# WordPlay: An Agent Framework for Language Learning Games

Anonymous ACL submission

## Abstract

We introduce a novel framework, WordPlay, for building language learning games tailored to a learner's proficiency level. WordPlay combines playful mini-puzzle games with large language models and text-to-image models to address the challenge of balancing engagement and effective language practice. We showcase the framework's adaptability by implementing a wide variety of language learning games with diverse learning objectives. We evaluate WordPlay's ability to target different proficiency level by conducting experimental sessions with English language learners. A fine-tuned BERT-based model rates the difficulty of both LLM-generated and user responses according to Common European Framework of Reference (CEFR) learning levels. Our results demonstrate that WordPlay successfully elicits learner output aligned with targeted proficiency levels.

## 1 Introduction

The importance of play in the learning process is well established (Gozcu and Çağda Kıvanç Çağanağa, 2016). However, modern language learning applications face a dilemma: typically, they either provide a playful environment without facilitating effective conversational practice or they provide structured language-based drills that lack the engaging elements essential for enthusiastic learning. This dichotomy is concerning because many learners experience an intermediate plateau in second language acquisition (SLA)—a phase where they perceive no progress regardless of their dedication and practice. This stagnation can be attributed to the increasing complexity of language and a decline in motivation (Mirzaei and Zoghi, 2017).

Game-play research demonstrates that adaptive experiences can counteract stagnation and boredom by maintaining player interest, especially when


Figure 1: WordPlay framework (left) and a hypothetical product experience (right) with our *Chicken crossing the road* puzzle.

challenges are scaled to the player's proficiency within their zone of proximal development (Cole et al., 1978; Hunicke and Chapman, 2004). Thus, integrating game-like elements in language learning can increase enjoyment (Vásquez and Ovalle, 2019) and potentially fortify motivation, yielding greater time-on-task.

Large language models (LLMs) are adept at shaping their outputs based on user input, often generating contextually pertinent—even if not strictly factual—responses. A widely adopted approach to mitigate LLMs' undesired outputs, like hallucinations, is to train them using human feedback or self-correction methods (Li et al., 2023) (Pan et al., 2023). This mirrors how humans often learn: through a cycle of trial, error, and correction. However, within the context of play, even LLM hallucinations can be channeled as creative assets rather than flaws.

LLMs have exhibited profound capabilities in reasoning and embodying characters. The power of prompting has been exemplified in recent works demonstrating a remarkable capability to simulate human-like behavior (Park et al., 2023). They excel as agents, proficiently executing tasks when given a reasoning structure and access to external resources, such as online search or calculators (Schick et al., 2023). Furthermore, they can assume the identity of specific characters, generating dialogue in alignment with that persona (Thoppilan et al., 2022). In the case of tutoring, past work has shown that intelligent tutoring systems are found to be as effective as human tutors (VanLehn, 2011). Recently, advances in LLMs have engendered abundant usage for language learning such as EnglishBot (Ruan et al., 2021) because they enable freeform interactions that leverage the persona effect (Lames et al., 1997).

Recent work has also shown text-to-image models to be capable of generating realistic and creative imagery, consistent with prompts (Ramesh et al., 2022) (Saharia et al., 2022). Images also provide a useful channel for learning, particularly when it comes to making conversation more comprehensible and aiding in acquisition of new terms/concepts. In fact, second language learners pay attention more to visual cues than native speakers (Felder and Henriques, 1995). However, due to their expense to generate by traditional means, images have often been underutilized in dynamic learning contexts.

In this work, we develop a novel framework, WordPlay, for authoring puzzle games that allow users rich conversational practice alongside acquisition of specific language structures. We use a tutor persona in each of our mini-puzzles and provide feedback on situational correctness of user responses. This framework uses an agentic approach and harnesses the recent capabilities of LLMs to orchestrate the puzzles, allowing authors to build new puzzles with only three prompts. WordPlay puzzles are intentionally tiered, catering to learners across a spectrum of proficiency levels, and are specifically designed to target specific learning criteria-all while being presented in digestible, easily accessible formats.

