

Article

Enhancing EFL Speaking Skills with AI-Powered Word Guessing: A Comparison of Human and AI Partners

Mondheera Pituxcoosuvorn , Midori Tanimura, Yohei Murakami  and Jeremy Stewart White

Faculty of Information Science and Engineering, Ritsumeikan University, Osaka 567-8570, Japan; m-tanimu@fc.ritsumei.ac.jp (M.T.); jwhite@fc.ritsumei.ac.jp (J.S.W.)

* Correspondence: mond-p@fc.ritsumei.ac.jp (M.P.); yohei@fc.ritsumei.ac.jp (Y.M.)

Abstract: This study explores the effects of interacting with AI vs. human interlocutors on English language learners' speaking performance in a game-based learning context. We developed *Taboo Talks*, a word-guessing game in which learners alternated between giving and guessing clues with either an AI or a human partner. To evaluate the impact of interaction mode on oral proficiency, participants completed a story retelling task, assessed using complexity, accuracy, and fluency (CAF) metrics. Each participant engaged in both partner conditions, with group order counterbalanced. The results from the retelling task indicated modest improvements in fluency and complexity, particularly following interaction with the AI partner. Accuracy scores remained largely stable across conditions. Post-task reflections revealed that learners perceived AI partners as less intimidating, facilitating more relaxed language production, though concerns were noted regarding limited responsiveness. Qualitative analysis of the gameplay transcripts further revealed contrasting interactional patterns: AI partners elicited more structured interactions whereas human partners prompted more spontaneous and variable interactions. These findings suggest that AI-mediated gameplay can enhance specific dimensions of spoken language development and may serve as a complementary resource alongside human interaction.

Keywords: AI-assisted learning; EFL speaking practice; game-based learning; language anxiety; human–AI interaction; taboo game



Academic Editors: Todorka Glushkova and Lyubka Doukovska

Received: 1 April 2025

Revised: 6 May 2025

Accepted: 20 May 2025

Published: 23 May 2025

Citation: Pituxcoosuvorn, M.; Tanimura, M.; Murakami, Y.; White, J.S. Enhancing EFL Speaking Skills with AI-Powered Word Guessing: A Comparison of Human and AI Partners. *Information* **2025**, *16*, 427. <https://doi.org/10.3390/info16060427>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Developing spoken fluency in a second language remains one of the most challenging aspects of language learning. Learners often struggle with limited opportunities for authentic conversation, anxiety about making mistakes, and inconsistent feedback, which can hinder the development of real-time communication skills [1].

Recent advances in artificial intelligence (AI) and natural language processing have enabled the creation of conversational agents capable of providing responsive and accessible speaking practice [2]. One promising approach is the integration of AI into game-based learning systems, which are known to improve student motivation and engagement [3].

In a previous work, we introduced *Taboo Talks* [4], an interactive language learning system that combines speech recognition, large language models (LLMs), and the structure of a word guessing game called the Taboo game to promote English speaking practice. The original version of *Taboo Talks* proposed the system architecture and card generation mechanism and reported a preliminary trial involving four participants and four game cards. The initial feedback suggested reduced learner anxiety and increased vocabulary acquisition, although limitations in pronunciation feedback and speech recognition accuracy

were noted. That version served as a proof of concept, laying the foundation for further empirical validation and design improvements. Nonetheless, questions remained about the effectiveness of AI compared to human interaction in more sustained language tasks.

This study aims to investigate how different partner types—human peers vs. AI partners—influence EFL learners’ speaking performance, communicative strategies, and perceptions during a structured word-guessing game. Specifically, we address the following research questions:

- RQ1.** How does the type of interaction partner (AI vs. human) affect explainer behavior and overall gameplay dynamics in a word guessing task?
- RQ2.** How does interaction with AI vs. human partners influence learners’ complexity, accuracy, and fluency of speech in a word guessing task?
- RQ3.** How do learners perceive and reflect on their experiences interacting with AI and human partners in the word guessing game?

By integrating large language models into a controlled game-based learning environment, we seek to examine how interaction partner type (human vs. AI) influences learners’ spoken language production across three core dimensions: complexity, accuracy, and fluency (CAF) [5,6]. By analyzing these dimensions in game-based speaking tasks, we aim to evaluate how different partner types affect both the effectiveness and efficiency of learner communication.

Our key contributions are as follows:

- An exploratory, comparison of learner performance in a word-guessing task, examining how AI functions as a gameplay partner relative to human peers using complexity, accuracy, and fluency (CAF) metrics;
- An analysis of interaction dynamics and communication strategies across human–human and human–AI gameplay conditions;
- Qualitative insights into learners’ perceptions of AI as a conversational partner, including its perceived supportiveness, challenge level, and impact on engagement.

This study was designed as an exploratory, formative evaluation—a type of early-stage research commonly used in usability testing, educational technology, and human–AI interaction [7]. Rather than aiming for broad generalizability, formative evaluations seek to uncover patterns, usability issues, and opportunities for system or design refinement through close observation of participant behavior.

2. Related Work

This study draws on five intersecting domains: (1) theoretical foundations in second language acquisition (SLA), (2) computer-assisted language learning (CALL) and artificial intelligence, (3) large language models (LLMs) and conversational AI in language education, and (4) game-based learning (GBL).

2.1. Interaction and Output in SLA

SLA theory emphasizes the critical role of interaction in language development. Long’s Interaction Hypothesis [8] and Swain’s Comprehensible Output Hypothesis [9] assert that conversational negotiation and effortful language production are essential to linguistic growth.

Empirical studies validate these claims in task-based settings; pre-task collaboration [10], corrective feedback [11], and collaborative mobile interaction [12,13] all enhance speaking fluency and motivation. These principles underpin our study design, which evaluates whether AI partners can support the same mechanisms of negotiation and pushed output as human interlocutors.

2.2. Technology-Enhanced Language Learning: From CALL to AI

The evolution from traditional CALL to LLM-enhanced platforms reflects a shift from static tools to interactive, adaptive learning systems. Early CALL research focused on multimedia tools, computer-mediated communication, and synchronous interaction to facilitate vocabulary and pronunciation practice [14,15].

However, in practice, implementation often remained teacher-centered [16], with successful outcomes heavily reliant on pedagogical integration [17,18]. For instance, Kim [16] examined the perceptions of ten ESL/EFL teachers who were concurrently enrolled in teacher education and educational technology programs. Using grounded theory methodology, the study found that despite the theoretical shift toward student-centered and constructivist CALL approaches, teachers largely perceived computers as supplementary instructional tools. Their beliefs continued to align with a teacher-centered paradigm, indicating a gap between CALL's pedagogical potential and its classroom application.

This underscores the importance of rethinking teacher training and the design of CALL environments to better align with student-centered learning goals. Modern AI-enhanced CALL builds on these foundations by offering more personalized, autonomous, and context-aware learning experiences, potentially bridging the gap between theoretical ideals and instructional realities.

2.3. Large Language Models and Conversational AI in Language Learning

LLMs such as ChatGPT by OpenAI (<https://openai.com/chatgpt>, accessed on 1 April 2025) have advanced the capabilities of language learning tools by offering conversational fluency, adaptive feedback, and content generation across a wide range of topics [19,20]. These systems show promise in boosting oral proficiency and learner autonomy, especially in speaking practice.

Yet limitations persist: hallucinated content and gaps in applied reasoning hinder reliability [21]. Techniques such as prompt engineering and few-shot learning are proposed to improve performance in educational contexts [21,22].

Despite enthusiasm for AI tools in education, learners express ambivalence. Benefits include flexibility and contextual responsiveness, but concerns about factual reliability and over-reliance remain [23,24]. Hybrid models that combine LLMs with human guidance are advocated to mitigate risks in high-stakes contexts such as medicine and education [25,26].

Ethical concerns, evaluation standards, and equitable access are equally important as these tools gain prominence. Responsible integration must foreground trust, transparency, and learner agency [27,28].

2.4. Game-Based Learning in Second Language Acquisition

Game-based learning (GBL) leverages interactivity and immersion to promote vocabulary development and communicative competence [29]. Studies show that digital games enhance motivation and lower learner anxiety while improving vocabulary retention and grammar accuracy [30–32].

Taboo, in particular, has proven effective in boosting vocabulary mastery, engagement, and peer collaboration [33–35]. Its structured output and need for negotiation align well with SLA principles.

AI-enhanced GBL expands these affordances further. Features such as adaptive feedback, chatbot partners, and speech recognition allow for real-time, personalized support [36–38]. However, concerns remain about the overemphasis on engagement at the expense of cognitive development [39,40].

Future research should prioritize rigorous evaluation, user-centered design, and the development of deeper interaction metrics to ensure AI-GBL contributes meaningfully to language acquisition [41,42].

3. Taboo Talks Platform

Taboo Talks is a custom-built, browser-based language learning platform designed to facilitate human and AI interaction in a word-guessing game. In this system, one participant assumes the role of either the explainer or the guesser, while the AI takes on the complementary role. The explainer describes a target word without using taboo clues, and the guesser attempts to identify the word based on the explanation. Although the platform was originally developed to assess learner performance in the explainer role, both roles were used in this study to simulate gameplay with a human partner.

3.1. Task Design

Taboo Talks is built around a modified version of the classic “Taboo” word guessing game, adapted as a structured speaking task for EFL learners. In each session, two participants take on alternating roles: the explainer, who describes a target word, and the guesser, who attempts to identify it. To increase the linguistic challenge, the explainer must avoid using a set of predefined taboo words, highly related terms that would otherwise make guessing too easy.

Gameplay proceeds in a turn-based manner. The explainer views the target word and its taboo words, while the guesser does not. After each clue, the guesser makes one guess. A round is successful if the target word is correctly guessed. This structure encourages focused speaking, real-time comprehension, and collaborative negotiation of meaning.

3.2. System Components

The Taboo Talks platform includes modular components to support both human–human (H–H) and human–AI (H–AI) gameplay modes. The user interface (UI) is intentionally minimal, placing emphasis on the language task rather than system navigation. It includes buttons to switch cards, a dataset selector, and real-time speech and response areas.

Two primary gameplay modes are supported: In Describe Mode, the learner acts as the explainer and receives a target word (e.g., “Bus”) along with taboo words that must be avoided. The learner provides a spoken explanation using a microphone button, and the utterance is transcribed using OpenAI’s Whisper-1 (OpenAI, San Francisco, CA, USA) [43]. GPT-4 (OpenAI, San Francisco, CA, USA) receives the transcription and replies with a single-word guess. The interface includes colored bubbles to show the user’s explanation and the AI’s guess, helping reinforce the interaction outcome.

In Guess Mode, the AI partner becomes the explainer and generates a natural language clue. The user, acting as the guesser, reads the clue and types in their guess. The interaction loop is simpler in this case, allowing learners to focus on comprehension and vocabulary retrieval. A robot icon is used to indicate AI involvement in both modes.

Figures 1 and 2 illustrate the user interfaces for these two gameplay modes.

3.2.1. Describe Mode Workflow

Figure 3 illustrates the system workflow for Describe Mode. The learner receives a Taboo card generated based on their English proficiency. Upon pressing the microphone button, the learner’s speech is captured and processed through Whisper, which converts it to text. GPT-4, acting as the guesser, receives the transcription and responds with a single-word guess. If the guess is incorrect, the learner may provide additional clues.

This cycle continues until the AI guesses correctly or the round ends. Compared to the original system [4], the updated version supports multi-turn guessing, feedback handling, and more human-like conversational flow.

