WordPlay: An Agent Framework for Language Learning Games

Suma Bailis
Google Research
sbailis@google.com

Lara McConnaughey
Google Research
laramcc@google.com

Jane Friedhoff
Google Research
jfriedhoff@google.com

Feiyang Chen
Google Research
feiyangc@google.com

Chase Adams
Google Research
chaseadams@google.com

Jacob Moon
Google Research
jacobm@google.com

Abstract

We introduce a novel framework, WordPlay, for building language learning games. 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. WordPlay allows content creators to quickly author bite-sized, personalized puzzles that cater to various proficiency levels, and uses generated images to aid comprehension and learning.

1 Introduction

The importance of play in the learning process is well established [1]. 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 [2].

Game-play research demonstrates that adaptive experiences can counteract stagnation and boredom by maintaining player interest, especially when challenges are scaled to the player's proficiency within their zone of proximal development [3, 4]. Thus, integrating game-like elements in language learning can increase enjoyment [5] 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 [6] [7]. 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 [8]. They excel as agents, proficiently executing tasks when given a reasoning structure and access to external resources, such as online search or calculators [9]. Furthermore, they can assume the identity of specific characters, generating dialogue in alignment with that persona [10]. In the case of tutoring, past work has shown that intelligent tutoring systems are found to be as effective as human tutors [11]. Recently, advances in LLMs have engendered abundant usage for

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

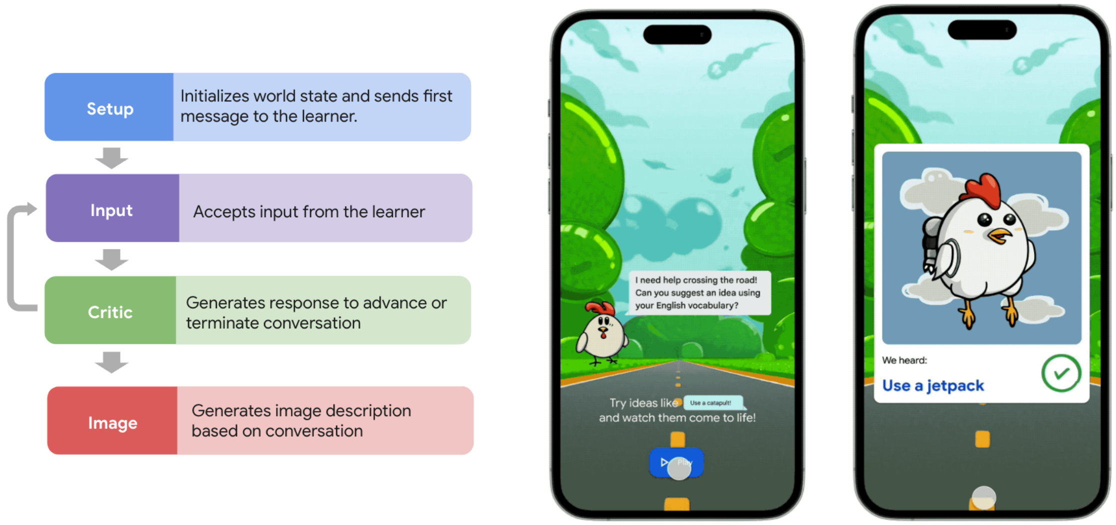
language learning such as EnglishBot [12] because they enable freeform interactions that leverage the persona effect [13].

Recent work has also shown text-to-image models to be capable of generating realistic and creative imagery, consistent with prompts [14] [15]. 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 [16]. 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. A hallmark feature of our approach lies in the integration of dynamically generated images that enhance language comprehension and elevate the overall learning experience.

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 [17]. The instructional content is relayed through a text-to-speech (TTS) system, with automatic speech recognition (ASR) [18] 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 [19].


Setup
Initializes world state and sends first message to the learner.
Input
Accepts input from the learner
Critic
Generates response to advance or terminate conversation
Image
Generates image description based on conversation
Use a catapult!
Play

I need help crossing the road! Can you suggest an idea using your English vocabulary?

Try ideas like Use a catapult! and watch them come to life!

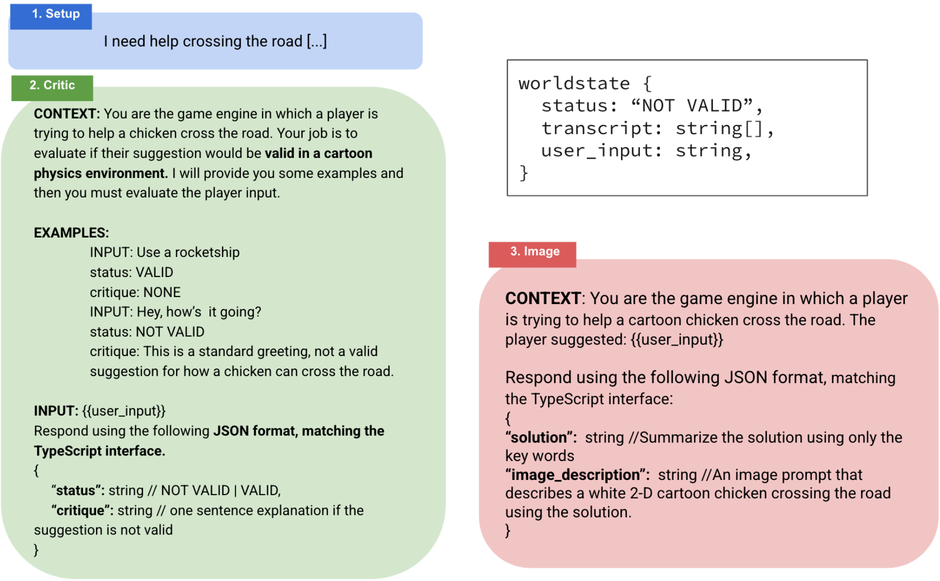
We heard:

Use a jetpack

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

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 [20] for language generation and Imagen [15] for image generation.