## 2 WordPlay Framework

The WordPlay framework generates engaging, adaptive mini-games using LLMs to judge situational proficiency and semantic acceptability, offering learners personalized and interactive experiences. We instantiate this framework across a variety of mini puzzle games that are aligned with diverse language function objectives and Common European Framework of Reference (CEFR) learning levels (CEF). The instructional content is relayed through a text-to-speech (TTS) system, with automatic speech recognition (ASR) (Zhang et al., 2023) accurately capturing and analyzing user response. Leveraging all these components creates an experience in which learners engage in listening, reading, and speaking, all of which are essential elements of language learning (Newton and Nation, 2020).

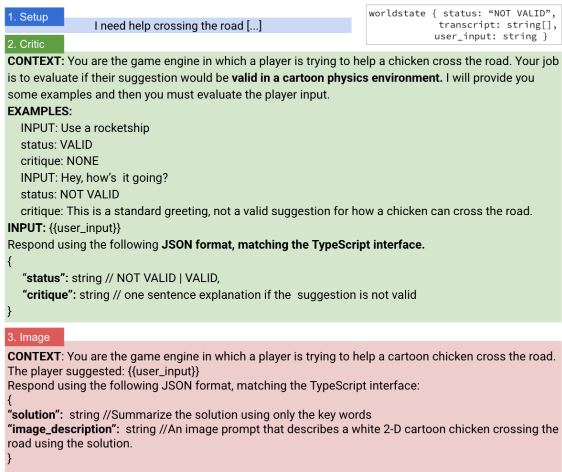



Figure 2: Verbatim prompts used in our *Chicken Crossing the Road* puzzle as well as the initial world-state at the beginning of the game.

WorldPlay consists of four agents guided by three prompts. The agents comprise a Setup agent, a Critic agent, an Input agent, and finally an Image agent shown in Figure 1. In the context of our running examples, our tutor assumes the persona of cartoon polar bear named Bearnard. We use PaLM (Chowdhery et al., 2022) for language generation and Imagen (Saharia et al., 2022) for image generation.

***Setup:*** The role of the Setup agent is to initialize the World-state with the necessary variables and to present the user with the first turn of dialogue. The dialogue can be predefined or generated by the LLM. This dynamic setup ensures replay-ability since each playthrough offers a unique experience.

***Input:*** The Input agent stops the execution of framework and waits for the user's input, which it then adds to the world-state.

***Critic:*** The Critic agent analyzes the conversation history to produce a response that either advances the dialogue or concludes the conversation under specific conditions, such as the occurrence of unsafe or explicit user utterances or the completion of the puzzle. This agent is responsible for the bulk of the puzzle orchestration, assessing the semantic appropriateness of the user's input by providing "critiques" for invalid inputs, determining the conclusion of the puzzle by setting the "status", and

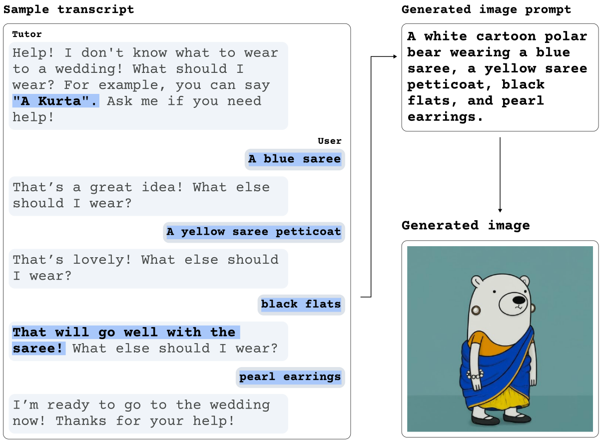



Figure 3: *Wedding* transcript, generated image prompt, and generated image.

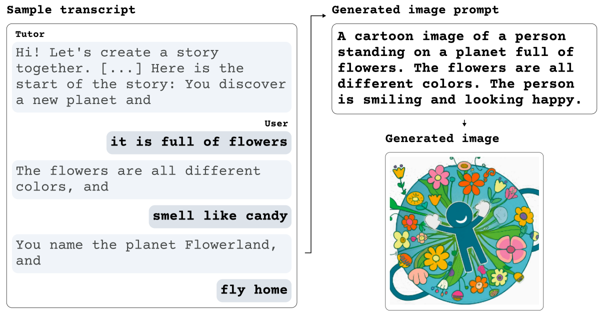



Figure 4: *Madlibs* transcripts, generated image prompt, and generated image.

responding to user inquiries.

