Tracing Interactional Change in Japanese EFL Learners' Interactions with a Teacher-Designed AI Voice Chatbot: A Pilot Study

Ken'ichi Yamazaki[1*]

[1] Akita Prefectural University, Japan
[*]Corresponding author's email: yamazaki@akita-pu.ac.jp
[*] https://orcid.org/0009-0005-3211-2765
https://doi.org/10.54855/ijaile.26331

® Copyright (c) 2026 Ken'ichi Yamazaki

Received: 15/05/2026 Revision: 03/09/2026 Accepted: 20/09/2026 Online: 09/11/2026

ABSTRACT

Although AI-mediated conversation practice is increasingly common in EFL teaching, qualitative analysis of learner–AI interaction has been insufficient. This exploratory pilot study examined Japanese EFL learners' interactional change. Ten university students completed three five-minute sessions with a teacher-designed chatbot on Google AI Studio. Logs were analyzed quantitatively and qualitatively. Effect sizes suggested a two-phase pattern of observed change: large effects (d = 0.88–1.13) from Sessions 1 to 2 marked an initial adjustment phase, while medium effects (d = 0.59–0.66) from Sessions 2 to 3 indicated slowed quantitative growth alongside continued qualitative elaboration, particularly in topic development. Qualitative analysis identified both constraints (vocabulary gaps, abstract questions) and affordances (explanatory expansions, analytical reflections, self-directed topic drift). Comparison with peer-to-peer data highlighted distinctive features: longer turns, near absence of backchannels, and more frequent language-related episodes. These preliminary findings offer initial guidance for teachers developing AI conversation tools using accessible platforms.

Keywords: AI conversation partner, conversation practice, teacher-designed chatbot, interactional features, EFL learners

Introduction

Conventional English conversation practice has involved interactions with native speakers or among non-native speakers. However, rapid advances in generative artificial intelligence (AI) technology have made AI-mediated conversation practice prominent in English language education. Research on performance outcomes has reported significant gains, particularly in speaking and writing skills. Another strand of research has focused on learner affect. Studies have shown that AI chatbots significantly reduce learners' anxiety about foreign-language conversations.

CITATION | Yamazaki, K. (2026). Tracing Interactional Change in Japanese EFL Learners' Interactions with a Teacher-Designed AI Voice Chatbot: A Pilot Study. *International Journal of AI in Language Education, 3*(3), 1-23. DOI: https://doi.org/10.54855/ijaile.26331

Although quantitative research on AI conversations is now widespread, only a few studies have examined the qualitative characteristics of second-language (L2) conversations. To better leverage the pedagogical potential of AI, users' utterances and responses in AI-mediated conversational tasks should undergo qualitative analysis, and areas for improvement need to be identified.

This study aims to quantitatively and qualitatively examine changes in the interactional characteristics of L2 conversations among Japanese learners of English as a foreign language (EFL) when they engage in repeated conversation practice with a teacher-designed AI chatbot. It focuses on patterns of change in discourse features and other interactional resources, especially topic development. Based on its findings, this study describes a two-phase pattern of observed change for L2 conversation practice with AI: (1) an initial "adjustment phase," marked by a sharp expansion of output quantity, followed by (2) an "elaboration phase," characterized by quantitative slowdown but continued qualitative elaboration, especially in topic development and language-related episodes (LREs).

Literature Review

Theoretical Foundations: Interaction and Second-Language Acquisition

The theoretical foundation for examining AI-mediated conversation practice lies in interaction-based approaches to L2 acquisition. According to Long's (1996) Interaction Hypothesis, negotiation of meaning during communicative interaction facilitates L2 development by drawing learners' attention to form-meaning mappings and providing opportunities for modified output (see also Gass & Varonis, 1994). Central to this framework is the claim that conversational adjustments—such as clarification requests, confirmation checks, and comprehension checks—promote the detection of linguistic gaps and allow learners to test hypotheses about target language structures (Ellis, 1994). Varonis and Gass (1985) extended this framework by demonstrating that even conversations between non-native speakers can generate productive negotiation sequences when interlocutors experience communication breakdown and engage in collaborative repair. Their model identifies trigger points (e.g., unfamiliar vocabulary, unclear pronunciation) that initiate negotiation routines, leading to modified input and pushed output.

Swain and Lapkin (1998) introduced the concept of LREs—instances in which learners discuss language while completing a task—as evidence of collaborative dialogue, building on Swain's earlier work on comprehensible output (Swain, 1985; Swain & Lapkin, 1995). Although LREs were not conceived as fluency measures, their occurrence suggests that learners engage in reciprocal interaction that involves negotiating meaning and attending to linguistic form. Swain's output hypothesis posits that producing language—particularly through negotiation and self-repair—drives learners from semantic to syntactic processing, thus consolidating interlanguage development (Swain, 2000). Supporting similar claims, interaction studies have shown that task-induced linguistic demands elicit pushed, more "comprehensible" output from learners (Pica et al., 1989). Tarone (1981) introduced the notion of communication strategies

(avoidance, paraphrase, transfer, appeal, mime) for managing communicative challenges when linguistic resources are insufficient.

Affective Factors in L2 Speaking

Alongside cognitive and interactional factors, affective variables play a critical role in L2 speaking performance. Horwitz et al. (1986) identified foreign-language anxiety (FLA) as a distinct construct characterized by communication apprehension, test anxiety, and fear of negative evaluation, and learners who answered the Foreign Language Classroom Anxiety Scale reported experiencing debilitating anxiety specifically in L2 contexts.

More recently, the notions of FLA and willingness-to-communicate (WTC) have been adopted in research on AI–human interaction. Zhang et al. (2024) reported that sustained interaction with a generative AI chatbot was associated with increased L2 learners' WTC, while Nadeem et al. (2025) demonstrated that AI-based speaking activities significantly reduced speaking-anxiety scores, with learners citing "not being judged by others" as a key source of comfort.

Technology-Mediated Language Learning: Evolution and Design Principles

Chapelle (2001, 2009) explained the relationship between second-language acquisition (SLA) theory and computer-assisted language learning (CALL) and proposed the following criteria for evaluating CALL tasks: language learning potential, learner fit, meaning focus, authenticity, positive impact, practicality, and feedback.

Levy and Stockwell (2006) distinguished teacher-designed from commercial CALL applications, noting that the former offer greater flexibility in aligning content with curriculum and learner proficiency but require technical expertise and time investment. Therefore, the current study employs a teacher-developed AI chatbot created with accessible tools (Google AI Studio), thereby providing practitioners with a model for designing their own conversational applications without extensive programming knowledge.