1. Setup

I need help crossing the road [...]

2. Critic

CONTEXT: You are the game engine in which a player is trying to help a chicken cross the road. Your job is to evaluate if their suggestion would be **valid in a cartoon physics environment.** I will provide you some examples and then you must evaluate the player input.

EXAMPLES:
INPUT: Use a rocketship
status: VALID
critique: NONE
INPUT: Hey, how's it going?
status: NOT VALID
critique: This is a standard greeting, not a valid suggestion for how a chicken can cross the road.

INPUT: {{user_input}}
Respond using the following **JSON format, matching the TypeScript interface.**
{
"status": string // NOT VALID | VALID,
"critique": string // one sentence explanation if the suggestion is not valid
}

```
worldstate {
  status: “NOT VALID”,
  transcript: string[],
  user_input: string,
}
```

3. Image

CONTEXT: You are the game engine in which a player is trying to help a cartoon chicken cross the road. The player suggested: {{user_input}}

Respond using the following JSON format, matching the TypeScript interface:
{
"solution": string //Summarize the solution using only the key words
"image_description": string //An image prompt that describes a white 2-D cartoon chicken crossing the road using the solution.
}

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.

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. In the *Chicken Crossing the Road* puzzle, this dialogue is predefined. However, in other puzzles like Taboo, as shown in Figure 9, the Setup agent prompt is sent to the LLM to generate a new “hidden word” and “taboo word” to initialize the game. 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 responding to user inquiries.

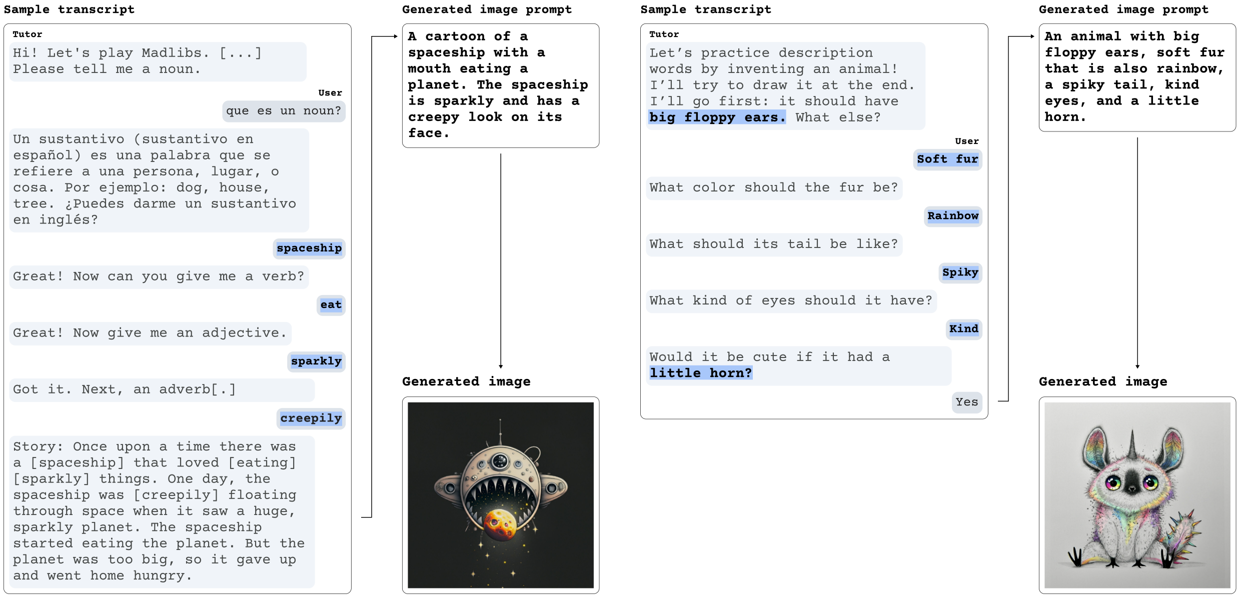
Image: The Image agent is responsible for generating prompts for image generation model and a summary of the solution to be displayed beneath the image. To maintain the safety of the image generation process, we employ the LLM to generate the image descriptions rather than allowing the user to do so. 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. Only the *transcript* updates are handled in a rule-based way, typically extended with each line of tutor or user dialogue, all other variables are replaced by the fields in the JSON object. 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 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 CEFR framework is stratified into six distinct levels: A1, A2, B1, B2, C1, and C2, with A1 denoting the most elementary user and C2 representing the highest level of proficiency. The examples illustrated in Figure 3 are tailored to accommodate learners at the foundational A1 and A2 levels.



Hi! Let's play Madlibs. [...] Please tell me a noun.

que es un noun?