***Image:*** The Image agent is responsible for generating prompts for image generation model and a summary of the solution to be displayed beneath the image. To maintain the safety of the image generation process, we employ the LLM to generate the image descriptions rather than allowing the user to do so. This introduces an additional layer of safety, leveraging the existing safeguards within the LLM concerning generated text. It's noteworthy that this prompt only uses zero-shot instructions to formulate the image description.

***World-state Management:*** Each agent returns its output in JSON format, circumventing the need for extensive parsing logic and enabling the direct updating of the world-state dictionary with the JSON object. This method serves as a streamlined approach for maintaining state and passing along only the essential information to downstream agents. During an evaluation of 69 user sessions across four puzzles, the system did not output any invalid JSON.

## 3 WordPlay Games

### 3.1 Beginner Puzzles

The CEFR framework is stratified into six distinct levels: A1/2, B1/2, and C1/2, with A1 denoting the most elementary user and C2 representing the highest level of proficiency. In our *Wedding* puzzle shown in Figure 3, designed for A2 to B1 learners, learners collaborate with a tutor character to select wedding attire, practicing contextual vocabulary recall and production. Appendix A.2 showcases similar puzzles created by changing the word "wedding" in the prompt to "opera" and "beach". This illustrates the scalability of the framework.

In each scenario, when the learner suggests attire or an accessory ill-suited for the chosen venue, the Tutor explains why the choice may not be appropriate. Apt suggestions are met with positive encouragement. The Critic agent, in Bearnard's role, determines when the outfit is complete, concluding the experience. This puzzle typically spans three to six turns of learner input.

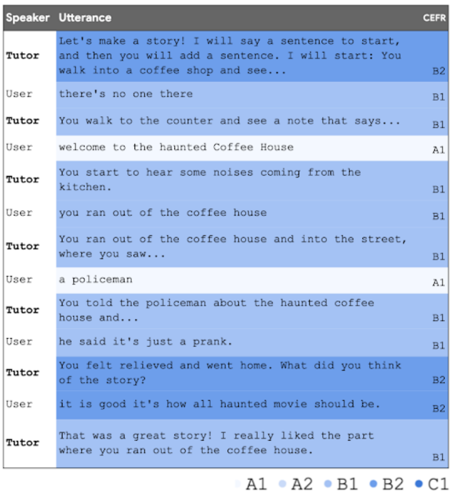
### 3.2 Intermediate Puzzles

Our intermediate puzzles are designed to induce longer spoken utterances from the learner, focusing on analytical and synthesis skills over recollection and reproduction of language. This focus is in line with the upper tiers of the Bloom Taxonomy, a hierarchical model of educational learning objectives (Anderson et al., 2001). These puzzles are tailored for learners at levels B1 through C1.

The *Messy Room Prepositions* puzzle focuses on practicing prepositions, which are essential grammatical structure at the intermediate CEFR level. Meanwhile, the *Finish the Story* shown in Figure 4 puzzle encourages learners to engage creatively by taking turns with our tutor to construct a narrative, thereby promoting the use of more complex sentence structures.

## 4 Evaluation

To evaluate the efficacy of our puzzles in aligning with the CEFR standards, we organized experimental sessions with native Hindi-speaking participants in India who were actively learning English. We hosted sessions involving several users, where each was tasked with solving a set of puzzles. We collected responses from 106 puzzle playthroughs. The responses—both from the participants and the model—were analyzed using a custom classification model. This model, an adaptation of a pre-



| Speaker | Utterance | CEFR |
|---|---|---|
| Tutor | Let's make a story! I will say a sentence to start, and then you will add a sentence. I will start: You walk into a coffee shop and see... | B2 |
| User | there's no one there | B1 |
| Tutor | You walk to the counter and see a note that says... | B1 |
| User | welcome to the haunted Coffee House | A1 |
| Tutor | You start to hear some noises coming from the kitchen. | B1 |
| User | you ran out of the coffee house | B1 |
| Tutor | You ran out of the coffee house and into the street, where you saw... | B1 |
| User | a policeman | A1 |
| Tutor | You told the policeman about the haunted coffee house and... | B1 |
| User | he said it's just a prank. | B1 |
| Tutor | You felt relieved and went home. What did you think of the story? | B2 |
| User | it is good it's how all haunted movie should be. | B2 |
| Tutor | That was a great story! I really liked the part where you ran out of the coffee house. | B1 |

A1 A2 B1 B2 C1

Figure 5: Utterance CEFR level predictions from a playthrough of a *Finish the Story* session.

existing BERT architecture (Devlin et al., 2019), uses a classification head on top of its initial layers and has been fine-tuned on a corpus of sentences, each associated with a CEFR level, as determined by experts in language education.