AI Chatbots in Language Education: Recent Developments

Early studies such as Fryer and Carpenter (2006) demonstrated that text-based chatbots could sustain extended learner engagement but are less suitable for advanced learners because of their linguistic limitations. Bibauw et al.'s (2019) synthesis of dialogue-based CALL research highlighted three key affordances—individualized practice opportunities, immediate feedback, and reduced anxiety—alongside persistent limitations, including restricted conversational flexibility, limited emotional sensitivity, and difficulty handling unexpected learner input. They suggested conducting more fine-grained qualitative analyses of learner–chatbot interaction to identify specific breakdown points and inform design improvements.

With the advent of large language models such as ChatGPT, studies have reported more promising results. Kohnke et al. (2023), who explored ChatGPT's potential for language teaching and learning, highlighted its ability to generate contextually appropriate responses, provide explanations, and engage in extended dialogue. Many recent studies have examined the learning outcomes of conversations with AI-based chatbots. Sok and Shin (2025) used ChatGPT with Korean university EFL learners and found that groups that used AI significantly outperformed a control group on post-test speaking scores while also reporting reduced

psychological barriers and positive evaluations of the tasks. Qualitative analysis showed that when learners struggled to convey their intentions, they drew on various interactive strategies, including attempts to negotiate meaning, paraphrase, and self-correct.

Similarly, Alenezi and Alenezi (2025) reported that Saudi secondary school students who underwent a three-week practice with an AI chatbot significantly outperformed a control group from a regular class on speaking tests. Their chatbot, based on OpenAI's GPT-3.5-turbo model, aligned with curriculum content and learner proficiency, showing how custom-built systems can serve as pedagogical partners that enhance learners' confidence and motivation. In a related study, Nguyen and Le (2025) examined 77 Vietnamese university students' three-week interactions with Call Annie, an AI conversational agent, and reported significant increases in learners' willingness to communicate and confidence, alongside persistent concerns about the naturalness of AI responses. Similarly, Nguyen (2025) reviewed the pedagogical affordances of TALKPAL.AI for adult L2 learners, highlighting its potential to support extended speaking practice while noting the need for empirical validation. Jeon's (2024) study on young EFL learners' experiences and perspectives on AI chatbots revealed that learners appreciated the opportunity for self-paced practice. Wang et al.'s (2024) discussion made clear the positive effects of AI chatbots with avatars on humans' tendency to engage in relationship-building behaviors with chatbots.

Notably, however, these findings are not uniformly positive. In some studies, AI-mediated practice is not necessarily superior to interaction with human interlocutors, with some focusing on fluency indicators (e.g., word and turn counts) and task completion and performing relatively few qualitative analyses of how learners negotiate meaning, develop topics, or self-repair. Addressing these limitations requires examining the micro-level organization of learner–chatbot interaction rather than only global outcomes.

Qualitative Analysis of L2 Interaction

Scholars have developed qualitative approaches to analyzing L2 interaction alongside quantitative methods. Markee (2000) and Seedhouse (2004) pioneered the application of conversation analysis (CA) to L2 classroom interaction, showing that interactional practices such as turn-taking, repair, and sequence organization shape language-learning opportunities. Matiienko-Silnytska et al. (2025) applied CA to interactions between Ukrainian university students and an AI chatbot, documenting self-repair, discourse-marker use, and discourse organization, including features related to discursive coherence. According to Yin and Satar's (2020) content analysis of learner–chatbot interaction logs, which aimed to identify the stages and trigger types of meaning negotiation, learners engaged in meaning-negotiation behaviors, and chatbot interaction could be recognized as dialogic negotiation rather than a sequence of mechanical responses. However, their analysis remained focused primarily on counting discrete negotiation routines rather than examining the emergent discourse-level features of learner–AI conversations.

Research Gap and the Present Study

Existing research on AI-mediated conversation practice presents an uneven picture. Quantitative and outcome-based studies have documented gains in learner affect and, in some

cases, measurable speaking outcomes, whereas the qualitative characterization of learner–AI interaction remains limited. Available qualitative analyses have generally focused on the frequency of discrete negotiation routines rather than tracing how interactional practices unfold and change across repeated conversational sessions. Recent reviews have also identified long-term effects and the temporal trajectory of AI-mediated speaking practice as an under-researched area (Nguyen, 2024), underscoring the value of tracing changes across repeated sessions rather than at a single time point.

Furthermore, analytical frameworks established in human–human interaction studies (Ellis, 1994; Swain & Lapkin, 1998; Tarone, 1981) have not been systematically applied to teacher-designed AI chatbots, and researchers have rarely examined the moment-by-moment organization of learner–chatbot interaction.

Finally, although commercial AI platforms have received substantial research attention, little is known about how teachers with limited programming expertise can design pedagogically tailored chatbots for L2 learners. These gaps motivate an examination of the micro-level organization of learners' interactional practices across short, repeated AI-mediated conversations with a teacher-designed chatbot.

The present study addresses these gaps through a mixed-methods analysis of three brief AI-mediated conversation sessions conducted by Japanese EFL learners. The teacher-designed chatbot used in this study was built on the freely available Google AI Studio platform and deployed on Vercel. The study offers practical directions for teachers intending to design their own conversational apps in the future.

Research Questions

To address these gaps, this study aims to answer two research questions:

1. Does repeated AI-mediated conversation practice lead to measurable changes in learners' observable interactional performance?
2. What interactional and discourse features characterize these changes across three brief AI conversation sessions?

Methods

Analytical Framework

This study employed a convergent parallel mixed-methods design (Creswell & Plano Clark, 2018), in which quantitative measures and qualitative discourse analyses were conducted concurrently on the same conversation logs and integrated in the Discussion to provide complementary evidence, allowing quantitative patterns to be interpreted in light of qualitative discourse features and vice versa. This study integrates three interaction-based frameworks that share a common commitment to interaction as a driver of L2 development but focus on different aspects: Long's (1996) Interaction Hypothesis foregrounds negotiation of meaning; Swain's (2000) output hypothesis emphasizes pushed output and language-related episodes; and Tarone's (1981) framework identifies communication strategies for managing linguistic gaps.

Together, these frameworks motivate the seven coding categories used in the present study: language-related episodes (LREs) and self-repairs derive from Swain's output-oriented work; topic development and reasoning derive from Long's Interaction Hypothesis and its concern with sustained meaningful exchange; questioning and formulaic sequences derive from Tarone's communication strategies; and examples represent a supplementary category capturing learners' elaborative moves. The researcher coded all transcriptions manually, in collaboration with an experienced native English teacher, using predetermined categories based on the theoretical frameworks above. The coding procedure involved four stages: (1) initial familiarization with the transcripts by the researcher; (2) provisional coding based on the predetermined categories derived from the theoretical frameworks (see Table 1); (3) independent re-coding of a 20% subsample by the second coder; and (4) resolution of any disagreements through discussion until consensus.

