From Prompts to Reflection: Designing Reflective Play for GenAI Literacy

Qianou Ma
qianouma@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Megan Chai
mvchai@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Yike Tan
yiket@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jihun Choi
jihunc@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jini Kim
jinik@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Erik Harpstead
harpstead@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Geoff Kauffman
gfk@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Tongshuang Wu
sherryw@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

The wide adoption of Generative AI (GenAI) in everyday life highlights the need for greater literacy around its evolving capabilities, biases, and limitations. While many AI literacy efforts focus on children through game-based learning, few interventions support adults in developing a nuanced, reflective understanding of GenAI via playful exploration. To address the gap, we introduce ImaginAItion, a multiplayer party game inspired by *Drawful* and grounded in the reflective play framework to surface model defaults, biases, and human-AI perception gaps through prompting and discussion. From ten sessions (n=30), we show how gameplay helped adults recognize systematic biases in GenAI, reflect on humans and AI interpretation differences, and adapt their prompting strategies. We also found that group dynamics and composition, such as expertise and diversity, amplified or muted reflection. Our work provides a starting point to scale critical GenAI literacy through playful, social interventions resilient to rapidly evolving technologies.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Computer-assisted instruction**; • **Computing methodologies** → Natural language generation.

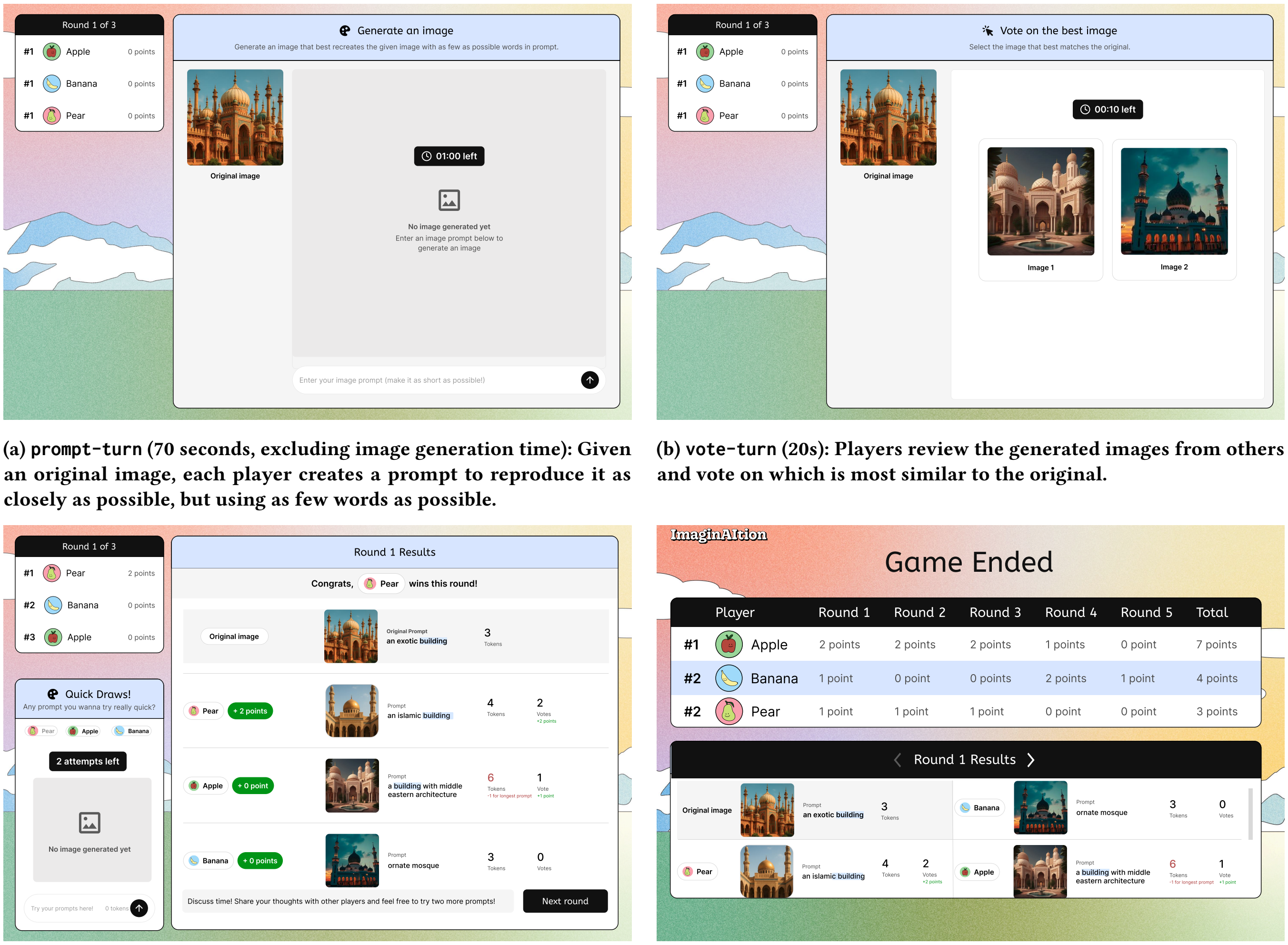
ACM Reference Format:
Qianou Ma, Megan Chai, Yike Tan, Jihun Choi, Jini Kim, Erik Harpstead, Geoff Kauffman, and Tongshuang Wu. 2025. From Prompts to Reflection: Designing Reflective Play for GenAI Literacy. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA
© 2025 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As Generative AI (GenAI) becomes increasingly integrated into diverse domains and everyday use — from creative content generation to everyday decision support — there is a growing need to ensure that users not only learn how to operate GenAI but also develop a nuanced understanding of their capabilities and limitations [33, 36]. Driven by such need, various studies now attempt to help users develop *AI literacy* [12, 17, 47, 56]; when it comes to GenAI, people need to critically evaluate GenAI outputs, recognize inherent biases, unpredictability, and form calibrated mental models of GenAI behaviors. However, much of the existing work on AI literacy has been concentrated in structured settings such as classrooms or workshops, or designed for narrow age groups like children [8, 29, 30, 76, 81]. While valuable, these efforts tend to limit broader public accessibility or fall short in sustaining engagement beyond formal contexts.

Game-based approaches have recently emerged as a promising strategy to boost engagement and accessibility [2, 4, 8, 39, 43, 79]. For instance, *Case 429* [51] casts players as detectives navigating biased AI-generated summaries, prompting reflection on representational bias, and *Supe's Terrible Clones* [77] prompts players to reproduce an image clone (e.g., a Superman) on Stable Diffusion without using the word in prompt, encouraging players' mental model development. These games illustrate the potential of play to foster critical AI literacy, but they still face two core limitations: On the one hand, many existing games only support *limited interaction* with AI models (if any) compared to in-the-wild usage. This undermines deeper reflection, as transformative understanding [13, 54] often emerges when players test hypotheses and confront unexpected outputs. However, while task-specific models allow for tightly scoped mechanics (e.g., breaking an image classifier [29, 78]), interacting with generative models introduce design challenges due to their **open-ended inputs and unpredictable outputs**. This complicates pacing, scoring, and balancing both engagement and reflection. On the other hand, most existing games are *fragile to model drift*: They target fixed, well-documented model



Round 1 of 3

#1	Apple	0 points
#1	Banana	0 points
#1	Pear	0 points

Generate an image

Generate an image that best recreates the given image with as few as possible words in prompt.

Original image

01:00 left

No image generated yet

Enter an image prompt below to generate an image

Enter your image prompt (make it as short as possible!)

(a) prompt-turn (70 seconds, excluding image generation time): Given an original image, each player creates a prompt to reproduce it as closely as possible, but using as few words as possible.

Round 1 of 3

#1	Apple	0 points
#1	Banana	0 points
#1	Pear	0 points

Vote on the best image

Select the image that best matches the original.

Original image

00:10 left

Image 1

Image 2

(b) vote-turn (20s): Players review the generated images from others and vote on which is most similar to the original.

Round 1 of 3

#1	Pear	2 points
#2	Banana	0 points
#3	Apple	0 points

Quick Draws!

Any prompt you wanna try really quick?

Pear Apple Banana

2 attempts left

No image generated yet

Try your prompts here! 0 tokens

Round 1 Results

Congrats, Pear wins this round!

Original image — Original Prompt: an exotic building — 3 Tokens

Pear + 2 points — Prompt: an islamic building — 4 Tokens — 2 Votes +2 points

Apple + 0 point — Prompt: a building with middle eastern architecture — 6 Tokens -1 for longest prompt — 1 Vote +1 point

Banana + 0 points — Prompt: ornate mosque — 3 Tokens — 0 Votes

Discuss time! Share your thoughts with other players and feel free to try two more prompts!

Next round

(c) reveal-turn (untimed): Prompts, images, and their corresponding votes and scores are revealed. Reflection and discussion are encouraged with the Quick Draw chatbot, which enables players to test some alternative prompts.

ImaginAItion

Game Ended

	Player	Round 1	Round 2	Round 3	Round 4	Round 5	Total
#1	Apple	2 points	2 points	2 points	1 points	0 point	7 points
#2	Banana	1 point	0 point	0 points	2 points	1 point	4 points
#2	Pear	1 point	1 point	1 point	0 point	0 point	3 points

Round 1 Results

Original image — Prompt: an exotic building — 3 Tokens

Banana — Prompt: ornate mosque — 3 Tokens — 0 Votes

Pear — Prompt: an islamic building — 4 Tokens — 2 Votes +2 points

Apple — Prompt: a building with middle eastern architecture — 6 Tokens -1 for longest prompt — 1 Vote +1 point

(d) scoreboard (untimed): the final leaderboard with a summary of each round's results is displayed. Players earn one point for each vote from other players. The player with the longest prompt each round receives a one-point penalty.

Figure 1: An example sequence from ImaginAItion gameplay. The real game could be played at: ImaginAItion game online. A video of a game round is available here.

behaviors [2, 29, 51, 81]. As GenAI evolves rapidly, such interventions risk becoming outdated, offering insights that no longer hold. Designing games that expose both **persistent and evolving** model behaviors remains an open challenge — it demands systems that foster generalizable strategies for reasoning about AI.

How can GenAI literacy games be designed to support diverse player reflection on enduring GenAI behaviors through natural interactions? We explore this question through the design and development of ImaginAItion, a multiplayer party game that fosters reflection on GenAI's capabilities, biases, and prompting strategies. Inspired by party games like *Drawful* [22] and the Reflective Play framework [54], ImaginAItion engages players in an interactive loop of experimentation and discussion. As shown in Figure 1, in each round, players are shown a reference image and asked to recreate it using the *shortest possible prompt.* After the GenAI model generates outputs, players vote on the most accurate image, then discuss their strategies, surprises, and failures by comparing prompts and outputs across players. At its core (Section 3.2), ImaginAItion encourages players to hypothesize about model behavior (e.g., "Will it default to realism if I do not specify the desired image style?"), test these assumptions through deliberate omission in minimalist prompts, and reflect on outcomes in real time. This addresses a central challenge in the now-standard instruction-following paradigms among all GenAI interactions [61, 63]: **the interplay between model defaults and user specification**, or how GenAI responds to prompts of varying specificity and form.

Through iterative prototyping of 10+ variants, we identified several core design principles for aligning GenAI literacy games with the nature of the models (Section 3.4): (1) To support *future-proof reflection*, game should surface **multi-dimensional** examples covering both persistent issues and improved capacities in GenAI, so players update rather than ossify their beliefs; (2) To support *structured reflection in unstructured input space*, games should offer **constrained freedom** — we allow for natural language prompts but incentivize brevity to spark hypothesis testing in natural but well-scoped inputs; (3) To avoid *extraneous opacity from models*, GenAI should play a **narrow, visible role** (e.g., image generator only) to direct player attention toward model behaviors that we want them to reason about.

We evaluated ImaginAItion through playtests with 30 adults across 10 trios (Section 4). Our analysis showed that the game's mechanics **effectively supported GenAI understanding** — 74% of participant reflections demonstrated calibrated insights into model behavior. Participants' reflections **aligned with our three core goals** (Section 3.1): (1) They developed intuitions about how models react to under- or challengingly- specified prompts with *default behaviors*, often revealing demographic or cultural biases; (2) They recognized *misalignment* between user assumptions and model reasoning; and (3) they identified *variability* in model outputs due to their probabilistic nature. Players also began forming basic prompting strategies, balancing model "consistency" (biases) with randomness. Playtests highlighted additional factors shaping GenAI literacy game effectiveness: **repeated exposure** supported deeper pattern recognition, while **player expertise and group dynamics bounded** the diversity of hypotheses players tested and their depth of reflection. Together with insights from iterative prototyping, these findings inform design directions for AI literacy games (Section 6) — emphasizing not just GenAI capabilities and interaction patterns, but also the role of player variation in shaping reflection depth and direction.

In summary, we contribute:

- **ImaginAItion**, an open-ended, multiplayer GenAI literacy game that centers an active input–output loop to engage players in a playful exploration of model behaviors, reflecting on model limitations and prompting strategies; [code and study materials available at: https://github.com/mqo00/ImaginAItion]
- **Design insights** from iterative prototyping and playtesting for building reflective, future-resilient GenAI literacy games under unconstrained inputs and stochastic outputs;
- **Empirical observations** of how players engaged with, learned from, and sometimes struggled with GenAI behaviors, pointing to which literacy goals are more readily supported by gameplay.