In Figure 5, we display the CEFR level annotations assigned by our model to a user's utterances from the *Finish the Story* puzzle. Notably, the bulk of the exchanges between the tutor and the learner remained within the B1-B2 range. This is indicative of the session's adherence to our design goal, which was to craft an intermediate-level puzzle suitable for learners at the B1 to C1 levels.

Figure 6 presents box plots that compare the predicted CEFR ratings of tutor and user utterances across three different puzzles: *wedding*, *messy room prepositions*, and *finish the story*. The data separates the language proficiency of tutors and users, categorizing them from A1 for beginners to C1 for advanced learners. The plots show a clear pattern in the tutor utterances, which consistently hit the B2 level, evidenced by the narrow interquartile ranges (IQR), indicating a targeted use of language that aligns with the intermediate level of language proficiency.

User utterances, however, show a much wider IQR, reflecting a greater variation in language proficiency levels. The *wedding* puzzle typically sees user responses at the A2 median level, while the *finish the story* puzzle displays a broader spread, with median between B1 and B2, suggesting that the latter puzzle challenges users with a more advanced language practice. This distinction underscores our puzzle design objectives: *wedding* is intended

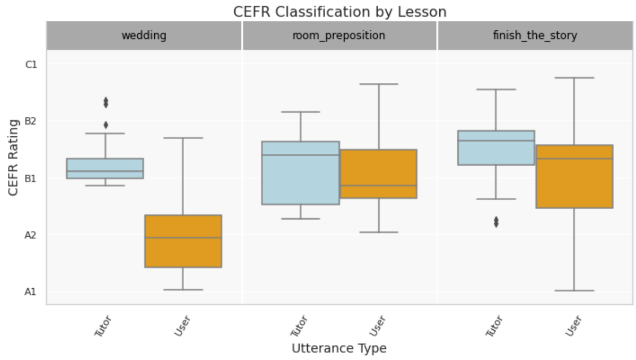



Figure 6: CEFR tutor and user utterances ratings for the *wedding*, *Messy room Prepositions*, and *Finish the Story* puzzles.

for beginners, whereas *finish the story* is aimed at engaging our most advanced learners.

Additionally, we include an annotated transcript from the *Taboo: Intermediate* puzzle depicted in Appendix Figure 16. A limitation of our classification model is highlighted when it comes to assigning accurate CEFR ratings to single words or brief phrases. For instance, the phrase "he's herbivorous" is categorized as B1, though it more likely corresponds to a C1 level. Such discrepancies can skew the median CEFR rating for user utterances downward, as shown in Appendix Figure 15, leading to potential inaccuracies in our assessment of the Taboo transcripts.

## 5 Conclusion

In conclusion, we present WorldPlay, a novel LLM-based framework for creating adaptive, engaging, mini-puzzles for second language acquisition. WordPlay addresses the dichotomy in SLA platforms, blending playful elements with structured, conversational practice to combat learning stagnation and boost motivation. By using an agentic framework, WorldPlay allows content creators to author engaging puzzles by customizing three simple prompts and adds a layer of delight and visual grounding by generating contextually relevant images. We demonstrate that WordPlay puzzles, when designed with language practice objectives, effectively promote language practice tailored to specific proficiency levels.

## 6 Limitations

WordPlay presents a promising framework, but there are inherent limitations to our analysis to consider. The fine-tuned BERT based CEFR classification model, while generally effective, may struggle with accurately rating single words or very short phrases. This highlights the challenge of automated language proficiency assessment. Furthermore the quality of and proficiency level alignment of WordPlay puzzles rely significantly on careful content curation and prompt engineering. Our success stems from collaboration with language learning experts; without this human-in-the-loop involvement, the resulting difficulty alignment might be less reliable. Finally, our experimental sessions were conducted with only Hindi-speaking participants in India who were actively learning English this may limit the generalizability of our evaluation. Future work should include learners from diverse linguistic backgrounds to help isolate factors that may be inherent to English learning from Hindi speakers.

## References


Common european framework of reference for languages. Online. Accessed on 09/24/2023.

Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, complete edition. Longman, New York.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Michael Cole, Vera P. John-Steiner, Sylvia Scribner, and Ellen Souberman. 1978. Mind in society : the development of higher psychological processes.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Richard M. Felder and Eunice Ribeiro Henriques. 1995. Learning and teaching styles in foreign and second language education. *Foreign Language Annals*, 28:21–31.

Emine Gozcu and Çağda Kıvanç Çağanağa. 2016. The importance of using games in efl classrooms. *Cypriot Journal of Educational Sciences*, 11:126–135.

Robin Hunicke and Vernell Chapman. 2004. Ai for dynamic difficulty adjustment in games. *Challenges in game artificial intelligence AAAI workshop*, 2.

Lames, C. Edwards Lester, Sharolyn, Andy tenor Converse, Susan, Edward Kahler, and Todd Barlow. 1997. The persona effect: affective impact of animated pedagogical agents. *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*.

Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism.

Mehdi Mirzaei and Masoud Zoghi. 2017. Understanding the language learning plateau: A grounded-theory study.

Jonathan M Newton and ISP Nation. 2020. *Teaching ESL/EFL listening and speaking*. Routledge.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *ArXiv*, abs/2304.03442.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Sherry Shanshan Ruan, Liwei Jiang, Qianyao Xu, Zhiyuan Liu, Glenn M. Davis, Emma Brunskill, and James A. Landay. 2021. Englishbot: An ai-powered conversational system for second language learning. *26th International Conference on Intelligent User Interfaces*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46:197 – 221.

Gonzalo Camacho Vásquez and Joan Camilo Ovalle. 2019. Video games: Their influence on english as a foreign language vocabulary acquisition. *GIST–Education and Learning Research Journal*, (19):172–192.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

# A More Wordplay Games

## A.1 Beginner Puzzles

The acquisition of specific language structures (e.g. past tense and conditional), occurs within a broad, linear progression. Accordingly, our beginner puzzles center on distinct language structures—below we focus on two puzzles aimed at parts of speech practice. The examples illustrated in Figures 7&8 are tailored to accommodate learners at the foundational A1 and A2 levels.

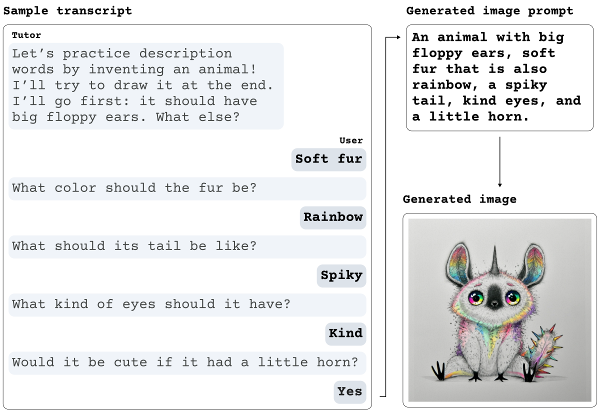



Figure 7: *Invent an Animal* beginner descriptive word elicitation puzzle transcripts, prompt, and generated image.

The *Invent an Animal* puzzle, shown in Figures 7, encourages learners to conceptualize a new animal in collaboration with the LLM tutor. This is achieved by eliciting descriptive words and adjectives from the users. In contrast, our *Madlibs* puzzle concentrates on more specific parts of speech, such as nouns and verbs. This puzzle exemplifies the capability of the Critic agent to assess semantic acceptability and to respond to users' inquiries regarding parts of speech.

For instance, in response to "what is a noun," the puzzle utilizes a technique known as *code-switching*. Here, the puzzle addresses the learner in their native language (i.e., Spanish) to explain the concept. This tactic enables the learning experience to remain within the A1 and A2 proficiency levels, facilitating understanding of abstract concepts using comparatively complex language. We enable this by adding the following sentence to the Critic agent prompt: *"If the player is struggling to understand you, you can code switch and respond in Spanish"*.