Participants

This exploratory pilot study involved 10 Japanese university students (7 female, 3 male; age range 19–21) who belonged to their English club, an English conversation practice group. This research used small samples to conduct an intensive analysis of learner–AI interaction through quantifiable measures. In fact, prior qualitative investigations into AI-assisted L2 conversation, such as Yin and Satar (2020) and Matiienko-Silnytska et al. (2025), used small groups for in-depth analysis of discourse-level data. They were broadly estimated to be at the Common European Framework of Reference for Languages (CEFR) A1–B1 level based on their self-reported Test of English for International Communication (TOEIC) scores and classroom performance. As club members, all participants had between one and three years of conversation practice at the time of the study. One student had participated in a 10-day study-abroad program. All 10 participants engaged in three AI-mediated conversation sessions over three weeks, with each session being held on a Monday. They visited a small classroom at a designated time and conversed with the app on a laptop provided by the instructor. The conversations mainly involved answering questions from the AI system about the local area where the university is located. Each session lasted five minutes.

Additionally, for reference, we collected a dataset from five pairs of Japanese EFL students (n = 10) from the same university population. These participants, who did not overlap with the main study cohort, engaged in a single five-minute peer conversation on the same topic as the AI-mediated task. We analyzed the reference dataset using the same coding framework as the main study. Because the two datasets differ in participants, number of sessions, and task structure, the peer data are used only as a descriptive reference rather than for systematic comparison. Observations are reported in the Discussion for interpretive purposes only.

The study received ethical approval from the university ethics committee, and all participants provided written informed consent before data collection. All participant names used in the excerpts are pseudonyms.

AI Chat Application and Task Design

Using Google AI Studio, the researcher developed Akita Review Chat, a conversation-practice chatbot, and deployed it as a web app on Vercel. The system ran on the Gemini 2.5 Flash model

(specifically, the gemini-2.5-flash-native-audio-preview-12-2025 model for real-time audio interaction). These platforms were selected because they require minimal programming expertise and are freely available to educators, highlighting the approach's replicability in typical classroom settings. In Google AI Studio, users can create applications by specifying instructions for the AI. For this study, the system was provided with the following core instructions: (1) ask the user questions based on uploaded documents; (2) when the user asks a question, respond using the AI's own knowledge; and (3) limit the AI's output to four sentences per turn, with each sentence containing at most 10 words. The researcher gave no instructions regarding the users' understanding, level, or attitude. To observe learners' spontaneous interactional behavior, no additional scaffolding was provided. The AI's speaking level can be set to four levels; this experiment used the easiest level (beginner). The instructor uploaded a document containing basic information about the prefecture where the university is located as the primary source text. Conversation time per session was fixed at five minutes.

Akita Review Chat has a built-in feature that sends feedback emails to students. After each five-minute conversation, users automatically received an email containing their estimated CEFR level, total number of words spoken, pronunciation score, and highlighted areas for correction so that they could immediately reflect on their performance. The participants were encouraged to read the feedback carefully and review their conversations. Teachers also receive feedback: after each practice session, written logs and audio files are saved to cloud storage and a spreadsheet containing learner errors, so teachers can provide timely guidance to students.

Data Collection and Transcription

The primary data source was the written logs collected for all sessions. Any unclear segment of the log was checked against the audio recordings or clarified directly with the involved student. We conducted analysis in collaboration with native-speaker instructors and applied necessary corrections and normalization. All learner utterances in the excerpts are transcribed verbatim, preserving grammatical errors, self-repairs, and pauses to maintain data authenticity.

The quantitative analysis used the following indicators: total word count, number of turns, number of words per turn, and the frequency of LREs, self-repairs, topic development, reasoning, questioning, formulaic sequences (FSs), and examples. We calculated mean values for each indicator across the 10 students. Because measurements were obtained at three time points, we examined differences across sessions using paired *t*-tests with a Bonferroni-adjusted significance level of $p < .0167$. Since the sample size in this study is small, the inferential findings are interpreted alongside effect sizes calculated as Cohen's d_z for paired samples (Lakens, 2013), computed as the mean of the difference scores divided by the standard deviation of the difference scores. Effect sizes are less influenced by sample size than *p*-values and provide a clearer indication of practical significance (Sullivan & Feinn, 2012). Results are therefore discussed as patterns rather than definitive findings.

Meanwhile, the qualitative analysis extracted characteristic exchanges and episodes of noticeable speech expansion or compressed talk, focusing on Sessions 2 and 3. Session 1 primarily reflected learners' initial adjustment to the AI chatbot format, with brief and often fragmented responses that offered limited material for discourse-level analysis; Sessions 2 and

3 produced more elaborated learner turns amenable to qualitative examination. The speech logs were read closely, with particular attention to "direct responses to questions" and "self-related expansions." From this, functional categories such as explanation, analysis, and personal reminiscence were identified. These functional categories were derived inductively during the qualitative reading, not predetermined. The researcher noted recurring functional patterns and subsequently confirmed them through a second reading by the collaborating native English teacher.

The reference peer-to-peer dataset was analyzed using the identical coding framework: total word count, turns, words per turn, and frequencies of LREs, self-repairs, topic development, reasoning, questioning, formulaic sequences, and examples. In addition, instances of L1 (Japanese, excluding the participants' names and other proper nouns) use were counted in this dataset. As in the analysis of the AI-mediated data, we qualitatively examined the same features in the peer conversation logs (negotiation, self-repair, topic development) to enable informal interpretive comparison.

Brief follow-up interviews (approximately 3–5 minutes each) were conducted with participants after selected sessions in which transcripts contained noteworthy interactional moments. These interviews clarified learners' perspectives on specific exchanges and are used as supplementary interpretive material for selected excerpts rather than as systematic data across all participants.

Coding Reliability

Table 1

Coding scheme for interactional features

Category	Operational Definition	Examples
LREs (Language-Related Episodes)	Learner explicitly asks the AI a language-related question (meaning, usage, or form).	e.g., "What do you mean by [word]?"
Self-repair	Learner corrects or reformulates their own utterance mid-turn or immediately after.	e.g., "I go, went there yesterday."
Topic development	Learner extends the current topic through additional information or elaboration; includes topic initiation shifts within the given topic.	e.g., adding context or examples to a topic beyond the AI's question.
Reasoning	Learner provides analytical, causal, or evaluative statements.	e.g., "because…", "so…" statements.
Questioning	Learner asks the AI a content question, treating the AI as an interlocutor with knowledge.	e.g., "What do you know about [topic]?"
Formulaic sequences (FSs)	Learner uses conventionalized multi-word expressions.	e.g., "I think," "for example," "you know."
Examples	Learner provides concrete instances to support a general statement.	e.g., naming specific items to illustrate a general claim.