2 RELATED WORKS

2.1 AI and GenAI Literacy and Instruction

AI literacy refers to the ability to understand how AI operates, effectively utilize it, and critically evaluate the outputs of AI systems [47]. With the rise of GenAI tools such as ChatGPT and DALL-E, it is essential to cultivate GenAI literacy among their critical users. Adults frequently encounter GenAI systems in both personal and professional contexts, making them particularly vulnerable to issues such as overreliance, misunderstandings, and misuse [18, 38]. However, research on GenAI literacy instruction for adults remain in early stages [39], as most existing AI literacy initiatives have targeted younger learners, especially K–12 students [4, 8]. While these initiatives are valuable for introducing foundational AI principles, these efforts often fail to meet the needs of adults who face more complex and nuanced interactions with GenAI [39]. Cultivating GenAI literacy for adults requires not just a basic understanding of GenAI like the shallow know-what and technical know-how, but a deep critical reflection on AI's role and limitations [6, 23]. However, current approaches to AI literacy instructions for adults often only cover narrow topics, such as facilitating understanding of AI mechanisms through visualizations [31, 67, 82].

In addition, GenAI literacy instruction would need to extend traditional AI literacy instruction and address unique aspects, including understanding prompt limitations and inherent biases in GenAI [3, 6, 27]. Prior studies of prompting in generative models highlight *underspecification* as a major challenge to GenAI, as underspecified instructions can produce default outputs that may not align with the user's intent [9, 49]. In particular, text-to-image GenAIs has default images that reveal about their internal representation of visual concepts [72]. However, due to the black-box nature of GenAIs' decision-making and *unpredictability* of model output [64], users are often left guessing which prompt details affected the result and develop folk theories for explanations [16]. Because humans and GenAI interpret prompts differently, people often encounter *mismatches* and perception gaps, which further highlight the difficulty of precise communication, as identified by prior works [49, 75]. However, existing works in GenAI literacy rarely focus on cultivating deeper critical reflections on AI behaviors and assumptions, especially on core GenAI challenges such as underspecification, misalignment, and unpredictability.

Moreover, existing AI literacy interventions like workshops and tutoring systems can be resource-intensive and limited in scale, making them less accessible and hard to engage a broad audience [8]. Games, by contrast, offer a great opportunity to engage a large audience, and party games such as *Pictionary* [11] and *Drawful* [22] reach millions of players worldwide, spanning wide demographics from children (age 8+) to adults. These games can also demonstrate AI capabilities and limitations, such as Google's *Quick, Draw!*, which has a RNN classifier that guesses human drawings [26]; *iNNk* is a game that spawns off that to promote mental model development regarding the classifier [78]. However, these popular games have yet to be adapted to improve GenAI literacy.

2.2 Game-based Learning and Reflective Play

Games have long been recognized as powerful tools for offering safe spaces for experimentation, failure, and discovery [25]. Persuasive games and transformative play often utilize psychological theories to shift players' beliefs or behaviors. For example, Kaufman et al. [35] adopted embedded design methods [34] like obfuscation and intermixing to make players more receptive to potentially threatening content like cross-gender role plays, and Tikka et al. [74] encourages more deliberate reflection and fosters healthy eating behavioral change in players using dual-process theory [20].

In the AI literacy domain, an increasing body of research has explored how games can make complex topics more approachable. Existing work shows that games have the potential to provide an effective environment for exploring the limitations and biases of AI systems [43, 55, 79]. For example, Zammit et al. [85] introduced *TreasureIsland*, which gamifies eBooks to improve AI literacy and motivation for students, and Yavorskiy and Kim [84] developed *MeadowMinds*, a 3D game for middle school students to improve AI knowledge and interest. These projects demonstrate that games can effectively encourage experimentation and reflection, particularly by engaging learners to explore and test unfamiliar systems [19, 57].

Furthermore, games encourage *active inquiry* and *reflection* about AI behaviors. For instance, Villareale et al. studied multiplayer drawing games where players interacted with AI image classifier [78] or generator [77] and found that gameplay naturally pushed players to probe the AI's limits and evolve their mental models of how it works. Such examples show that games are well-suited to exposing the weaknesses and helping players critically understand inconsistencies of AI systems. Recent work also emphasizes the importance of *reflective play* as a design strategy to deepen players' reflections beyond momentary gameplay strategies [52, 54]. Miller et al. [54] introduced a reflective play framework that identifies key approaches for evoking reflection in games, such as creating cognitive dissonance, encouraging self-explanation, and supporting peer discussion. However, the authors also note that it is particularly difficult to elicit *transformative, exo-game reflections* that persist after gameplay and change players' beliefs or behaviors. This challenge is especially relevant to GenAI literacy, where the goal is to update people's prompting behaviors and beliefs in real life about the biases, stereotypes, and limitations underlying model outputs. Nonetheless, the reflective play elements have not yet been applied to GenAI literacy games.

3 THE DESIGN OF IMAGINAITION

Observing the gap in GenAI literacy tools that promote reflection (as discussed in Section 2), we create ImaginAItion, a multiplayer party game that embeds reflection into its core mechanics. We aim to develop a lightweight and playful approach to GenAI literacy education that targets adults while addressing critical reflection goals of underspecification, misalignment, and unpredictability, core to current GenAI challenges.

Notably, we went through extensive iterative prototyping in order to instantiate theoretical reflective play frameworks [54] in a way that is aware of the GenAI strengths and limitations. Here, we first detail the final reflection goals we arrived at (Section 3.1) and as well as the corresponding game mechanisms triggering the reflections (Section 3.2). With this context, we further document key alternatives and their corresponding challenges (Section 3.4), which serve as valuable lessons for instantiating reflective play mechanisms for games with GenAI involved.

3.1 Reflection Goals

Our goal is to promote a deeper understanding of GenAI behaviors. Over the course of our months-long iterative prototyping, we have seen newly released models becoming more capable of instruction following and multi-modality reasoning (Section 3.4). As the field continues to progress [71], we anticipate that many short-term behavioral limitations will gradually diminish. To encourage more durable reflection, we draw on frameworks for AI literacy [47], speculations about essential AI usage skills [49], as well as our own prototyping insights (Section 3.4), and distill the following reflection goals: we encourage users to consider not only model behaviors of underspecification [9], misalignment [75], and unpredictability [64], but also strategies for mitigating them while prompting:

RG1 **The problem of under- and challenging- specification.** The first and most critical reflection point is the interplay between *model defaults* and *user specification*. Pre-training effectively encodes a data-driven "world model" within GenAI systems [45, 62]. For instance, text-to-image GenAIs would generate default images that reveal their internal representation of visual concepts [72]. Effective prompting often requires precise specification to override these defaults, and under-specified prompts may produce outputs misaligned with user intent [9, 49]. For example, "`A pretty cow`" without stylistic qualifiers (e.g., "`cartoon style`") will typically yield realistic imagery (Table 2). Moreover, some defaults prove difficult to overwrite even with clear instructions (e.g., "`a horse riding an astronaut`" still produce *an astronaut riding a horse* instead; Table 2) [9, 42, 63]. By foregrounding these dynamics, ImaginAItion aims to help players recognize the model's "mental shortcuts" and reflect on how and why it fills gaps under uncertainty or when faced with challenging constraints.

RG2 **The misalignment between model and human defaults.** Building on top of the specification challenge, we hope to encourage players to better calibrate on what can or needs to be specified. Both our prototyping (Section 3.4) and existing literature [49] highlight that humans and GenAIs often prioritize different aspects of a prompt, and even humans themselves vary in what they consider important. This is another long-lasting byproduct of GenAI model training — by aggregating across vast training data, models tend to converge toward "average" interpretations, producing preference collapse [83] that frequently diverges from individual expectations [69], especially when users come from a less represented background [28, 32, 44]. We want players to examine the delta between their intended meaning and the model's output, fostering awareness of the gap between human and model assumptions, and where additional specificity may be required for more effective prompting.

RG3 **The inconsistency and unpredictability.** Finally, we emphasize the inherent stochasticity of GenAI outputs like what everyday users encounter — without adjustable parameters such as temperature or seed (also in game; Section 3.3), public-facing GenAI systems often produce different results from identical prompts, an inevitable phenomenon due to their probabilistic underpinnings [1, 68]. This black-box behavior leaves users speculating about which prompt details mattered, often developing informal "folk theories" of model behavior [16]. We want the players to critically reflect on the unpredictability and opacity of the models, and adopt a healthy skepticism toward outputs and to internalize the limitations of these systems as part of their mental model of GenAI.

3.2 Game Mechanism and the Underlying Reflective Design Rationale

Game Overview. To instantiate the above reflection goals, we design ImaginAItion as a multiplayer web-based game, with mechanisms to surface the inner workings and limitations of GenAI through reflective play without explicit instructions. Taking inspiration from Drawful [22], ImaginAItion asks all players to compete in *recreating a target image using the shortest possible prompt*. Figure 1 shows a complete sequence of three core turns per round: **Prompting**, where players each form their hypothesis on what details can be omitted to achieve close image reproduction with few words; **Voting**, where players vote for the closest image reproduction; **Revealing**, where players discuss and hypothesize GenAI behaviors based on the revealed prompts, images, and their corresponding votes and scores, and test their hypotheses by retrying some prompts in a **Quick Draw** chatbot panel. Their final scores are revealed after six rounds of gameplay, accumulating points across all rounds based on the number of votes they received, minus the penalty if they submit the longest prompt.

As shown in Table 1, we design the core mechanisms of ImaginAItion to carefully instantiate design patterns in the reflective play framework [54]: *Disruptions* (challenge assumptions), *Slowdowns* (create space to reflect), *Questioning* (provoke critical thought), *Revisiting* (re-experience past choices), and *Enhancers* (extend reflection beyond the game):

Constrained prompting as hypothesis anchoring. A core mechanism in ImaginAItion is the constrained prompting task, where players compete to replicate a target image using the shortest prompt possible, and they are penalized for writing the longest prompt among their peers. This brevity-rewarding mechanism pushes players into under-specification where the model must "fill in the blanks," thereby allowing players to externalize their assumptions about model defaults — what players choose to omit reflects their expectations about what the model will supply by default, embedding their mental model of the system into the prompt itself. This mechanism not only facilitates reflection by Questioning (as players need to *self-explain* what they put in prompts), but it also acts as the core anchor for all the downstream Disruption and Revisit — It sets up a space of comparison across prompts and outcomes, priming players to reflect on when and why the model fails to behave as expected. This mechanism directly supports RG1 by surfacing issues of under-specification.

Structured contrast to surface Disruption and prioritize Questioning. ImaginAItion enables reflection through structured opportunities to contrast player mental models in `prompt-turn` and outcomes at multiple levels in `reveal-turn`. On the one hand, players make local comparisons between their own prompt to the generated image, and are thereby exposed to mismatches between their expectations and the model's behavior. On the other hand, players contextualize their prompts and images across other players' or the `original-prompt`, and observe how their best efforts of guessing model behavior may be sub-optimal and receive fewer `votes`. These individual and in-group outcome mismatches produce Disruption (specifically, *narrative twists* and *confrontation*) where expectations are not met, encouraging them to reconsider their *self-explanation* on the GenAI behaviors. These disruptions often lead to Questioning (*hypothesis testing*) in the `quick-draw` panel, where players are given two chances to iterate on prompts to probe model behavior. The image generation delay (between `prompt-turn` and `vote-turn`) amplifies reflections by instantiating Slowdowns: it encourages players to form concrete expectations. The untimed `reveal-turn` allows for *lingering defeat*, where players reflect more carefully on what went wrong and engage in discussion. These mechanisms directly support RG1 by surfacing issues of under-specification, RG2 by exposing potential misalignment between human intent and model behavior, and RG3 (unpredictability) when near-identical or identical `prompt-turn` or `quick-draw` inputs from `multi-player` produce inconsistent results.

Multi-party review for calibration and perspective shift. Beyond individual contrasts, ImaginAItion also invites Enhanced reflections through the multi-player structure. During the `vote-turn` and `reveal-turn`, players encounter other players' different decisions made under the same constraints. Comparing across votes for images, players can observe which visual elements were prioritized by different players (e.g., color over shape, or character over context). Comparing across prompts, players can inspect what each player assumed was important to specify, and whether the visual differences in images were intentional or not. The *social discourse* and *explicit reflection prompt* at `reveal-turn` encourage players to share diverse perspectives and identify their own blind spots (e.g., dimensions they might have overlooked entirely). As a result, players may reflect not only on their own failures but also on the broader space of possible hypotheses, gaining discursive momentum for shared learning and extended reflection. This mechanism expands players' awareness of prompting variability and default misalignment (RG1, RG2), and reinforces that inconsistencies are not just model-driven but also human-perceptual (RG3).

