning Environments* 32(7), 3240-3256.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Dizon, G. 2020. Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology* 24(1), 16-26.

Dörnyei, Z. 2005. *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Lawrence Erlbaum Associates.

Ellis, R. 2018. *Reflections on Task-Based Language Teaching*. Multilingual Matters.

Hanafiah, W., M. Aswad, H. Sahib, A. Yassi and M. Mousavi. 2022. The impact of CALL on vocabulary learning, speaking skill, and foreign language speaking anxiety: The case study of Indonesian EFL learners. *Education Research International 2022*, 1-13.

Hew, K. F., W. Huang, J. Du and C. Jia. 2022. Using chatbots to support student goal setting and social presence in fully online activities: Learner engagement and perceptions. *Journal of Computing in Higher Education* 35*,* 40-68.

Horwitz, E. K., M. B. Horwitz and J. Cope. 1986. Foreign language classroom anxiety. *Modern Language Journal*, 70, 125-132.

Jeon, J. 2023. Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning* 36(7), 1338-1364.

Jeon, J. 2024. Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning* 37(1), 1-26.

Ji, H., I. Han and Y. Ko. 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.

Kim, J. O. 2018. Ongoing speaking anxiety of Korean EFL learners: Case study of a TOEIC intensive program. *The Journal of Asia TEFL* 15(1), 17-31.

Kim, A. and Y. Su. 2024. How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. *System* 122, 103256.

Kim, H., H. Yang and D. Shin. 2022. Design principles and architecture of a second language learning chatbot. *Language Learning & Technology* 26(1), 1-18.

Krashen, S. D. 1982. *Principles and Practice in Second Language Acquisition*. Pergamon Press.

Lee, J. H., H. Yang and D. Shin. 2020. Chatbots – Technology for the language teacher. *ELT Journal* 74(3), 338-344.

Lee, J. H., D. Shin and Y. Hwang. 2024. Investigating the capabilities of large language model-based task-oriented dialogue chatbots from a learner's perspective. *System* 127, 103538.

Liu, G. L., R. Darvin and C. Ma, 2024. Exploring AI-mediated informal digital learning of English (AI-IDLE): A mixed-method investigation of Chinese EFL learners' AI adoption and experiences. *Computer Assisted Language Learning*. Available online at https://doi.org/10.1080/09588221.2024.2310288

MacIntyre, P. D. 2017. An overview of language anxiety research and trends in its development. In C. Gkonou, M. Daubney and J.-M. Dewaele, eds., *New Insights into Language Anxiety: Theory, Research and Educational Implications*, 11-30. Multilingual Matters.

MacIntyre, P. D., Z. Dörnyei, R. Clément and K. A. Noels. 1998. Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal* 82(4), 545-562.

Mukminin, A., M. Masbirorotni, N. Noprival, S. Sutarno, N. Arif and M. Maimunah. 2015. EFL speaking anxiety among senior high school students and policy recommendations. *Journal of Education and Learning* 9(3), 217-225.

Öztürk, G. and N. Gürbüz. 2014. Speaking anxiety among Turkish EFL learners: The case at a state university. *Journal of Language and Linguistic Studies* 10(1), 1-17.

Peng, J. E. 2022. Willingness to communicate. In S. Li, P. Hiver and M. Papi, eds., *The Routledge Handbook of Second Language Acquisition and Individual Differences*, 159-171. Routledge.

Peng, J. E. 2019. The roles of multimodal pedagogic effects and classroom environment in willingness to communicate in English. *System* 82, 161-173.

Peng, J. E. and L. Woodrow. 2010. Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language learning*, *60*(4), 834-876.

Reinders, H. and S. Wattana. 2014. Can I say something? The effects of digital game play on willingness to communicate. *Language Learning & Technology* 18(2), 101-123.

Shadiev, R., X. Wang, Y. Halubitskaya, and Y. M. Huang. 2022. Enhancing foreign language learning outcomes and mitigating cultural attributes inherent in Asian culture in a mobile-assisted language learning environment. *Sustainability* 14, Article 8428.

Tai, T.Y. and H.H.J. Chen. 2024. The impact of intelligent personal assistants on adolescent EFL learners' speaking proficiency. *Computer Assisted Language Learning* 37(3), 1-28.

Willis, J. 1996. A flexible framework for task-based learning. *Challenge and Change in Language Teaching* 52(1), 52-62.

Yang, C. T. Y., Lai, S. L and H.J. Chen. 2024. The impact of intelligent personal assistants on learners' autonomous learning of second language listening and speaking. *Interactive Learning Environments* 32(5), 2175-2195.

Yang, H., H. Kim, J. H. Lee and D. Shin. 2022. Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL* 34(3), 327-343.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Appendix A. Speaking Test

1. **Task**: Describe the picture in as much detail as you can. *(Image not shown here, so here's a verbal alternative)*
   Sample Description Prompt:
   You see a picture of a group of people having a meeting in an office. Some are sitting around a table, and one person is presenting something using a projector.
2. **Task**: Express your opinion and support it with reasons or examples.
   Question: What is your favorite place to visit on weekends? Describe it and explain why it is your favorite place to go.