To assess the reliability of the coding scheme, a random subsample comprising 20% of the data (2 participants × 3 sessions = 6 recordings, comprising 42 category-level observations) was

independently re-coded by the collaborating second coder (an experienced native English teacher) after a two-week interval. Agreement was quantified using the Intraclass Correlation Coefficient (ICC(2,1)) for count data (Koo & Li, 2016), together with exact percentage agreement and percentage agreement within a tolerance of ±1 count. Overall agreement was excellent: exact agreement was 83.3% (35/42 observations), agreement within ±1 count was 100%, and ICC(2,1) = .957, 95% CI [.92, .98], $p < .001$. All observed disagreements were of a single count and occurred exclusively in higher-frequency categories (formulaic sequences and reasoning), reflecting the boundary judgments inherent in counting rather than categorical assignment. This level of agreement is conventionally interpreted as excellent reliability (Koo & Li, 2016).

Quantitative Results

Overall Changes Across Three Sessions

Given the small sample size and short practice period, the quantitative findings should be interpreted as exploratory and pattern-oriented rather than confirmatory. Table 2 shows that learner output changed substantially across the three sessions. From Session 1 to Session 2, total word count and words per turn showed statistically significant changes with large effect sizes (d = 0.90–1.13, $p < .01$), while the decrease in turns showed a large effect size (d = 0.88) but did not reach significance at the Bonferroni-adjusted threshold (p = .019). From Sessions 2 to 3, the direction of change continued at a more modest rate with medium effect sizes (d = 0.59–0.66), but none of the three measures reached statistical significance at this stage (total words: p = .078; turns: p = .064; words per turn: p = .038, significant only at the unadjusted α = .05 level). Notably, the number of turns steadily dropped across sessions, while words per turn nearly doubled from Session 1 (M = 8.7) to Session 3 (M = 18.1), indicating a shift from brief, fragmented responses to longer, more elaborated utterances. This asymmetric pattern of change—significant initial gains in output and turn-taking efficiency, followed by a plateau without further statistical confirmation—offers preliminary, pattern-based evidence for a possible two-phase pattern of AI-mediated conversation practice.

Table 2

Descriptive statistics and effect sizes for learner output across three sessions (n = 10)

	Session 1 *M* (*SD*)	Session 2 *M* (*SD*)	Session 3 *M* (*SD*)	S1→S2 *d*	S2→S3 *d*	S1→S3 *d*
Total words	115.60 (47.8)	149.30 (56.4)	175.10 (56.9)	1.13**	0.64	1.51**
Turns	13.80 (1.7)	12.00 (2.9)	10.60 (2.1)	0.88	0.66	1.61**
Words/turn	8.70 (4.4)	14.00 (8.3)	18.10 (8.6)	0.90**	0.59	1.21**

Note. **$p < .0167$ (Bonferroni-adjusted). Effect sizes (Cohen's d): small = 0.2, medium = 0.5, large = 0.8 (Cohen, 1988).

To further explore this pattern of change, Table 3 shows the frequency of LREs and other interactional features in learner speech across the three sessions. In this study, LREs are instances in which learners explicitly ask the AI language-related questions. Topic development pertains to utterances that further develop, or attempt to develop, the current conversation topic. Learners in a broad free-conversation or task-based setting may also initiate new topics and shift away from the current theme, called topic initiation. In the present task, however, the AI's main role was to ask questions, and because the content was largely constrained by the uploaded document, topic initiation was not expected to occur frequently. Nevertheless, some students attempted to introduce different topics, whereas others tried developing the given topic after receiving a question from the AI. Therefore, for this study, we combined topic initiation and topic development and counted both as topic development. The mean frequency increased for all categories.

Table 3

Mean frequency of interactional features per learner across three sessions

	LREs	Self-repairs	Topic development	Reasoning	Questions	FSs	Examples
Session 1	0.2	1.2	0.4	0.1	0.1	1.9	0.0
	(0.42)	(1.03)	(0.70)	(0.32)	(0.32)	(1.29)	(0.00)
Session 2	0.6	2.8	1.8	0.6	0.2	3.2	0.4
	(0.84)	(1.03)	(0.92)	(0.70)	(0.42)	(1.87)	(0.70)
Session 3	1.5	3.1	3.0	1.3	1.1	3.8	0.7
	(1.18)	(1.60)	(0.82)	(1.42)	(1.37)	(1.32)	(0.95)

Note. Values indicate *M* (*SD*) per learner ($n = 10$).

Table 4

Pairwise *t*-test results for interactional features across sessions (*p*-values)

	LREs	Self-repairs	Topic development	Reasoning	Questions	FSs	Examples
S1 to S2	.269	< .001**	< .001**	.015**	.591	.103	.104
S2 to S3	.029	.591	.001**	.132	.095	.260	.520
S1 to S3	.009**	.001**	< .001**	.018	.023	.008**	.045

Note. **$p < .0167$ (Bonferroni-adjusted). Complete test statistics (including *t* values, degrees of freedom, and 95% confidence intervals) for the major output measures are provided in Table 5.

Table 4 summarizes the pairwise *t*-test results for each feature in Table 3. Despite the overall increase in the descriptive means, the inferential results showed significant increases in self-repairs, topic development, and reasoning from Session 1 to Session 2. From Session 2 to Session 3, only topic development increased significantly. LREs and FSs displayed significant increases only between Sessions 1 and 3. Questions and examples did not reach significance at any comparison. Thus, with each additional practice session, topic development continued to improve whereas the pattern for the other indicators was "a sharp increase from Session 1 to

Session 2, followed by a plateau." This suggests that repeated AI interaction may initially stimulate output expansion and then qualitatively restructure interactional behavior.

Table 5

Complete statistical details for paired comparisons of major output measures

Variable	Comparison	$t(9)$	p	Mean difference [95% CI]
Total words	S1→S2	3.74	.005**	33.70 [13.30, 54.10]
Total words	S2→S3	1.99	.078	25.80 [−3.53, 55.13]
Total words	S1→S3	5.02	<.001***	59.50 [32.71, 86.29]
Turns	S1→S2	−2.87	.019*	−1.80 [−3.22, −0.38]
Turns	S2→S3	−2.11	.064	−1.40 [−2.90, 0.10]
Turns	S1→S3	−5.09	<.001***	−3.20 [−4.62, −1.78]
Words/turn	S1→S2	3.39	.008**	5.24 [1.75, 8.73]
Words/turn	S2→S3	2.42	.038*	4.10 [0.27, 7.93]
Words/turn	S1→S3	4.87	<.001***	9.34 [4.99, 13.69]

Note. $df = 9$ for all comparisons. *$p < .05$, **$p < .0167$ (Bonferroni-adjusted), ***$p < .001$. Cohen's d values for these comparisons are reported in Table 2.