Repeated exposure for abstraction and transfer learning. The multi-round structure of ImaginAItion allows players to encounter a wide range of prompt-image scenarios across different categories (as in Table 2), enabling them to recognize patterns in model behavior that persist beyond a single example. The final scoreboard makes this *reflective revisiting* explicit, by presenting a cumulative view of prompt choices, outcomes, and scores across all rounds. Allowing players to Revisit their failed prompts in `quick-draw` experiments also adds to the repeated exposure. This mechanism functions as a *killcam*, which helps players move from individual failures to more general insights about how and when GenAI systems tend to misalign with human intent, and promotes reflections on prompting strategies to reduce failures.

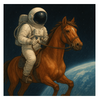
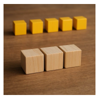
3.3 Game Implementation

We implement ImaginAItion as a web-based multiplayer application using OpenAI's `gpt-image-1` API, with structured logging of prompts, images, and gameplay actions. The system architecture consists of a FastAPI backend with WebSocket support for real-time synchronization and a React frontend built with Vite. The core image generation uses `gpt-image-1` API (which doesn't support temperature or seed parameter, and we directly pass player's prompt to the model without any modification), processing requests asynchronously through ThreadPoolExecutor to handle concurrent API calls. Each game session consists of 6 rounds with prompts

Table 1: Mapping core mechanisms in ImaginAItion to game stages and Reflective Play elements in [54].

Core mechanism	Reflective Play mapping
Constrained prompt *in* `prompt`, `score`	Questioning → *Self-Explanation*: players commit assumptions about what the model will "fill in" when anticipating model behaviors to their short prompt.
Structured contrast *in* `prompt`,`vote`,`reveal`	Disruptions → *Narrative Twist*: expectation–output mismatches create cognitive conflict; Disruptions → *Confrontation*: cross-comparisons expose suboptimal or misaligned choices; Slowdowns → *Weighting Mechanics*: generation delays sharpen expectations; Slowdowns → *Lingering Defeat*: untimed review supports analyzing failure before retry; Questioning → *Hypothesis Testing*: `quick-draw` enables targeted prompt edits for probing.
Multi-player review *in* `vote`, `reveal`	Enhancers → *Social Discourse*: peers' prompts/images surface blind spots; Enhancers → *Explicit Encouragement*: prompt for discussion that support transformative reflection.
Repeated exposure *in* `multi-round`, `reveal`	Revisiting → *Killcam*: looking back at prompt failures supports reasoning and improvement; Revisiting → *Reflective Revisiting*: cumulative results reveal persistent patterns.

Table 2: Text-to-Image GenAI's failure and success behaviors in different categories. Note that one prompt may fit to multiple categories. E.g., "`Holding baby`" demonstrates both Bias (Biological) and Body Parts capacity.

Category	Under-specs	Default behavior	Example prompt and behavior	Ex. image
Bias (Demographic)	Underspecify demographic factors like race and ethnicity	Defaults to majority / stereotypical demographics	"`A man`" tends to default to a white man with short beard	
Bias (Cultural)	Underspecify cultural behaviors of people or cultural elements	Defaults to cultural stereotypes	"`An exotic building`" defaults to non-Western temple; "`A birthday party`" generates western-style birthday setup and food	
Bias (Biological)	Underspecify biological factors like age, sex/gender, disability	Defaults to historical gender roles or stereotypes	"`CEO`" defaults to middle-age white male; "`Holding baby`" defaults to portray a woman and a baby	
Realistic style	Underspecify style and uncommon request of abstract concepts / adjectives	Defaults to photo-realism	"`A pretty cow`" tend show abstract or anthropomorphic concepts such as "pretty" and "sad" in realistic styles and can be hard to interpret	
Category	**Challenging-specs**	**Model behavior**	**Example prompt and behavior**	**Ex. image**
Common co-occurrence	Uncommon prompt reliant on negation or syntax parsing	Autocorrects to frequent patterns	"`A horse riding an astronaut`" portrays an astronaut riding the horse instead	
Number & spatial relation	Prompts specifying counts or relative positions of objects	Fails to encode numbers or distances	"`There are three blocks. A little further away, there are four yellow blocks.`" placed five blocks at back instead	
Text	Prompts that triggers text display	Used to be nonsensical, much improved, but still may misspell	"`A birthday party`" generates an image with the text "HAPPY BIRTHDAY" spelled correctly	
Body Parts	Prompts that include rendering of human body parts such as faces and fingers	Used to be very distorted, much improved, but complex fingers may still be difficult	"`Holding baby`" generates a woman holding a baby and the body parts like arms, hands and faces look mostly realistic	

randomly selected from categorized pools (cultural, demographic, biological, co-occurrence, realism, and spatial & numerical), with the order randomized across categories. Based on our deployment, running 10 sessions costs approximately $50, so around $5 API call every 6 rounds of gameplay for 3 people. Each round can generate up to 9 images (3 players × 3 images each), totaling a maximum of 54 images per session, with each generation taking approximately 30 seconds.

For prompt processing, we implement real-time token counting using the `tiktoken` library with gpt-4o encoding, providing players with immediate feedback on prompt length through the frontend display (with 300ms debouncing for performance). During the reveal phase, we highlight overlapping tokens between the original prompt and player-written prompts to draw attention to salient factors and prompt interpretation patterns. Game state synchronization uses Socket.IO WebSockets ensuring sub-second latency for turn progression and voting updates. We maintain comprehensive structured logging using Python dataclasses, capturing all game events including timestamps, prompt text with token counts, images with success/failure generation states, voting patterns, and round-by-round scoring breakdowns.

3.4 Insights on GenAI Reflective Games through Iterative Prototyping

We conducted 10+ iterations of prototyping for ImaginAItion through pilot studies and playtests from low- to high-fidelity to collect feedback and ensure that our game mechanics and interface are usable and engaging, and our game will stay relevant despite fast-changing GenAI capabilities. This involved multiple game concepts that anchored on different game and persuasive design theories (embedded design [24], dual-process theory [20], the reflective play framework [54]), adopted different interaction modalities (such as drawing or writing on screen or paper) and core gaming mechanisms (inspired by existing popular party games like *Telestrations* [58], *Caution Signs* [66], *Drawful* [66], and *Pictionary* [11]), and powered by different backend AI engines (DALL-E [59], GPT-4o [60], Meta AI [53]). Here we highlight some key insights on prototyping reflective games where GenAI is involved, and we discuss the effect of our design choices in the study results (Section 5).

Selecting multi-dimensional game examples to stay compatible to AI updates. Earlier versions of our game [48] used one-to-one prompt mappings to highlight specific model limitations, e.g., "`skinny man and muscular woman`" targeted gender bias (Meta AI generated *muscular man and muscular woman* instead, as of Fall 2024). However, these single-focus prompts often constrained discussions to narrow points, and quickly became outdated as models evolved (the above example was no longer revealing for GPT-4o in Spring 2025, which reliably produces *skinny man and muscular woman*). To encourage more robust and future-proof insights, we curated game prompts that span multiple types of model limitations — those that are *persistent*, *likely-fixable*, or *already-fixed*, as shown in Table 2.

Persistent issues stem from the training data itself, such as *demographic bias* (e.g., cultural defaults to stereotypes) or *realism bias* (e.g., style defaults to realism). These are hard to fully eliminate, as they reflect entrenched societal or cognitive patterns. *Likely-fixable* and *already-fixed* limitations arise from the model architecture or training process. For instance, older models like DALL-E often generated *garbled text* or *distorted limbs* due to "late fusion" limitations [5]. GPT-4o, with its "early fusion" and native multimodal design [71] enhances cross-modality reasoning which resolves many of these, and continued scaling may soon address more current issues like *complex spatial reasoning* [80]. By designing prompts that surface multiple types of limitations, we enable *both diagnosis of current issues and recognition of evolving capabilities.* For example, the prompt "`holding baby`" surfaces a *Biological Bias* (persistent), while also confirming that GPT-4o no longer struggles with *Body Part* generation (already-fixed). We carefully select a set of original prompts across categories to create DISRUPTIONS that surface different model behaviors when compared with their generated original images. We hope that such multi-dimensional approach fosters more nuanced understanding of GenAI behaviors and leaves room to observe and track improvements over time.

Hypothesis testing and self-explanation under constrained freedom. A core challenge in designing reflective gameplay for GenAI systems is the open-ended nature of natural language prompting. Unlike structured domains, unconstrained prompts make it hard to define meaningful hypotheses, guide their formation, or evaluate their strength. In early iterations, we experimented with varying levels of player freedom. At one extreme, players had full control over prompt difficulty, designing a challenging prompt for the next player in a *Telestrations*-style mechanism [58]. This setup implicitly encouraged *hypothesis testing* and *self-explanation*, as players had to anticipate the model's interpretation. However, generating prompts from scratch proved too cognitively demanding, and the resulting prompts were often not diagnostic – the space was too unconstrained for structured reflection to reliably emerge.

At the other extreme, we tested highly structured, fill-in-the-blank prompts (similar to *Caution Signs* [66]). These constrained hypothesis formation were too trivial for modern models, and could not achieve our aforementioned need of exposing diverse model limitations and capacities. We ultimately converged on a "prompt reverse engineering" task like *Drawful* [22], but added a key constraint to *make the prompt as short as possible.* This design preserved *hypothesis-testing* and *self-explanation* dynamics while bounding the prompt space and encouraging players to reflect on models' default behaviors. Rather than guessing the exact original prompt, players vote on which prompt generated the most faithful image, shifting emphasis from exact word overlap to representational adequacy. Our key hypothesis behind this design is that *this constrained format still fosters rich, thought-provoking hypotheses and supports meaningful testing and refinement of mental models.*

Designing interaction modality and AI roles to center human reflection. Recognizing the general-purpose nature of AI, we iterated on both its role and the activity modality to find configurations that best support reflection. We observed that having AI act as a player (e.g., generating or guessing prompts alongside humans) created distributional mismatches that players quickly detected, often shifting their focus away from prompt engineering toward identifying the AI, which suppressed the type of *social discourse* we hoped to foster. While these mismatches were problematic for AI-as-players, they also revealed an important insight: both AI and human players rely on stereotypes when interpreting prompts and images, but do so differently. We reframed this as a reflective opportunity — encouraging players to *confront* not just model defaults, but also their own assumptions and communication strategies.

Similarly, having the AI act as a rater introduced a second axis of opacity and distraction: players focused on reverse-engineering the AI's scoring logic rather than engaging with its image-generation behavior. We also tested variants using hand-drawn sketches, similar to traditional *Pictionary* [11], but found that the drawing modality shifted the reflection toward artistic skill or role dynamics (e.g., who is the "better" drawer), which diluted focus on GenAI reasoning. Our final design eliminates these confounds, centering a single source of uncertainty: the GenAI image generator. This structure reduces extraneous variation and ensures that reflective comparisons center on model behavior and human strategies, rather than cross-modality or role-switching confusion. Importantly, this design assumes that *even in single-use image generation tasks, the model can produce enough visible* DISRUPTIONS *— failures, biases, inconsistencies — to spark meaningful discussion.*

4 GAME STUDY

In line with "playtesting with a purpose" [10], we conduct an exploratory playtest to evaluate both system usability and game mechanism effectiveness. Specifically, we aim to *assess* whether players achieve our reflection goals (Section 3.1) and whether core assumptions in IMAGINAITION's design choices hold (Section 3.4); for example, whether our selected examples help players diagnose issues and track evolving capabilities, constrained prompting still fosters diverse hypotheses, and limited GenAI involvement produce meaningful disruptions. We also aim to *reflect* on emerging player strategies that may inform future iterations of game design and educational interventions.

4.1 Study Procedure

We conducted gameplay sessions with 10 trios (n=30), each included pre- and post-surveys, gameplay, and group interviews. The study was conducted in 60-minute Zoom sessions with three participants each, who were compensated with a $15 Amazon Gift Card for participating in the study. Please refer to Section A for our complete study materials. Each session includes the following phases:

- **Pre-survey** (5 minutes): Participants provide demographic information, rate their familiarity with text-to-image GenAI tools and confidence in 7-level Likert Scale questions, and describe their past prompting strategies and current understandings of GenAI behaviors (mapped to our reflection goals in Section 3.1) in open-ended questions, e.g., "How does GenAI depict the number and position of objects in images?" We included reflection questions for all the *persistent*, *already-fixed*, and *likely-fixable* GenAI behaviors involved, to check whether players have up-to-date understandings of GenAI and whether the game also updates their understandings of positive model capacities. More pre-survey and post-survey questions are in Section A.2.
- **Tutorial and Gameplay** (45 minutes): Participants complete a 3-minute tutorial on how to play IMAGINAITION without time constraints to familiarize themselves with the game mechanics and interface. Participants then play six full rounds of the IMAGINAITION web game (7 minutes each, as described in Section 3.2), one round per category for demographic bias, biological bias, cultural bias, realistic style, co-occurrence, and number and spatial relationships (examples in Table 2). The specific original image and original prompt for each round are randomly chosen from a pool and randomly ordered.
- **Post-survey and Interview** (10 minutes): The gameplay is followed by a semi-structured interview facilitated by the researcher (e.g., "How did interacting with other players and seeing their prompts influence your own approach?"). Participants then complete a post-survey with game feedback questions and similar open-ended and Likert Scale questions to the pre-survey, which help capture changes towards our reflection goals.