Un sustantivo (sustantivo en español) es una palabra que se refiere a una persona, lugar, o cosa. Por ejemplo: dog, house, tree. ¿Puedes darme un sustantivo en inglés?

spaceship

Great! Now can you give me a verb?

eat

Great! Now give me an adjective.

sparkly

Got it. Next, an adverb[.]

creepily

Story: Once upon a time there was a [spaceship] that loved [eating] [sparkly] things. One day, the spaceship was [creepily] floating through space when it saw a huge, sparkly planet. The spaceship started eating the planet. But the planet was too big, so it gave up and went home hungry.

A cartoon of a spaceship with a mouth eating a planet. The spaceship is sparkly and has a creepy look on its face.

Let's practice description words by inventing an animal! I'll try to draw it at the end. I'll go first: it should have **big floppy ears.** What else?

Soft fur

What color should the fur be?

Rainbow

What should its tail be like?

Spiky

What kind of eyes should it have?

Kind

Would it be cute if it had a **little horn?**

Yes

An animal with big floppy ears, soft fur that is also rainbow, a spiky tail, kind eyes, and a little horn.

Figure 3: *Madlibs* (left) and *Invent an Animal* (right) transcripts, generated image prompt, and generated image.

The *Invent an Animal* puzzle 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.

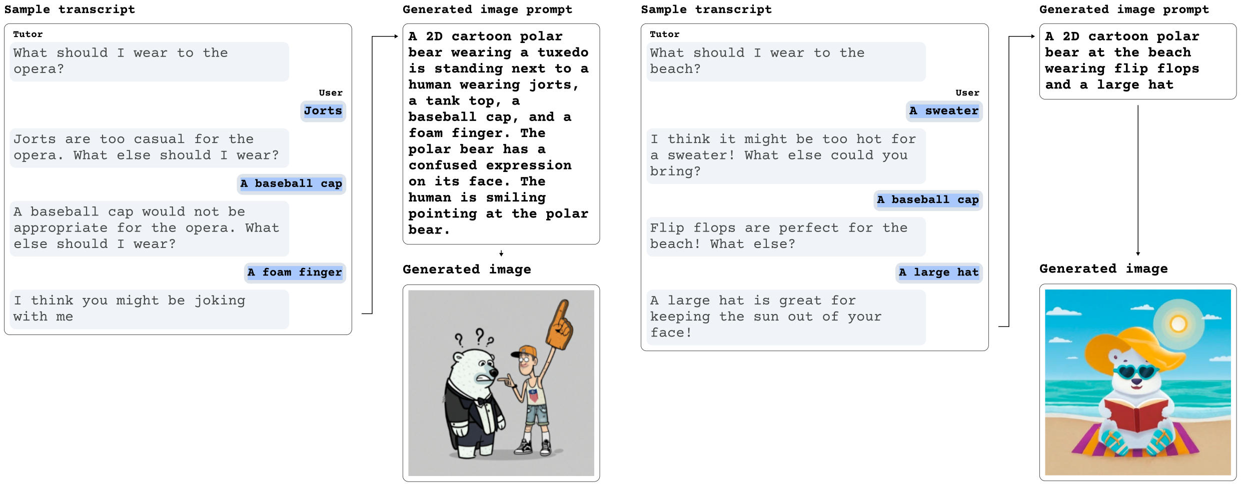
For instance, in response to "what is a noun," the puzzle utilizes a technique known as *code-switching*. Here, the puzzle addresses the learner in their native language (i.e., Spanish) to explain the concept. This tactic enables the learning experience to remain within the A1 and A2 proficiency levels,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

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.

3.2 Handling Incorrect Responses

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 are the words "opera" and "beach," illustrating the scalability of this approach.

In each scenario, when the learner suggests attire or an accessory unsuitable for the chosen venue, the Tutor explains why that might not work. 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.



Sample transcript

Tutor
What should I wear to the opera?

User
Jorts

Jorts are too casual for the opera. What else should I wear?

A baseball cap

A baseball cap would not be appropriate for the opera. What else should I wear?

A foam finger

I think you might be joking with me

Generated image prompt

A 2D cartoon polar bear wearing a tuxedo is standing next to a human wearing jorts, a tank top, a baseball cap, and a foam finger. The polar bear has a confused expression on its face. The human is smiling pointing at the polar bear.

Generated image

Sample transcript

Tutor
What should I wear to the beach?

User
A sweater

I think it might be too hot for a sweater! What else could you bring?

A baseball cap

Flip flops are perfect for the beach! What else?

A large hat

A large hat is great for keeping the sun out of your face!