Qualitative Findings: Cases of Stalled Learner Talk

Lack of Vocabulary Comprehension and Ask-Back Strategies

In Excerpt 1, the AI asks a question about traditional culture. Although the question is not particularly difficult, the student, Shin, cannot respond immediately, and the conversation stalls. Later, during the interview, it became clear that the student did not catch the word "well-known" and was unsure how to answer. The utterance stops at "… culture is…," and no clarification request (e.g., "What did you say?") is produced. In a human–human conversation, the student would find it relatively easy to ask the interlocutor to repeat the question, and a human partner would also infer confusion from the learner's facial expression and prosody.

Excerpt 1

Shin, Session 1

5 AI: That brand is very famous. Forestry is important too. What is a well-known traditional culture?

6 Student: Uh,… traditional,… culture is….

Stopping at Abstract Questions

In Excerpt 2, the student responds to the AI's question "What traditions do you like?" only with the word "Tradition…," and the interaction momentarily stalls. The student reportedly understood the AI's words but lacked sufficient vocabulary and conceptual resources to respond to the abstract topic of "tradition." However, when the next AI question provided concrete examples, the student could answer immediately, even though it still involved an abstract notion

("heritage"). In a follow-up interview, the student mentioned that they would have been able to respond if the AI had asked about "traditional food" or "traditional events." These patterns suggest that beginners find abstract topics particularly challenging.

Excerpt 2

Taku, Session 1

9 AI: Namahage is special tradition. It happens in Oga. Akita has many traditions. What traditions do you like?

10 Student: Tradition….

11 AI: Akita Kanto Festival. Kakunodate samurai houses. Akita has much heritage. Which do you like?

12 Student: Well, I like Kanto festival.

Minimal Responses on Uninteresting Topics

As one might expect, learners also tended to give only minimal responses to topics they found uninteresting. In Excerpt 3, the student provides relatively short and hesitant answers when asked about skiing, a topic that is not particularly appealing to them. Meanwhile, when the conversation shifts to "future plans," a topic of personal relevance, this student's response is longer and more detailed. Because the AI cannot observe the learner's facial expressions or other nonverbal cues, it is less able than a human interlocutor to adjust its questions to the learner's degree of interest.

Excerpt 3

Tomo, Session 3

19 AI: Skiing is great. Akita has many snowy mountains. Perfect for skiing. Do you ski often?

20 Student: Yes I often do it,… I often go, went skiing.

21 AI: Akita has great ski resorts. You must visit Akita again. What is your next plan?

22 Student: Next spring vacation, I'm going to go to Nikaho city, because and then I, I want to, I want to go to Mount Chokai.

Qualitative Findings: Cases of Extended Learner Talk

Explanatory Expansions

This section presents several instances of students' successful engagement in extended talk and topic development. In Excerpt 4, the student responds to the AI's question "What makes Kamakura special?" by attempting to explain its characteristics, albeit somewhat hesitantly. The AI already "knows" what "Kamakura" is, so in principle, a human-like explanation is not necessary. Nevertheless, in this exchange, the participant carefully describes "Kamakura" as though they were talking to another person. Despite being fully aware that the interlocutor is an AI, the student still undertakes this "explanatory" work, which can be construed as a kind of "monologue in the presence of an interlocutor," which goes beyond a minimal, machine-

oriented response.

Excerpt 4

Tomo, Session 3

14 Student: Yes, I like it. Uh,… personally, especially uh,… that festival yes, I, I like it.

15 AI: Kamakura festival is special. Yokote has much snow. Winter brings beautiful scenes. What makes Kamakura special?

16 Student: Uh,… so it is so big house, so it seems so big house. Uh, yeah, I so I got so excited. It's so big.

Analytical Reflections

In Excerpt 5, the student responds to a "what" question by offering a developed example. The AI does not explicitly request multiple examples in its question about nature, but the learner contrasts seasonal differences in the beauty of Mount Chokai ("in summer," "in autumn," and "in winter"). This narrative transcends simple description and can be regarded as analytical and evaluative, as the learner temporally organizes the response and highlights different aspects of the same natural object.

Excerpt 5

Taku, Session 3

7 AI: Mount Chokai is very beautiful. Akita values such nature. What nature do you like?

8 Student: I like mountains because we can see,… uh,… Mount Chokai, which has different types of beauty of mountain in a in the year. For example, uh,… in summer season we can enjoy mountain, and in fall season we can see colorful changing,… we can enjoy watching changing leaves and in winter season we can see really beautiful snow mountain.

9 AI: Akita has many snowy mountains. Winter is beautiful. Skiing is popular. Have you tried it?

10 Student: No, yeah, actually. I will, would like to try something.

Questioning After an Explanation

Excerpt 6 illustrates how a learner explains a traditional event. As shown in Excerpt 2, a potential reason for this successful expansion is that the AI first provides model response examples before asking the highly human-like basic question "What do you like?" The student then follows up with a question that further expands the topic. After describing Namahage, the student asks the AI, "What do you know about Akita?," thus treating the AI as a conversational partner with its own "knowledge state." The use of "you" to address the AI suggests that, although the learner knows the interlocutor is a machine, they still orient to it as a legitimate interactional partner and actively co-construct the dialogue.

Excerpt 6

Erina, Session 3

3 AI: Nature, culture, and events are attractive. Shirakami and Tazawa are beautiful. Meet Akita dogs.

What do you like?

4 Student: I'm interested in Namahage, which is a kind of demon-like figure. They visit local houses and teach some manner lessons to kids. What do you know about Akita?

5 AI: Akita has rich nature. Rice, forestry are key industries. Culture is very unique. What natural spot you like most?

6 Student: Nature, uh,… lake Tazawa. I know this is the deepest lake in Japan, and very beautiful.

Similarly, Excerpt 7 shows that the student provides a detailed description of one sightseeing spot. After the AI's next question, the student can ask about the meaning of an unfamiliar word. Again, the student talks as if addressing a human listener.

Excerpt 7

Reina, Session 3

12 Student: It has some famous spots. One of them is Lake Tazawa. It is said that,… no, it is a deepest lake in Japan and then which is very popular for….

13 AI: Lake Tazawa is Japan's deepest. Kakunodate has Samurai residences. What is the administrative capital of Akita?

14 Student: Administrative? I don't know. What do you mean administrative capital?

15 AI: Administrative capital is the location of government. Akita City is the capital. What is one more important point from the material?