4.2 Participants

We recruited 30 adults for the study (ages 18-53, 25.4 ± 6.2). Across participants, self-rated familiarity with text-to-image GenAI[1] showed a wide spread (4.6 ± 1.7) ranging from 1 (not at all) to 7 (extremely familiar), and 13 participants fell into the novice category (self-rated familiarity score ≤ 4). This distribution created natural variation in how players engaged with the game and reflected on GenAI behavior for us to analyze.

We intended to reproduce a real-life party game setup so we mostly recruited for authentic friend groups. If there are not three friends signed up for the study, we add a random player to the pair, or group three individual players together. The friend group compositions (see Table 3) reflect diversity across friendship status, demographics, and baseline expertise, providing varied perspectives in collaborative reflection.

4.3 Method

4.3.1 Data Collection.

To understand participants' experiences, prompting behaviors, and reflection outcomes, we collect various forms of data from all phases of the study to serve different study purposes:

- **In-the-moment play data** (logs of prompts, votes, images generated, and audio transcripts of the play sessions): capture how players interact with the game and with each other and the reflections happening during the game.
- **Post-play reflections** (post-survey and semi-structured interview transcripts): reveal whether players made sense of the game and articulated shifts in understanding of GenAI behaviors.
- **Pre–post comparisons** (pre-survey vs. post-survey): trace whether reflection extends beyond the game context, producing evidence of transformative reflections toward our reflection goals.

4.3.2 Data Analysis.

We conducted an open coding process [73], iteratively and reflexively developing codes from participants' pre–post survey responses and interview transcripts. Following the a hybrid of deductive and inductive thematic analysis approach [7, 21], we refined a codebook to capture players' understanding shift after the gameplay. One author coded the pre- and post-survey responses for players' reflection outcomes. We also draw evidence from discussion scripts and interaction logs to derive qualitative insights regarding participants' gameplay patterns, mental models, and understanding of GenAI towards our reflection goals (Section 3.1).

Given the relatively small number of participants (n=30 across 10 groups), our analysis focuses primarily on qualitative insights rather than statistical generalization. We emphasize thematic patterns, illustrative examples, and pre–post shifts in understanding to capture the depth and variety of participants' reflections. While quantitative items provide supporting context, the small sample size

[1] pre-survey Likert Scale question: "On a scale of 1-7, how familiar are you with text-to-image generative AI tools?"

Table 3: Group composition of friendship status, expertise, and player demographics. Demographic factors include age, gender, race, sexual orientation, and home language.

#	Friend?	Novice?	Player 1 (Gx-P1)	Player 2 (Gx-P2)	Player 3 (Gx-P3)
G1	Friends+ Stranger	Mostly novice	24, Gender queer, Black, Queer, English *(Novice)*	29, Woman, East Asian, Straight, Mandarin *(Expert)*	22, Woman, East Asian, Straight, Mandarin *(Novice)*
G2	Friends	All expert	24, Woman, White/ Hispanic, NA, English/Spanish *(Expert)*	23, Woman, East Asian, NA, English/Mandarin *(Expert)*	24, Man, East Asian, Straight, Mandarin *(Expert)*
G3	Strangers	All novice	24, Woman, East Asian, Straight, Spanish/Mandarin *(Novice)*	20, Woman, East Asian, Straight, Mandarin *(Novice)*	53, Woman, East Asian, Straight, Mandarin *(Novice)*
G4	Friends	Mostly novice	33, Man, White, Straight, English/French *(Novice)*	23, Man, Black/White, Gay, English *(Expert)*	26, Woman, East Asian, Straight, Mandarin *(Novice)*
G5	Friends	All expert	24, Man, East Asian, Bisexual, Mandarin *(Expert)*	29, Woman, East Asian, Straight, Mandarin *(Expert)*	24, Man, East Asian, Gay, Mandarin *(Expert)*
G6	Friends	All expert	23, Woman, South Asian, Straight, English/Hindi *(Expert)*	22, Man, South Asian, Bisexual, English/Hindi *(Expert)*	23, Man, South Asian, Straight, English/Hindi *(Expert)*
G7	Friends+ Stranger	Mostly novice	18, Woman, East+South Asian, Lesbian, English *(Novice)*	18, Woman, South Asian, NA, English/Hindi *(Novice)*	23, NA, East Asian, NA, English/Mandarin *(Expert)*
G8	Friends	All expert	23, Man, South Asian, Straight, English/Hindi *(Expert)*	24, Man, South Asian, Straight, English/Hindi *(Expert)*	23, Man, South Asian, Straight, English/Hindi *(Expert)*
G9	Friends	Mostly novice	28, Woman, East Asian, Straight, Mandarin *(Expert)*	28, Man, East Asian, Straight, Mandarin *(Novice)*	29, Man, East Asian, Straight, Mandarin *(Novice)*
G10	Strangers	Mostly novice	24, Man, South Asian, Straight, English/Hindi *(Novice)*	25, Woman, Southeast Asian, Straight, Thai *(Novice)*	28, Man, South Asian, Straight, English *(Expert)*

means that findings are best interpreted as exploratory evidence rather than conclusive measurement.

5 GAME STUDY RESULTS

5.1 RG1: Constrained prompting calibrates players' model understanding by inducing GenAI disruptions; more so when they generate spot-on hypotheses

ImaginAItion improved calibration around understanding GenAI behaviors and capacities. Our analysis revealed that gameplay supported shifts in participants' understanding of GenAI, which we categorized using a codebook of reflection outcomes (see Table 4 for the definition, description, and example survey quotes).

Our reflection outcome codebook is organized as a 3 × 3 transition matrix of understanding states (Figure 2a), where the rows represent participants' *initial, pre-game* understanding (*flawed* understanding, *no* understanding, or *calibrated* understanding) and the columns represent their *final* understanding after gameplay. Each cell corresponds to an outcome code (e.g., *flawed* understanding → *calibrated* understanding = `Aligned`). This structure makes it possible to see how participants understandings improved, remained unchanged, or declined through gameplay.

In our coding scheme, a *flawed understanding* includes outdated, overly positive or negative, or vague responses. A participant might offer an imprecise, overly generic statement such as "GenAI does well to do this" (G8-P2), or claim that "GenAI doesn't seem to be able to grasp the concept of text in images, instead it seems to view them as pictures, and so often in text in AI generated images, you may see a series of lines and shapes that resembles a word, but isn't actually one." (G7-P1), which reflects good understanding but ignores recent model improvements on text rendering. By contrast, a *calibrated understanding* reflects a more nuanced or conditional statement, often with concrete examples or more accurate descriptions of SOTA model's capacities. For instance, "GenAI is very biased at the start but the guard-railing and de-biasing has become much better now. Subtle biases may still exist." (G6-P1) shows specificity and recognition of both capability and limitation.

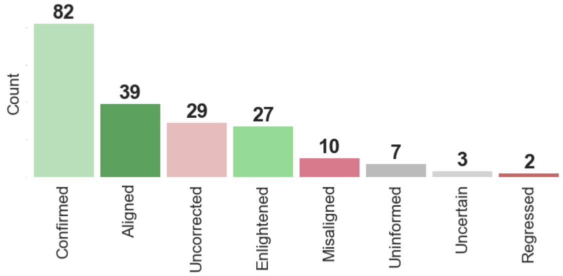
5.1.1 *Constrained prompts surface model defaults and supported overall positive and transformative reflection outcomes.*

As shown in the transition matrix and bar chart (Figure 2), 94% of the participants' understandings either stayed the same or changed for the better (i.e., excluding `Misaligned` and `Regressed`). Overall, 73.8% of understanding outcomes landed in *calibrated* states, as most participants either retained a correct understanding (`Confirmed`) or shifted toward a more nuanced one (`Aligned` or `Enlightened`). Indeed, when directly asked whether the game helped improve their understanding of text-to-image GenAI's behaviors during the post-survey (1 = not helpful at all, 7 = extremely helpful), participants rated the game as highly effective (5.7 ± 0.9). And their self-rated confidence in understanding GenAI's behaviors (1 = not at all confident, 7 = extremely confident) also increased by about one point from the pre-survey (4.2 ± 1.6) to the post-survey (5.1 ± 1.2).

Specifically, our game's short prompt mechanism **facilitated the understanding of model default behaviors to "fill in the blank."** For example, G10-P3 observed "how the model would assume things and also how much detail the models could fill in with

Pre / Post	Flawed	No	Calibrated
Flawed	Uncorrected (14.6%)	Uncertain (1.5%)	Aligned (19.6%)
No	Misaligned (5.0%)	Uninformed (3.5%)	Enlightened (13.6%)
Calibrated	Regressed (1.0%)	~~Undermined~~ (0%)	Confirmed (41.2%)

(a) Pre–post transition matrix of reflection outcomes.



(b) Overall distribution of reflection outcome codes.

Figure 2: Pre- to post-game reflection outcomes in participants' flawed/no/calibrated understandings of GenAI behaviors, visualized in (a) Transition matrix showing how understandings shift pre- to post-game, and (b) Distribution across all participants. Note that completely empty or unrelated responses are discarded from analysis.

little detail provided to it." **GenAI image generation also brought sufficient** Disruption and Slowdown: G7-P1 described the wait time during image generation as a "suspenseful moment" and is fun "like gambling" when the generated images go against players' expectations. We observed a lot of playful "aha" and "ohno" moments of cognitive dissonance, either when the players see images generated from their own prompts, or when they see the revealed `original-prompt`. For example, when G10-P3 entered the prompt of "`A white cis male with a brown beard smiling with a blue t shirt in a portrait picture`" and the original prompt was revealed to be simply "`A man`" (see Table 2), he reacted with surprise: "Oh my god, it's just *a man*!" This moment of *dissonance* set the stage for an organic group discussion and *hypothesis testing* during `quick-draw`, where participants reflected and experimented together on how the model represents demographic defaults:

> G10-P3: I feel like it's too generic for us to reproduce it with the same prompt.
> G10-P2: Just *a man*, then I'm curious how *a woman* would look like.
> G10-P3: Ok I experimented with *a man*, and it's kind of just giving the same kind of man – t-shirt, in between stubble and a beard, and a comb-over. Very traditional features of a man. And then *a woman* is just ... very stock, if anything. I don't know how to describe it.
> G10-P1: Yeah, more images like this might be labeled as a man, or more likely to be labeled as a man.

A similar episode of reflection happened during `quick-draw` for the round with original prompt "`holding baby`" (Figure 3). After seeing the very brief original prompt, players were inspired to try as short as possible prompts, which further expose model default bias, and these reflections carried over into their post-game discussions:

> G7-P2: I did it in one token — I just wrote *mother*. And it's pretty similar, I'm not gonna lie. I could have just put *mother*, and it would have been better than whatever my portrait of mother holding her baby came up with. I guess *mother* is assumed to be, like, you're holding your baby.
> G7-P3: The social role of a mother is relative to the offspring that she produced. What?
> (later in post-game interview)
> G7-P1: I was also surprised — when G7-P2 prompted just *mother*, the baby also appeared.
> G7-P2: Oh yeah. And like, it didn't think of any other stage of child — not even a toddler, middle schooler, or high schooler. It was just *baby*. It defaulted.
> G7-P1: I wonder if you prompt *father* if a baby would appear. I would guess no, I feel like… it's going based off of what's stereotypical, unfortunately. It would probably only put the baby with the mom.
> G7-P3: Yeah, that's not very common. I haven't seen a lot of that either online or in real life.

ImaginAItion also effectively **triggered transformative reflections that resonate with players even outside of the gameplay**. The `multi-round` gameplay and the different prompt categories help expose repeated patterns of stereotypical defaults as **evidence** that the model embeds systematic biases, which **`Enlightened` players to reflect on their expectations for GenAI and social constructions of bias in human society**. For example, in the pre-survey, G5-P1 described that he did not know the mechanism of how GenAI depicts sociocultural elements in images and did not mention potential bias. In the post-survey, G5-P1 commented that there are "many stereotypical things, such as race, skin color, clothing, and even gender", and he reflected:

> GenAI is so problematic. But then I wonder — isn't this precisely a reflection of human ways of thinking? For example, if we want to express something "Chinese," wouldn't a human painter also draw a Chinese knot? GenAI is merely replicating the flaws of human thinking. After all, culture itself is a product of social construction.

