**Evaluator: MN**
MN is an American-born Russian speaker who grew up in a Russian and English-speaking family. She has experience learning Japanese at UCLA and is a non-native Japanese speaker. Her background in multiple languages and formal education in Japanese provides her with a unique perspective on the tool's effectiveness for language learners, particularly those studying Japanese.

**User Experience:**
During her evaluation, MN encountered several challenges with the tool. She found the UI confusing, particularly the "Continue" and "Title" buttons, and noted the inability to use a username. After selecting Japanese as her target language, she struggled to differentiate between "roleplay" and "conversation" modes, suggesting that short descriptions for each would help clarify their purposes. The Japanese audio failed to play, and despite repeatedly clicking the sound button, she was unable to hear the pronunciation. The AI's responses felt repetitive and unnatural, often repeating phrases like "I am a dragon," which made the conversation feel stiff and unengaging. Additionally, the AI had difficulty understanding her inputs, reducing the interactivity of the experience. While MN, as a more advanced speaker, did not use the suggestion box much, she acknowledged its potential usefulness for others. She also found the lesson section less noticeable and suggested making it more prominent for better accessibility.

**Assessment Portion:**
MN rated the conversation and roleplay experience a **2 out of 5** for naturalness and quality, citing the AI's repetitive and unnatural responses as a major drawback. However, she rated the tool's potential helpfulness for language learning a **4 out of 5**, recognizing its promise despite the current limitations. She emphasized the need for improvements such as adding descriptions for modes, fixing audio playback, enhancing the AI's understanding and responsiveness, improving the UI for better navigation, and making the lesson section more accessible. Overall, while MN found the tool promising, she believes these changes would significantly enhance its usability and effectiveness for language learners.

**Evaluator: AP**

AP is a heritage speaker who evaluated the tool with the intent of testing its robustness and effectiveness. AP approached the tool with a focus on pushing its limits, including attempting to break it, testing its handling of inappropriate content, and evaluating its ability to provide natural and high-quality conversation practice.

**User Experience:**

During the evaluation, AP actively tried to break the system by inputting NSFW (not safe for work) content and testing the microphone feature, which she noted was not fully updated. She also explored the suggestion box and searched for various words to assess the tool's responsiveness and accuracy. AP appreciated the tool's ability to censor inappropriate content effectively, which she found to be a strong point. She also reviewed the lessons, describing them as "rather on point" but felt they were too general and could benefit from more specificity. Overall, AP found the tool to be functional but identified areas where it could be improved, particularly in terms of naturalness and handling edge cases.

**Assessment Portion:**

AP rated the conversation and roleplay experience a **4 out of 5**, noting that while the quality was very good, the interactions could feel slightly unnatural at times. She believed the tool would be highly effective for conversation practice and language immersion. For language learning helpfulness, AP gave the tool a **5 out of 5**, emphasizing its potential to provide valuable practice and immersion in a target language. She suggested that improving the naturalness of conversations and refining the lessons to be more specific would further enhance the tool's effectiveness.

**Evaluator: KB**

KB is a non-native learner of Chinese who evaluated the tool with a focus on its usability for beginners and intermediate learners. As someone still developing proficiency in Chinese, KB's feedback highlights the challenges faced by learners who may not yet have a strong grasp of the language, particularly in terms of accessibility, ease of use, and the tool's ability to cater to different skill levels.

**User Experience:**

KB encountered several issues during the evaluation. At the beginning, the Chinese language feature appeared to be broken, and the audio autoplay did not function as expected. While KB appreciated the dictionary feature, they noted that it lacked word-to-word translation, offering only sentence-level translations, which made it less useful for beginners. They also found the Chinese word definitions too lengthy and technical, requiring a level of expertise that made it difficult for less proficient learners to understand. KB emphasized the need for Pinyin in Chinese and Romaji in Japanese to aid pronunciation and comprehension. They found the conversations overly difficult, often lacking context and failing to answer their questions directly. Additionally, the tool unexpectedly switched to English at times, which was confusing. KB did not use the suggestion box much and expressed frustration at the lack of lessons and the inability to save progress, which they felt limited the tool's usefulness for ongoing learning.

**Assessment Portion:**

KB rated the conversation and roleplay experience a **3 out of 5**, citing their limited Chinese proficiency and the tool's unexpected switch to English as major drawbacks. However, they gave the language learning helpfulness a **4 out of 5**, acknowledging the potential of the dictionary feature while suggesting improvements such as word-to-word translation, Pinyin/Romaji support, and simpler conversation options for beginners. KB's feedback underscores the need for the tool to be more accessible and user-friendly for learners at lower proficiency levels, with clearer context, better language consistency, and additional features like lessons and progress-saving capabilities.

**Evaluator: JF**
JF is a heritage Chinese speaker with strong proficiency in the language. During the evaluation, she focused on testing the tool's ability to handle real-world scenarios, particularly in a restaurant setting. Her feedback highlights the importance of role clarity, scenario depth, and the tool's ability to maintain natural and engaging conversations.

