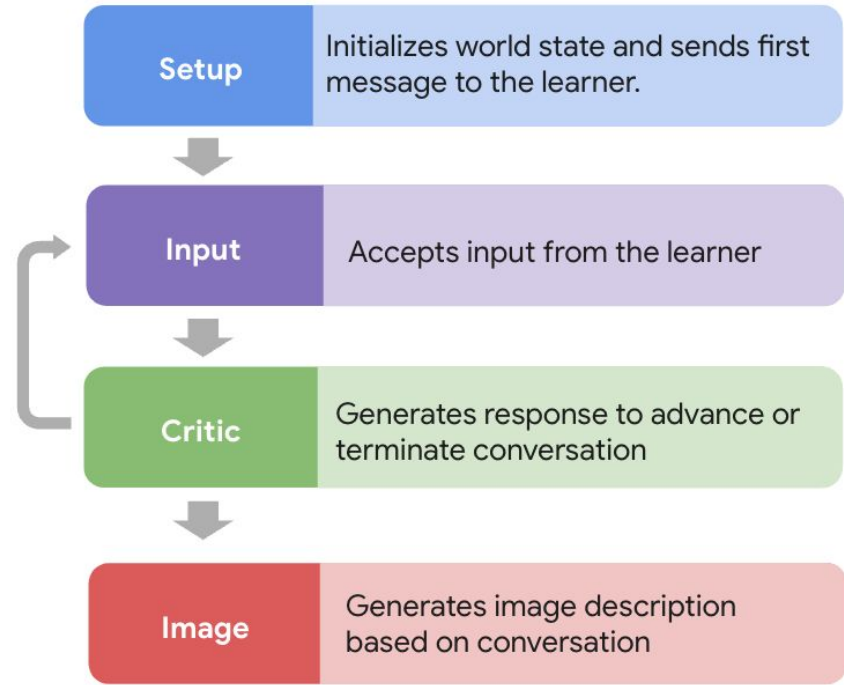




# WordPlay: An Agent Framework for Language Learning Games

Suma Bailis, Lara McConaughy, Jane Friedhoff, Feiyang Chen, Chase Adams, Jacob Moon

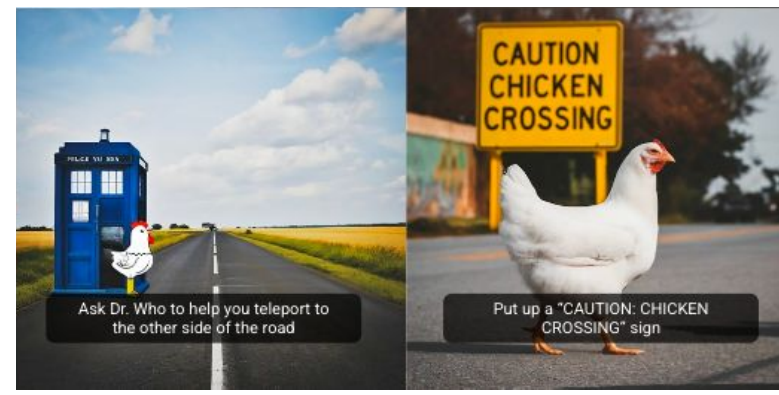
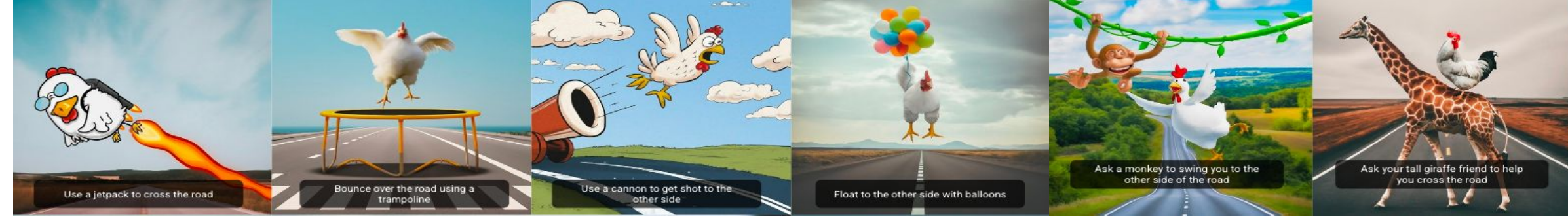
## 1 WordPlay Framework