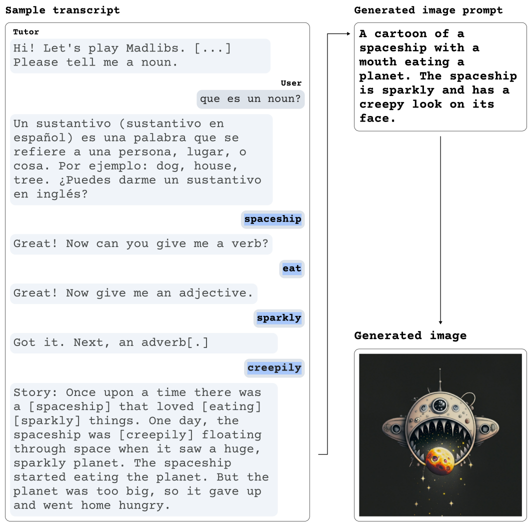



Figure 8: *Madlibs* beginner parts of speech practice puzzle transcripts, prompt, and generated image. The transcript shows an example of the LLM choosing to code switch in response to the user using their L1.

Ultimately, the composed story incorporates the words supplied by the learner, placed within brackets, enabling the learner to understand the context of usage. Even though the story may surpass their language proficiency, the accompanying image visually grounds the narrative, offering extra support for comprehension.

## A.2 Handling Incorrect Responses

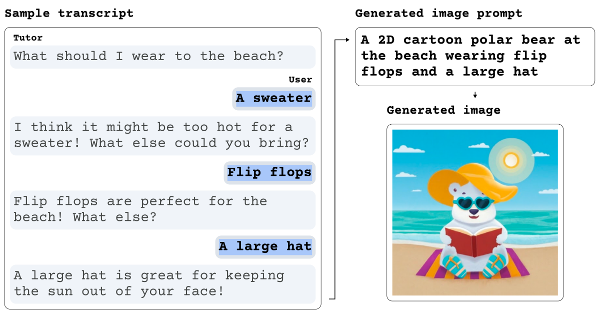



Figure 9: What to wear to the *beach* transcripts, generated image prompt, and generated image. The transcript shows an example of the LLM analyzing the semantic acceptability of the user's suggestions.

In puzzles tailored for A2 to B2 proficiency levels, learners collaborate with a tutor character to find suitable attire for either the beach or the opera, focusing on recalling and producing contextual vocabulary. The only differences between the prompts of these puzzles and the *Wedding* puzzle prompt are the words "opera" and "beach".

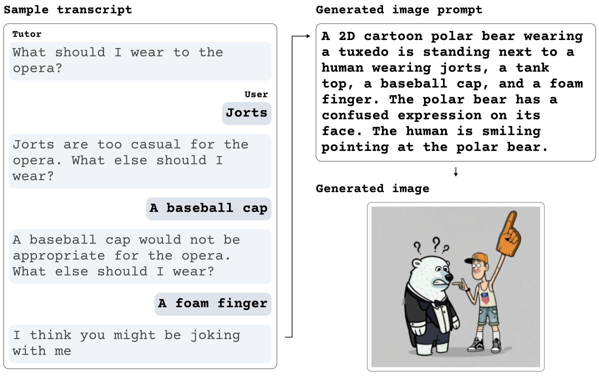



Figure 10: What to wear to the *opera* transcripts, generated image prompt, and generated image. The image still incorporates the user's unsuitable suggestions.

In Figure 10, although the learner's suggestions for attire to the opera are inappropriate, the generated image still provides visual context to the dialogue. It depicts a second character wearing the learner's inappropriate suggestions—"jorts," "foam finger," and "baseball cap"—contrasted with Bearnard dressed in a tuxedo with a puzzled expression on this face. This visual contrast can generate a surprising and delightful moment for the learner, enriching the learning experience.

### A.3 Messy Room Prepositions Playthrough

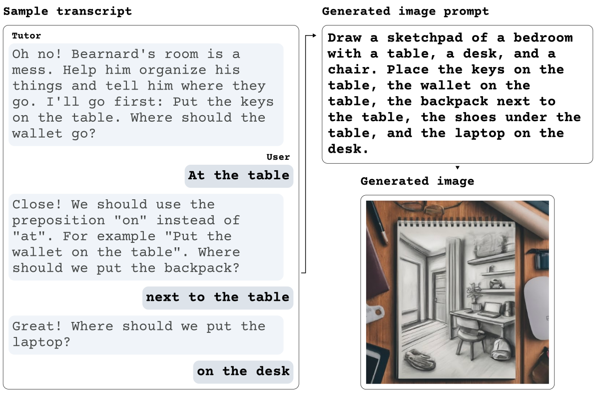



Figure 11: The *Messy Room Preposition* transcripts, generated image prompt, and generated image. The transcript shows an example of the LLM correcting the user's incorrect use of a preposition.