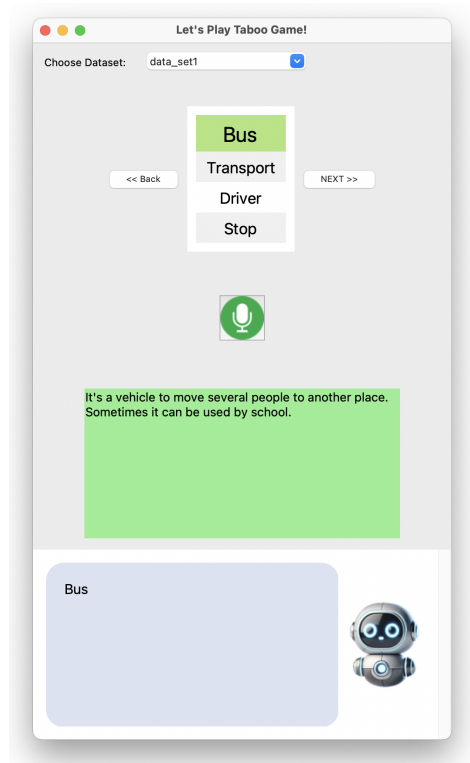


Figure 1. Interface for Describe Mode.

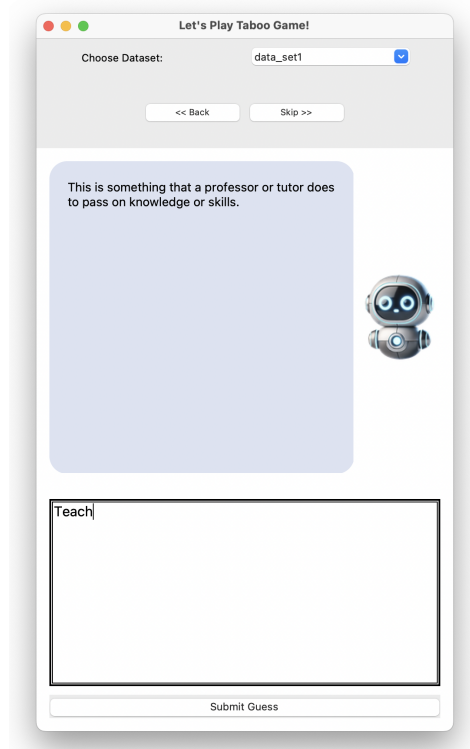


Figure 2. Interface for Guess Mode.

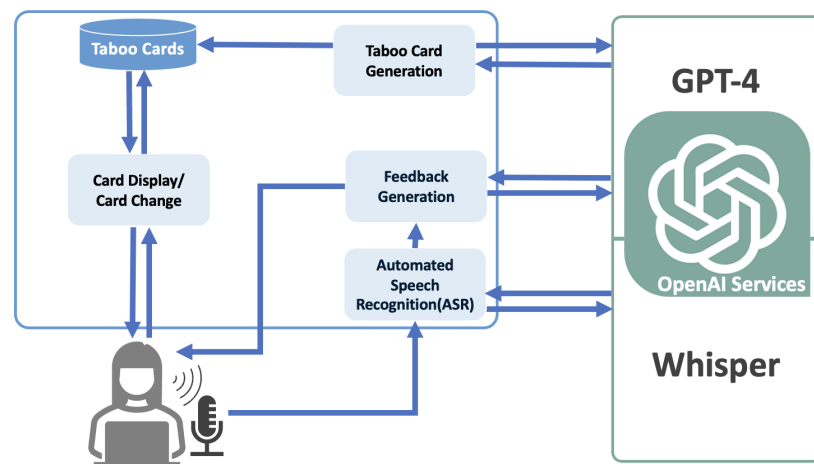


Figure 3. System Workflow for Describe Mode.

3.2.2. Guess Mode Workflow

Figure 4 shows the Guess Mode process. In this mode, the AI acts as the explainer. A target word and associated taboo words are drawn from the card set. GPT-4 generates a clue while avoiding the taboo terms, and the learner types in their guess. The system returns correctness feedback and continues the interaction as needed. This Guess Mode was not available in the earlier prototype and represents a novel extension toward simulating complete two-way gameplay.

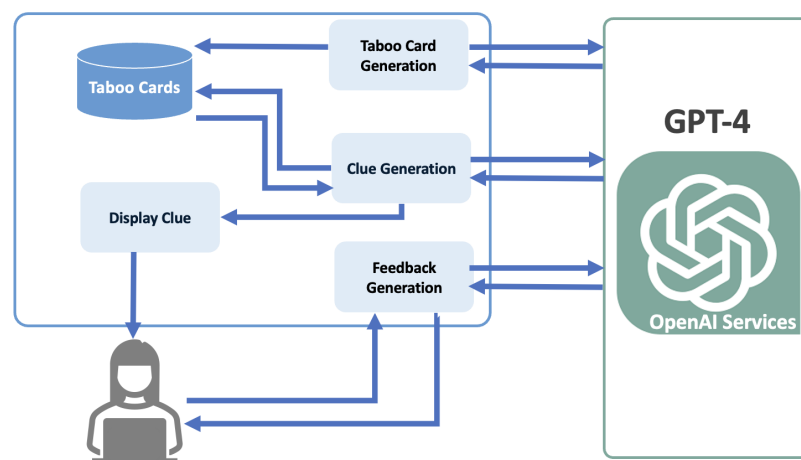


Figure 4. System Workflow for Guess Mode.

3.3. LLM Integration

The AI partner leverages GPT-4 for natural language generation, enabling fluid interaction in both explainer and guesser roles. When acting as the guesser, the LLM receives Whisper-transcribed user input and is prompted to make a single-word guess. If the guess is incorrect, it uses accumulated clues to inform subsequent guesses, simulating human-style reasoning.

The system prompt instructs the LLM to adhere strictly to the game's rules and maintain brevity. In the Guess Mode, when the AI plays the explainer, it generates sequential clues while avoiding the taboo words. It can also adapt its explanation strategy in response to failed user guesses, providing revised or simplified hints akin to human behavior.

These capabilities represent a significant enhancement over the previous version [4], which featured only one-way interaction from user to AI. The current implementation

allows for full two-way, multi-turn communication, resulting in a more authentic game-play experience.

LLM Prompt Design

To ensure that the AI behaves consistently with human gameplay, dedicated prompts were used to control GPT-4's behavior in each role.

When acting as the guesser in Describe Mode, the system used the following prompt:

You are a Taboo game guesser. You should try to guess the word from the given clue. Only give the word you think is the answer. Do not answer as a sentence or add something else that is not your answer.

When acting as the explainer in Guess Mode, the system used the prompt:

You are playing a Taboo game. Your task is to give a helpful clue for the word '{key_word}' without using the following taboo words: {taboo_words}. Provide a single sentence that helps someone guess the keyword without breaking the rules.

These engineered prompts helped align the AI's behavior with game expectations while keeping its output predictable and user-friendly.

4. Experimental Design

4.1. Story Retelling as an Evaluation Task

A range of task types are commonly employed to assess spoken language proficiency, including structured interviews, picture descriptions, role-plays, and free-form conversations. Each of these tasks elicits distinct aspects of language use: interviews facilitate the exploration of personal opinions, picture descriptions tend to prompt lexical retrieval and syntactic descriptions of static content, and role-plays are well-suited for evaluating pragmatic and interactive competencies [44,45].

In the present study, which aims to investigate learners' abilities to convey and reformulate meaning, the story retelling task was selected due to its distinctive advantages. Story retelling elicits extended and coherent speech by requiring learners to reconstruct a narrative sequence based on prior input. This format engages both linguistic and cognitive faculties, such as memory, sequencing, and discourse organization, thereby offering a comprehensive sample of language use.

As demonstrated by Koizumi and Hirai [46], the story retelling speaking test (SRST) constitutes a practical and valid instrument for assessing spoken language performance. Compared to more rigid, automated assessments such as Versant, or resource-intensive formats like the standard speaking test (SST), the SRST provides a balanced approach that combines standardization with elicitation of naturalistic language. It is especially suitable for classroom and research contexts where logistical efficiency and communicative authenticity are both desired.

Furthermore, retelling tasks are well-suited for measuring key dimensions of spoken proficiency—complexity, accuracy, and fluency (CAF)—as they simultaneously require lexical retrieval, syntactic structuring, and discourse-level planning. For these reasons, story retelling was adopted as the primary evaluation task in this study, serving as a robust benchmark for learners' spoken performance alongside the experimental Taboo-style explanation task.

4.2. Participants and Conditions

A total of 18 Japanese university students participated in this study. All were non-native English speakers with mixed proficiency levels ranging from A2 to B2 on the CEFR scale.

The experiment used a within-subject design: all participants experienced both human–AI and human–human versions of the Taboo game. To counterbalance order effects, participants were randomly assigned to one of two settings:

- Setting 1 (AI-first): Human–AI game followed by human–human game
- Setting 2 (Human-first): Human–human game followed by human–AI game

4.3. Experiment Setting

Each participant was assigned a laptop with the custom-built program installed. During the H–AI sessions, participants worked individually with Taboo Talks through a graphical interface that supported speech input and automatic transcription. All voice data and game logs were captured and stored digitally during gameplay.

In contrast, the H–H game was conducted face-to-face using printed Taboo word cards. Each pair of students shared the cards and took turns playing as explainer and guesser. Their conversations were recorded using external microphones and later manually transcribed for analysis.

4.4. Session Flow

Each session lasted approximately 90 min and followed a fixed sequence of activities. Table 1 summarizes the two counterbalanced task orders. This will be explained below.

Table 1. Counterbalanced Task Orders for the Two Conditions.

Phase	Condition 1 (AI-First)	Condition 2 (Human-First)
1	Explanation	Explanation
2	Retelling 1 (Pre)	Retelling 1 (Pre)
3	Human–AI Taboo Game	Human–Human Taboo Game
4	Retelling 2 (Mid)	Retelling 2 (Mid)
5	Human–Human Taboo Game	Human–AI Taboo Game
6	Retelling 3 (Post)	Retelling 3 (Post)
7	Free Talk Session (Peer Discussion)	Free Talk Session (Peer Discussion)
8	Questionnaire	Questionnaire

4.5. Speaking Proficiency Assessment: Retelling Task

To assess participants’ speaking proficiency and monitor changes over the course of the experiment, a structured retelling activity was conducted at three points: before, between, and after the Taboo game sessions. This activity served as a diagnostic tool to evaluate learners’ fluency, accuracy, and complexity in spoken English.

Participants completed three retelling tasks, modeled after the Eiken Pre-2, roughly equivalent to the A2 level in the CEFR, speaking test format. Each task followed this procedure:

1. The participant was shown a short passage (about four sentences) on a screen or printed sheet.
2. They had 1 min to read and silently memorize the content.
3. After turning over the text, they verbally retold the content in English using their own words, as if explaining it to someone unfamiliar with the topic.

The passages covered familiar topics such as “Virtual Reality Games”, “Pet Cafés”, and “Food Trucks”, and were written at a EIKEN Pre-2nd. Each spoken retelling was recorded and later evaluated using three core criteria: complexity, accuracy, and fluency. Complexity referred to the lexical and syntactic richness of the spoken output, while accuracy considered the grammatical correctness and appropriateness of the expressions

used. Fluency was assessed based on the smoothness of delivery, speech rate, and the presence of hesitations or pauses.