Generated image prompt

A 2D cartoon polar bear at the beach wearing flip flops and a large hat

Generated image

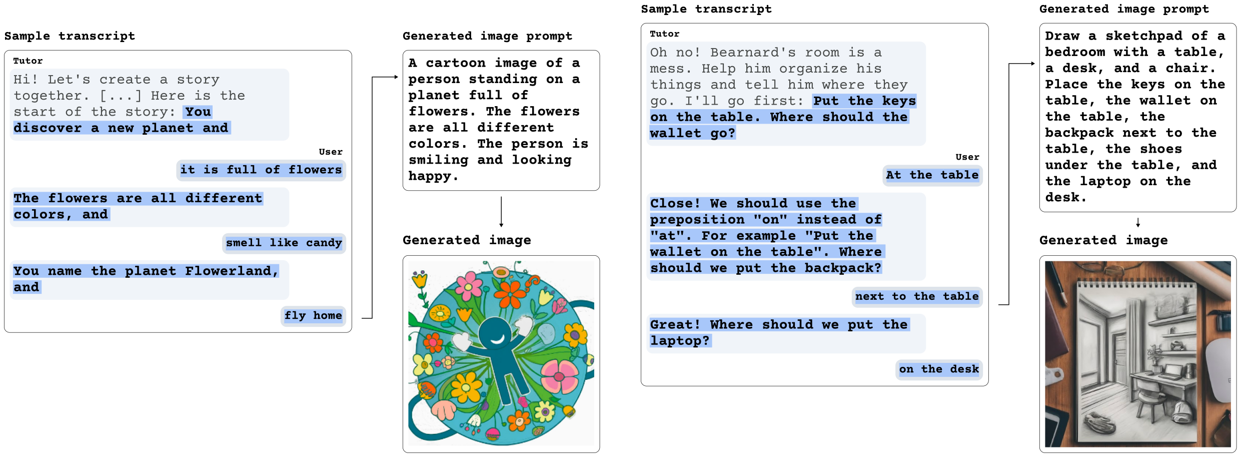
Figure 4: What to wear to the opera (left) and beach (right) mini puzzle transcript, image prompt, and generated image.

On the left-hand side, 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.

3.3 Few-shot Learning Induces Difficulty Control

LLMs are known for their ability to learn new tasks using only a handful of examples [21]. 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 Appendix A.1 we provide the Taboo prompt which is populated by our external few-shot dictionary. This demonstrates how a puzzle can be created to satisfy levels from A2 through C2.

3.4 Intermediate Puzzles



Sample transcript

Tutor
Hi! Let's create a story together. [...] Here is the start of the story: You discover a new planet and

User
it is full of flowers

The flowers are all different colors, and

smell like candy

You name the planet Flowerland, and

fly home

Generated image prompt

A cartoon image of a person standing on a planet full of flowers. The flowers are all different colors. The person is smiling and looking happy.

Generated image

Sample transcript

Tutor
Oh no! Bearnard's room is a mess. Help him organize his things and tell him where they go. I'll go first: Put the keys on the table. Where should the wallet go?

User
At the table

Close! We should use the preposition "on" instead of "at". For example "Put the wallet on the table". Where should we put the backpack?

next to the table

Great! Where should we put the laptop?

on the desk

Generated image prompt

Draw a sketchpad of a bedroom with a table, a desk, and a chair. Place the keys on the table, the wallet on the table, the backpack next to the table, the shoes under the table, and the laptop on the desk.

Generated image

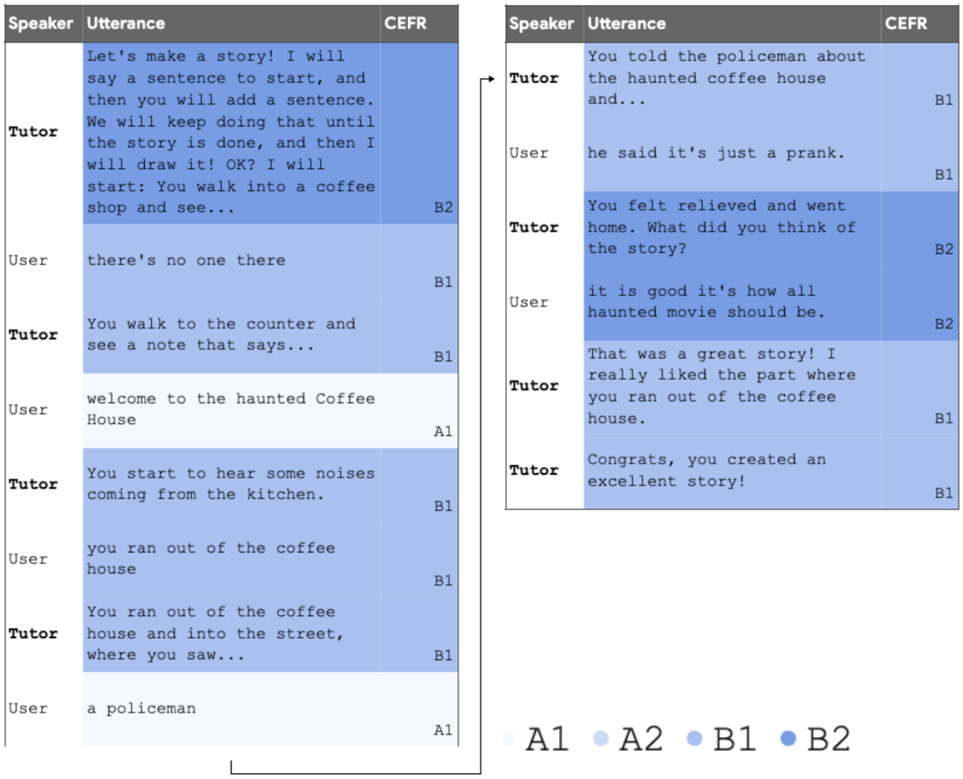
Figure 5: *Finish the Story* (left) and *Messy Room Prepositions* (right) mini puzzle transcript, image prompt, and generated image.

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[22]. These puzzles are tailored for learners at B1 through C1 proficiency levels.