Topic Drift to Self-Interest

In Excerpt 8, the AI poses a fairly basic question about snow. After providing a brief answer, the student quickly shifts the focus to personal concerns, such as not owning a car and the difficulty of moving around in winter. This illustrates how a question-driven task can lead to the learner's own "talkable territory," drawing on everyday experiences and values. This interaction demonstrates how, even within a tightly structured, AI-led question sequence, learners redirect the topic to issues that matter to them and hence produce more extended, affectively rich talk.

Excerpt 8

Taku, Session 3

11 AI: Akita winters are perfect. Deep snow is abundant. Do you like snow?

12 Student: Not really. Because I don't have cars, so when it comes to go somewhere, I somewhere,… in winter season, I have to go there on my own foot. So, if I have snow, it makes me so annoying. It's so annoying. It is really hard for me to go somewhere in the season.

13 AI: Heavy snow is difficult. Akita has winter events. Kamakura snow domes are famous. Do you know Kamakura culture?

14 Student: Yes. Of course.

Survey Results

The questionnaire was developed for this exploratory pilot study and has not been formally validated; the results are therefore reported as supplementary indicators of learner perceptions rather than as psychometrically validated measures. It was included to provide additional insight into learners' perceptions of the AI-mediated practice. After three weeks of practice, the participants completed a post-study questionnaire. Table 6 summarizes 9 of the 10 items, which were all rated on a five-point Likert scale from "strongly agree" to "strongly disagree." Higher mean scores generally indicate more positive evaluations; however, the items differ in focus, ranging from perceived skill gains (Q1–Q3) and affective responses (Q4–Q7) to willingness to continue and preferences regarding wider implementation (Q8–Q9). As such, the questionnaire results are interpreted descriptively rather than as a unified index of chatbot effectiveness.

Overall, the learners reported moderate improvement in speaking skills and vocabulary acquisition, but only a few perceived listening ability gains. They also expressed mixed feelings about interacting with the AI: some felt relaxed, while others still felt anxious when communicating with the chatbot. Although the learners did not consistently perceive the conversation as equivalent to human–human interaction, many of them were willing to continue practicing with the AI and agreed that these chatbots should be implemented more widely at universities. Although the participants generally had positive impressions of talking with the AI, it should be noted that no participant selected "strongly agree" for any item, which may reflect a limitation of AI-mediated conversation practice, or of the specific chatbot used in this study. An additional question (Questionnaire Item 10) assessed the time when learners began to feel comfortable using the AI chatbot. Six participants reported feeling comfortable by the second session, two by the third session, and two reported that they did not become comfortable during the study period.

Table 6

Questionnaire Results (n = 10)

	SA	A	N	D	SD	Mean
Q1 My English speaking skills improved through AI training.		5	3	2		3.30
Q2 I was able to acquire new vocabulary and expressions.		7	3			3.70
Q3 My listening skills improved.		1	8	1		3.00
Q4 I felt relaxed while practicing.		5	2	3		3.20
Q5 The conversation flowed as naturally as with a human interlocutor.		4	1	5		2.90
Q6 I was able to express everything I wanted to say during the session.		5	2	2	1	3.10
Q7 I could understand everything the AI chatbot said.		2	5	3		2.90
Q8 I would like to continue practicing with the AI chatbot.		8	1	1		3.70
Q9 AI chatbots should be implemented in more universities.		6	4			3.60

Note. SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree (n = 10).

Discussion

Two-Stage Effects of AI-Mediated Conversation

Regarding RQ 1, the findings show a two-phase pattern of observed change in learners' observable interactional performance. Large effects emerged between Sessions 1 and 2, followed by more moderate growth between Sessions 2 and 3, with total word count no longer statistically significant. Notably, however, this quantitative plateau was accompanied by continued growth in topic development, suggesting a shift from output expansion to qualitative elaboration.

This pattern highlights two phases: an initial adjustment phase (Sessions 1 to 2) marked by rapid behavioral adjustment, and an elaboration phase (Sessions 2 to 3) marked by slowed quantitative growth but continued qualitative elaboration. In the initial stage, learners moved from fragmented responses to substantially longer turns, with significant gains in both self-repair and topic development. Once accustomed to the AI context, learners began engaging more actively in discourse-level expansion. Questionnaire item 10 also showed that while several learners reported feeling comfortable by Session 2, a substantial minority did not become comfortable within the study period, suggesting individual differences in adaptation to AI-mediated interaction.

During the second stage, raw output measures slowed to medium effect sizes, while topic development continued to grow significantly and LREs more than doubled in frequency. Qualitatively, learners produced more explanatory elaborations, analytical reflections, and follow-up questions, and initiated shifts to personally relevant topics. These behaviors suggest that, after affective and procedural barriers were reduced, learners took greater control of the conversation and engaged more actively in meaning-focused communication.

AI-mediated practice enabled the participants to produce substantially longer turns after each AI turn than they might in typical peer conversations. An informal comparison with referential peer-to-peer conversation data (n = 10; 5 pairs) helps contextualize this quantitative pattern. While per-participant word counts were broadly similar between the two conditions, a striking structural contrast emerged: learners produced substantially more turns in human–human interaction, and this contrast led to considerably shorter words-per-turn values. This structural difference likely reflects the "social safety net" of backchannels and short acknowledgment tokens (e.g., "yeah," "I see") that routinely separate human conversation. In peer exchanges, the burden of communication is shared; however, in the AI-mediated environment examined here, these scaffolds were largely absent. Neither the chatbot nor the learner produced such minimal responses; consequently, the participants in this study had to carry the discourse alone.

A complementary explanation may lie in learners' perception of the AI. Unlike a human partner who can instinctively "read between the lines" or infer meaning or emotional subtext from contextual cues, the chatbot may be perceived as a literal interlocutor that requires more explicit and self-contained information within each turn. This pushes learners to elaborate their thoughts rather than relying on the mutual intuition that characterizes human–human conversation.

Pedagogically, the findings suggest that teachers should expect one or two initial adjustment sessions before they observe qualitative development. A plateau in quantitative measures should

not be construed as stagnation, since discourse-level growth may continue beyond initial output expansion. Hence, AI-mediated conversation tasks may be more effective when implemented across several sessions rather than as one-off activities.

Interactional Features and Learner Agency

Regarding RQ 2, the qualitative analysis identified three patterns: the limited application of clarification strategies, difficulty responding to abstract questions, and topic drift toward personally meaningful content. These highlight both constraints and opportunities in AI-mediated interaction.

First, learners often failed to implement basic ask-back strategies when faced with unfamiliar vocabulary, allowing conversations to stall. This suggests that explicit pretraining in clarification and repair methods may strengthen the effectiveness of AI-based practice.

