

Can They Dixit? Yes they Can!

Dixit as a Playground for Multimodal Language Model Capabilities

Nishant Balepur
University of Maryland
nbalepur@umd.edu

Dang Nguyen
University of Maryland
dangmn@umd.edu

Dayeon Ki
University of Maryland
dayeonki@umd.edu

Abstract

Multi-modal large language models (MLMs) are often assessed on static, individual benchmarks—which cannot jointly assess MLM capabilities in a single task—or rely on human or model pairwise comparisons—which is highly subjective, expensive, and allows models to exploit superficial shortcuts (e.g., verbosity) to inflate their win-rates. To overcome these issues, we propose game-based evaluations to holistically assess MLM capabilities. Games require multiple abilities for players to win, are inherently competitive, and are governed by fix, objective rules, and makes evaluation more engaging, providing a robust framework to address the aforementioned challenges. We manifest this evaluation specifically through Dixit, a fantasy card game where players must generate captions for a card that trick some, but not all players, into selecting the played card. Our quantitative experiments with five MLMs show Dixit win-rate rankings are perfectly correlated with those on popular MLM benchmarks, while games between human and MLM players in Dixit reveal several differences between agent strategies and areas of improvement for MLM reasoning.¹

1 Introduction

Multi-modal large language models (MLMs) have made significant progress, showing impressive performance in tasks requiring both image and text inputs, such as image captioning, classification, and visual understanding (Zheng et al., 2023b). As a testament to this, several MLMs with diverse architectures have been introduced in the past few years, each aiming to excel in such tasks (Chen et al., 2024b; Bai et al., 2023; Agrawal et al., 2024). As the number of MLMs continues to grow, the need for robust and comprehensive evaluation frameworks becomes increasingly important (White et al., 2024; Saxon et al., 2024). These

frameworks are key for reliably comparing models, enabling users to choose those that best align with their needs (Li et al., 2024).

Such MLM benchmarks fall into two main categories. The first approach evaluates models on individual tasks—such as image captioning, classification, or understanding—pinpointing abilities in isolation (Duan et al., 2024; Lee et al., 2024b). While useful, this precludes jointly testing model capabilities allowing for direct comparisons between models. The second involves sampling outputs from two models and using pairwise comparison judgments from humans or other models to identify which model is better (i.e. win-rate) (Wu et al., 2023; Chiang et al., 2024). While this better compares models, such evaluations are inherently subjective (Lopez et al., 2021), as they rely on external human or model judgments without objective rules (Zheng et al., 2023a), and such judges are prone to biases like verbosity and position (Wang et al., 2023; Shi et al., 2024; Ye et al., 2024). They are also expensive, requiring large-scale human annotation or computational resources for model-based judging (Lee et al., 2024a; Chen et al., 2024a).

To address these issues, we draw from Natural Language Processing (NLP) research showing multi-agent LLM evaluations can be conducted through games (Xu et al., 2024; Ye et al., 2024; Hu et al., 2024), and propose game-based evaluations as a framework for evaluating MLMs. Games inherently involve competition (Lebed, 2006), enabling direct comparison of model capabilities, with rankings serving as a proxy for model performance. Further, games are governed by fixed rules (DeVries, 2015), eliminating the need for external judges and ensuring objective, efficient evaluations. The rules of games are also designed in a manner to prevent players, in our case models, from exploiting superficial shortcuts to inflate their win-rates (Webb and Soh, 2007). As an added benefit, games are designed to be fun (Prensky,