The *Messy Room Prepositions* puzzle focuses on practicing prepositions, which are essential grammatical structure at the intermediate CEFR level. Meanwhile, the *Finish the Story* 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. The responses—both from the participants and the model—were analyzed using a custom classification model. This model, an adaptation of a pre-existing BERT architecture [23], 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.



Speaker	Utterance	CEFR
Tutor	Let's make a story! I will say a sentence to start, and then you will add a sentence. We will keep doing that until the story is done, and then I will draw it! OK? 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
Tutor	Congrats, you created an excellent story!	B1

Figure 6: Utterance CEFR level predictions from a playthrough of a *Finish the Story* session.

In Figure 6, 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 7 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.

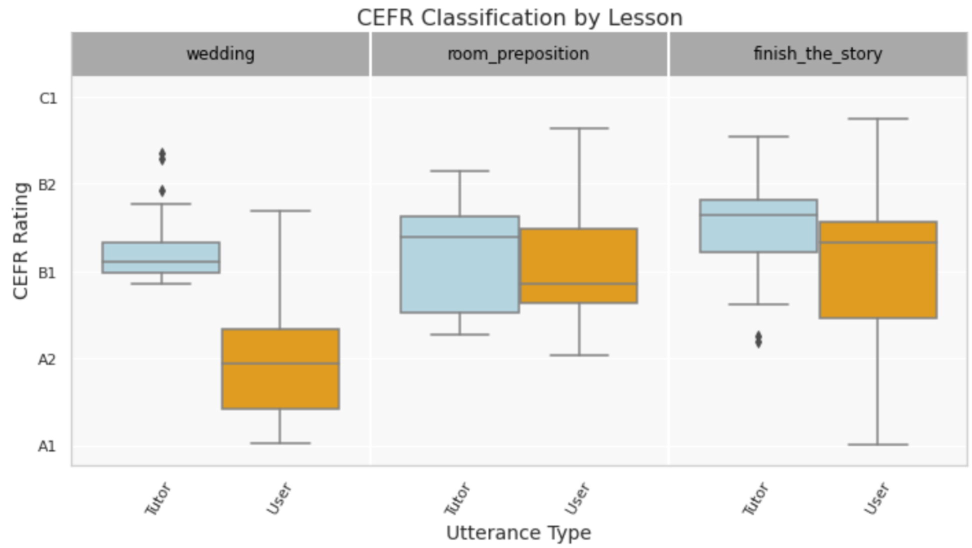


Figure 7: CEFR ratings for tutor and user utterances for the *wedding*, *Messy room Prepositions*, and *Finish the Story* puzzles.

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 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 10. 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 11, 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.

WordPlay is designed to have puzzles that are somewhat ‘on-rails’, offering structured conversation as opposed to freeform, generalized dialogue. This may limit exploring conversational depth and breadth, but our structured approach has optimized for a safety guardrail around the tutor’s output.

In future work, we seek to automatically generate puzzles from higher-level descriptions and also explore puzzles that use generated images on a turn-by-turn basis. We are also interested in ways to find a balance between enabling a more generalized conversation and maintaining the integrity and safety of the learning experience.

Acknowledgments

Special thanks to Chloe Bellows and Melissa Smith, our experts in language learning and content development. Their expertise and dedication greatly enriched the quality of our work. We would also like to thank Jared Bernstein for their constructive feedback and pedagogical insights during the development of this paper.

References

[1] Emine Gozcu and Çağda Kıvanç Çağanağa. The importance of using games in efl classrooms. *Cypriot Journal of Educational Sciences*, 11:126–135, 2016. URL `https://api.semanticscholar.org/CorpusID:54216627`.

[2] Mehdi Mirzaei and Masoud Zoghi. Understanding the language learning plateau: A grounded-theory study. 2017. URL `https://api.semanticscholar.org/CorpusID:67095866`.

[3] Michael Cole, Vera P. John-Steiner, Sylvia Scribner, and Ellen Souberman. Mind in society : the development of higher psychological processes. 1978. URL `https://api.semanticscholar.org/CorpusID:140858445`.

[4] Robin Hunicke and Vernell Chapman. Ai for dynamic difficulty adjustment in games. *Challenges in game artificial intelligence AAAI workshop*, 2, 01 2004.

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

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

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

[8] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *ArXiv*, abs/2304.03442, 2023. URL `https://api.semanticscholar.org/CorpusID:258040990`.

[9] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761, 2023. URL `https://api.semanticscholar.org/CorpusID:256697342`.

[10] 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. Lamda: Language models for dialog applications, 2022.

[11] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46:197 – 221, 2011. URL `https://api.semanticscholar.org/CorpusID:16188384`.

[12] Sherry Shanshan Ruan, Liwei Jiang, Qianyao Xu, Zhiyuan Liu, Glenn M. Davis, Emma Brunskill, and James A. Landay. Englishbot: An ai-powered conversational system for second language learning. *26th International Conference on Intelligent User Interfaces*, 2021. URL `https://api.semanticscholar.org/CorpusID:233224176`.

[13] Lames, C. Edwards Lester, Sharolyn, Andy tenor Converse, Susan, Edward Kahler, and Todd Barlow. The persona effect: affective impact of animated pedagogical agents. *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 1997. URL `https://api.semanticscholar.org/CorpusID:3351600`.

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

[15] 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. Photorealistic text-to-image diffusion models with deep language understanding. 2022.

[16] Richard M. Felder and Eunice Ribeiro Henriques. Learning and teaching styles in foreign and second language education. *Foreign Language Annals*, 28:21–31, 1995. URL `https://api.semanticscholar.org/CorpusID:2461070`.