**WordPlay Framework** generates engaging, adaptive mini-games using LLMs and text-to-image models to address the challenge of balancing engaging and effective language practice.

## 2 Motivation

- Allows content creators to quickly author bite-sized, puzzles that cater to various CEFR proficiency levels, and uses generated images to aid comprehension.



## 4 WordPlay Games

Sample transcript

**Tutor:** 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?

**User:** Soft fur

What color should the fur be?

**User:** Rainbow

What should its tail be like?

**User:** Spiky

What kind of eyes should it have?

**User:** Kind

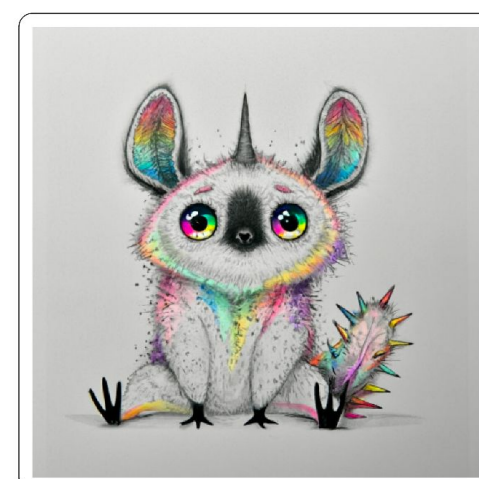
Would it be cute if it had a little horn?

**User:** Yes

Generated image prompt

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

Generated image



- Invent an Animal** puzzle encourages learners to conceptualize a new animal in collaboration with the LLM tutor

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

**User:** smell like candy

You name the planet Flowerland, and

**User:** 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



- 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.

## 3 Prompts

### 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
}
  
```

- Wordplay consists of only three prompts (or agents) we use *PaLM* for language generation and *Imagen* for image generation. All output is in JSON format and the game's world-state is updated accordingly.