Second, learners tended to respond minimally to abstract questions unless accompanied by concrete examples. Example-rich prompts were associated with more elaborated topic development, possibly by offering learners more accessible entry points into discussion, though the present observational data cannot support causal claims.

Another interesting trend observed in the reference peer-to-peer data concerns language-related episodes (LREs)—instances where students directly raise questions about language use. LREs were virtually absent in the reference peer-to-peer data (0 instances across the 5 pairs), whereas they were observed in each of the three AI-mediated sessions and the session mean increased from 0.2 in Session 1 to 1.5 in Session 3 (see Table 3). This stark contrast may reflect differing learner orientations toward human versus AI interlocutors. Students appear more than ready to treat the AI chatbot as a valuable source of language information, possibly orienting to it as a knowledgeable interlocutor rather than a conversational peer. In contrast, such metalinguistic questions rarely occurred in peer interaction, possibly due to face concerns, assumptions about the partners' limited English skills, or the implicit expectation that they should keep the conversation flowing rather than interrupt for language clarification. Within the scope of this study, this finding would suggest that AI-mediated practice may uniquely facilitate explicit attention to linguistic form—a mechanism central to interaction-based accounts of SLA (Swain & Lapkin, 1998)—in ways that peer interaction among same-L1 learners does not. Additionally, learner-to-learner conversation in the reference data featured frequent code-switching into Japanese (L1), which was largely absent in AI-mediated sessions.

Third, learners sometimes redirected conversations toward their interests, generating longer and more elaborated talk. They did not remain passive respondents even within a question-driven format and occasionally exercised agency by shaping the direction of interaction. These findings suggest that AI-mediated conversation can support emerging agency among learners when topics resonate with their experiences.

Comparison with Prior Empirical Studies

The present findings both align with and extend prior empirical work in three related areas.

Speaking outcomes. Sok and Shin (2025) and Alenezi and Alenezi (2025) reported high post-test speaking gains after learners' engagement with AI chatbots. Since the present study was

not designed to test outcome gains, it cannot corroborate or refute these findings. However, the observed increases in words per turn and in the frequency of topic-development moves are consistent with the kinds of interactional gains that may underlie such improvements.

Learner–AI discourse. Although Yin and Satar (2020) identified negotiation-for-meaning sequences in text-based chatbot interactions, they limited their analysis to frequency counts of discrete routines within single sessions. By tracing how interactional features change across repeated sessions, the present study extends this line of work. Additionally, it reveals a two-phase pattern that single-session frequency analyses cannot identify. Matiienko-Silnytska et al. (2025) reported that Ukrainian university learners produced more self-repairs and discourse markers in their interactions with an AI chatbot. Similarly, the significant increase in self-repairs observed here between Sessions 1 and 2 is consistent with their finding. This study adds cross-linguistic corroboration from a Japanese EFL context.

Affective outcomes (WTC and anxiety). Zhang et al. (2024) and Wang et al. (2024) reported that AI conversation practice was associated with increased willingness to communicate. Nadeem et al. (2025) found reductions in speaking anxiety. The present study did not measure WTC or anxiety directly, but responses to Questionnaire Item 10—where six learners reported feeling comfortable by Session 2, two by Session 3, and two not at all—nuance these largely positive findings by highlighting notable individual variability in affective adaptation to AI interlocutors.

Pedagogical and Design Implications

This study deployed a teacher-designed AI conversation app in which the AI primarily asks questions, responds when students themselves initiate questions, moves on if learners do not answer, and keeps its turns relatively short. From the above-mentioned findings, several pedagogical and design improvements can be proposed.

Design Principle 1: Repair-Responsive AI

First, AI systems should respond adaptively to signs of learner difficulty. When a learner remains silent for several seconds after being asked a question, the AI could initiate a brief repair sequence, such as asking, "Do you have any questions?" or "Was that difficult to understand?" instead of simply proceeding to the next prompt. These follow-ups could encourage learners to use clarification requests and other basic communication strategies, thus transforming moments of nonunderstanding into opportunities for interactional development. This type of repair-responsive design may help promote meaning negotiation, which has been identified as a key mechanism in interaction-based accounts of SLA (Long, 1996).

Design Principle 2: Example-Scaffolded Abstract Prompting

Second, AI systems should provide scaffolded support when learners encounter abstract conversational prompts. This involves routinely giving concrete examples calibrated to the learner's proficiency level when posing questions involving abstract concepts. Listing options (e.g., naming specific festivals, foods, or places) can reduce learners' cognitive load and help them identify accessible entry points for topic development. This type of scaffolded prompting may extend learner output by lowering conceptual and linguistic barriers, thus supporting the

production of more elaborated discourse, which has been associated with opportunities for pushed output in interaction-based SLA research (Swain, 2000).

Design Principle 3: Interest-Responsive Topic Prompting

Third, AI systems should dynamically adapt to the topical interests of learners, who tend to produce more extended and engaged speech when discussing topics they find personally meaningful. One observable indicator of such interest is increased learner output, characterized by longer turns or multiple consecutive utterances on a single topic. AI systems can use these signals to identify learner engagement areas and respond by asking follow-up questions that deepen or extend the topic rather than shifting prematurely to unrelated prompts. This interest-responsive design may help sustain learner engagement and extend discourse production, which contributes to developing L2 interactional competence.

Design Principle 4: Explicit Social-Interaction Support

Fourth, AI-mediated conversation systems should be supported by instruction that directly helps learners develop the necessary linguistic resources for managing interaction. Unlike human interlocutors, AI systems cannot interpret facial expressions, gestures, and other nonverbal cues, which means learners need to rely entirely on verbal means to express confusion, signal interest, or request clarification. Providing explicit instruction and example phrases (e.g., clarification requests, paraphrasing, or expressions of uncertainty) can help learners engage more effectively in AI-mediated interactions.

These design principles emphasize that the effectiveness of AI-mediated conversation practice depends not only on providing interaction opportunities but also on structuring and supporting interaction itself. Rather than functioning merely as a question generator, AI systems should adapt to learner difficulty, provide scaffolded support for abstract expression, sustain learner engagement through interest-responsive topic development, and support learners' use of explicitly taught interactional strategies. These findings suggest that AI conversation tools are most effective when designed around interaction-based perspectives on L2 development.