[17] Common european framework of reference for languages. Online. URL `https://www.coe.int/`. Accessed on 09/24/2023.

[18] 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. Google usm: Scaling automatic speech recognition beyond 100 languages. 2023.

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

[20] 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. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022. URL `https://api.semanticscholar.org/CorpusID:247951931`.

[21] 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. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL `https://api.semanticscholar.org/CorpusID:218971783`.

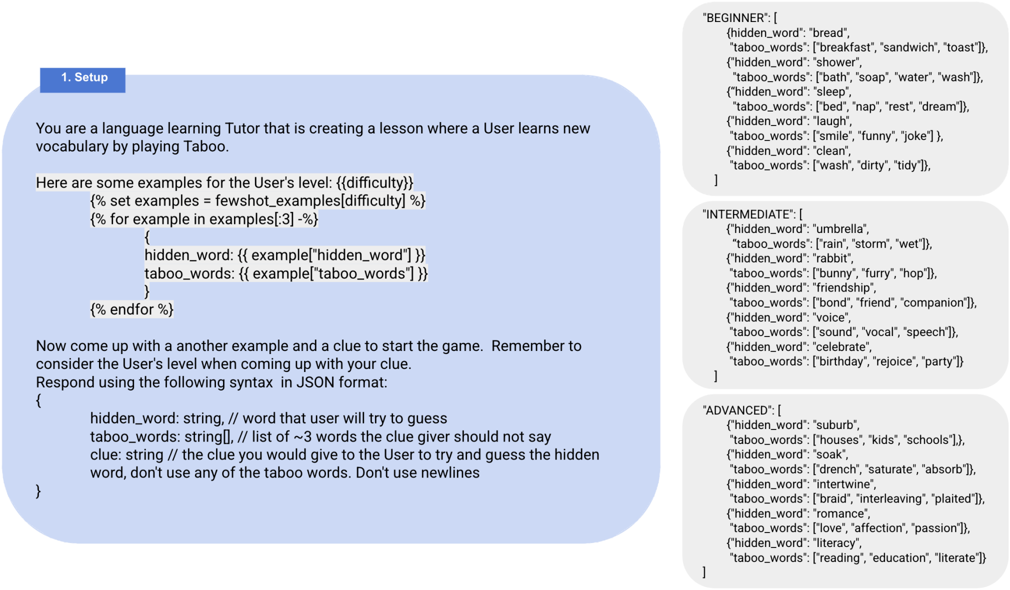
[22] Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, complete edition, 2001.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL `https://api.semanticscholar.org/CorpusID:52967399`.

A Taboo Deep-dive

A.1 Prompt Template

As shown in Figure 8 the few-shot examples are provided for beginner, intermediate, and advanced learner play. These few-shot examples were curated by an expert in the field of English Language Learning to fit at the intended CEFR level.



1. Setup

```
You are a language learning Tutor that is creating a lesson where a User learns new
vocabulary by playing Taboo.

Here are some examples for the User's level: {{difficulty}}
        {% set examples = fewshot_examples[difficulty] %}
        {% for example in examples[:3] -%}
                {
                hidden_word: {{ example["hidden_word"] }}
                taboo_words: {{ example["taboo_words"] }}
                }
        {% endfor %}

Now come up with a another example and a clue to start the game.  Remember to
consider the User's level when coming up with your clue.
Respond using the following syntax  in JSON format:
{
        hidden_word: string, // word that user will try to guess
        taboo_words: string[], // list of ~3 words the clue giver should not say
        clue: string // the clue you would give to the User to try and guess the hidden
        word, don't use any of the taboo words. Don't use newlines
}
```