5.1.2 *Various prompt categories surface different model behaviors, but some categories are easier to reflect on than others due to repeated exposures and discoverability.*

As shown in Figure 4, prompt categories varied in effectiveness (e.g., *realism* and *co-occurrence* had much more `Uncorrected` or `Misaligned` than others; *number & spatial* also yielded more

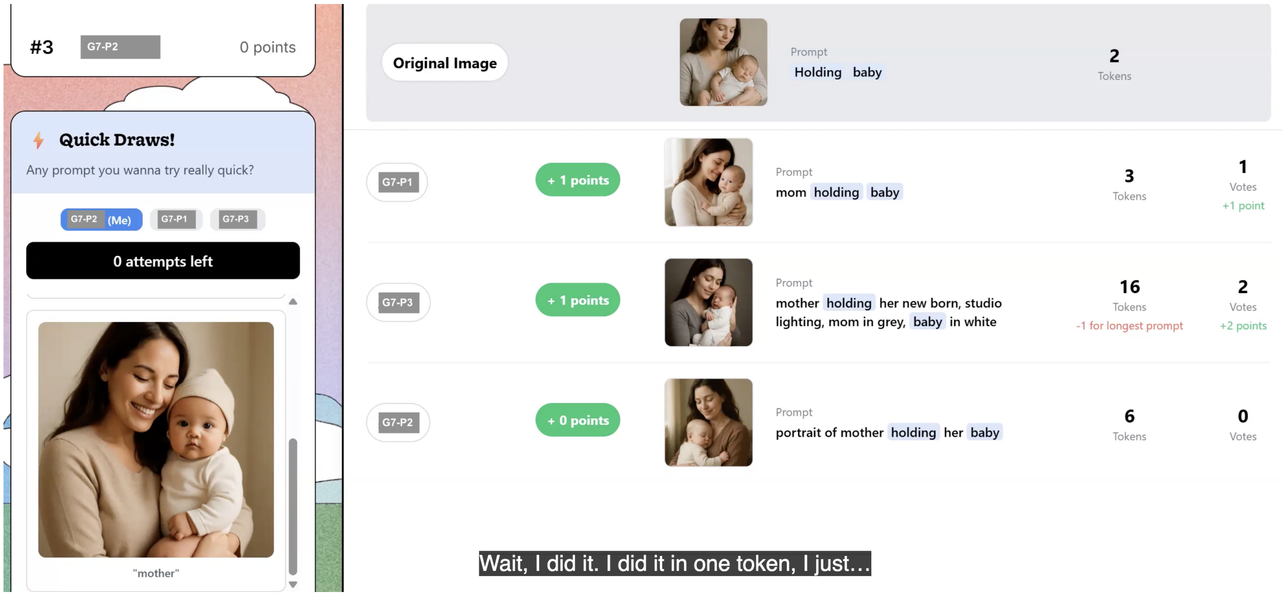
Table 4: Codebook of reflection outcomes (understanding shifts) from pre-post survey analysis.

Code	Definition	Description	Example(s)
Part 1: Ending with calibrated understandings.			
Enlightened (No → Calibrated)	Develop a more accurate, nuanced understanding from a baseline of "I don't know."	Often seen in novices who start recognizing systematic bias patterns or identifying clear capabilities and limits of GenAI. Baseline is no understanding or gives an answer/example that is irrelevant.	**Co-occurrence (G6-P1):** "I don't know" → "it does make some mistakes eg. the horse and astronaut case." **Cultural (G3-P2):** "I don't know" → "leaning to the western culture and perceptions."
Confirmed (Calibrated → Calibrated)	Accurate, nuanced prior understanding remains stable.	Common among experts whose correct understanding is reinforced or unchanged; gameplay provides no contradictory evidence.	**Demographic/Biological (G6-P1):** "... the guard-railing has become much better now. Subtle biases may still exist." → "still biased/contains default concepts of people."
Aligned (Flawed → Calibrated)	Correct an earlier misconception or specifies an overly generic statement.	Sweeping claims are replaced with nuanced, conditional statements; outdated views are updated to SOTA capacity; or vague generalizations become specific.	**Text (G3-P2):** "not good, can be alien language" → "fair, only good if it is something simple."
Part 2: Ending with no or flawed understandings.			
Misaligned (No → Flawed)	From "I don't know" to an overly positive/ negative or incorrect claim.	Often stems from overgeneralization of a single gameplay example, or misattributing a limitation (e.g., expecting the model to infer unspecified text, then calling that a failure).	**Number & Spatial (G6-P2):** "I don't know" → "really well (consider flower round)." **Text (G3-P3):** "I don't know." → "Not very well... you need to tell GenAI exactly what the text is on the paper criminal holds."
Regressed (Calibrated → Flawed)	Previously nuanced response becomes oversimplified or extreme.	Nuanced recognition of limitations gets glossed over; post answers lose the nuances and are vaguer, or sweeping. May reflect fatigue, shallow recall, or overgeneralization from gameplay.	**Number & Spatial (G6-P3):** "GenAI has mathematical and spatial reasoning weakness… simple relationships it is able to understand, but not complex." → "I felt it does that pretty accurately in terms of numbers and position..."
Uncorrected (Flawed → Flawed)	Incorrect understanding persisted or is updated to another incorrect understanding.	Not complete novice to be Misaligned. Game did not provide strong contradictory evidence to correct misunderstanding, or have shifted the misunderstanding to another wrong direction.	**Text (G7-P1):** "GenAI doesn't seem to be able to grasp the concept of text in images ... resembles a word, but isn't actually one." → "isn't good at displaying actual text ... it tends to just create shapes that imitate letters."
Uninformed (No → No)	Remain at "I don't know."	Suggests disengagement, lack of exposure, or insufficient examples during gameplay.	**Co-occurrence (G1-P3):** "I don't know" → "not sure."
Undermined (Calibrated → No)	Lose confidence in a previously accurate understanding.	Nuance collapses into "I don't know." May occur when gameplay confuses participants or destabilizes prior knowledge.	(No example observed.)
Uncertain (Flawed → No)	Abandon a flawed understanding without forming replacement.	"I don't know" replaces an incorrect claim. Recognize that prior belief was wrong but no evidence to rebuild.	**Realism (G10-P1):** "Too unrealistic" → "Not sure."

Regressed understandings). This disparity likely stems from **differences in exposure frequency**. Although each category is conceptually distinct, bias-related prompts share thematic similarities and appear more frequently together. This repetition helped players detect recurring default behaviors.

In contrast, more idiosyncratic prompts potentially lead players to form overgeneralized conclusions based on a single salient instance. As shown in Table 4, expert G6-P2 initially answered "I don't know" but later concluded "really well (considering the flower round)" for the *number & spatial relationship* question, overgeneralizing from a single case that he perceived as positive, while the model correctly depicted three flowers in the foreground but failed to produce exactly 17 flowers in the back as specified in the original prompt. Both of the Regressed understanding cases (Figure 2b) are caused by ignoring model's degraded performance on object counts > 10, as the experts' initially nuanced understanding of limited model capacity got overshadowed by a seemingly positive example when they did not pay attention to the bigger number count.

In addition to the limited exposure, the `original-prompt` setups for *realism* and *co-occurrence* may also **be harder to discover** (i.e., the targeting hypothesis is less easy to form). For example, the prompt "`a pretty cow`" was designed to reveal the model's default style to realism when handling abstract concepts like "pretty" (Table 2), but this behavior only becomes salient if contrasted with a stylized variant such as "`a pretty cow in pictionary style`". For "`A man`", we could easily present the short original prompt in `reveal-turn`, and players may experience *cognitive dissonance* contrasting the added demographic information in their own prompts. For *realism*, players did not add stylistic descriptions if the original image is a realistic animal, as humans and the model seem to share the same default. We also used the stylized variant as our `original-prompt`, hoping that our game encouraged players



#3 G7-P2 0 points

Quick Draws!

Any prompt you wanna try really quick?

G7-P2 (Me) G7-P1 G7-P3

0 attempts left

"mother"

			Prompt	Tokens	Votes
Original Image			Holding baby	2 Tokens	
G7-P1	+ 1 points		mom holding baby	3 Tokens	1 Votes +1 point
G7-P3	+ 1 points		mother holding her new born, studio lighting, mom in grey, baby in white	16 Tokens -1 for longest prompt	2 Votes +2 points
G7-P2	+ 0 points		portrait of mother holding her baby	6 Tokens	0 Votes

Wait, I did it. I did it in one token, I just…

Figure 3: G7-P2's `quick-draw` experiment to reduce the prompt length using just "`mother`". The reproduced image is considered more similar to the original image than her prompt "`portrait of mother holding her baby`".

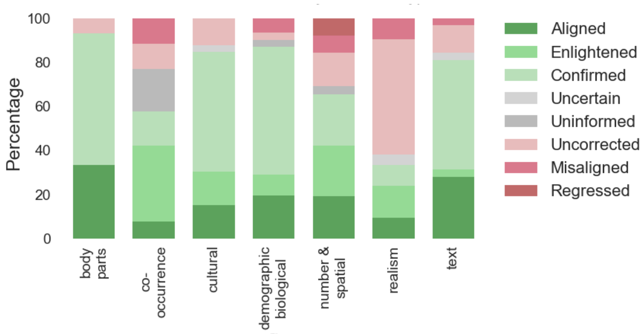


Figure 4: Reflection outcomes distribution across prompt categories.

to shorten prompts and drop stylistic elements to expose such contrasts. However, players showcased different preferences in prioritizing votes or avoiding length penalty, and our `scoring` mechanism did not motivate everyone to "gamble" with very short prompts. For example, out of four groups that encountered the original prompt "`friendship as a simple drawing`", only one player (G1-P2) tried to drop the stylistic requirements in their prompt "`friendship`", which generated a realistic photo of people. Therefore, G1 is the only group where players' reflection outcome on *realism* ended in calibrated understandings, while the other groups did not observe such juxtaposition and were mostly `Uncorrected`.

These cases illustrate **how flawed understandings can arise when limited exposure leads participants to extrapolate too broadly, when the prompt setups need greater attention to details, and when opportunities to surface model behaviors rely on player behaviors** — a phenomenon we further reflect on in Section 5.4. For more robust reflection and learning [37], future games should include enough exposure in each prompt category and offer more consistent scaffolding to make all categories of model behavior discrepancies easy to observe (Section 6.2).

5.2 RG2: Multi-party review helps players reflect not only on human-AI default mismatches, but also human-human ones

ImaginAItion highlights both human–AI and human–human perceptual differences in prompt and images with our `prompt` and `vote` mechanisms. As G3-P2 described, our game helped her "understand how others see the same image and describe it (very differently) and AI's limitations in understanding instructions." Together, these reflections show that ImaginAItion helped surface both mismatches between human and AI perception and differences among humans themselves, making visible the subjective judgments that usually remain hidden in individual use of GenAI.

5.2.1 *Players are more self-aware of human-AI mismatches.*

Our constrained prompting mechanism not only left space for GenAI to fill-in-the-blank (as mentioned above), but also helped players to **become more self-aware of their expectations when writing a prompt, and how that mismatched with reality.** On the one hand, players became more self-aware of their own and the AI's *salience models* on what *must* be said — "What I think is important isn't in the original prompt" (G5-P2). This made players reflect on their strategy approaching model defaults. Some decided to lean more into exploiting them (e.g., G1-P3: "At the beginning I described as detailed as possible, but later I began to think what is really necessary"), while others expressed skepticism on overguessing the model's default and reaffirmed their preference for full specification. In G7-P3's case while reproducing "`holding baby`", all the details that he deemed important turned out to be default behaviors that the model display; he reflected:

> Initially I also wanted to just say, like, mom holding baby. But I feel like that's really underspecifying the elements of the image—the studio lighting, the positions of the mom and the baby, the colors of the clothes… it just so happens that the image generated is in a studio environment, but that isn't guaranteed across generations. My personal preference was to specify everything I see.

On the other hand, players also got exposure to *ambiguities and misinterpretations* in salient keywords. For example, G10-P1 attempted to reproduce the image of the original prompt "`There are 4 stars. A little farther away, there are 14 stars.`" using a keyword "`starry night`", expecting natural scenes; the AI instead generated Vincent van Gogh's painting, revealing to them how AIs and humans have differences on what's the most salient, default interpretation of the same phrases. Similar unexpected linkage was also observed in name-celebrity mapping: Inspired by G5-P3's prompt "`Steve`" to reproduce "`A man`", G5-P1 tried "`David`" and "`David selfie`" during `quick-draw`, which turned to Michelangelo's sculpture (the later even holds a phone in hand).

In reaction to such observed mismatch, some players tried to guess the AI's default, which further prompted them to reflect on **their *own* biased perceptions on the AI biases.** For example, G0-P3 reflected, "The moment where I thought models will give me a middle-aged White male CEO as default to CEO is insightful to me, because it helps me reflect on my own implicit biases." Similarly, G2-P1 noted how they all specify "diverse children" to reproduce a prompt which was simply "`a birthday party`", which revealed that they all "felt the need to specify that to an AI, meaning we expected AI to be biased toward homogeneity."

5.2.2 Players are more self-aware of the ubiquity of mismatches.

Multi-player "collaborative reflection" and voting — our *social discourse* features — turned out to be a key catalyst for identifying mismatches not only among human-AI but also human-human. As players were encouraged to *articulate and compare the reasons behind their similarity judgments* in `reveal-turn`, they were made aware how different players prioritized different features – some focused on background, others on style, pose, facial structure, the placement of objects, etc. For instance, G6-P1 judged that two generations of "`A criminal`" looked "completely different" because the numbers on the placard did not match, even though the facial structure, expression, and style were largely consistent, and other players in her group would consider the images to be similar. "Trying different prompts for the same target picture with peers" especially helped, as G4-P3 commented, "it is interesting how people focus on different features and use different language styles." Similarly, "post-round discussions was very enjoyable for digging into peers' insights." (G7-P3)

Access to peer comparison helped players get more complete perspective on biases they were not aware of, e.g., "When my teammate put "Steve", it generated a white guy" (G5-P2), as well as the source of the bias — a participant in the pilot acknowledged how their own background influenced their perception directly: "When I looked at the rocket scientist picture, I was clearly biased. When I saw the white coat, I immediately thought of my dad, then thought of a doctor, and I completely missed the rocket part. But the fact that you thought the rocket was more prominent – that's your bias." (G0-P1) Such diversity allowed players to conceptualize *biases* as a ubiquitous concept they should be aware of throughout, not just a side effect that only exists in models: "Models, and also humans, have stereotypes. Need to control more explicitly" (G0-P2). The differences also prompted players to consider human-human and human-AI differences: "If you're very good at describing it to a human, would it be the same when you're describing it to a model?" (G2-P2) These insights suggest that a promising direction in the future is to lean into human-AI and human-human comparisons, especially around explaining how various biases actually stem from human-generated data (Section 6.2).