Limitations and Future Research

This exploratory pilot study has several methodological limitations. First, with only ten learners engaged in three brief (five-minute) sessions, the findings have limited generalizability, and the statistical power of inferential tests is correspondingly reduced. The reported effects should therefore be treated as exploratory patterns rather than confirmed differences. Second, the authors did not establish a matched control condition, so observed changes cannot be attributed exclusively to interaction with the chatbot. Third, the reference peer-to-peer dataset (n = 10) is drawn from a different cohort and differs from the main dataset in participants, session count, and task structure. Its role is therefore limited to descriptive contextualization, and the observed divergences between the two datasets—particularly the virtual absence of LREs and the higher turn frequency in peer interaction are suggestive rather than definitive. Future research adopting a between-subjects design with matched control conditions, larger cohorts, and extended session durations would substantially strengthen the inferential basis for the two-phase pattern reported here. In addition, the app design was constrained to a single teacher-designed chatbot and a structured question–answer format. Thus, future research should compare multiple

teacher-designed systems with varying degrees of scaffolding, adaptivity, and multimodal support to identify optimal configurations for interactional development. Fourth, the post-study questionnaire was developed for this study and has not been formally validated, limiting the inferential weight of the perception data.

Finally, this study relied on a specific analytic framework. To deepen the understanding of AI-mediated interaction, further studies could incorporate longitudinal CA, multimodal approaches, or identity-focused perspectives. Expanding task types, app designs, and analytic lenses will help accumulate knowledge about how AI can both support and shape L2 interactional development.

Conclusion

In this study, the interactional behavior of Japanese EFL learners when interacting with a teacher-developed AI chatbot was investigated. Applying both quantitative and qualitative approaches to analyzing the learners' discourse, the current study uncovered changes in both output and organization of the participants' speaking behavior.

The findings can be summarized as follows. First, the participants got used to the AI after one or two lessons and managed to have a rather smooth communication during the short period of practice. Second, while raw output growth slowed in subsequent sessions—with total word count no longer reaching statistical significance—topic development continued to expand, suggesting qualitative elaboration in later sessions. Third, abstract questions, vocabulary gaps, and a lack of shared interest in the subject appeared to disrupt conversations, revealing important aspects to consider in future chatbot design. Fourth, the informal comparison with peer-to-peer data highlighted some notable features of AI-mediated interaction, including longer turns, the near absence of backchannels, and language-related episodes that rarely occurred in peer conversations. In addition, the current study contributes to the literature on AI-assisted language learning by demonstrating that a language instructor with minimal programming expertise can create an AI chatbot using accessible platforms such as Google AI Studio and Vercel to help facilitate meaningful conversation practice for L2 learners.

Collectively, these exploratory findings suggest that communication via AI may function as an alternative context, which may call for different discourse strategies from learners. For practitioners, the current research presents a methodology for creating an AI-mediated language learning chatbot, based on interaction-based SLA principles, which educators can emulate moving forward. Such principled design can support the educational potential of AI-mediated language learning.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number JP26K04066.

References

Alenezi, A., & Alenezi, A. (2025). Evaluating the effectiveness of chatbot-assisted learning in enhancing English conversational skills among secondary school students. *Education Sciences, 15*(9), 1136. https://doi.org/10.3390/educsci15091136

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning, 32*(8), 827–877. https://doi.org/10.1080/09588221.2018.1535508

Chapelle, C. A. (2001). *Computer applications in second language acquisition*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524681

Chapelle, C. A. (2009). The relationship between second language acquisition theory and computer-assisted language learning. *The Modern Language Journal, 93*, 741–753. https://doi.org/10.1111/j.1540-4781.2009.00970.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

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

Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition, 16*(3), 283–302. https://doi.org/10.1017/S0272263100013097

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Jeon, J. (2024). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning, 37*(1–2), 1–26. https://doi.org/10.1080/09588221.2021.2021241

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

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer-assisted*

language learning. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203708200

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press. https://doi.org/10.1016/B978-012589042-7/50015-3

Markee, N. (2000). *Conversation analysis*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410606471

Matiienko-Silnytska, A., Mikava, N., Savranchuk, I., Tkhor, N., & Poliakova, H. (2025). Conversational analysis of learner–AI chatbot interactions in developing spoken fluency. *Arab World English Journal, 16*(4), 224–238. https://doi.org/10.24093/awej/vol16no4.13

Nadeem, S., Iftikhar, S., & Rasheed, A. (2025). Addressing the fear of speaking: The role of AI in reducing speaking anxiety in ESL learners. *Journal of Applied Linguistics and TESOL, 8*(2), 825–833. https://jalt.com.pk/index.php/jalt/article/view/693

Nguyen, C. V. (2025). Employing TALKPAL.AI to enhance speaking proficiency for Vietnamese adult learners: A literature review. *International Journal of AI in Language Education*, *2*(2). 40–54. https://doi.org/10.54855/ijaile.25223

Nguyen, H. A. (2024). Harnessing AI-based tools for enhancing English speaking proficiency: Impacts, challenges, and long-term engagement. *International Journal of AI in Language Education*, *1*(2), 18–29. https://doi.org/10.54855/ijaile.24122

Nguyen, L. A. D., & Le, T. T. P. (2025). Exploring the effects of an AI chatbot on emotional engagement in English speaking lessons: Insights from Call Annie. *International Journal of AI in Language Education*, *2*(2), 79–99. https://doi.org/10.54855/ijaile.25225

Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition, 11*(1), 63–90. https://doi.org/10.1017/S027226310000783X

Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective*. Blackwell Publishing.

Sok, S., & Shin, H. W. (2025). Do interactions with ChatGPT influence L2 learners' oral speaking ability, summarization ability, and perceptions of generative AI tasks? *TESOL Quarterly, 59*, S19–S51. https://doi.org/10.1002/tesq.70001

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the *p* value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

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

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language*

learning (pp. 97–114). Oxford University Press.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16*(3), 371–391. https://doi.org/10.1093/applin/16.3.371

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82*(3), 320–337. https://doi.org/10.1111/j.1540-4781.1998.tb01209.x

Tarone, E. (1981). Some thoughts on the notion of communication strategy. *TESOL Quarterly, 15*(3), 285–295. https://doi.org/10.2307/3586754

Varonis, E. M., & Gass, S. M. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics, 6*(1), 71–90. https://doi.org/10.1093/applin/6.1.71

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

Yin, Q., & Satar, M. (2020). English as a foreign language learner interaction with chatbots: Negotiation for meaning. *International Online Journal of Education and Teaching, 7*(2), 390–410. https://www.iojet.org/index.php/IOJET/article/view/707

Zhang, D., Wu, J. G., & Fu, Z. (2024). From shy to fly: Facilitating EFL learners' willingness to communicate with an AI chatbot and an intelligent tutoring system. *System, 127*, 103501. https://doi.org/10.1016/j.system.2024.103501

Biodata

Ken'ichi Yamazaki is an associate professor of English education at Akita Prefectural University, Japan. With 20 years of teaching experience, his research interests include conversation analysis, the pedagogical effects of AI chatbots, and effective methods for enhancing Japanese EFL learners' speaking abilities.