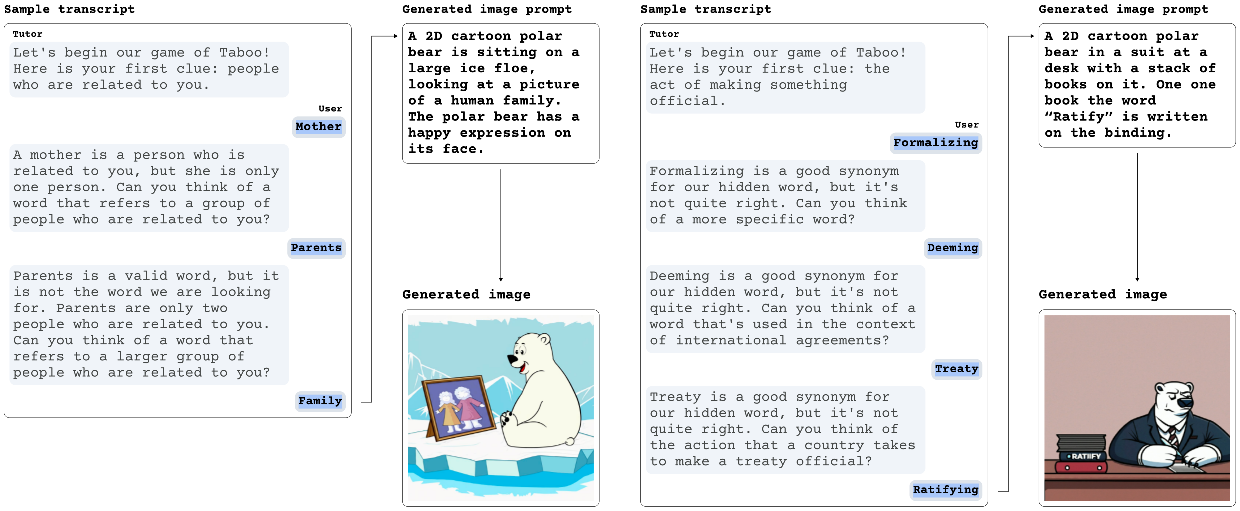
```
"BEGINNER": [
      {hidden_word": "bread",
       "taboo_words": ["breakfast", "sandwich", "toast"]},
      {"hidden_word": "shower",
        "taboo_words": ["bath", "soap", "water", "wash"]},
      {“hidden_word": "sleep",
        "taboo_words": ["bed", "nap", "rest", "dream"]},
      {"hidden_word": "laugh",
       "taboo_words": ["smile", "funny", "joke"] },
      {"hidden_word": "clean",
       "taboo_words": ["wash", "dirty", "tidy"]},
  ]
```

```
"INTERMEDIATE": [
      {"hidden_word": "umbrella",
        “taboo_words": ["rain", "storm", "wet"]},
      {"hidden_word": "rabbit",
       "taboo_words": ["bunny", "furry", "hop"]},
      {"hidden_word": "friendship",
       "taboo_words": ["bond", "friend", "companion"]},
      {"hidden_word": "voice",
       "taboo_words": ["sound", "vocal", "speech"]},
      {"hidden_word": "celebrate",
       "taboo_words": ["birthday", "rejoice", "party"]}
  ]
```

```
"ADVANCED": [
      {"hidden_word": "suburb",
       "taboo_words": ["houses", "kids", "schools"],},
      {"hidden_word": "soak",
       "taboo_words": ["drench", "saturate", "absorb"]},
      {"hidden_word": "intertwine",
       "taboo_words": ["braid", "interleaving", "plaited"]},
      {"hidden_word": "romance",
       "taboo_words": ["love", "affection", "passion"]},
      {"hidden_word": "literacy",
       "taboo_words": ["reading", "education", "literate"]}
]
```

Figure 8: 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.

A.2 Example Playthroughs



Sample transcript

Tutor

Let's begin our game of Taboo! Here is your first clue: people who are related to you.

User

Mother

A mother is a person who is related to you, but she is only one person. Can you think of a word that refers to a group of people who are related to you?

Parents

Parents is a valid word, but it is not the word we are looking for. Parents are only two people who are related to you. Can you think of a word that refers to a larger group of people who are related to you?

Family

Generated image prompt

A 2D cartoon polar bear is sitting on a large ice floe, looking at a picture of a human family. The polar bear has a happy expression on its face.

Generated image

Sample transcript

Tutor

Let's begin our game of Taboo! Here is your first clue: the act of making something official.

User

Formalizing

Formalizing is a good synonym for our hidden word, but it's not quite right. Can you think of a more specific word?

Deeming

Deeming is a good synonym for our hidden word, but it's not quite right. Can you think of a word that's used in the context of international agreements?

Treaty

Treaty is a good synonym for our hidden word, but it's not quite right. Can you think of the action that a country takes to make a treaty official?