4.6. Taboo Gameplay

Participants played two Taboo-style games:

- H–AI mode: Using the laptop program, participants interacted with a language model. The AI guessed words based on spoken clues, and participants also responded to AI explanations. All utterances and transcriptions were automatically logged.
- H–H mode: Two students played using printed cards. They alternated roles and followed the same gameplay rules. Audio was recorded and manually transcribed later.

The analyses of the game interactions focused on several dimensions. These included the number of attempts required to correctly guess each word, the total number of words and utterances used during each session, and the overall success rates in both modes.

4.7. Exploring User Experience

Before the questionnaire, participants took part in a free talk session, where they discussed their impressions of the games with a peer. These conversations were conducted in Japanese and served as a brief focus-group-style reflection. Recordings were transcribed and thematically analyzed.

Finally, participants completed a post-task questionnaire, using 5-point Likert scales to evaluate:

- Clarity, challenge, and enjoyment of each mode
- Language learning impact
- Emotional responses, such as confidence and anxiety

Additional short and informal interviews were conducted in Japanese to further explore their experiences.

5. Retelling Task Analysis

To address RQ2, we evaluated how interaction partner type influenced learners' speech across the dimensions of complexity, accuracy, and fluency.

To evaluate language development during retelling tasks, we designed a three-stage procedure comprising pre-task, game play, and post-task trials. Participant pairs were divided into two experimental groups: AI First and Human First. Each participant received both AI and human feedback in alternating order, enabling a within-subject comparison while preserving a between-subject game play sequence.

5.1. CAF Evaluation Framework

Scoring Procedure

To evaluate learners' language development, we employed the complexity, accuracy, and fluency (CAF) framework—a widely adopted method in second language acquisition research [5,6]. Each retelling was rated on a 1–5 scale for each CAF dimension, as detailed in Table 2.

Audio recordings of learners' retellings were first transcribed using OpenAI's Whisper speech recognition system, a state-of-the-art model known for its multilingual and domain-agnostic performance. Transcripts were reviewed for accuracy and manually corrected where needed.

Given recent studies demonstrating that LLMs, such as GPT-4, align closely with human evaluators in both writing and speaking assessments [47] and achieve inter-rater reliability comparable to expert instructors [48,49], we employed GPT-4o as a co-rater in

this study. Specifically, GPT-4o was used to assign CAF scores to each explanation turn based on the standardized five-point rubric.

Final scores for each dimension were used to calculate treatment-based deltas across gameplay sessions, allowing us to compare performance trends in the AI vs. human feedback conditions.

Table 2. CAF Evaluation Rubric (1–5 scale).

CAF Dimension	Descriptor	
Complexity (Lexical and Grammatical Range)	5:	Wide range of vocabulary and complex sentence structures (e.g., subordination, embedded clauses)
	4:	Good range of vocabulary and some complex structures
	3:	Moderate variety; mostly simple structures with some variation
	2:	Limited vocabulary and repetitive patterns
	1:	Very basic vocabulary; minimal variation in sentence form
Accuracy (Grammatical and Lexical Correctness)	5:	Virtually no errors; message consistently clear
	4:	Minor errors that do not hinder meaning
	3:	Occasional errors; mostly understandable
	2:	Frequent errors that obscure meaning
	1:	Persistent errors; difficult to comprehend
Fluency (Flow, Pausing, Self-correction)	5:	Smooth delivery with minimal hesitation
	4:	Generally fluent with minor disruptions
	3:	Noticeable pauses or repetition, but message remains followable
	2:	Frequent hesitation or reformulation
	1:	Very disfluent; hard to follow due to disrupted flow

5.2. CAF Score Change Analysis for Retelling Task

For the retelling CAF analysis, we evaluated participants' performance across three trials of the story retelling task. Out of 54 total transcripts, 47 were retained for scoring. Seven transcripts were excluded due to incomplete or inaudible recordings. These excluded items came from five participants, each of whom had at least one unusable trial. As a result, 13 participants had complete and analyzable data across all three trials.

To compute CAF score changes, we measured CAF deltas across adjacent trials. For participants in the *AI First* group, the change in the H–AI score was defined as Trial 2 minus Trial 1, while the change in the H–H score was Trial 3 minus Trial 2. The order was reversed for the Human First group to ensure the game effect comparisons were consistent across conditions.

Only participants with complete CAF scores (complexity, accuracy, and fluency) across both windows were included in the final analysis. This filtering step ensured robust within-subject comparisons.

5.2.1. Statistical Analysis of Treatment-Based Score Changes

To analyze the differential effects of the two treatments—AI-assisted retelling (H_{AI}) and human-assisted retelling (H_H)—on learners' speaking performance, we computed score changes in three CAF dimensions. For each participant, we defined two post-treatment deltas:

- Δ_{H-AI} = score after H_{AI} – score before H_{AI}
- Δ_{H-H} = score after H_H – score before H_H

The exact trial boundaries depended on the participant's group assignment. For each metric, we calculated paired-sample t -tests and Cohen's d to assess effect size.

As shown in Table 3 and Figure 5, **Complexity** showed the most pronounced improvement following H–AI, with a mean increase of +0.60 and a large effect size ($d = 0.82$), despite the lack of statistical significance. **Fluency** also showed a favorable trend toward H–AI ($\Delta = +1.00$ vs. +0.64, $d = 0.29$). **Accuracy** improvements were identical in both conditions (mean $\Delta = +0.40$).

Table 3. Paired-Subset Comparison of CAF Score Changes.

Metric	Mean Δ_{H-AI}	Mean Δ_{H-H}	p -Value	Cohen's d
Complexity	+0.60	−0.40	0.142	0.82
Accuracy	+0.40	+0.40	1.000	0.00
Fluency	+1.00	+0.64	0.551	0.29

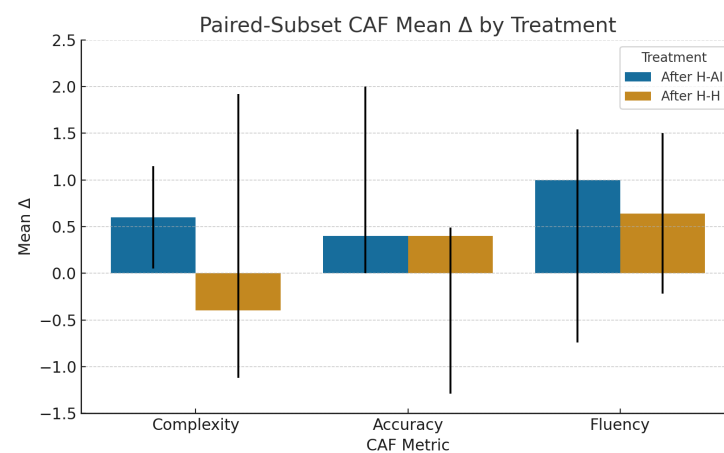


Figure 5. CAF Mean Δ Scores After H–AI and H–H (Bar Chart with SD Error Bars).

5.2.2. Examples of Complexity and Fluency Changes

To ground our statistical findings in actual learner output, we examined the retelling transcripts of the participants. Below we present exemplar before/after excerpts that illustrate the marked gains in Complexity and Fluency after AI-assisted (H_{H-AI}) vs. human-assisted (H_{H-H}) retelling tasks.

Complexity

After H_{AI} , learners frequently embedded subordinate clauses and produced more structurally complex sentences, supporting the observed large effect size ($d = 0.82$). For example:

- From User9 (Pet Cafe Theme)
 - Trial 1 (Pre-test): “Now, I will introduce about pet cafes in Japan. In pet cafe, people can enjoy drink and spend time with animals like cats, dogs and even owls. Pet cafe is popular with people who loves animals but cannot get pet in house. People can enjoy relaxing time with animals and take cute picture.” (Complexity = 3)
 - Trial 2 (After H_H): “Pet cafes are unique places in Japan where people can enjoy drinks and spend time with animals like cats, dogs, and even owls. It is popular with people who loves animals but cannot have pets in house. People can enjoy playing with animals and take cute pictures. Now I will introduce about robot guides in this museum.” (Complexity = 3,

$\Delta = 0$)

No notable increase in syntactic variety.

- Trial 3 (After H_{AI}): “Pet cafe is a unique place in Japan where people can enjoy drinks and spend time with animals like cats, dogs, and even owls. These cafes are popular with people who love animals but cannot have pets at house. People can feel relaxed and happy as they play with animals and take cute pictures.” (Complexity = 4, $\Delta = +1$)
Uses relative clause (“who love animals”), coordination, and descriptive elaboration.

Fluency

Fluency gains (mean $\Delta = +0.96$ after H_{AI}) manifested as longer runs and fewer filled pauses. For instance:

- From User16 (Pet Cafe Theme)
 - Trial 1 (Pre-test): “It is said about pet cafe. Pet cafe, people can enjoy to drinking and seeing animals in pet cafe. People who come to pet cafe like animals but they can not have animals in their house so these people enjoy touching animals.” (Fluency = 3)
 - Trial 2 (After H_H): “It is about, it is say about pet cafe in Japan. People who like pet but cannot have a pet at home is coming to this pet cafe. Pet cafe can enjoy animals, and drinking. People can enjoy that animals, and taking a picture.” (Fluency = 3, $\Delta = 0$)
Some clause chaining, but limited natural flow.
 - Trial 3 (After H_{AI}): “It is said about pet cafe in Japan. People who like animals but cannot have a animal in their house is coming to this one. They can enjoy taking cute pictures and playing with animals.” (Fluency = 4, $\Delta = +1$)
Improved fluency and cohesion across clauses.

These examples confirm the quantitative results: H_{AI} substantially enhanced structural complexity and spoken fluency, whereas H_H often led learners back to simpler, more segmented speech patterns. The large effect size in complexity ($d = 0.82$) underscores the pedagogical potential of AI-assisted retelling for promoting advanced syntactic experimentation.

5.3. Word Count Analysis for Retelling Task

We compared word count of retelling tasks after H–AI and H–H. Average counts were higher post-H–AI ($M = 106.2$ vs. 99.3 ; $p = 0.58$), with a small effect size ($d = 0.24$). This trend may reflect AI’s ability to lower learners’ anxiety and promote longer utterances.

6. Efficiency and Clarity in Clue-Giving Across Partner Types

This section addresses RQ1 by analyzing explainer behavior and gameplay dynamics across two conditions: human–AI (H–AI) and human–human (H–H). We report metrics including clue success rate, number of attempts, and linguistic efficiency.

To evaluate the impact of interaction partner type on language production quality, we analyzed participants’ clue-giving performance in the Taboo-style game under two conditions: human–AI (H–AI) and human–human (H–H). We assessed the speech samples using three dimensions of the CAF framework, along with utterance length, measured by average word count.

6.1. Game Play CAF Score Comparison: H–AI vs. H–H

In this analysis, M denotes the mean score, p the p -value indicating statistical significance, and d Cohen’s d effect size. Figure 6 shows side-by-side boxplots of the three CAF dimensions. While syntactic and lexical complexity remained virtually unchanged between the two conditions ($M_{textH-AI} = 1.04$, $M_{textH-H} = 1.05$, $p = 0.61$, $d = -0.06$), there was a noticeable difference in fluency. Participants were significantly more fluent in the

H–AI condition ($M = 3.74$) than in the H–H condition ($M = 1.99$), with a large effect size ($p < 0.001$, $d = 1.17$). This suggests that speaking to an AI partner reduced hesitation, fillers, or self-corrections. This may reflect lower social pressure in the AI condition, as some participants later reported feeling more relaxed interacting with a non-human partner (see Section 9.2).