## B Taboo Deep-dive

### B.1 Few-shot Learning Induces Difficulty Control

LLMs are known for their ability to learn new tasks using only a handful of examples (Brown et al., 2020). We leverage few-shot learning in our *Taboo* puzzle to control the difficulty of the `hiddenWord` based on the learner's proficiency level. For three levels of 'beginner', 'intermediate', and 'advanced' we have 5-6 examples of hidden words and taboo words. In Figure 12 we provide the Taboo prompt which is populated by our external few-shot dictionary. This demonstrates how a puzzle can satisfy levels from A2 through C2. These few-shot examples were curated by an expert in the field of English Language Learning to fit at the intended CEFR level.

### B.2 Example Playthroughs

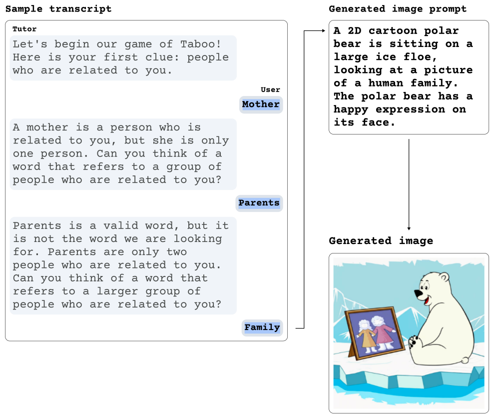



Figure 13: *Taboo* transcripts, generated image prompt, and generated image from a beginner mode play through where the word is set to "family".

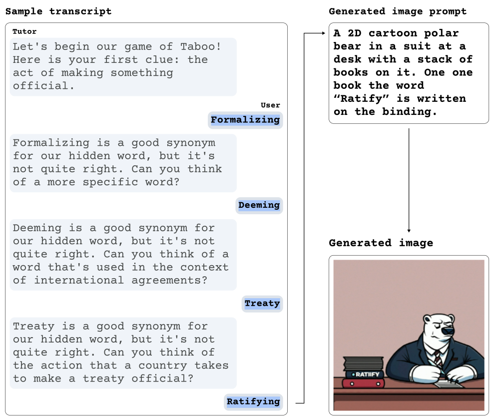



Figure 14: *Taboo* transcripts, generated image prompt, and generated image from a advanced mode play through where the word is set to "ratify".

### B.3 CEFR Classification: Beginner and Intermediate User Transcripts

Solving the *Taboo* puzzle provides an opportunity for a learner to practice both comprehension skills and recall of vocabulary. Upon choosing a difficulty level, the learner must comprehend the lan-

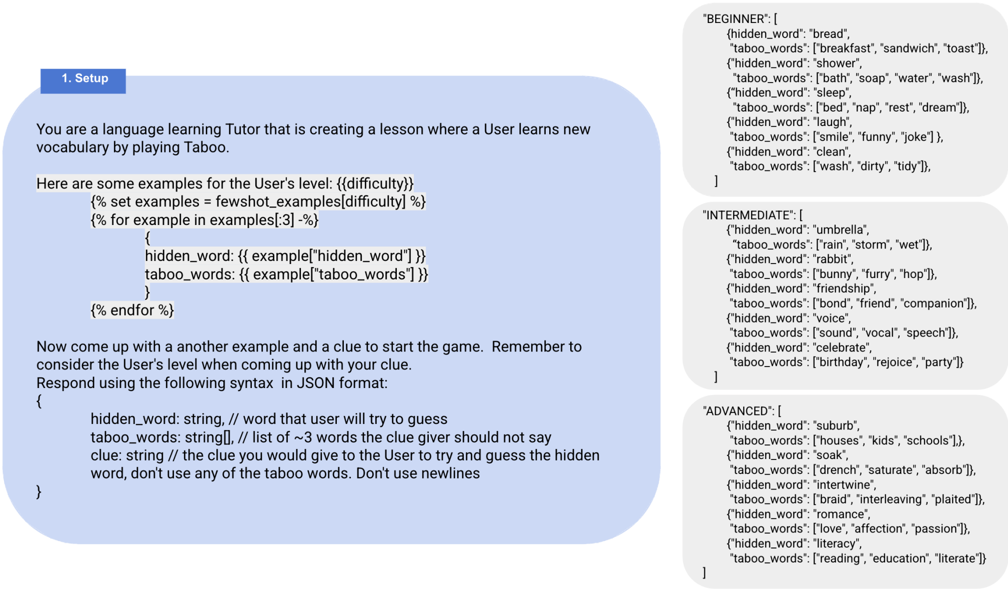



Figure 12: Setup prompt for the *Taboo* mini puzzle, demonstrating how the hidden word and associated taboo words (right) of varying difficulty are dynamically integrated into the few-shot template (left) specific to the user's selected level of difficulty.

guage generated at that level in order to solve the puzzle.

