WordPlay: An Agent Framework for Language Learning Games

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 \cite{12} because they enable freeform interactions that leverage the persona effect \cite{13}.

Recent work has also shown text-to-image models to be capable of generating realistic and creative imagery, consistent with prompts \cite{14} \cite{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 \cite{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 \cite{17}. The instructional content is relayed through a text-to-speech (TTS) system, with automatic speech recognition (ASR) \cite{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 \cite{19}.

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

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 \cite{20} for language generation and Imagen \cite{15} for image generation.
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

```json
worldstate {
  status: "NOT VALID",
  transcript: string[],
  user_input: string,
}
```
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.

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.
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

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.

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\cite{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\cite{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.

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.
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


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.

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

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 hidden word is “family” and for an advanced level the hidden word is “ratify.”
### A.3 Utterance Level CEFR

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
<th>CEFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutor</td>
<td>We're all set. Here is your first clue: This large animal lives in Africa and Asia and has big ears.</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>help</td>
<td>A2</td>
</tr>
<tr>
<td>Tutor</td>
<td>Do you want to know its habitat? It eats leaves and fruit from trees.</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>it is big and white color</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>You are on the right track. This animal is big and white. It is a land animal.</td>
<td>B1</td>
</tr>
<tr>
<td>User</td>
<td>he eat fish</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>The animal I'm thinking of does not eat fish. Do you want to know its habitat?</td>
<td>B1</td>
</tr>
<tr>
<td>User</td>
<td>yes</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>The animal I'm thinking of does not eat fish. It lives in savannas, grasslands, and forests.</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>he's herbivorous</td>
<td>B1</td>
</tr>
<tr>
<td>Tutor</td>
<td>You're right, it's herbivorous. Do you want to know its size?</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>what is its size</td>
<td>A2</td>
</tr>
<tr>
<td>Tutor</td>
<td>His animal is the largest land animal. It can be up to 13 feet tall and weigh up to 15,000 pounds.</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>he's dangerous</td>
<td>A2</td>
</tr>
<tr>
<td>Tutor</td>
<td>The animal I'm thinking of is not dangerous. It's actually quite gentle. Do you want to know its lifespan?</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>yes</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>The animal I'm thinking of has a lifespan of about 70 years. Do you want to know its social behavior?</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>Yes</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>Tell me more about it</td>
<td>B1</td>
</tr>
<tr>
<td>User</td>
<td>Where he lives</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>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.</td>
<td>C1</td>
</tr>
<tr>
<td>User</td>
<td>he's panda</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>The animal I'm thinking of lives in savannas, grasslands, and forests. It is found in Africa and Asia.</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>yes</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>Pandas are not the largest land animals, and they are not found in Africa. Would you like another clue?</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>Bear</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>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.</td>
<td>C1</td>
</tr>
<tr>
<td>User</td>
<td>yes</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>Bears are not the largest land animals, and they are not used in circuses. Do you want another clue?</td>
<td>B2</td>
</tr>
<tr>
<td>User</td>
<td>yes</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>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.</td>
<td>C1</td>
</tr>
<tr>
<td>User</td>
<td>elephant</td>
<td>A1</td>
</tr>
<tr>
<td>Tutor</td>
<td>Great job!</td>
<td>A1</td>
</tr>
</tbody>
</table>

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.

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.