Accuracy also showed a mild upward trend in the H–AI condition ($M = 3.05$) compared to H–H ($M = 2.77$), though this difference did not reach statistical significance ($p = 0.10$). The small effect size ($d = 0.18$) suggests a limited but observable tendency, which may reflect participants' efforts to craft more precise or targeted clues when interacting with AI partners—perhaps anticipating more literal or restricted interpretations. In line with the exploratory nature of the study, we interpret this trend as indicative of possible interactional adjustments rather than conclusive evidence of improved performance.

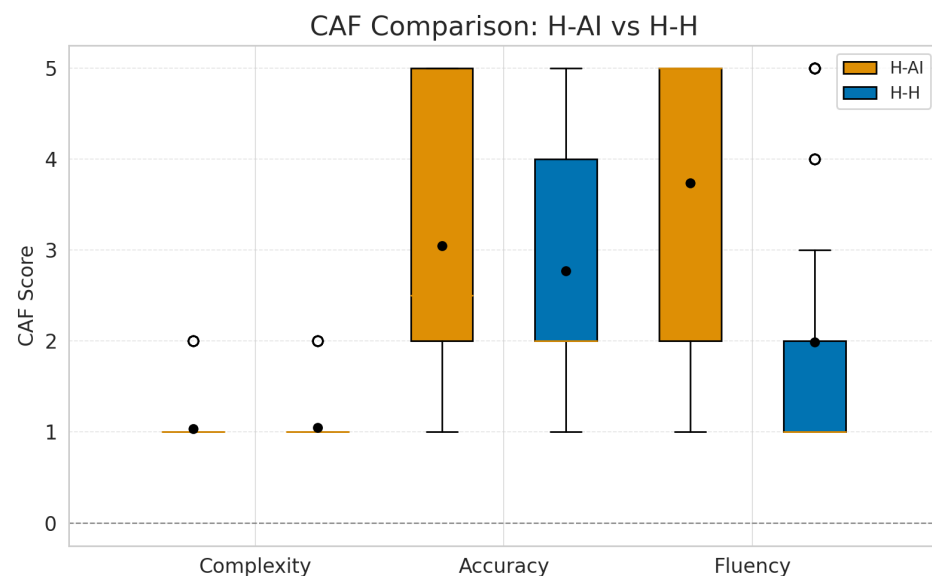


Figure 6. CAF score comparison between H–AI and H–H Conditions. Fluency shows a noticeable upward trend in the H–AI condition. Accuracy appears slightly higher, while complexity remains relatively stable across conditions.

Interpretation and Implications

These findings highlight a compelling benefit of using AI partners in communicative tasks. The AI partner likely offers a more forgiving and low-pressure environment, encouraging participants to speak with greater confidence and fluency. The reduced word count, paired with improved fluency and maintained or improved accuracy, suggests that interaction with AI supports efficient and clear language use—a valuable feature for language learning, training, or assessment.

Overall, the combination of quantitative metrics and qualitative insights supports the potential of AI-facilitated interaction to promote more effective spoken communication.

6.2. Qualitative Differences in Clue-Giving

In addition to quantitative differences, several transcripts demonstrate how fluency and efficiency vary by partner type. Table 4 provides sample utterances from both conditions.

In the H–H condition, utterances tend to be longer, but filled with repetition, hesitation, or fragmented thoughts—e.g., one participant said “No, no, no, no, no, no. Touch. Ah. Ah, so. No...”, which, despite being 42 words long, received a fluency score of just 1.0. Such

utterances suggest speakers are struggling to clarify or correct themselves in real-time, possibly due to pressure or unpredictability of the human partner.

By contrast, H–AI utterances were typically shorter but much more fluent. For instance, “A car which can ride many person” (7 words, Fluency = 5.0) effectively conveys a concept despite imperfect grammar. These cases support the interpretation that AI partners elicit clearer, more composed explanations—likely because users feel less socially evaluated.

Table 4. Clue-Giving Utterances by Condition (Excerpt).

User	Partner	Words	Fluency	Transcript (Excerpt)
user16	H–H	52	1.0	<i>This, this, this, this question, this question, okay? This question. I say this ...</i>
user12	H–H	44	1.0	<i>This is a kind of act. Often this is used for windows. Window. And... Small... M...</i>
user16	H–H	42	1.0	<i>No, no, no, no, no, no. Touch. Ah. Ah, so. No, no, no, no, no, no. No, no, no, n...</i>
user1	H–AI	7	5.0	<i>A car which can ride many person.</i>
user11	H–AI	9	5.0	<i>This is a thing to clear the brown environment.</i>
user13	H–AI	6	5.0	<i>We can carry this more power.</i>

7. Interactional Patterns in Taboo Games with AI and Humans

In our study, participants engaged in a turn-based Taboo-style guessing game. Each user played in two interactional conditions: H–H and H–AI. Each user alternated roles across multiple games. This paper focuses on the **explainer role**, analyzing how users adapted their communication when interacting with different partner types.

7.1. Performance Comparison

7.1.1. Methodology

To assess explainer performance, we analyzed 92 human–AI (H–AI) games and 61 human–human (H–H) games, each representing a unique instance in which a participant acted as the explainer. These game counts served as the basis for computing clue success rate (CSR), defined as whether the target word was successfully guessed within each game. For the other two metrics—*attempts* and *linguistic efficiency*—we evaluated a total of 263 explanation turns in the H–AI condition and 170 turns in the H–H condition. These values correspond to the number of valid explanation attempts (i.e., non-empty transcripts), with each turn treated as a distinct data point in the analysis.

Clue success rate, denoted as CSR, captures the proportion of games in which the explainer’s clues led to a correct guess by the guesser. It is calculated as

$$CSR = \frac{N_{\text{success}}}{N_{\text{total}}}$$

where N_{success} is the number of successful games, and N_{total} is the total number of games played by the user in that condition.

Attempts refer to the average number of utterances or turns the explainer used before the guesser arrived at the correct answer. Let U_i be the number of utterances in the i -th game and n be the number of successful games. Then the average number of attempts, A , is given by

$$Attempt = \frac{1}{n} \sum_{i=1}^n U_i$$

Linguistic efficiency, LE , measures verbosity by calculating the average number of words per attempt. Let W_i be the number of words used in the i -th game. Then,

$$LE = \frac{\sum_{i=1}^n W_i}{\sum_{i=1}^n U_i}$$

Each metric was computed at its appropriate unit of analysis. CSR and attempts were calculated at the game level. CSR indicates whether the explainer's clues led to a correct guess in a given game. Attempts refers to the number of utterances the explainer produced before the guesser arrived at the correct answer and was computed only for games that ended successfully. In contrast, LE was computed at the utterance level, as it reflects the average number of words per utterance across all valid explanation turns.

Since the number of data points differs across metrics, with one per game for CSR and attempts and one per utterance for LE, independent-sample *t*-tests were used to compare mean performance between the H-AI and H-H conditions. Although a paired design was initially considered, the final analysis was conducted using independent comparisons to account for the uneven number of games and explanation turns contributed by each participant across conditions.

We interpret *p*-values below 0.05 as statistically significant and values between 0.05 and 0.1 as marginal trends.

7.1.2. Performance Comparison Results

To visualize how interaction mode impacts task performance, we present three box-plots corresponding to key outcome measures (Figures 7–9). Figure 7 displays the distribution of clue success rate, capturing the proportion of successfully guessed target words per game. This measure reflects how effectively participants conveyed meaning to their partners under each condition. Figure 8 shows the number of attempts per game, representing the communicative effort required to achieve a correct guess. A lower number of attempts suggests more efficient clue delivery. Finally, Figure 9 illustrates linguistic efficiency, which quantifies the amount of information conveyed relative to utterance length. This metric is particularly relevant in evaluating how concisely participants communicated. Together, these figures provide a comparative overview of performance across modes in terms of accuracy, effort, and efficiency.

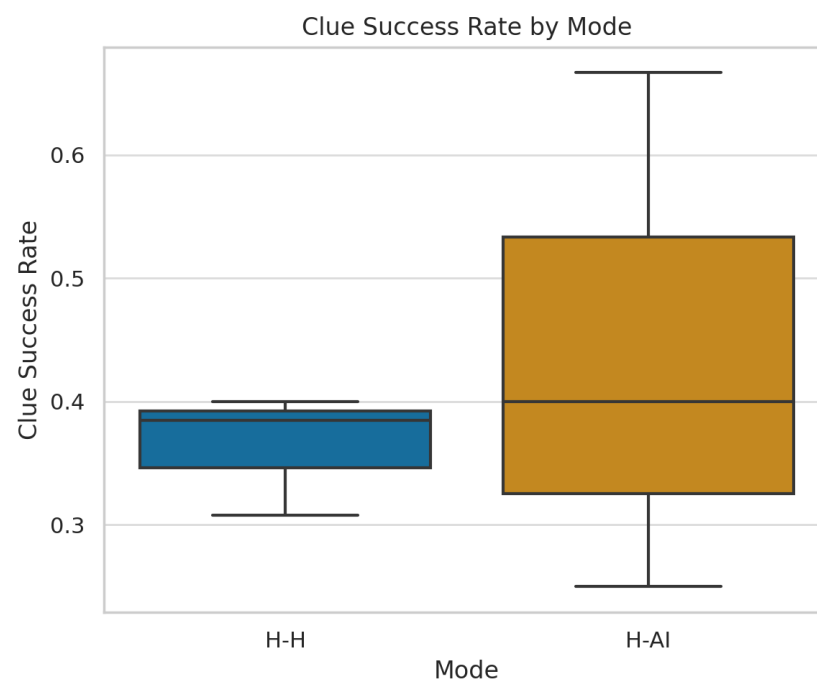


Figure 7. Clue Success Rate by Mode.

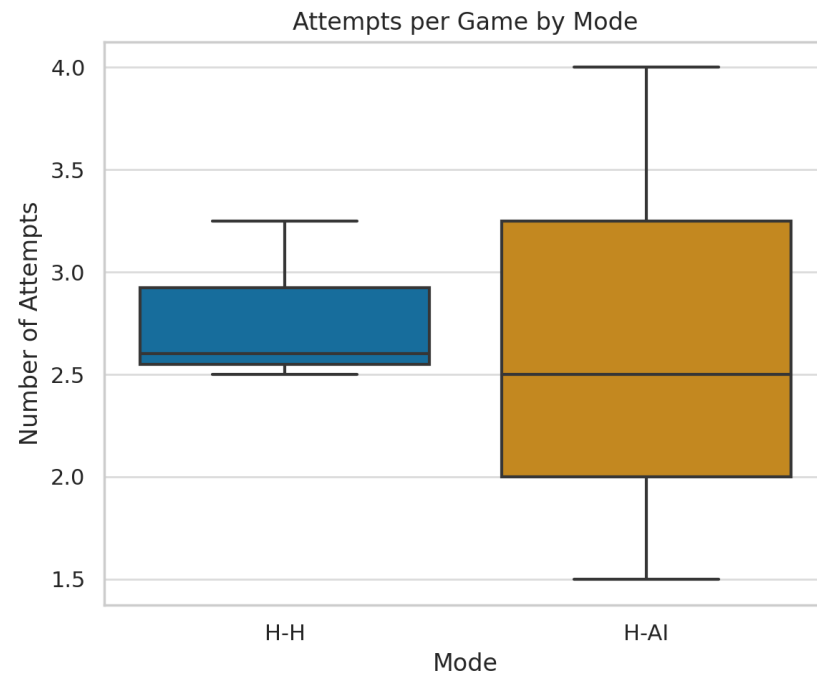


Figure 8. Attempts per Game by Mode.