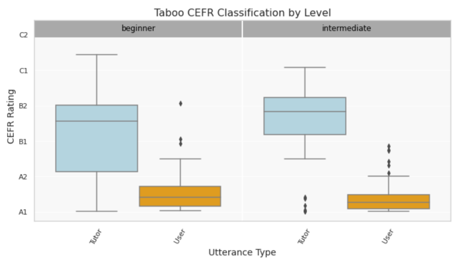



Figure 15: CEFR ratings across all 28 Taboo user sessions.

Illustrated in Figure 15 the text generated by the LLM consistently aligns with the learner's selected level. Our fine-tuned BERT model assigns CEFR ratings to the beginner level tutor utterances within the range of A2-B1, while the intermediate level tutor utterances falls between B1-C1. However, the user CEFR rating remains similar for both the beginner and intermediate levels. As shown in Figure 16 the *Taboo* puzzle elicits mostly single word or short sentence answers from the user. The CEFR classification model employed is not inherently suitable for single words; consequently, the ratings may not be consistently accuracy for the *Taboo* user utterances.

| Speaker | Utterance | CEFR |
|---|---|---|
| **Tutor** | We're all set. Here is your first clue: This large animal lives in Africa and Asia and has big ears. | B2 |
| User | help | A2 |
| **Tutor** | Do you want to know its habitat? It eats leaves and fruit from trees. | B2 |
| User | it is big and white color | A1 |
| **Tutor** | You are on the right track. This animal is big and white. It is a land animal. | B1 |
| User | he eat fish | A1 |
| **Tutor** | The animal I'm thinking of does not eat fish. Do you want to know its habitat? | B1 |
| User | yes | A1 |
| **Tutor** | The animal I'm thinking of does not eat fish. It lives in savannas, grasslands, and forests. | B2 |
| User | he's herbivorous | B1 |
| **Tutor** | You're right, it's herbivorous. Do you want to know its size? | B2 |
| User | what is its size | A2 |
| **Tutor** | his animal is the largest land animal. It can be up to 13 feet tall and weigh up to 15,000 pounds. | B2 |
| User | he's dangerous | A2 |
| **Tutor** | The animal I'm thinking of is not dangerous. It's actually quite gentle. Do you want to know its lifespan? | B2 |
| User | yes | A1 |
| **Tutor** | The animal I'm thinking of has a lifespan of about 70 years. Do you want to know its social behavior? | B2 |
| User | Yes | A1 |

| Speaker | Utterance | CEFR |
|---|---|---|
| **Tutor** | The animal I'm thinking of is social and lives in herds. It is also known for its long memory. | B2 |
| User | tell me more about it | B1 |
| **Tutor** | This animal is a symbol of good luck in some cultures. It is also used as a work animal in some parts of the world. | C1 |
| User | Where he lives | A1 |
| **Tutor** | The animal I'm thinking of lives in savannas, grasslands, and forests. It is found in Africa and Asia. | B2 |
| User | he's panda | A1 |
| **Tutor** | Pandas are not the largest land animals, and they are not found in Africa. Would you like another clue? | B2 |
| User | yes | A1 |
| **Tutor** | Pandas are not the largest land animals, and they are not found in Africa. Here is another clue: This animal is often used in circuses. | C1 |
| User | Bear | A1 |
| **Tutor** | Bears are not the largest land animals, and they are not used in circuses. Do you want another clue? | B2 |
| User | yes | A1 |
| **Tutor** | Bears are not the largest land animals, and they are not used in circuses. Here is another clue: This animal is used for transportation in some parts of the world. | C1 |
| User | elephant | A1 |
| **Tutor** | Great job! | A1 |

A1 • A2 • B1 • B2 • C1

Figure 16: Transcript from a user session of Taboo at the Intermediate level.