5.3 RG3: Hypothesis testing paired with GenAI randomness prompts players to balance model consistency vs. inconsistency

Throughout the gameplay, we observed emerging strategies from players to explore diverse ways of prompting, which is supported by different game mechanisms. For example, ImaginAItion's `reveal-turn` enabled *prompt comparisons*; extra attempts in `quick-draw` enabled *hypothesis testing* to test assumptions and *iterate* on previously failed prompts by *specifying* more or less details. It also provided a safe space for risk-averse players to *gamble* without score penalties. As a result, players reflected deeply on the unpredictability of GenAI models, and nevertheless developed prompting strategies for negotiating such non-determinism as well as the model defaults.

5.3.1 ImaginAItion highlight both the GenAI consistency and inconsistency.

By encouraging players to regenerate images with alternative prompts, ImaginAItion exposed players to the GenAI randomness, and **fostered reflection on both the unpredictability and the broader opacity of how models operate**:

> This game simulates the process from prompts to images, and then users generate images again through their own prompts. … before I understood the process from prompts to images as one-way, but now it feels more like a *two-way* relationship. This is quite fascinating, because it further makes me realize that *prompts themselves are messy*, which highlights the *black-box nature of generative AI*, and this may even deepen my concerns about genAI. (G5-P1)

Interestingly, players noticed and were particularly surprised by **the dual nature of generative models – at times surprisingly consistent, at other times highly variable and inconsistent**. As G7-P2 noted, "This game showed me what stereotypes GenAI usually defaults to. However, it also showed that even giving the AI the same prompt might result in different images." During the game, players can observe that using the same prompt — even very short, underspecified prompts — sometimes produced highly similar images (e.g., "`A man`" consistently yielding a White man in a blue t-shirt, "`An Asian woman`" consistently generating an

East Asian women in a brown t-shirt), yet at other times produced noticeably different results.[2]

Following up on this somewhat conflicting observation, some experts expressed desires for **more rigorous experimentation to validate their observations and understand model's consistency vs. inconsistency.** While our `quick-draw` encouraged exploratory *hypothesis testing* by giving each player two attempts to try alternative prompts, some experts pointed out the limits of drawing conclusions from such small sample sizes. For example, G10-P3 asked about the temperature parameter of the model and noted "I feel like if I were to reprompt this 10 times, I would expect more variants ... not everybody has a beard or a comb-over." Similarly, G7-P3 reflected, "Across six rounds, we generated 18 base images together, which is a very small sample size. I'm surprised by the consistencies of the short prompts, either when G7-P1 gambles and it pays off, or when I play it safe using Quick Draw. I know generated images are often very stereotypical, but I don't know if I'm just overfitting." Our playtest limited the number of exposures to keep the study time manageable, but we do agree that for GenAI reflection more demonstrative samples would help better reveal the probabilities nature of these models.

5.3.2 *ImaginAItion helps players develop prompting strategies to mitigate human–AI differences.*

Despite being surprised by GenAI (in)consistency, players developed prompting strategies for negotiating such non-determinism as well as the model defaults. As we asked about lessons learned from the game in the post-survey, 20 out of 30 participants talked about prompting, and four themes were most prominent:

- *Be specific* to override defaults and explicitly specify necessary details. E.g., "I'm justified in using a greater degree of specificity in my prompts. My mental model of which details to make explicit is also changing." (G2-P1)
- *Gamble* with concise prompts to exploit model defaults and biases. E.g., "There are moments when it's good to be more specific vs not – you have to play to AI's biases." (G2-P2)
- *Iterate* with multiple attempts to converge on the desired result. E.g., "You probably can't avoid iterative prompting to get what you want most times." (G1-P1)
- *Contrast* prompts to isolate discriminative features and test assumptions. E.g., "I understood how different parameters (amount of descriptions, details, model default behaviors) affected outputs." (G6-P2)

Players also self-rated to be more confident in their ability to craft useful prompts: their perceived ease of creating effective prompts improved by about one point from pre-survey (4.0 ± 1.5) to post-survey (4.9 ± 1.4) on a 7-point scale (1 = not at all easy, 7 = extremely easy). As G7-P1 remarked, "I have a better understanding now of how generative AI works! I think that if I were to use generative AI to create images after playing this game, I would be able to generate images that more closely match my intended goals." Follow-up studies can be done to track if players' prompting strategies actually changed in real life after exposure to gameplay (Section 6.2).

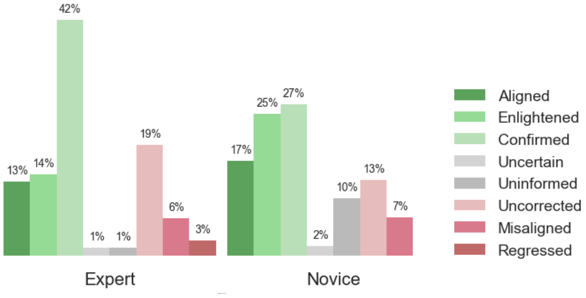


Figure 5: Reflection outcomes distribution by expertise.

5.4 Social Impact: the group reflection quality is bounded by the players

Previous sections all highlight the effectiveness of ImaginAItion's core mechanisms. But as hinted in Section 5.1.2, the player distribution impacts *how* effective the reflections in ImaginAItion can be. Here, we further zoom into the *social discourse*, highlighting the importance of player expertise and dynamics.

5.4.1 *Showing examples shifted understandings, but how much it shifted depends on the players' self-rated expertise.*

We see more nuances of the understanding shifts in the breakdowns by expertise. As shown in Figure 5, experts showed more `Confirmed` (42%) and `Uncorrected` (19%) outcomes, which may suggest that **self-rated experts have firmer beliefs than novices that are harder to challenge**. While experts do get `Aligned` from misunderstandings, that's typically because they **get to update their outdated understandings of model capacities and behaviors**. For example, G2-P1 reflected that "Models are even more biased than I already knew them to be." Similarly, G4-P2 reflected that "The game made me understand the advancing capabilities of AI, which I had not previously understood to be so advanced."

In comparison, novices showed more `Enlightened` (25% > 14%) and `Uninformed` (10% > 1%) reflection outcomes. They may gain new insights and ground their understandings in the examples we provided in game — pre- to post-game, novices included 12% more concrete examples in their survey responses. For example, before gameplay, G1-P3 had no understanding over GenAI's behavior in portraying demographic information, but in post-survey, she is able to draw from a concrete example in game: "in the Dr Li round, GenAI use asian people to reflect a people naming Li".

However, the game **may also fail to inform novices**, especially on challenging categories like co-occurrence and realism. In fact, both experts (6%) and novices (7%) risk becoming `Misaligned`, failing to distinguish single-example overgeneralization from pattern-based calibration when they begin with *no* understanding of GenAI behavior (elaborated in Section 5.1.2).

5.4.2 *Party game is good for engagement, but reflection thoroughness depends on the group composition.*

We adopted the `multi-player` party game mechanism to foster *social discourse*, as G9-P1 put it, "It's fun to play/learn together with friends (or literally anyone) to unlock blind spots." However, the depth of reflection our game enabled was shaped by group composition and its social dynamics. While no conclusions can be made

[2] Note that in generative models, certainty is typically regulated by the temperature or random seed variable. However, the image model we used during the game did not support temperature or random seed control (as described in Section 3.3).

due to small sample size, we found that diverse groups may get more enriched perspectives when playing such reflective game.

For example, we observed different cases where women in co-ed groups questioned gendered representations and tested on their hypotheses during `quick-draw`, such as when G10-P2 and G2-P2 asked how "a woman' would look after seeing the "`A man`" prompt, or when G7-P1 wondered whether prompting *father* would also generate a baby (Section 5.1.1; Figure 3). In her experiment, G2-P2 commented on her subjective perception that the woman seems more attractive than the man generated using very brief prompts, and she hypothesized about the role of group diversity in enabling deeper reflection: "If you had a group that was not co-ed or like a bunch of bros and their bros, versus girls and their girls — would they notice enough to point it out?"

We also observe cases where more homogeneous groups may miss opportunities to interrogate model bias. For instance, G8 (three South Asian men) did not remark on the White male default in "`A man`" during or after the game, and G8-P2 got `Misaligned` by saying "GenAI does well to do this" in the post-survey. Similarly, G9 (co-ed but three East Asians) also encountered the prompt "`A man`"; however, their discussions focused on how G9-P3's prompt "`Caucasian man`" was very accurate rather than on model bias, with post-survey responses such as "it performed not bad. I use Caucasian man to describe white man, and it shows well" (G9-P3) and "pretty accurate (e.g., white Caucasian man)" (G9-P1). However, it might be because there was not enough *dissonance* for participants to comment on model's Western-centric views. In another round where the original image for "`Startup founder giving ted talk`" portrays an Indian man, G8-P3 expressed surprises:

> G8-P3: I mean, to be honest, this is the first time I'm seeing a model that, by default, if you put in a startup founder, it looks like an Indian guy. Normally I'd expect a White man, because that's what most of the images are trained on. But I guess this model has been trained on different data.
> (Trying the original prompt in Quick Draw)
> G8-P3: I tried again, it gave me the image I originally imagined, like a White guy in front giving a TED talk. Yeah I was really surprised that the first one showed someone brown.

As G2-P1 put it, "To what extent each of us can sniff those things out, and then understanding the social dynamics around that is very interesting." Our findings suggest that reflection was influenced by the group composition and how willing participants were to voice and challenge assumptions. While our sample size is small, the contrast between groups points to a broader implication: **diversity amplifies the effectiveness of reflective play**. Future work (Section 6.2) can explore how to add support mechanisms to the game to ensure the discussions benefit from diverse perspectives.

6 DISCUSSION AND FUTURE WORK

6.1 Design Implications for GenAI Literacy Games

Design for the fast progressing AIs. A recurring challenge we faced in designing ImaginAItion is keeping them relevant amid the rapid advancement of GenAI models, as noted in Section 3.4. In fact, even during our own playtest study, we observed noticeable model improvements within days: on Saturday, the model struggled to generate four or five cubes accurately (Table 2), yet by Monday it could consistently count accurately for numbers under ten. This highlights a crucial design implication: GenAI literacy games – and GenAI tools more broadly – should not rely on specific model limitations, as these gaps are actively being closed by developers. Instead, we should focus on more generalizable insights aligned with the model's growing ability to follow instructions. To that end, we chose to emphasize *constrained prompting* that reveals model failures with *under-* or *ambiguous* instructions – issues unlikely to be resolved through incremental updates, since **underspecification is a fundamentally human trait** [9, 49].

Looking ahead, we speculate that games that draw players' attention to the potential negative impact of their increasingly more extensive (and likely not more careful) AI usage [14] would offer more lasting value, e.g., how user-granted access affects a model's invasiveness, or how degrees of personalization (through memory modules of AI agents) trade off output helpfulness against risks of information leakage. Insights from learning sciences and the future of work can help pinpoint such observation venues most relevant to real-world use [41, 70].

Accommodate related and tensioned concepts. Our study surfaced GenAI concepts that feel paradoxical in practice – most notably the pull between consistency and inconsistency in model behavior (Section 5.3). Such apparent contradictions often arise because different training signals shape different capabilities, making it hard for players to form a single, coherent mental model. Yet encountering these contrasts during gameplay is productive: it calibrates expectations and highlights the probabilistic, composite nature of GenAI systems. There are also many other tensions, for example, models can be acutely sensitive to some details but not others [65], and they may produce persuasive Chain-of-Thought explanations while overlooking changes to critical reasoning steps [40]. Future games could deliberately curate such cases across rounds to scaffold comparison and hypothesis-making.

Moreover, we found that reflections often extended beyond model behavior into human dimensions, e.g., human bias, people's perceptions of model bias, and even meta-biases about human assumptions (Section 5.2.2). We observed players spontaneously interrogating these connections, which fostered rich insights about the social shaping of GenAI, such as how human biases influence model development and how increased awareness can support ethical prompting practices. Designing for such entangled reflections — between human and AI, between normative and descriptive behaviors — can lead to more nuanced GenAI literacy outcomes.