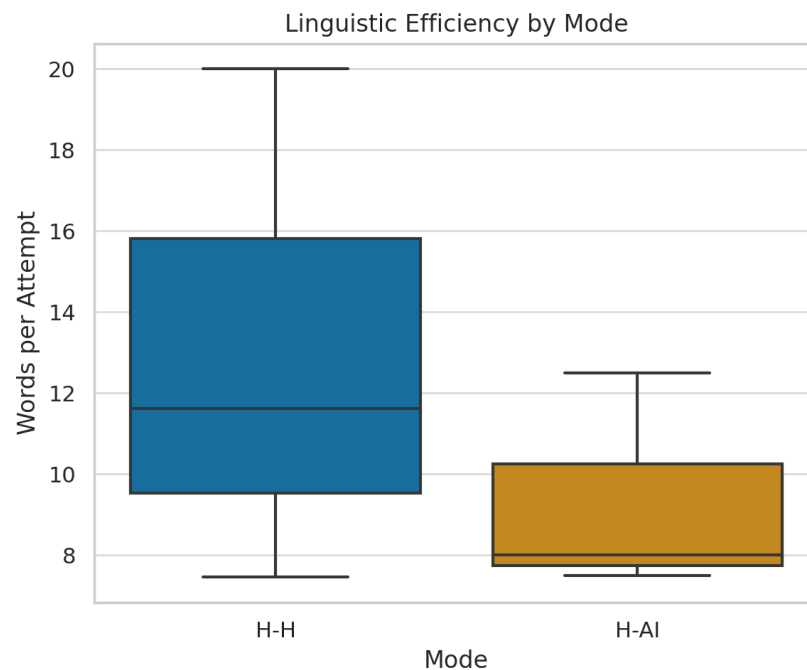


Figure 9. Linguistic Efficiency by Mode.

7.1.3. Performance Patterns

The performance metrics reveal a consistent pattern across interaction modes. First, participants achieved a significantly higher *clue success rate* when interacting with AI partners compared to human partners. This suggests that either the AI was more receptive to the style of clues provided or that participants unconsciously adjusted their explanation strategies to better suit the AI's predictable response behavior. This aligns with findings from prior work on H-AI communication, where users often simplify and adapt their speech when interacting with artificial agents, leading to more successful task completion [50,51].

Second, users required fewer *attempts*, defined as utterances per game, when collaborating with AI. This reduced need for iterative clarification or repetition indicates faster convergence and potentially smoother interaction flow in the H–AI condition. Similar trends have been observed in computer-supported collaborative tasks, where AI can facilitate more efficient task resolution through timely and consistent feedback.

Finally, regarding *linguistic efficiency*, users tended to use fewer words per attempt when interacting with AI. However, this reduction did not reach statistical significance ($p = 0.205$, from Table 5). Although the trend suggests more concise communication with AI, the variability among participants and task instances may have diluted the overall effect. Nonetheless, this pattern is consistent with earlier studies [50,51] showing that AI partners often encourage more streamlined linguistic output, particularly when users adapt their language to accommodate perceived limitations in the system’s comprehension.

Table 5. Comparison of explainer performance across interaction modes.

Metric	Mean (H–H)	Mean (H–AI)	<i>t</i> -Test <i>p</i> -Value
Clue Success Rate	0.36	0.44	0.044
Attempts	2.78	2.67	0.038
Linguistic Efficiency	13.03	9.33	0.205

7.2. Utterances Until Success

We also calculated the number of utterances (or turns) made by the explainer before the guesser arrived at the correct answer. This metric reflects the amount of back-and-forth interaction required to reach a successful outcome. In this subsection, we focus exclusively on successful games completed without skipping, rather than analyzing the entirety of the gameplay data.

Table 6 summarizes the average number of utterances per game until a correct guess, comparing H–H and H–AI interactions.

Table 6. Average number of utterances until successful guess.

Metric	Mean (H–H)	Mean (H–AI)	<i>t</i> -Test <i>p</i> -Value
Utterances	2.62	1.98	0.033

Figure 10 provides a visual comparison of utterance counts per game for each interaction mode.

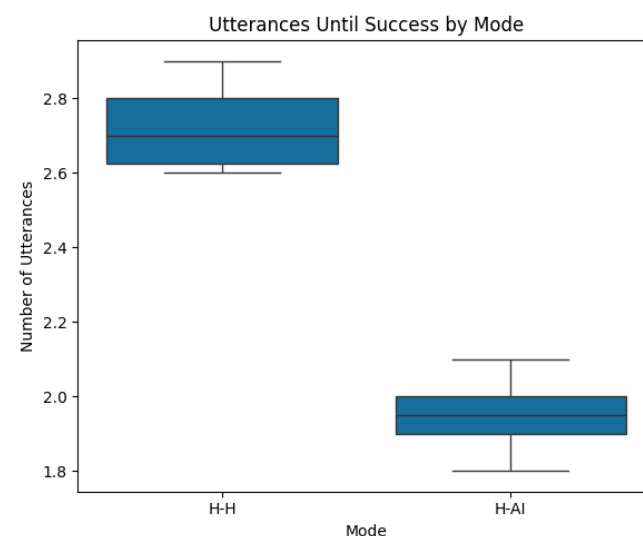


Figure 10. Utterances per game until success by mode.

The reduction in the number of utterances in the H–AI condition was statistically significant. This indicates that successful guesses occurred with fewer turns in H–AI games, reflecting a faster interaction pattern between explainer and AI guesser.

7.3. Partner-Specific Communication Strategies

A qualitative examination of clue-giving transcripts reveals a striking difference in the types of strategies participants employed depending on their interaction partner. In the H–H condition, explainers occasionally invoked shared cultural or contextual knowledge—what Clark [52] refers to as “common ground”. Rather than providing fully descriptive clues, some participants leveraged mutual understanding, using references to people, places, or experiences familiar only to their human partner.

By contrast, such partner-specific cues were entirely absent from H–AI interactions. Instead, participants adopted more literal and generalized phrasing, likely driven by an awareness that AI lacks shared cultural or situational grounding. These shifts suggest that explainers adapt their communicative strategies based on their assumptions about the listener’s knowledge state.

Table 7 provides illustrative examples of this contrast. Whereas H–H clues often relied on implicit shortcuts rooted in shared context, H–AI clues were more explicit and elaborative—designed for interpretability rather than relational nuance. This highlights how interaction partner type can shape not only language complexity but also the pragmatic framing of speech.

Table 7. Partner-Specific Clue Strategies: Shared vs. Explicit References.

User	Condition	Transcript
user17	H–H	Miss K. and Mr. T. like this.
user16	H–H	People ride this when they go to BKC campus.
user1	H–AI	A car which can ride many person.
user4	H–AI	It is used when people play baseball and to catch balls.

8. Effect of Gameplay Order

To investigate how the sequence of interaction partners (AI or human) affects language performance, we analyzed participants’ CAF scores across both H–AI and H–H games. We split participants by their assigned condition order: those who played with the AI first (H–AI First) and those who started with a human partner (H–H First).

8.1. CAF Score Summary by Game Type and Condition Order

Table 8 presents the CAF scores broken down by both game type and the order in which participants experienced the human and AI conditions. The data suggest that the order in which participants interacted with AI or human partners had subtle but meaningful effects on their language output. In the H–AI games, those who encountered the AI first exhibited higher fluency ($M = 3.85$). Meanwhile, those who played H–AI after engaging with a human partner demonstrated slightly higher accuracy ($M = 3.13$).

In the H–H games, we observe a small boost in fluency and complexity when participants had previously interacted with AI. Participants in the H–AI First condition achieved a fluency score of 2.07 and complexity of 1.10, compared to 1.95 and 1.03, respectively, for the H–H First group. This pattern suggests a potential transfer benefit, where the experience of explaining concepts to a consistent and patient AI listener may help participants to speak more fluently in later peer interaction.

Table 8. CAF Scores by Game Type and Condition Order.

Game Type	Condition Order	Complexity	Accuracy	Fluency
H–AI	H–AI First	1.02	2.95	3.85
H–AI	H–H First	1.05	3.13	3.64
H–H	H–AI First	1.10	2.68	2.07
H–H	H–H First	1.03	2.80	1.95

8.2. Qualitative Comparison of H–H Games by Condition Order

To better understand the numerical trends in fluency and complexity, we examined specific utterances from H–H games. Table 9 presents real examples from participants in both condition groups: those who interacted with AI first and those who began with a human partner.

Table 9. Sample H–H Game Transcripts by Condition Order.

User	Condition	Transcript	C	A	F
user6	H–AI First	Ok. We can use the AI in that way. To eat bread, to toast...	2.0	3.0	4.0
user7	H–AI First	It’s used in Thailand. This is ride. Not place.	1.0	3.0	4.0
user17	H–H First	place	1.0	1.0	1.0
user11	H–H First	No. Uh... Give in... Uh... Body... Inside.	1.0	1.0	1.0

8.3. Transfer Benefits from AI to Human Interaction

These qualitative examples provide further support for the observed differences in CAF scores by condition order. Participants who completed H–AI first produced longer and more structured utterances when subsequently engaging in H–H gameplay. For instance, user6 constructed a multi-part sentence with clear intent, while user7 combined location, action, and exclusion logic—despite grammatical imperfections. This suggests that interacting with the AI may have helped participants rehearse or internalize clearer communicative strategies before switching to a human partner.

In contrast, participants who began with H–H often displayed lower fluency and limited elaboration, as shown by user17’s single-word clue and user11’s fragmented, hesitant delivery. These patterns reinforce the idea that AI interaction can scaffold more confident and structured speaking, potentially due to its predictable and non-threatening feedback environment.

Such evidence strengthens the pedagogical implication that AI-based warm-up tasks might enhance learners’ readiness for peer interaction. Sequencing partner types not only influences linguistic performance quantitatively but also shapes the qualitative nature of student explanations.

Design Implication

Taken together, these findings imply that gameplay order can influence linguistic outcomes. Starting with AI may boost learner fluency and expressive confidence, whereas beginning with a human may sharpen accuracy for later interaction. These patterns carry practical relevance for classroom design, suggesting that AI-supported warm-up sessions could serve as a productive scaffold before peer-based speaking tasks. Taken together, these findings imply that gameplay order can influence linguistic outcomes. Starting with AI may boost learner fluency and expressive confidence, whereas beginning with a human may sharpen accuracy for later interaction. These patterns carry practical relevance

for classroom design, suggesting that AI-supported warm-up sessions could serve as a productive scaffold before peer-based speaking tasks.

9. Learner Perceptions of AI and Human Partners

In response to RQ3, this section analyzes participant perceptions and comments to better understand how learners perceived the AI and human partners.

To better understand learner experiences during the experimental tasks, we administered a post-experiment questionnaire to all 18 participants. The instrument was designed to assess perceptions of both AI-based and human-based language learning. Each item was presented as a paired statement—one referring to AI interaction and the other to human interaction—covering topics such as comfort, vocabulary acquisition, perceived benefits, and learning confidence.

Participants responded using a five-point Likert scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (5). Table 10 summarizes the questionnaire items used in the study.

Table 10. Paired questionnaire items comparing AI-based and human-based language learning.