Ratifying

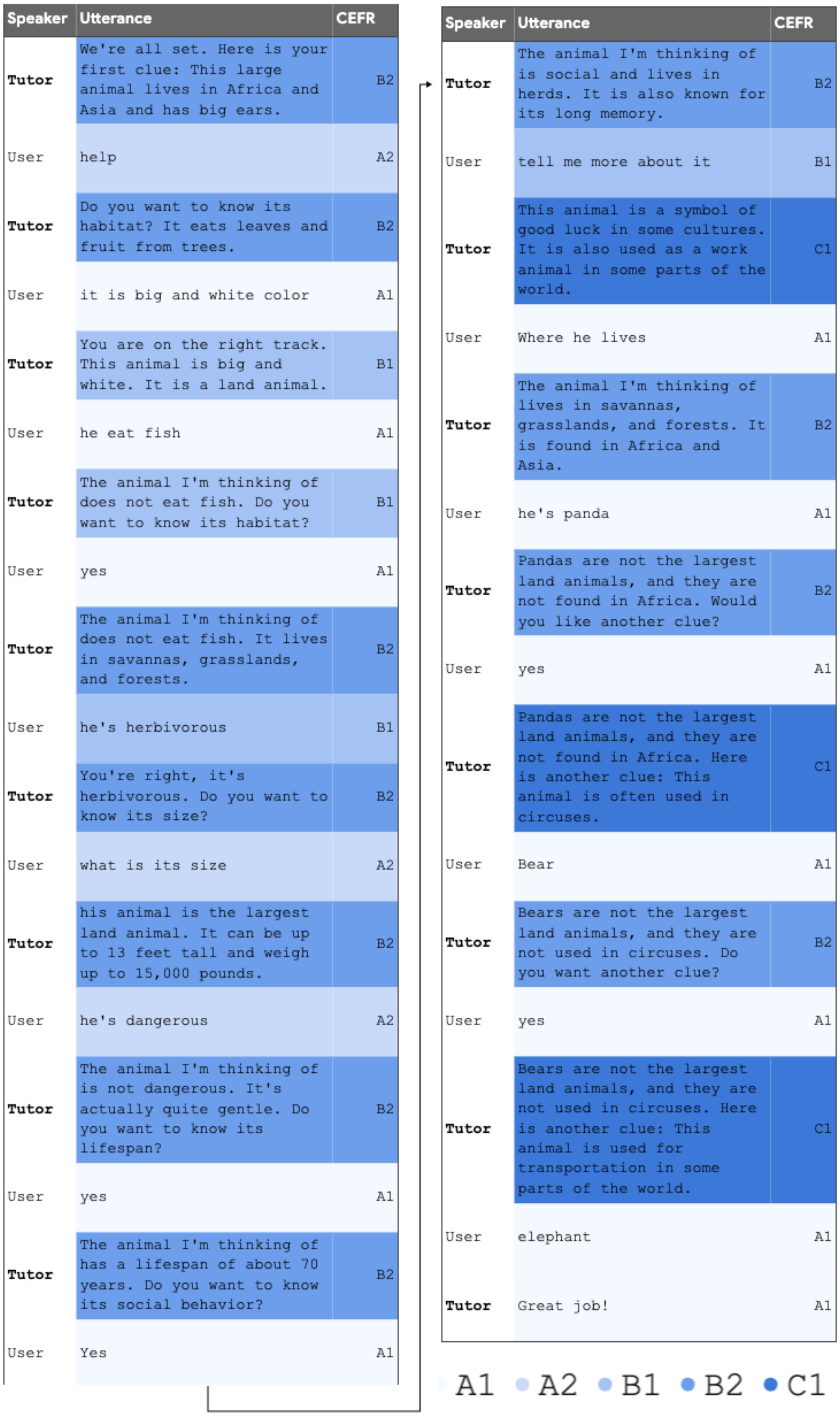
Generated image prompt

A 2D cartoon polar bear in a suit at a desk with a stack of books on it. One one book the word “Ratify” is written on the binding.

Generated image

Figure 9: *Taboo* in beginner mode (left) and *Taboo* in advanced mode (right) mini puzzle transcript, image prompt, and generated image. When the level is specified as beginner the `hiddenWord` is "family" and for an advanced level the `hiddenWord` is "ratify".

A.3 Utterance Level CEFR



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

A.4 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 language generated at that level in order to solve the puzzle.

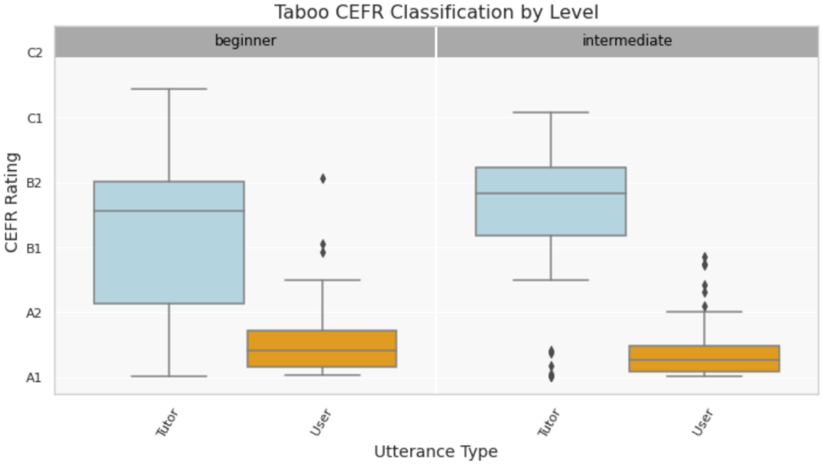
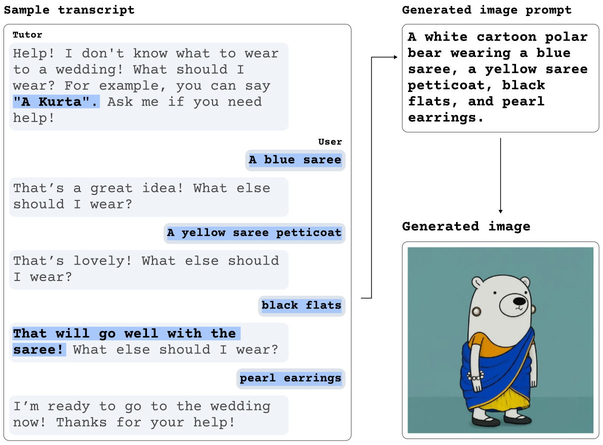


Figure 11: CEFR ratings across all 28 Taboo user sessions

Illustrated in Figure 11 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 10 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.

B Wedding Puzzle

For learner evaluation, we make a few simple modifications to the *Opera* beginner level puzzle turning it into a *Wedding* puzzle that is more relevant to the learner. Additionally, the *Wedding* puzzle content is also localized to make the puzzle more engaging. An example *Wedding* play-through shown in Figure 12.



Tutor

Help! I don't know what to wear to a wedding! What should I wear? For example, you can say **"A Kurta".** Ask me if you need help!

User

A blue saree

That's a great idea! What else should I wear?

A yellow saree petticoat

That's lovely! What else should I wear?

black flats

That will go well with the saree! What else should I wear?

pearl earrings

I'm ready to go to the wedding now! Thanks for your help!

A white cartoon polar bear wearing a blue saree, a yellow saree petticoat, black flats, and pearl earrings.

Figure 12: *Wedding* transcript, generated image prompt, and generated image.