**User Experience:**
JF chose the **Scenario Roleplay** option, specifically a restaurant ordering scenario. She noted that while the tool was helpful for practicing specific tasks like ordering food, there were several areas for improvement. JF found the AI's replies too long, which felt unnatural for a casual conversation. She also mentioned that the lack of auto-play for audio meant she didn't use the sound feature. JF expressed confusion about the difference between "conversation" and "scenario" modes, especially since the user wasn't assigned a clear role in the scenario. She felt that without a defined role or additional context (like a menu for a restaurant scenario), the conversation felt directionless and often ended too quickly. JF also tried customizing the scenario but found that typing in a different language broke the roleplay. While the suggestion feature was available, it didn't always work as expected. JF suggested that scenarios should be more fine-grained and that the conversation should have the option to end naturally when appropriate.

**Assessment Portion:**
JF rated the **naturalness and quality** of the conversation a **4 out of 5**, noting that while the quality of the responses was good, the lack of context and repetition made the conversation feel less natural over time. However, she gave the **helpfulness for language learning** a **5 out of 5**, praising the ability to create custom scenarios. JF found the tool particularly useful for practicing specific, goal-oriented tasks like ordering food, which she felt prepared her for real-world situations. She emphasized that the tool works well for emergent scenarios where the user knows what they want to practice, but improvements in role clarity, scenario depth, and natural conversation flow would enhance the experience further.

**Evaluator: NZ**
NZ is a heritage Chinese speaker with strong proficiency in the language. During the evaluation, he focused on testing the tool's conversational capabilities, particularly in casual and topic-specific scenarios like discussing Pokémon. His feedback highlights the importance of emotional engagement, topic focus, and the tool's ability to maintain context during conversations.

**User Experience:**
NZ began with conversation practice and tried using the microphone feature, but it didn't work for Chinese, so he couldn't test it. He used the suggestion box occasionally but felt it made the conversation seem like "two AIs talking to each other," which reduced the naturalness. NZ attempted to discuss Pokémon but found the AI's responses lacking emotion and depth—for example, it didn't express a favorite Pokémon, which made the conversation feel flat. He also noted the inability to send pictures, which could have enhanced the discussion. NZ found the lessons unsatisfying and was disappointed that progress couldn't be saved, especially when switching between scenarios and lessons. He observed that the AI often drifted from the current scenario during roleplay, breaking the immersion. Additionally, when he returned to the lessons, he found that his progress hadn't been saved, which was frustrating.

**Assessment Portion:**
NZ rated the **naturalness and quality** of the conversation a **3 out of 5**, citing the lack of emotional engagement, topic focus, and the AI's tendency to drift from the scenario as key issues. He gave the **helpfulness for language learning** a **3 out of 5**, noting that the tool didn't effectively help with pronunciation or vocabulary building. NZ suggested that the tool could improve by allowing users to select their language proficiency level, which would help the AI adjust its responses accordingly. He also recommended providing users with a list of words or topics to learn, rather than offering generic suggestions. Overall, NZ felt the tool had potential but needed significant improvements in maintaining context, emotional engagement, and providing structured learning opportunities.

### Evaluator: TL

TL is a non-native learner of Japanese with intermediate proficiency. She has been studying Japanese for two years and is particularly interested in improving her conversational skills for everyday interactions.

**User Experience:**
TL chose Japanese as her target language and started with the **conversation mode**. She appreciated the suggestion box, which helped her when she struggled to find the right words, but felt that the AI's responses were often too formal and not reflective of casual, everyday Japanese. TL tried discussing hobbies and daily routines but found the AI's replies repetitive and lacking depth. She also noted that the audio playback for Japanese was inconsistent. Sometimes it worked, but other times it didn't, which made it difficult to practice pronunciation. She explored the lesson section but felt it was too basic and didn't align with her intermediate level. TL suggested that the tool could benefit from more varied conversation topics, better audio functionality, and lessons tailored to different proficiency levels.

**Assessment Portion:**
TL rated the **naturalness and quality** of the conversation a **3 out of 5**, citing the AI's overly formal and repetitive responses as a drawback. She gave the **helpfulness for language learning** a **4 out of 5**, acknowledging the potential of the suggestion box but emphasizing the need for better audio functionality and more advanced lessons. TL recommended adding more casual conversation examples and improving the microphone feature to better support speaking practice.

### Evaluator: RK

RK is a native Chinese speaker with beginner-level proficiency in English. He is using the tool to improve his English for work-related conversations and travel.

**User Experience:**

RK selected English as his target language and opted for **scenario roleplay**, choosing a travel-related scenario. He found the AI's responses helpful initially but felt they were too fast-paced for his beginner level, making it difficult to keep up. RK appreciated the dictionary feature but was frustrated that it didn't provide word-to-word translations, only sentence-level translations, which made it hard for him to understand individual words. He also noted that the AI didn't always correct his grammar or pronunciation, which he felt was a missed opportunity for learning. He explored the lesson section but felt it was too generic and didn't address his specific needs, such as learning work-related vocabulary. RK suggested that the tool could improve by offering slower-paced conversations for beginners, word-to-word translations, and more targeted lessons for specific contexts like work or travel.

**Assessment Portion:**

RK rated the **naturalness and quality** of the conversation a **3 out of 5**, citing the AI's fast-paced responses and lack of grammar corrections as major issues. He gave the **helpfulness for language learning** a **3 out of 5**, acknowledging the potential of the scenario roleplay but emphasizing the need for slower-paced conversations, better translation features, and more context-specific lessons. RK recommended adding beginner-friendly options and improving the microphone's ability to recognize different accents.