No.	AI-Based Learning	Human-Based Learning
Q1	AI-based language learning required no concern for politeness.	Human-based language learning required no concern for politeness.
Q2	I want to continue AI-based language learning.	I want to continue human-based language learning.
Q3	AI-based language learning was relaxing.	Human-based language learning was relaxing.
Q4	Using the AI game helped me acquire related vocabulary.	Learning with a person helped me acquire related vocabulary.
Q5	AI-based language learning was beneficial.	Human-based language learning was beneficial.
Q6	I was not embarrassed to make mistakes during AI-based learning.	I was not embarrassed to make mistakes during human-based learning.
Q7	The AI game was easier than expected.	Learning with a person was easier than expected.
Q8	I gained confidence in learning language through the AI game.	I gained confidence in learning language through the human partner.
Q9	I became better at explaining through the AI game.	I became better at explaining through learning with a person.
Q10	I felt a sense of accomplishment from the AI game.	I felt a sense of accomplishment from learning with a person.

9.1. Quantitative Findings

Because the ten questionnaire items are single 5-point Likert responses (ordinal), we analyzed paired differences with the *Wilcoxon signed-rank test*; continuous performance measures reported elsewhere were kept under *t*-tests after normality checks.

Participants completed ten matched Likert-scale items that probed comfort, perceived benefit, and affect when learning with an *AI partner* vs. a *human partner*. Table 11 reports descriptive means and Wilcoxon results.

Learners reported significantly *less social pressure* and felt *less embarrassed* when practicing with the AI partner ($p < 0.05$). A marginal trend indicated they also felt more relaxed

($p = 0.050$). All other aspects showed no statistically significant difference, although mean scores were similar across modes.

Table 11. Mean Ratings and Wilcoxon Signed-Rank Results for AI- vs. Human-Partner Learning Items ($N = 18$).

Question (Q1–Q10)	AI Mean	Human Mean	W	p-Value
Q1. No social pressure	4.33	3.44	15	0.036
Q2. Want to continue	4.06	4.17	39.5	0.701
Q3. Feel relaxed	4.50	3.67	17	0.050
Q4. Vocabulary acquisition	4.22	4.22	9	0.834
Q5. Beneficial	4.33	4.33	9	0.834
Q6. Not embarrassing	4.78	3.61	2	0.007
Q7. Easier than expected	3.17	2.94	23	0.398
Q8. Built confidence	3.72	3.78	9	0.834
Q9. Better at explaining	3.89	3.83	12	0.441
Q10. Felt accomplished	4.06	4.22	12	0.441

Figure 11 confirms the pattern observed in Table 11. The largest mean gaps favoring the AI partner occur on *No social pressure* ($\Delta M = 0.89$) and *Not embarrassing* ($\Delta M = 1.17$), the only items reaching statistical significance. A smaller but noticeable gap appears for *Feel relaxed*; its error bars overlap slightly, reflecting the marginal $p = 0.050$. For the remaining seven items, the AI and human means are nearly identical, and their error bands heavily overlap, indicating no reliable difference in perceived benefit, vocabulary gain, confidence, or task ease.

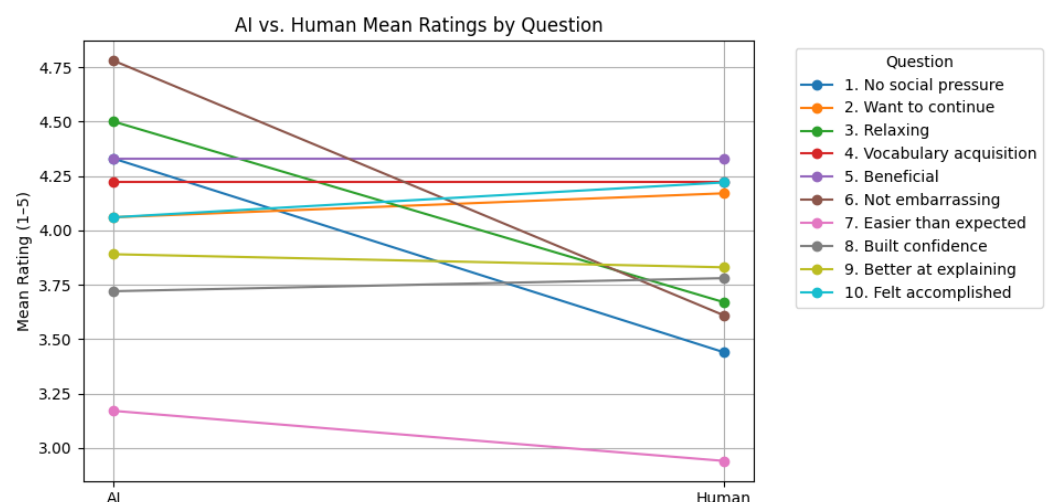


Figure 11. Mean ratings for each item comparing AI and human learning partners. Error bars represent ± 1 SD.

9.2. Open-Ended Feedback and Thematic Analysis

For the open-ended responses, we asked participants to reflect on their experience by answering the following two questions:

1. How did you feel about participating in the AI game? Please describe in detail.
2. Compared to learning a language with a person, how did you feel about learning with the AI game? Please describe in detail.

Both the original questions and the free-form responses were written in Japanese.

9.2.1. Thematic Analysis

Thematic analysis is a widely used method in HCI research for analyzing free-writing style questionnaire feedback. According to Bowman et al. [53], thematic analysis (TA) refers to a range of flexible and evolving approaches for qualitative data analysis, and its use has been increasing in HCI research.

To analyze participants' open-ended responses, we employed a four-theme coding scheme. The categories—*cognitive*, *emotional*, *social*, and *technical*—were selected to capture key dimensions of learner experience in AI-mediated speaking tasks. This framework allowed us to identify how learners reflected on language strategies and grammar awareness (cognitive), expressed enjoyment or frustration (emotional), commented on interpersonal dynamics with AI vs. human partners (social), and raised issues related to AI performance such as speech recognition and feedback quality (technical).

9.2.2. Results of Thematic Analysis

To report the results of the thematic analysis, we examined all 36 open-ended responses provided by 18 participants—each responding to two questions. Responses were coded using the four-theme scheme: *cognitive*, *emotional*, *social*, and *technical*. The results revealed that **cognitive** themes were the most frequently mentioned overall, appearing in 44.4% of all responses, reflecting participants' attention to grammar, vocabulary use, and metacognitive reflection. **Social** themes followed closely (41.7%), particularly in Q2, where participants compared AI to human interlocutors in terms of communication style and interaction quality. **Emotional** themes were also prominent (38.9%), with learners expressing enjoyment, reduced anxiety, and motivational aspects. **Technical** themes (33.3%) appeared across both questions, often relating to speech recognition limitations and system feedback. The table below summarizes the distribution of themes across both questions.

Table 12 presents the thematic breakdown of responses to Question 1, which asked participants how they felt about participating in the AI game. The examples provide insight into learners' cognitive reflections, emotional reactions, social comparisons, and technical feedback.

Table 12. Theme Analysis of Q1 Responses with Representative Examples (N = 18).

Theme	Explanation and Representative Comments (English)	# of Users
Cognitive	Learners reflected on language learning (grammar, vocabulary, clarity, thinking time, noticing).	7
	• “I tried harder to explain in sentences with the AI.” (User 5)	
	• “The AI used vocabulary well, which helped me learn grammar.” (User 11)	
	• “The AI used concise and clear structures. Just memorizing those helps.” (User 12)	
Emotional	Emotions such as fun, surprise, shock, enjoyment, or embarrassment.	8
	• “It felt like a game, so it was fun.” (User 3)	
	• “It felt relaxed, but I was shocked at how poorly I performed.” (User 17)	
	• “I was surprised how smart the AI was.” (User 1)	

Table 12. Cont.

Theme	Explanation and Representative Comments (English)	# of Users
Social	Social comparison with human interaction, freedom from embarrassment, communication freedom.	5
	• “I could speak freely without worrying.” (User 10) • “I felt bad when I couldn’t lead a human to the answer.” (User 13) • “It felt like I was chatting.” (User 18)	
Technical	Comments on AI accuracy, speech recognition, system praise/feedback.	6
	• “It didn’t catch my pronunciation well.” (User 8) • “The AI understood even my incorrect English.” (User 16) • “Its responses were harder to understand than a human’s.” (User 4)	

Table 13 shows the thematic analysis for responses to Question 2, which asked participants to compare language learning with the AI game to learning with a human partner. The responses highlight participants’ perceptions of cognitive demand, emotional comfort, social engagement, and technical affordances of AI-mediated interaction.

Table 13. Theme Analysis of Q2 Responses with Representative Examples (N = 18).

Theme	Explanation and Representative Comments (English)	# of Users
Cognitive	Learners reflected on grammar awareness, sentence planning, learning strategies, and skill development.	9
	• “With the AI, I had to form correct sentences. That made me think more carefully.” (User 7) • “To convey meaning precisely, grammar is really important.” (User 12) • “From the second round, I wanted to increase the number of sentences.” (User 13)	
Emotional	Users expressed enjoyment, pressure, embarrassment, motivation, or low anxiety.	6
	• “It was easier with AI and felt less stressful.” (User 3) • “With humans, I felt nervous around strangers.” (User 4) • “I felt guilty when my English wasn’t understood.” (User 18)	
Social	Comparison between human–human and AI interactions; comments on shared understanding, gestures, feedback, and naturalness.	10
	• “With people, we share gestures and common knowledge, which made it easier.” (User 16) • “Facial expressions helped when playing with humans.” (User 18) • “AI lacked the sense of accomplishment I got with people.” (User 8)	

Table 13. Cont.

Theme	Explanation and Representative Comments (English)	# of Users
Technical	Feedback on AI prediction, hints, understanding, or rigid behavior.	6
	• “AI guessed easily, which made it convenient.” (User 3)	
	• “Sometimes the AI didn’t give strong enough hints.” (User 2)	
	• “AI needed accurate language, while people could understand even rough speech.” (User 9)	

9.3. Post-Task Peer Debriefing

After completing all Taboo game tasks, participants engaged in a reflective, unmoderated conversation with their partner. These *post-task peer debriefings* were audio-recorded, transcribed, and analyzed using inductive qualitative content analysis. As participants were not given any specific prompts, they were free to reflect on any aspect of the experience. Through open coding of the transcripts, we identified recurring patterns and grouped them into three emergent dimensions: **(1) communication experience**, **(2) learning perception**, and **(3) emotional response**. Each dimension is described below with illustrative quotes from participants.

9.3.1. Dimension 1: Communication Experience

Theme: Difficulty Expressing in English

Many participants reported difficulty expressing their thoughts in English during the game, even when the ideas were clear in Japanese. This gap between receptive understanding and productive output led to frequent hesitation.

“I had to think a lot. I could come up with Japanese, but I couldn’t come up with English at all.”

“Even if I understand it in English, when it comes time to speak, I end up thinking in Japanese.”

Theme: Differences Between AI and Human Partners

Participants also reflected on the nature of interaction with their two different partners. AI was often described as clear and structured, while human peers were considered more intuitive, fun, and flexible.

AI as structured but rigid:

“The way the AI explains how to guide someone to a word they don’t know was pretty easy to understand.”

“AI guessed ‘salad’ from ‘this is on dressing.’ That surprised me.”

Humans as intuitive and emotionally supportive:

“It was fun playing the guessing game with a friend.”

“Since we know each other, we can kind of guess what the other person is trying to say.”

AI supported clarity and vocabulary modeling, while human partners fostered emotional safety and flexible collaboration.