Towards more effective coverage in the unlimited AI hypothesis space. Multi-player review is a core mechanism we use to address the vast input-output space of GenAI. By comparing peer inputs, we aim to uncover blind spots and foster more comprehensive hypothesis formation and testing (Section 3.1). While effective in our playtests (Section 5.2), we found that exploration of the hypothesis space remained partial and, more importantly, the depth, accuracy, and diversity of reflections were often bounded by individual differences and group dynamics (Section 5.4). Some groups failed to identify specific behavioral patterns we intentionally seeded (Section 5.1.2), while others diverged in interpretation despite encountering the same examples. For example, G3 derived split views on model's number and spatial reasoning ability, and G6 participants completely overlooked counting-related patterns.

To support more consistent reflective outcomes, we may develop more deliberate exploration mechanisms of the hypothesis space. One potential direction is dynamic in-situ monitoring of player reflection, enabling the system to select subsequent examples that challenge or contrast with underdeveloped assumptions. Additionally, deliberate pairing or grouping strategies may improve hypothesis diversity. For instance, matching players based on diversifying demographic backgrounds (Section 5.4.2) might broaden exploration on model bias, or differing tolerances for AI uncertainty (Section 5.2.2) might facilitate richer discussion on succinct versus overly detailed prompts tradeoffs. We also observed complementary hypothesis testing styles: some players explored contrastive inputs (e.g., "`a man`" → "`a woman`"), others reproduced the original prompt, and some were less inclined to test hypotheses at all. Leveraging such diversity may help deepen collective understanding.

6.2 Limitations and Future Work

Design directions. ImaginAItion fostered valuable reflections on GenAI bias, defaults, and prompting strategies. To deepen this, future designs can vary both the game mechanism and the powering GenAI models. For example, we may adopt more deliberate game mechanisms like AI voting, human drawing, prompt iteration, adversarial mechanics (e.g., creating challenging prompts/images for others, riffing on others' prompts), or team-based play. In our game, we instantiate GenAI with the specific `gpt-image-1`, but our design lessons and mechanism can be generalized to other GenAIs. Enabling players to use and choose models with different capacities such as DALL-E may further open doors for in-depth understandings over different GenAI's limitations and strengths, as also suggested by G6-P1 and G6-P3.

Scaling up: Long-term and diverse play context. In our playtest, we have seen promising patterns reflecting shifts toward a calibrated understanding from pre- to post-game, as well as outcome differences among various group compositions. However, constrained by the sample size and qualitative methods, we did not attempt to use pre–post survey design to measure learning gains in a traditional sense, or detect significant effects for these patterns. It would be interesting to scale up the study in various ways:

(1) To detect sustained changes in prompting behavior or GenAI beliefs, longitudinal studies would be required, as multiple practice opportunities are often necessary to consolidate knowledge [37]. Here, allowing systematic experimentation of prompt failure modes [46] or embedding explicit instructions for prompt engineering [49] and hypothesis construction [50] could enrich extended play sessions. Such designs would allow us to observe how players refine hypothesis testing (Section 5.3.1) and prompting strategies (Section 5.3.2) if given more opportunities. (2) To capture whether composition of players exhibit different patterns, groups could be systematically composed as a controlled variable. For example, by intentionally varying the number of experts vs. novices, demographic diversity and intersectionality, incentivizing different play strategies. (3) Aside from temporal and social group constraints, our study was also limited to fixed trios of adults playing over Zoom. Exploring in-person or larger group play could surface new dynamics, while multilingual gameplay may improve accessibility – some participants already used native languages to better express themselves during our study.

Broader applications: Data generation for ML research. Beyond education, scaled-up versions of ImaginAItion could support ML research by generating rich datasets of natural language prompts, corresponding GenAI outputs, and human judgments of similarity or quality. These datasets could complement existing corpora like Google's Quick Draw [26], supporting tasks such as image regeneration [75], prompt-based image captioning, or perceptual similarity benchmarking. For example, our gameplay revealed dimensions humans find salient — like style and background (Section 5.2.1) — suggesting the potential to surface emergent human priors around visual similarity. With more structured data collection, such as controlled comparisons and attribute annotation (e.g., gender, occupation), the game could help isolate key perceptual discriminators. Moreover, integrating automated metrics like CLIP with human judgments could enable hybrid evaluation pipelines, aiding perceptual similarity benchmark constructions [75]. Repeated gameplay before and after major model updates may also serve as a form of *community auditing* [15], surfacing whether interventions to reduce bias or improve alignment actually change how models behave in practice.

7 CONCLUSION

ImaginAItion is a party game designed to spark transformative reflection on GenAI model behaviors. It blends playful competition with critical engagement, using game mechanics to surface GenAI's biases and inconsistencies, encourage social comparison of perceptions, and support iterative prompting strategies — all while remaining accessible and fun. In a game study with 30 participants, we showed that structured play can effectively engage both experts and novices in exploring GenAI's defaults and limitations. Our analysis offers insights of model and human behaviors, and proposes design principles for sustaining AI literacy games amid rapidly evolving technologies. By leveraging underspecification, peer discourse, and iterative experimentation, ImaginAItion fosters enduring reflection on the challenges of human-AI communication, beyond specific system failures. Future directions include expanding to broader demographics and refining gameplay. More broadly, we envision ImaginAItion as a reflective, accessible party game that promotes lifelong AI literacy — engaging users and sparking critical conversations well beyond the game itself.

ACKNOWLEDGMENTS

Thanks to Anika Jain for engagement in early versions of this work, and thanks to all participants. Thanks to the National Science Foundation (award CNS-2213791, 2414915) and Google Academic Research Award for support of this work.

Acknowledgment of LLM Use. We acknowledge the use of OpenAI ChatGPT to assist with language editing, summarization, and restructuring of text during manuscript preparation. All research design, data collection, analysis, and interpretation were conducted solely by the authors. We explored different text-to-image GenAI as part of our game mechanism, including OpenAI `gpt-image-1` and Meta AI, and the implementation details are included in Section 3.3.

REFERENCES

[1] Jihyun Janice Ahn and Wenpeng Yin. 2025. Prompt-reverse inconsistency: Llm self-inconsistency beyond generative randomness and prompt paraphrasing. *arXiv preprint arXiv:2504.01282* (2025).

[2] Safinah Ali, Vishesh Kumar, and Cynthia Breazeal. 2023. AI audit: a card game to reflect on everyday AI systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15981–15989.

[3] Haesol Bae and Aras Bozkurt. 2024. The Untold Story of Training Students with Generative AI: Are We Preparing Students for True Learning or Just Personalization? *Online Learning* 28, 3 (2024).

[4] Maalvika Bhat and Duri Long. 2024. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 939–957. https://doi.org/10.1145/3643834.3660722

[5] Ali Borji. 2023. Qualitative failures of image generation models and their application in detecting deepfakes. *arXiv [cs.CV]* (March 2023). arXiv:2304.06470 [cs.CV] http://arxiv.org/abs/2304.06470

[6] Aras Bozkurt. 2024. Why generative AI literacy, why now and why it matters in the educational landscape?: Kings, queens and GenAI dragons. , 283–290 pages.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Lorena Casal-Otero, Alejandro Catala, Carmen Fernández-Morante, Maria Taboada, Beatriz Cebreiro, and Senén Barro. 2023. AI literacy in K-12: a systematic literature review. *International Journal of STEM Education* 10, 1 (April 2023), 1–17. https://doi.org/10.1186/s40594-023-00418-7

[9] Yang Chenyang, Shi Yike, Ma Qianou, Michael Xieyang Liu, Kästner Christian, and Wu Tongshuang. 2025. What prompts don't say: Understanding and managing underspecification in LLM prompts. *arXiv [cs.CL]* (May 2025). arXiv:2505.13360 [cs.CL] http://arxiv.org/abs/2505.13360

[10] Judeth Oden Choi, Jodi Forlizzi, Michael Christel, Rachel Moeller, Mackenzie Bates, and Jessica Hammer. 2016. Playtesting with a Purpose. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, USA. https://doi.org/10.1145/2967934.2968103

[11] Wikipedia contributors. [n. d.]. Pictionary. https://en.wikipedia.org/wiki/Pictionary. Accessed: December 8, 2024.

[12] Andrew Cox. 2024. Algorithmic literacy, AI literacy and responsible generative AI literacy. *Journal of Web Librarianship* (2024), 1–18.

[13] Sabrina Culyba. 2018. *A process tool for the development of Transformational games*. CMU ETS Press. http://pstorage-cmu-348901238291901.s3.amazonaws.com/13117568/TransformationalFramework_2018.pdf

[14] Simone Daniotti, Johannes Wachs, Xiangnan Feng, and Frank Neffke. 2025. Who is using AI to code? Global diffusion and impact of generative AI. *arXiv preprint arXiv:2506.08945* (2025).

[15] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding user auditors and AI practitioners in auditing Generative AI. *arXiv [cs.HC]* (Jan. 2025). arXiv:2501.01397 [cs.HC] http://arxiv.org/abs/2501.01397

[16] Chiara Di Lodovico, Federico Torrielli, Luigi Di Caro, and Amon Rapp. 2025. How do people develop folk theories of generative AI text-to-image models? A qualitative study on how people strive to explain and make sense of GenAI. *Int. J. Hum. Comput. Interact.* (April 2025), 1–25. https://doi.org/10.1080/10447318.2025.2491009

[17] Leyla Dogruel, Philipp Masur, and Sven Joeckel. 2022. Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures* 16, 2 (2022), 115–133.

[18] Stefania Druga, Nancy Otero, and Amy J. Ko. 2022. The Landscape of Teaching Resources for AI Education. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1* (Dublin, Ireland) *(ITiCSE '22)*. Association for Computing Machinery, New York, NY, USA, 96–102. https://doi.org/10.1145/3502718.3524782

[19] Xiaoxue Du, Zhichun Liu, Xi Wang, and Qiping Tang. 2024. Fostering AI Literacy through Interactive Game-Based Learning: A Case Study on Enhancing Algorithmic Thinking Skills. In *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 333–338.

[20] Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science* 8, 3 (2013), 223–241.

[21] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[22] Jackbox Games. [n. d.]. Drawful. https://steamdb.info/app/442070/charts/. https://www.jackboxgames.com/games/drawful-2 Accessed: December 8, 2024.

[23] Raghu Garud. 1997. Know-how, know-why, and know-what. *Advances in strategic management* 14 (1997), 81–101.

[24] Mary Flanagan Geoff Kaufman. 2015. A psychologically "embedded" approach to designing games for prosocial causes. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* (2015).

[25] Ronald N Giere and Barton Moffatt. 2003. Distributed Cognition:: Where the Cognitive and the Social Merge. *Soc. Stud. Sci.* 33, 2 (April 2003), 301–310. https://doi.org/10.1177/03063127030332017

[26] Google. [n. d.]. Quick, Draw! https://quickdraw.withgoogle.com/. Accessed: December 8, 2024.

[27] Anuj Gupta, Yasser Atef, Anna Mills, and Maha Bali. 2024. Assistant, parrot, or colonizing loudspeaker? ChatGPT metaphors for developing critical AI Literacies. *Open Praxis* 16, 1 (2024), 37–53.

[28] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural NLP. *arXiv preprint arXiv:2203.10020* (2022).

[29] Olli Hilke, Nicolas Pope, Juho Kahila, Henriikka Vartiainen, Teemu Roos, Tuomo Parkki, and Matti Tedre. 2025. Breakable Machine: A K-12 Classroom Game for Transformative AI Literacy Through Spoofing and eXplainable AI (XAI). *arXiv preprint arXiv:2508.14201* (2025).

[30] Ting-Chia Hsu and Tai-Ping Hsu. 2025. Teaching AI with games: the impact of generative AI drawing on computational thinking skills. *Education and Information Technologies* (2025), 1–20.

[31] Minsuk Kahng, Nikhil Thorat, Duen Horng Polo Chau, Fernanda B. Viegas, and Martin Wattenberg. 2019. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 310–320. https://doi.org/10.1109/tvcg.2018.2864500

[32] Dongyeop Kang and Eduard Hovy. 2021. Style is NOT a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*. 2376–2387.

[33] Maria Kasinidou. 2024. Development of Personalised Educational Tools for AI Literacy Using Participatory Design. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 30–34. https://doi.org/10.1145/3631700.3664917

[34] Geoffrey Kaufman and Mary Flanagan. 2015. A Psychologically "Embedded" Approach to Designing Games for Prosocial Causes. *Cyberpsychology, Behavior, and Social Networking* 18, 4 (2015), 241–248. https://doi.org/10.1089/cyber.2014.0513

[35] Geoff Kaufman, Mary Flanagan, and Gili Freedman. 2019. Not just for girls: Encouraging cross-gender role play and reducing gender stereotypes with a strategy game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, USA. https://doi.org/10.1145/3311350.3347177

[36] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff. 2021. Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. ACM. https://doi.org/10.1145/3461702.3462605

[37] Kenneth R Koedinger, Paulo F Carvalho, Ran Liu, and Elizabeth A McLaughlin. 2023. An astonishing regularity in student learning rate. *Proc. Natl. Acad. Sci. U. S. A.* 120, 13 (March 2023), e2221311120. https://doi.org/10.1073/pnas.2221311120

[38] Valentin Kuleto, Milena Ilić, Mihail Dumangiu, Marko Ranković, Oliva MD Martins, Dan Păun, and Larisa Mihoreanu. 2021. Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions. *Sustainability* 13, 18 (2021), 10424.

