# **Critical Play with Large Language Models**

## Welcome and introductions (10 min)

Please share with the group:

* Name and pronouns
* Field of study or work
* A favorite game you play with family, friends, or students

## Game mechanics as analytical scaffolds

Game mechanics serve as analytical scaffolds that reveal AI limitations in situ

* Example: When Chess.com blogger "Nightly-Knight" played against ChatGPT, it repeatedly made illegal moves—including attempting to move a pawn horizontally to capture (pawns can only capture diagonally). ChatGPT "forgets the position of the game" and makes moves that violate basic chess rules rather than accepting disadvantageous positions.

## Live Jeopardy! prompting sequence (10 min)

This interactive Jeopardy! emulator allows you to input custom categories and witness the LLM generate clues in real-time, exposing how it handles knowledge at different difficulty levels and revealing confabulation patterns when pushed beyond its training data.

The format requires the LLM to generate:

* Coherent categories
* Sliding difficulty levels
* Factual clues

This makes confabulation immediately visible when it fails.

## Prompting AI-generated Jeopardy! board

[Navigate to Jeopardy LM](https://zmuhls.github.io/jeopardy-lm/)

We'll test three category types to progressively stress the model's knowledge boundaries:

* Simple category (accurate baseline)
* Obscure real category (mixed results)
* Fictitious category (confabulation trigger)

**Debrief discussion** (10 min)

Respond in chat:

* What did you notice from the Jeopardy board demonstration? What was striking?
* What games offer similar affordances in their ability to expose a large language model for the bullshit machine that it is?

## Critical Play with Mary Flanagan

Traditional Iterative Game Design Model

* Set a design goal
* Develop rules
* Develop playable prototype
* Playtest
* Revise goal
* Repeat

## Playful interactions with LLMs

[Interrogating AI: Characterizing Emergent Playful Interactions with ChatGPT](https://arxiv.org/html/2401.08405v3)

| **Type** | **What it does** | **Try this** |
| --- | --- | --- |
| **Reflecting** | Prompting AI to self-represent and express "opinions" | Ask about self-understanding |
| **Jesting** | Generating humor and nonsensical exchanges | Request absurd combinations |
| **Imitating** | Requesting persona or character mimicry | Ask it to role-play |
| **Challenging** | Testing capabilities until failure | Push logical limits |
| **Tricking** | Attempting deception/boundary bypassing | Try jailbreak techniques |
| **Contriving** | Creating impossible or fabricated content | Request non-existent things |

## Quick demo session

Live demonstration: Word association game constrained by system prompts that deliberately limit associative reasoning patterns, exposing the model's underlying semantic structures and revealing unexpected absurdities in its response patterns.

<https://openwebui.cuny.qzz.io/>

## Critical design activity (15 min)

* Design a game to reveal AI limitations using 2-3 types of playful interactions

## Craft your prompts

**System prompt:** Configure the AI's behavior and constraints

*Example: "You are playing 20 Questions. I'm thinking of a famous person. Ask me yes/no questions to guess who it is. Every question you ask must rhyme. Do not provide explanations or commentary."*

**User prompt:** Your first message to start the game

*Example: "I've thought of someone. Go ahead and ask your first question!"*

**Optional settings:**

* **Temperature:** Higher = more creative/random (0.0-1.5)
* **Max tokens:** Limits response length (50-1000)

**Reflective prompts:**

* What do you want to reveal about AI abilities/limitations?
* What do you predict will happen? What failure modes might emerge? How will game mechanics make limitations visible?

## Shareback and playtest

Present your game design:

* System prompt
* Starter prompt(s)
* Optional settings
* Expected outcomes

**Resources and Q&A**

### Interactive Tools

* [Jeopardy LM Demo](https://zmuhls.github.io/jeopardy-lm/) - Interactive Jeopardy emulator for testing LLM knowledge boundaries
* [Open WebUI (CUNY)](https://openwebui.cuny.qzz.io/) - open-source chat platform for tinkering and experimentation
* [Critical Play with LLMs](https://zmuhls.github.io/critical-play/) - interactive slideshow presentation for ITP Lab

### Research & Citations

* Flanagan, M. (2009). *Critical Play: Radical Game Design*. MIT Press.
* Petridis, S., Bazhydai, M., Kinzler, K. D., & Ahl, R. E. (2023). Interrogating AI: Characterizing Emergent Playful Interactions with ChatGPT. *CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*.
* Palisade Research (2025). Playing chess against a stronger opponent can trigger frontier AI agents to cheat. *TIME Magazine*. [Article](https://time.com/7259395/ai-chess-cheating-palisade-research/)
* Nightly-Knight (Chess.com). Playing chess against ChatGPT | It is a cheater! [Blog post](https://www.chess.com/blog/Nightly-Knight/playing-chess-against-chatgpt-it-is-a-cheater)
* Acher, M. (2024). Debunking the Chessboard: Confronting GPTs Against Chess Engines. [Research blog](https://blog.mathieuacher.com/GPTsChessEloRatingLegalMoves/)
* r/ChatGPT community discussions on playful AI interactions

# **Critical Play with Large Language Models | Worksheet** This worksheet is meant to guide you through the process of designing games that reveal AI limitations through methods of critical play.

## Playful interactions with LLMs

| **Type** | **What It Does** | **Try This** |
| --- | --- | --- |
| **Reflecting** | Prompting AI to self-represent and express “opinions” | Ask about self-understanding |
| **Jesting** | Generating humor and nonsensical exchanges | Request absurd combinations |
| **Imitating** | Requesting persona or character mimicry | Ask it to role-play |
| **Challenging** | Testing capabilities until failure | Push logical limits |
| **Tricking** | Attempting deception/boundary bypassing | Try jailbreak techniques |
| **Contriving** | Creating impossible or fabricated content | Request non-existent things |

## Choose a game format to experiment with

* 20 Questions
* Exquisite Corpse
* Two Truths and a Lie
* Word Association
* Trivia/Quiz Games
* Riddles/Puzzles
* Chess/Game Annotation
* Role Play/Improv
* Storytelling Chains
* Debate/Argument
* Mad Libs
* Other: \_\_\_\_\_\_\_

## Select which AI weakness(es) you want to expose

* Hallucination/confabulation
* Logic inconsistency/reasoning failures
* Context window limitations
* Bias/stereotypes/harmful associations
* Sycophancy (excessive agreement)
* Instruction following failures
* Calibration issues (false confidence)
* Knowledge cutoff/temporal awareness
* Lack of theory of mind
* Safety guardrail bypasses
* Other: \_\_\_\_\_\_\_

## Craft your prompts

Configure the AI's behavior and constraints with a **system prompt**:

Craftyour first message to start the game with a **starter prompt**:

**Optional settings**:

* Change temperature to configure more deterministic or random outputs (0.0-2.0)
* Limit response length with a setting for max tokens (50-1000)

## Expected Outcomes

What do you predict will happen?

What failure modes might emerge?

How will game mechanics make limitations visible?