9.3.2. Dimension 2: Learning Perception

Theme: Learning Through Repetition

Many participants felt that their fluency improved through repeated retelling. Successive attempts helped them retain vocabulary and gain confidence in output.

“I learned it well the more I did retelling.”

“After doing it two or three times, I felt like I could say a little more.”

Theme: Learning from AI’s Explanation Style

Several learners noted that the way the AI gave clues helped them reflect on how to better structure their own speech.

“I realized that if I explained things like that, I could make myself better understood.”

AI served not only as a conversational partner but also as a model of communicative strategy that participants sought to imitate.

9.3.3. Dimension 3: Emotional Response

Theme: Frustration and Self-Doubt

Some participants expressed negative emotions such as embarrassment, self-criticism, or shame when they felt unable to express themselves as clearly as they wanted.

“It’s pathetic. If this goes public, I’ll be ashamed.”

“I feel sorry for the AI.”

These emotional reflections underscore the vulnerability learners experience in real-time speaking tasks, even in game-based environments.

10. Discussion

The findings from this study reveal a nuanced picture of how learners interact with AI and human partners in speaking tasks and how those interactions shape both linguistic performance and user experience. This section synthesizes key insights across the retelling, game, and perception tasks to offer broader pedagogical and design implications.

10.1. Differential Benefits of AI and Human Interaction

Results across both the retelling and Taboo game tasks suggest that AI and human partners offer distinct yet complementary benefits. These learners demonstrated greater gains in fluency and conciseness when engaging with AI, likely due to the reduced social pressure and the AI’s predictable, turn-based response style. Meanwhile, human partners seemed to afford richer scaffolding for clue success rate, including clarification requests and mutual adjustment during breakdowns.

These patterns suggest a hybrid learning approach: learners can warm up or build fluency in low-pressure AI interactions, then shift to human-peer tasks to deepen accuracy and collaborative meaning-making.

10.2. Partner-Specific Adaptation and Shared Context

The study also highlights how learners adapt their communicative strategies based on the nature of their interaction partner. In human–human (H–H) games, participants often relied on shared cultural references, linguistic shortcuts, or physical context (e.g., pointing to objects in the room or referencing the experimental setting). For instance, some learners used locally relevant terms—such as “Fuji” to explain the word “mountain”—or drew on concepts embedded in their shared background to guide the guesser. This flexibility

enabled more creative and contextually grounded clue-giving, even when it occasionally deviated from the target vocabulary domain.

In contrast, interactions with the AI partner prompted more literal and structured explanation styles. Learners tended to avoid informal references or indirect hints, instead adopting generalized, textbook-like phrasing. This partner-specific adaptation reflects not only linguistic awareness but also metacognitive control over strategy selection—an important skill in second language acquisition.

10.3. *H–AI Game as a Learning Opportunity*

While much emphasis is placed on learners benefiting from AI-generated feedback, our findings suggest that AI can also serve as a learning tool when it plays the *guesser* role. As participants attempted to make their clues understood by the AI, they had to carefully calibrate their language—choosing clear, unambiguous words and avoiding culturally-specific shortcuts. Through this process, learners reported discovering more effective ways to explain, which might support the result that playing with AI first make them fluent for the later games, comparing to the learners who played with human first. They also reported that they became more aware of their own communication habits. This aligns with peer teaching literature, which shows that explaining concepts to others reinforces one's own understanding.

10.4. *System Design Insights: Turn-Taking and Multimodal Cues*

While the overall system supported engaging gameplay, user reflections revealed several areas for refinement. Participants noted that AI responses felt overly mechanical or rigid, especially in longer exchanges. This highlights the importance of simulating *human-like turn-taking*, including pauses, backchanneling, and clarification requests—elements that make conversation feel interactive rather than transactional.

A promising avenue for future development lies in multimodal feedback. Several learners attempted gestures or eye contact during human interactions, suggesting that visual cues were integral to meaning making. Integrating gesture recognition or visual avatars may enable the AI to more fully participate in embodied communication, aligning with how language is used in authentic contexts.

10.5. *Emotional and Motivational Factors*

The affective dimension of interaction played a significant role in shaping outcomes. Learners reported feeling less judged and more willing to make mistakes with AI partners, contributing to increased fluency and longer retellings. This supports the use of AI as a psychologically safe practice partner, particularly for shy or anxious learners. On the other hand, several participants expressed greater satisfaction and emotional engagement when playing with human peers, especially when shared laughter or successful guessing occurred. These experiences underscore the role of emotional reward in sustaining motivation [38].

10.6. *Toward Context-Aware Language Learning Systems*

Taken together, these insights suggest that language learning tools should not treat human and AI interaction as interchangeable. Instead, systems should be designed to amplify the unique strengths of each partner type. One direction is to create context-aware systems that adjust difficulty or feedback style based on user behavior, emotional state, or task history. Another is to build multi-agent interfaces that combine AI guidance with peer collaboration, leveraging social dynamics for engagement while retaining AI's reliability and scalability.

Overall, the findings reaffirm that AI-powered tools like Taboo Talks can serve not only as substitutes for human interaction but as strategic supplements. When thoughtfully

designed, they can encourage experimentation, reduce anxiety, and guide learners toward more effective communication strategies in English.

10.7. Limitations and Future Directions

While this study provides promising insights into integrating a Taboo-style word guessing game with speech recognition and LLMs for EFL speaking practice, several limitations must be acknowledged, which, in turn, inform directions for future research and system development.

First, although the sample size is modest, it aligns with accepted standards in user experience (UX) research and mixed-method studies in educational technology. Prior work in language learning system evaluation has shown that sample sizes of 5 to 20 participants are sufficient for identifying usability challenges and producing rich qualitative insights [54,55]. However, to rigorously assess learning outcomes and improve generalizability, future work should include a larger and more diverse participant cohort, enabling statistical comparisons between human–human (H–H) and human–AI (H–AI) interaction conditions.

Second, this study did not control for interpersonal dynamics between paired participants. Approximately half of the pairs in the H–H condition were acquainted before the study. Familiarity among peers may influence engagement, turn-taking, and explanation strategies. Additionally, gender pairing (e.g., male–female vs. same-gender dyads) may affect communication style and comfort level. Prior studies have found that both gender composition and social familiarity can influence language use, particularly in collaborative or communicative tasks [56]. Future work should systematically examine these sociolinguistic variables to determine their impact on learner interaction and outcomes.

Third, while the speech recognition system (Whisper) performed well in recognizing Katakana-influenced English, this raises a question of target intelligibility. Although such recognition benefits Japanese learners by reducing frustration, it may inadvertently reinforce pronunciation patterns that are not easily understood by native or international English speakers. This tension highlights the need to balance local comprehensibility with broader goals of global intelligibility [57].

Looking ahead, several system enhancements are envisioned:

- **Expanded participant demographics:** Future studies will include a larger and more demographically varied sample, capturing metadata such as gender, prior acquaintance, and communication preferences.
- **Pronunciation support features:** Incorporating visual or audio pronunciation models and feedback mechanisms to assist learners in refining their articulation.
- **Lexical and fluency evaluation tools:** Providing automated feedback on learners' choice of vocabulary, clarity of explanation, and fluency.
- **AI with facial expressivity:** Testing whether embodied agents or those with facial expressions foster higher engagement and emotional comfort compared to text-only or neutral avatars.

By addressing these limitations and building upon the observed user needs, future iterations of the Taboo system aim to deliver more effective, inclusive, and pedagogically robust speaking practice tools for EFL learners.

11. Conclusions

This study explored the impact of interacting with AI vs. human partners on English learners' speaking performance, interactional behavior, and user perceptions through a game-based learning task. Based on a within-subject design combining retelling evaluations and gameplay transcript analysis, three key insights emerged.

First, regarding speaking performance (RQ1), learners demonstrated modest gains in fluency and complexity in their story retelling after engaging with the AI partner, particularly among those who experienced AI interaction first. In contrast, accuracy scores remained relatively stable across trials, suggesting that while AI-mediated game-play may support more fluent and structurally varied output, it has limited influence on grammatical precision.

Second, the interactional analysis (RQ2) revealed clear differences in conversational structure between partner types. Human–AI sessions were characterized by more structured exchanges. Human–human sessions, on the other hand, displayed more spontaneous and collaborative patterns, including overlaps and backchannels. These contrasting dynamics reflect the differing affordances of AI and human interlocutors in facilitating language use.

Finally, learner reflections (RQ3) highlighted mixed attitudes toward AI partners. While many appreciated the reduced anxiety and judgment-free atmosphere when speaking with AI, they also noted its limited adaptability and responsiveness. Human partners were perceived as more dynamic and context-aware, though some learners found the social pressure more intimidating.

Together, these findings suggest that AI-mediated interaction can serve as a complementary tool in speaking practice—particularly for building fluency and structural range—while human interaction remains essential for richer communicative engagement and responsiveness. Future work may consider optimizing AI scaffolding strategies to more closely emulate the flexibility and feedback of human partners.

Author Contributions: Conceptualization, M.P., M.T., Y.M. and J.S.W.; data curation, M.P.; formal analysis, M.P.; funding acquisition, J.S.W. and M.P.; investigation, M.P. and M.T.; methodology, M.P., M.T. and Y.M.; project administration, M.T.; resources, J.S.W.; software, M.P.; supervision, Y.M.; validation, M.P.; visualization, M.P.; writing—original draft preparation, M.P.; writing—review and editing, M.P., M.T., Y.M. and J.S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Early-Career Scientists, Project Number 21K17794, and by the AY2024 Grassroots Practice Support Program of Ritsumeikan University.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the nature of the research, which involved voluntary participation by university students in a non-invasive language learning activity. The study did not collect sensitive personal data and posed minimal risk to participants. It was conducted in accordance with institutional guidelines at Ritsumeikan University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Participants received written information and consented to data collection and analysis. They were compensated at the standard hourly minimum wage of Osaka Prefecture in the form of gift cards.

Data Availability Statement: The data supporting the findings of this study—including annotated Taboo game logs, CAF scores, and questionnaire responses—are available from the corresponding author upon reasonable request. Due to participant privacy considerations, public sharing is restricted.