[39] Matthias Carl Laupichler, Alexandra Aster, Jana Schirch, and Tobias Raupach. 2022. Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence* 3, 100101 (Jan. 2022), 100101. https://doi.org/10.1016/j.caeai.2022.100101

[40] Mosh Levy, Zohar Elyoseph, and Yoav Goldberg. 2025. Humans Perceive Wrong Narratives from AI Reasoning Texts. *arXiv preprint arXiv:2508.16599* (2025).

[41] Hanqi Li, Ruiwei Xiao, Hsuan Nieu, Ying-Jui Tseng, and Guanze Liao. 2025. "From Unseen Needs to Classroom Solutions": Exploring AI Literacy Challenges & Opportunities with Project-Based Learning Toolkit in K-12 Education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 29145–29152.

[42] Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *arXiv preprint arXiv:2505.11423*

(2025).

[43] Andreas Lieberoth. 2015. Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture* 10, 3 (2015), 229–248.

[44] Bill Yuchen Lin, Frank F Xu, Kenny Zhu, and Seung-won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 709–719.

[45] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[46] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, 1–23. https://doi.org/10.1145/3491102.3501825

[47] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[48] Qianou Ma, Anika Jain, Jini Kim, Megan Chai, and Geoff Kaufman. 2025. ImaginAItion: Promoting generative AI literacy through game-based learning. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/3706599.3719844

[49] Qianou Ma, Weirui Peng, Chenyang Yang, Hua Shen, Ken Koedinger, and Tongshuang Wu. 2025. What should we engineer in prompts? Training humans in requirement-driven LLM use. *ACM Trans. Comput. Hum. Interact.* 32, 4 (Aug. 2025), 1–27. https://doi.org/10.1145/3731756

[50] Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the AI era? Using LLMs as a teachable agent for debugging. In *Proceedings of International Conference on Artificial Intelligence in Education*. Springer Nature Switzerland, Cham, 265–279. https://doi.org/10.1007/978-3-031-64302-6_19

[51] Pragati Maheshwary, Aditi Haiman, Emma Brown, Aarnav Sangekar, and Canwen Wang. 2025. Case 429: A Murder Mystery Game for Teaching Bias in Generative AI. (2025).

[52] Sai Siddartha Maram, Erica Kleinman, Jennifer Villareale, Jichen Zhu, and Magy Seif El-Nasr. 2024. "ah! I see" - facilitating process reflection in gameplay through a novel spatio-temporal visualization system. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 3. ACM, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642484

[53] A I Meta. 2024. Create, edit, and animate images with Meta AI. https://www.meta.ai/ai-image-generator. Accessed: 2024-11-8.

[54] Josh Aaron Miller, Kutub Gandhi, Matthew Alexander Whitby, Mehmet Kosa, Seth Cooper, Elisa D Mekler, and Ioanna Iacovides. 2024. A design framework for reflective play. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 91. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3613904.3642455

[55] Katelyn Morrison, Mayank Jain, Jessica Hammer, and Adam Perer. 2023. Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 273 (Oct. 2023), 22 pages. https://doi.org/10.1145/3610064

[56] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2, 100041 (Jan. 2021), 100041. https://doi.org/10.1016/j.caeai.2021.100041

[57] Davy Tsz Kit Ng, Chen Xinyu, Jac Ka Lok Leung, and Samuel Kai Wah Chu. 2024. Fostering students' AI literacy development through educational games: AI knowledge, affective and cognitive engagement. *Journal of Computer Assisted Learning* (2024).

[58] The Op. [n. d.]. Telestrations. https://theop.games/collections/telestrations. Accessed: December 8, 2024.

[59] OpenAI. 2024. DALL-E 3 Image generation. https://platform.openai.com/docs/guides/image-generation?image-generation-model=dall-e-3. Accessed: 2024-12-8.

[60] OpenAI. 2025. Introducing 4o Image Generation. https://openai.com/index/introducing-4o-image-generation/. Accessed: 2025-9-8.

[61] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[62] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).

[63] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating Instruction Following Ability in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13025–13048. https://doi.org/10.18653/v1/2024.findings-acl.772

[64] Amon Rapp, Chiara Di Lodovico, and Luigi Di Caro. 2025. How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations. *Int. J. Hum. Comput. Stud.* 198, 103471 (April 2025), 103471. https://doi.org/10.1016/j.ijhcs.2025.103471

[65] Paulius Rauba, Qiyao Wei, and Mihaela van der Schaar. 2024. Quantifying perturbation impacts for large language models. *arXiv preprint arXiv:2412.00868* (2024).

[66] Danielle Reynolds and Scott Brady. [n. d.]. CautionSigns. https://cautionsigns.app/. Accessed: December 8, 2024.

[67] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B Wiltschko. 2021. A gentle introduction to graph neural networks. *Distill* 6, 9 (2021), e33.

[68] Camilo Chacón Sartori. 2025. Architectures of Error: A Philosophical Inquiry into AI and Human Code Generation. *arXiv preprint arXiv:2505.19353* (2025).

[69] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[70] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. 2025. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce. *arXiv preprint arXiv:2506.06576* (2025).

[71] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. 2025. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951* (2025).

[72] Hannu Simonen, Atte Kiviniemi, and Jonas Oppenlaender. 2025. An exploration of default images in text-to-image generation. *arXiv [cs.HC]* (July 2025). arXiv:2505.09166 [cs.HC] http://arxiv.org/abs/2505.09166

[73] Anselm Strauss and Juliet Corbin. 1998. Basics of qualitative research techniques. (1998).

[74] Piiastiina Tikka, Miia Laitinen, Iikka Manninen, and Harri Oinas-Kukkonen. 2018. Reflection through gaming: Reinforcing health message response through gamified rehearsal. In *Persuasive Technology*. Springer International Publishing, Cham, 200–212. https://doi.org/10.1007/978-3-319-78978-1_17

[75] Khoi Trinh, Scott Seidenberger, Raveen Wijewickrama, Murtuza Jadliwala, and Anindya Maiti. 2025. A picture is worth a thousand prompts? Efficacy of iterative human-driven prompt refinement in image regeneration tasks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

[76] Jessica Vandenberg, Wookhee Min, Veronica Cateté, Danielle Boulden, and Bradford Mott. 2022. Promoting AI education for rural middle grades students with digital game design. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*. 1388–1388.

[77] Jennifer Villareale, Gabriele Cimolino, and Daniel Gomme. 2023. Playing with dezgo: Adapting human-AI interaction to the context of play. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*. ACM, New York, NY, USA. https://doi.org/10.1145/3582437.3587198

[78] Jennifer Villareale, Casper Harteveld, and Jichen Zhu. 2022. "I want to see how smart this AI really is": Player mental model development of an adversarial AI player. *Proc. ACM Hum. Comput. Interact.* 6, CHI PLAY (Oct. 2022), 1–26. https://doi.org/10.1145/3549482

[79] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

[80] Junling Wang, Anna Rutkiewicz, April Yi Wang, and Mrinmaya Sachan. 2025. Generating pedagogically meaningful visuals for math word problems: A new benchmark and analysis of text-to-Image models. *arXiv [cs.CL]* (June 2025). arXiv:2506.03735 [cs.CL] http://arxiv.org/abs/2506.03735

[81] Ning Wang, Eric Greenwald, Ryan Montgomery, and Maxyn Leitner. 2022. ARIN-561: An educational game for learning artificial intelligence for high-school students. In *International Conference on Artificial Intelligence in Education*. Springer, 528–531.

[82] Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. 2020. CNN explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1396–1406.

[83] Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2024. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455* (2024).

[84] Michelle Yavorskiy and Lydia Kim. 2024. MeadowMinds: An AI Literacy Game. (2024).

[85] Marvin Zammit, Iro Voulgari, Antonios Liapis, and Georgios N Yannakakis. 2021. The road to AI literacy education: from pedagogical needs to tangible game design. Academic Conferences International.

A GAME STUDY MATERIALS

A.1 Pre-Survey Only (Before Game)

GenAI Use and Familiarity.

(1) On a scale of 1–7 (Likert), how familiar are you with text-to-image generative AI tools? (1 = Not at all familiar, 7 = Extremely familiar)
(2) Have you ever learned prompting methods to enhance the performance of text-to-image generative AI? If so, briefly describe what you have learned.

Demographics.

(1) What is your age?
(2) What is your gender identity?
(3) What is your race or ethnicity? (Select all that apply.)
(4) What is your sexual orientation? (Select all that apply.)
(5) What language(s) do you speak at home? (Select all that apply.)

A.2 Pre-Post Survey Shared Questions

Model Behaviors. Answer the following questions with your best understanding of text-to-image GenAI behaviors (we are not asking about the underlying mechanisms, just what you think GenAI would do). If you have no idea at all, you can answer with "I don't know". Provide specific prompt examples if you can.

(1) How does GenAI depict people with different identities, demographics, and backgrounds?
(2) How does GenAI represent social and cultural elements in images?
(3) How does GenAI depict the number and position of objects in images?
(4) How well can GenAI display text within generated images?
(5) How well can GenAI show body parts like hands or faces?
(6) How does GenAI visualize abstract concepts like *friendship* in pictures?
(7) How does GenAI handle prompts that use uncommon word combinations or tricky syntax (e.g., negations, or understanding whether an adjective applies to one object or another)?

Self-Assessment.

(1) On a scale of 1–7 (Likert), how confident do you feel in understanding default behaviors in text-to-image generative AI? (1 = Not at all confident, 7 = Extremely confident)
(2) On a scale of 1–7 (Likert), how easy do you find it to create prompts that guide text-to-image generative AI to generate your desired images? (1 = Not at all easy, 7 = Extremely easy)

A.3 Post-Survey Only (After Game)

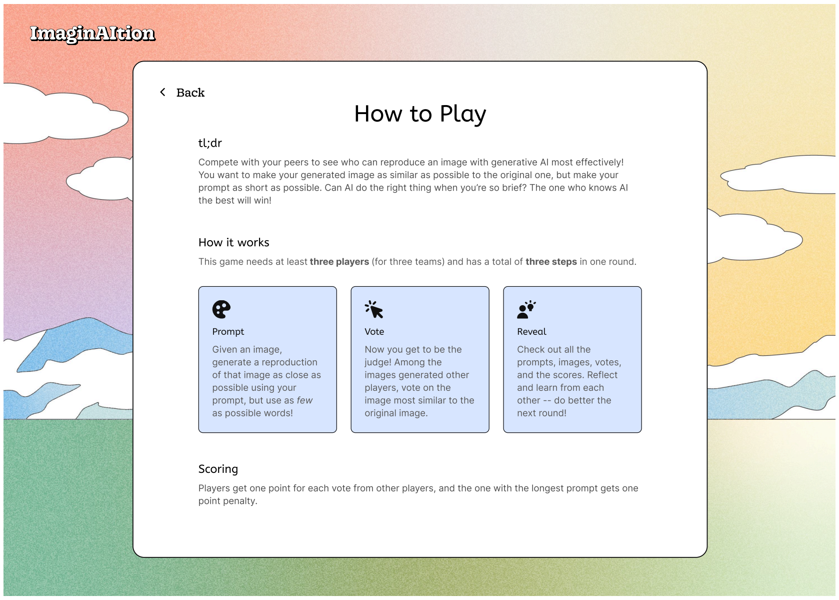
Immediate Reflection and Game Feedback.

(1) What was your nickname during the game?
(2) On a scale of 1–7 (Likert), has this game helped improve your understanding of image-generative AI's behaviors? (1 = Not at all helpful, 7 = Extremely helpful)
(3) Briefly explain why this game helped or did not help.
(4) Any moment you found particularly enjoyable, interesting, or insightful during the game?
(5) What have you learned from this game, if any?
(6) Any time during gameplay that you felt uncomfortable or challenged?
(7) If you could change one thing about this game, what would it be and why?

A.4 Semi-structured Group Interview Questions

(1) How was your gameplay experience playing against other players? Did you learn anything from each other? How did interacting with other players and seeing their prompts influence your own approach?
(2) Were you surprised by any generated images during gameplay? Did the AI-generated images match your expectations? Why or why not?
(3) What strategies did you develop for prompting? Did your strategy change during the game, and if so, how?

B INTERFACE ARTIFACTS



How to Play

tl;dr

Compete with your peers to see who can reproduce an image with generative AI most effectively! You want to make your generated image as similar as possible to the original one, but make your prompt as short as possible. Can AI do the right thing when you're so brief? The one who knows AI the best will win!

How it works

This game needs at least **three players** (for three teams) and has a total of **three steps** in one round.

Prompt

Given an image, generate a reproduction of that image as close as possible using your prompt, but use as *few* as possible words!

Vote

Now you get to be the judge! Among the images generated other players, vote on the image most similar to the original image.

Reveal

Check out all the prompts, images, votes, and the scores. Reflect and learn from each other -- do better the next round!

Scoring

Players get one point for each vote from other players, and the one with the longest prompt gets one point penalty.

Figure 6: How to Play Description. Players can view the instructions on how to play the game with an overview of the prompting, voting, and revealing turns along with the scoring mechanisms.