```

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.
}
  
```

## 5 Evaluation

- To assess our puzzles' alignment to CEFR standards, we conducted experimental sessions with native Hindi-speaking individuals in India learning English.

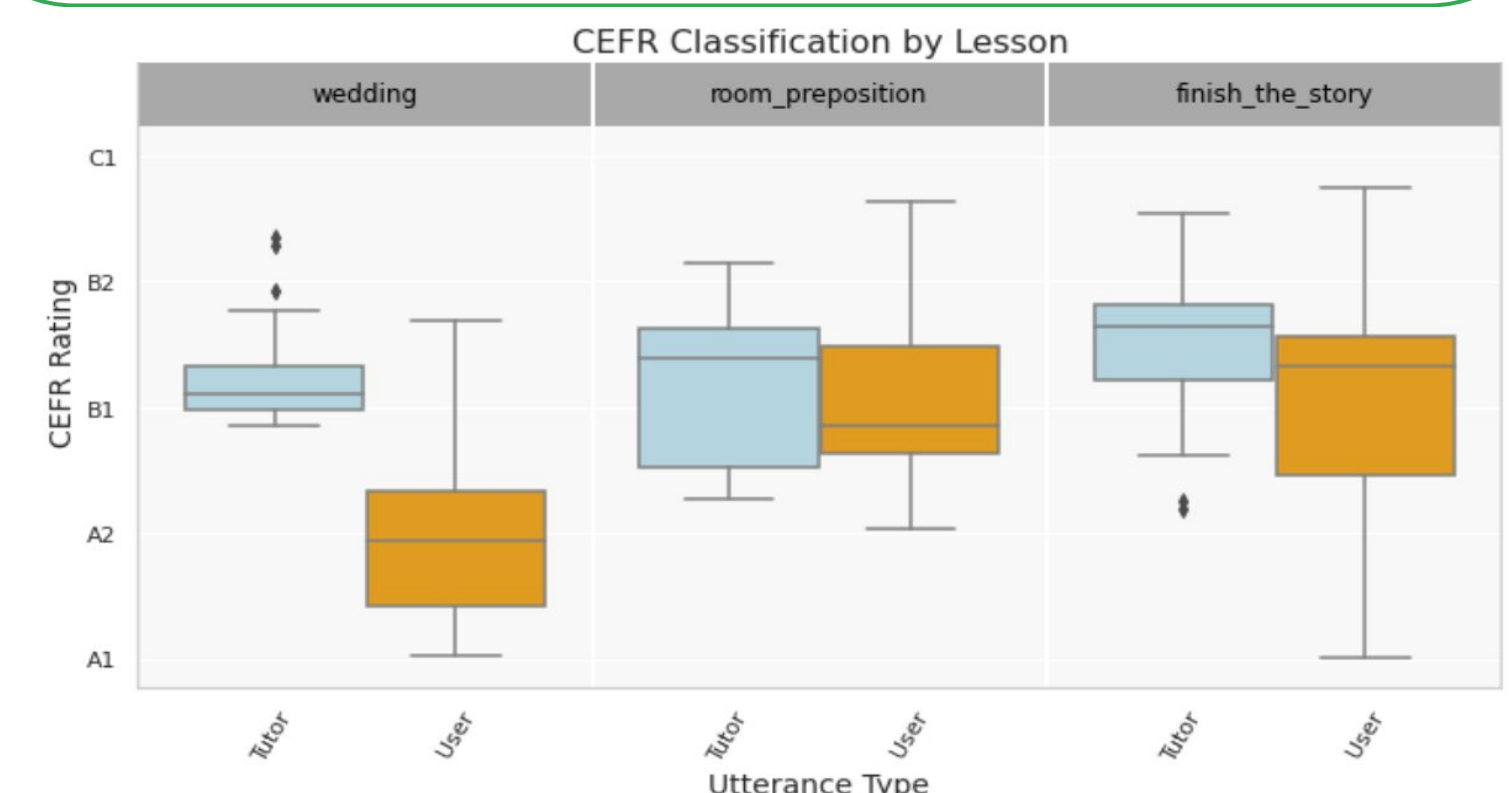
Speaker	Utterance	CEFR
<b>Tutor</b>	Let's make a story! I will say a sentence to start, and then you will add a sentence. We will keep doing that until the story is done, and then I will draw it! OK? I will start: You walk into a coffee shop and see...	B2
<b>User</b>	there's no one there	B1
<b>Tutor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

A1 • A2 • B1 • B2 • C1

Utterance CEFR level predictions from a playthrough of a *Finish the Story*.

- The CEFR level of responses from both participants and the model underwent analysis through a custom CEFR classification model.

- Puzzles evaluated:**
  - Wedding** - beginner puzzle asking users for suitable wedding attire
  - Finish the Story** - intermediate puzzle asks users to take turns with the tutor to construct a narrative
  - Room Prepositions** - intermediate puzzle to practice prepositions



- Tutor utterances consistently hit the B2 level, evidenced by the narrow interquartile ranges (IQR)
- User utterances have a wide IQR proficiency range, with the **wedding** puzzle centering around A2 and the **finish the story** puzzle spanning B1 to B2

## 6 Conclusion & Next Steps

- WorldPlay allows content creators to author engaging and level appropriate puzzles by customizing three simple prompts.
- The framework design prioritizes structured conversation over freeform dialogue, emphasizing safety in the tutor's output.
- Future work aims to automatically generate puzzles from higher-level descriptions and also explore puzzles that use generated images on a turn-by-turn basis.