Acknowledgments: Special thanks to Barry Condon, at Ritsumeikan University, for refining and adapting the taboo word dataset used in this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
CALL	Computer-assisted language learning
ESL	English as a second language
H-AI	Human–AI (Human interacting with AI)
H-H	Human–human (Human interacting with human)
CAF	Complexity, accuracy, and fluency
LLM	Large language model
TAM	Technology acceptance model
TOEIC	Test of English for international communication
UI	User interface

References

1. Elbashir, E.E. The Challenges of Spoken English Fluency among EFL Learners in Saudi Universities. *Sch. Int. J. Linguist. Lit.* **2023**, *6*, 280–282. [\[CrossRef\]](#)
2. Xiao, F.; Zhao, P.; Sha, H.; Yang, D.; Warschauer, M. Conversational agents in language learning. *J. China-Comput.-Assist. Lang. Learn.* **2024**, *4*, 300–325. [\[CrossRef\]](#)
3. Nadeem, M.; Oroszlanyova, M.; Farag, W. Effect of Digital Game-Based Learning on Student Engagement and Motivation. *Computers* **2023**, *12*, 177. [\[CrossRef\]](#)
4. Pituxcoosuvann, M.; Radhapuram, S.C.T.; Murakami, Y. Taboo Talks: Enhancing ESL Speaking Skills through Language Model Integration in Interactive Games. *Procedia Comput. Sci.* **2024**, *246*, 3674–3683. [\[CrossRef\]](#)
5. Skehan, P. *A Cognitive Approach to Language Learning*; Oxford University Press: Oxford, UK, 1998.
6. Ellis, R. *Task-Based Language Learning and Teaching*; Oxford University Press: Oxford, UK, 2003.
7. Tessmer, M. *Planning and Conducting Formative Evaluations*; Routledge: Oxfordshire, UK, 2013.
8. Long, M.H. The Role of the Linguistic Environment in Second Language Acquisition. In *Handbook of Second Language Acquisition*; Ritchie, W.C., Bhatia, T.K., Eds.; Academic Press: Cambridge, MA, USA, 1996; pp. 413–468.
9. Swain, M. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In *Input in Second Language Acquisition*; Gass, S., Madden, C., Eds.; Newbury House: Cork, Ireland, 1985; pp. 235–253.
10. Leeming, P.; Aubrey, S.; Lambert, C. Collaborative Pre-Task Planning Processes and Second-Language Task Performance. *RELJ* **2022**, *53*, 534–550. [\[CrossRef\]](#)
11. Rassaei, E. The Interplay between Corrective Feedback Timing and Foreign Language Anxiety in L2 Development. *Lang. Teach. Res.* **2023**. [\[CrossRef\]](#)
12. Aubrey, S.; King, J.; Almukhailid, H. Language Learner Engagement During Speaking Tasks: A Longitudinal Study. *RELJ* **2022**, *53*, 519–533. [\[CrossRef\]](#)
13. Elverici, S.E. Mobile Assisted Language Learning: Investigating English Speaking Performance and Satisfaction. *Rumelide Dil Edeb. Araştırmaları Derg.* **2023**, *Ö13*, 1305–1317. [\[CrossRef\]](#)
14. Wachowicz, K.A.; Scott, B. Software That Listens. *CALICO J.* **2013**, *16*, 253–276. [\[CrossRef\]](#)
15. Tsai, S.C. Implementing Interactive Courseware into EFL Business Writing: Computational Assessment and Learning Satisfaction. *Interact. Learn. Environ.* **2018**, *27*, 46–61. [\[CrossRef\]](#)
16. Kim, H.K. Beyond Motivation. *CALICO J.* **2013**, *25*, 241–259. [\[CrossRef\]](#)
17. Parmaxi, A.; Zaphiris, P. Web 2.0 in Computer-Assisted Language Learning: A Research Synthesis and Implications for Instructional Design and Educational Practice. *Interact. Learn. Environ.* **2016**, *25*, 704–716. [\[CrossRef\]](#)
18. Mohsen, M.A. The Use of Help Options in Multimedia Listening Environments to Aid Language Learning: A Review. *Br. J. Educ. Technol.* **2015**, *47*, 1232–1242. [\[CrossRef\]](#)
19. Yigci, D.; Eryilmaz, M.; Yetisen, A.K.; Tasoglu, S.; Ozcan, A. Large Language Model-Based Chatbots in Higher Education. *Adv. Intell. Syst.* **2024**, *7*, 2400429. [\[CrossRef\]](#)
20. Zou, B.; Li, Q.; Luo, W. Supporting Speaking Practice by Social Network-Based Interaction in Artificial Intelligence (AI)-Assisted Language Learning. *Sustainability* **2023**, *15*, 2872. [\[CrossRef\]](#)
21. Gao, Y.; Nuchged, B.; Li, Y.; Peng, L. An Investigation of Applying Large Language Models to Spoken Language Learning. *Appl. Sci.* **2023**, *14*, 224. [\[CrossRef\]](#)
22. Rusmiyanto, R.; Huriati, N.; Fitriani, N.; Tyas, N.; Rofi'i, A.; Sari, M. The Role Of Artificial Intelligence (AI) In Developing English Language Learner's Communication Skills. *J. Educ.* **2023**, *6*, 750–757. [\[CrossRef\]](#)

23. Lucas, H.C.; Singh, A.; Kim, M.J.; Al-Twijri, R.; Chuang, T.H. A Systematic Review of Large Language Models and Their Implications in Medical Education. *Med. Educ.* **2024**, *58*, 1276–1285. [\[CrossRef\]](#)
24. Park, Y.J.; Pillai, A.; Deng, J.; Guo, E.; Gupta, M.; Paget, M.; Naugler, C. Assessing the Research Landscape and Clinical Utility of Large Language Models: A Scoping Review. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 72. [\[CrossRef\]](#)
25. Choudhury, A.; Chaudhry, Z. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskillling of Health Care Professionals. *J. Med. Internet Res.* **2024**, *26*, e56764. [\[CrossRef\]](#)
26. Jošt, G.; Taneski, V.; Karakatič, S. The Impact of Large Language Models on Programming Education and Student Learning Outcomes. *Appl. Sci.* **2024**, *14*, 4115. [\[CrossRef\]](#)
27. Haltaufderheide, J.; Ranisch, R. The Ethics of ChatGPT in Medicine and Healthcare: A Systematic Review on Large Language Models (LLMs). *NPJ Digit. Med.* **2024**, *7*, 183. [\[CrossRef\]](#)
28. Campbell, H.; Bluck, T.; Curry, E.; Harris, D.; Pike, B.; Wright, B. Should We Still Teach or Learn Coding? A Postgraduate Student Perspective on the Use of Large Language Models for Coding in Ecology and Evolution. *Methods Ecol. Evol.* **2024**, *15*, 1767–1770. [\[CrossRef\]](#)
29. Plass, J.L.; Homer, B.D.; Kinzer, C.K. Foundations of game-based learning. *Educ. Psychol.* **2015**, *50*, 258–283. [\[CrossRef\]](#)
30. Chowdhury, M.; Dixon, L.Q.; Kuo, L.J.; Donaldson, J.P.; Eslami, Z.; Viruru, R.; Luo, W. Digital game-based language learning for vocabulary development. *Comput. Educ. Open* **2024**, *6*, 100160. [\[CrossRef\]](#)
31. Zhou, S. Gamifying language education: The impact of digital game-based learning on Chinese EFL learners. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1518. [\[CrossRef\]](#)
32. Esteban, A.J. Theories, principles, and game elements that support digital game-based language learning (DGBLL): A systematic review. *Int. J. Learn. Teach. Educ. Res.* **2024**, *23*, 1–22. [\[CrossRef\]](#)
33. Lestari, S.D.; Damanik, E.S.D. Students' Perceptions on the Effectiveness of the Taboo Game in Enhancing English Vocabulary Acquisition. *Elsya J. Engl. Lang. Stud.* **2024**, *6*, 172–184. [\[CrossRef\]](#)
34. Agung, W.K.S. The Effectiveness of Taboo Game to Improve Students' Vocabulary Mastery. *ELTALL Engl. Lang. Teaching, Appl. Linguist. Lit.* **2023**, *4*, 88–98.
35. Yaacob, A.; Alsaraireh, M.Y.; Suryani, I.; Yulianeta, Y.; MdHussin, H. Effectiveness of Taboo Word Game on Augmenting Business Vocabulary Competency Through Reflective Action Research. *Arab. World Engl. J. (AWEJ)* **2024**, *15*, 267–281. [\[CrossRef\]](#)
36. Abusahyon, A.S.E.; Alzyoud, A.; Alshorman, O.; Al-Absi, B.A. AI-driven technology and chatbots as tools for enhancing English language learning in the context of second language acquisition: A review study. *Int. J. Membr. Sci. Technol.* **2023**, *10*, 1209–1223. [\[CrossRef\]](#)
37. Dennis, N.K. Using AI-powered speech recognition technology to improve English pronunciation and speaking skills. *IAFOR J. Educ. Technol. Educ.* **2024**, *12*, 107–123. [\[CrossRef\]](#)
38. Du, J.; Daniel, B.K. Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100230. [\[CrossRef\]](#)
39. Zhan, Z.; Tong, Y.; Lan, X.; Zhong, B. A systematic literature review of game-based learning in Artificial Intelligence education. *Interact. Learn. Environ.* **2024**, *32*, 1137–1158. [\[CrossRef\]](#)
40. N, M.; Kumar, P.N.S. Investigating ESL Learners' Perception and Problem towards Artificial Intelligence (AI)-Assisted English Language Learning and Teaching. *World J. Engl. Lang.* **2023**, *13*, 290. [\[CrossRef\]](#)
41. Flores, J. Using Gamification to Enhance Second Language Learning. *Digit. Educ. Rev.* **2015**, *27*, 32–54.
42. Klimova, B.; Pikhart, M.; Al-Obaydi, L.H. Exploring the potential of ChatGPT for foreign language education at the university level. *Front. Psychol.* **2024**, *15*, 1269319. [\[CrossRef\]](#)
43. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
44. Mcnamara, T.; May, L.; Hill, K. 12. Discourse and Assessment. *Annu. Rev. Appl. Linguist.* **2002**, *22*, 221–242. [\[CrossRef\]](#)
45. Nippold, M.A.; Frantz-Kaspar, M.W.; Vigeland, L.M. Spoken Language Production in Young Adults: Examining Syntactic Complexity. *J. Speech, Lang. Hear. Res.* **2017**, *60*, 1339–1347. [\[CrossRef\]](#)
46. Koizumi, R.; Hirai, A. Comparing the Story Retelling Speaking Test with Other Speaking Tests. *JALT J.* **2012**, *34*, 35–56. [\[CrossRef\]](#)
47. Uchida, S. Evaluating the Accuracy of ChatGPT in Assessing Writing and Speaking: A Verification Study Using ICNALE GRA. *Learn. Corpus Stud. Asia World* **2024**, *6*, 1–12.
48. Huang, Q.; Willems, T.; Wang, P.K. The application of GPT-4 in grading design university students' assignment and providing feedback: An exploratory study. *arXiv* **2023**, arXiv:2409.17698.
49. Hirunyasiri, D.; Thomas, D.R.; Lin, J.; Koedinger, K.R.; Aleven, V. Comparative Analysis of GPT-4 and Human Graders in Evaluating Praise Given to Students in Synthetic Dialogues. *arXiv* **2023**, arXiv:2307.02018.

50. Luger, E.; Sellen, A. “Like having a really bad PA”: The gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 5286–5297. [\[CrossRef\]](#)
51. Clark, L.; Radziwill, N.M. What makes a good conversation? Challenges in designing truly conversational agents. In Proceedings of the 2019 IEEE International Conference on Artificial Intelligence Testing (AITest), Newark, CA, USA, 4–9 April 2019; pp. 101–107. [\[CrossRef\]](#)
52. Clark, H.H. *Using Language*; Cambridge University Press: Cambridge, UK, 1996.
53. Bowman, R.; Nadal, C.; Morrissey, K.; Thieme, A.; Doherty, G. Using thematic analysis in healthcare HCI at CHI: A scoping review. In Proceedings of the Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–18.
54. Faulkner, L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behav. Res. Methods Instruments Comput.* **2003**, *35*, 379–383. [\[CrossRef\]](#)
55. Nielsen, J. Why You only Need to Test with 5 Users. Nielsen Norman Group. 2000. Available online: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/> (accessed on 30 March 2025).
56. Burleson, B.R.; Kunkel, A. Gender differences in communication: Implications for salespeople. *J. Pers. Sell. Sales Manag.* **2003**, *23*, 371–387.
57. Derwing, T.M.; Munro, M.J. What do ESL students say about their accents? *TESOL Q.* **2005**, *39*, 379–397. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.