¹<https://github.com/nbalepur/dixit>

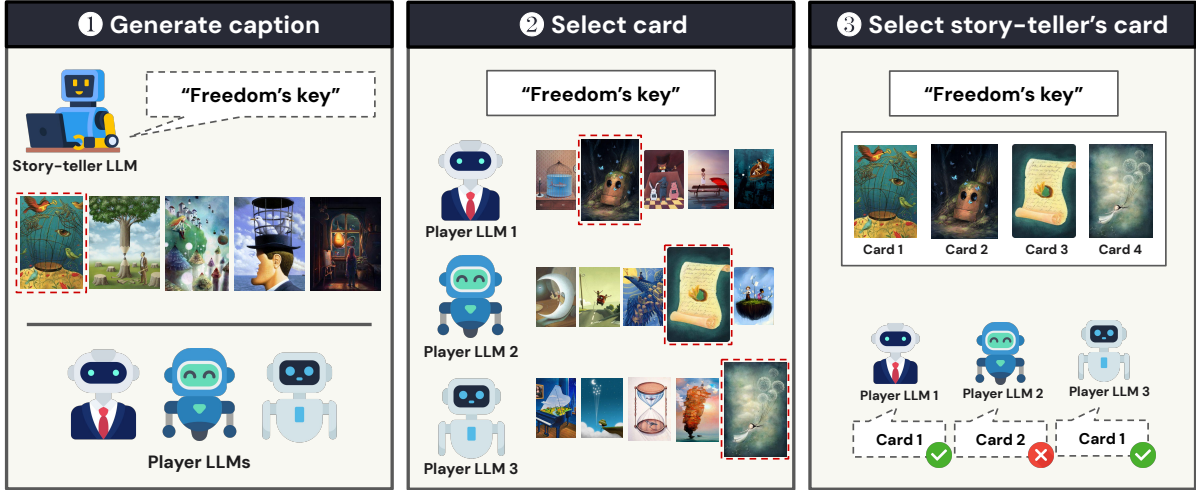


Figure 1: Overview of playing Dixit with MLMs. **Step 1:** The story-teller model generates a caption “*Freedom’s key*” for one of its cards, which is shown to all players. **Step 2:** Each player model select a card from their hand that they think corresponds to the caption. **Step 3:** From the pool of selected cards, all player models select the card they believe was the story-teller’s.

2001), so practitioners can directly intervene in game-based evaluations to assess their models in a more engaging manner.

To advocate for this approach, we adopt Dixit (Kunda and Rabkina, 2020a; Hsu et al., 2019)—a fantasy card game—as our game-based evaluation suite for MLMs, as it evaluates several diverse MLM capabilities (Figure 1). In Dixit, one player called the *storyteller* selects a card from their hand and generates a caption for it. The other players then choose cards from their hands that they believe best match the caption. All of these cards are pooled, and players vote on which card (not their own) they think is the storyteller’s. Non-story teller players score points if they select the correct card, but the storyteller faces a unique challenge to score points: their caption must be sufficiently relevant to guide other players to the correct card but not so obvious such that all players guess correctly; the storyteller does not score points if either no one or everyone guesses their card.

Dixit provides a unified and nuanced evaluation suite for multiple MLM capabilities. A strong player must be adept at: **1)** generating creative captions that allow only some users to pick the correct card (testing creativity, image comprehension, and calibration in image captioning); and **2)** identifying cards that align with what players believe the storyteller to have played and what players believe other players may select (testing accuracy, image comprehension, and theory-of-mind in image classification). Dixit captures diverse model abilities

in a single, cohesive task, making it an appealing game for our analysis.

We build our own evaluation interface for Dixit and have five MLMs compete as players: GPT-4o (Achiam et al., 2023), Claude-3.5 Sonnet (Anthropic, 2023), Intern-VL2 (Chen et al., 2024b), Qwen2VL (Bai et al., 2023), and Molmo (Deitke et al., 2024). We find that all players significantly outperform a player who randomly generates captions and selects cards, showing that MLMs that are not directly trained on Dixit can still demonstrate above-random Dixit capabilities. Further, ranking players by Dixit performance perfectly matches the rankings on popular leaderboards like ChatbotArena (Chiang et al., 2024) and the Open VLM Leaderboard (Duan et al., 2024), suggesting that Dixit comprehensively tests MLM capabilities within a single task. We then have humans play against one MLM to study if MLMs can surpass humans in Dixit. GPT-4o Mini, a relatively weak MLM, was able to surpass one human player, showing that MLMs possess strong Dixit capabilities that rival and may even surpass human abilities. Finally, we assess the game logs between MLM players in Dixit to understand differences between model and human Dixit strategies; we find when humans are story-tellers, they tend to generate more abstract captions that reference external knowledge, while MLM captions are easier as they literally describe the input image, suggesting future models can improve in Dixit by enhancing their common-sense reasoning capabilities.

We advocate for game-based frameworks, like Dixit, to evaluate MLM capabilities. Such games allow us to jointly compare several model abilities in a single unified task without relying on expensive and biased external judges. It also allows us to make evaluations more engaging, as humans can directly intervene in games as players to evaluate their models.

2 Related Work

2.1 Playing Games with MLMS

Multimodal LLMs (MLMs) have demonstrated their potential in tasks involving text and image input, such as image captioning and classification (Zheng et al., 2023b). Beyond these conventional tasks, MLMs have been increasingly applied as agents to perform complex tasks, where they have exhibited impressive generalization capabilities in dynamic environments (Tan et al., 2024; Wang et al., 2024b). This progress has sparked significant interest in the use of MLMs in game playing, a domain that naturally integrates reasoning, strategy, and multimodal understanding.

Previously, MLMs have been applied to popular games like Minecraft (Lifshitz et al., 2024) to navigate and interact within intricate environments and as low-level controllers in Atari video games to assess their ability to interpret complex visual scenes and formulate strategies (Waytowich et al., 2024). Similarly, game-based benchmarks like MatchIt (Chalamalasetti et al., 2023; Hakimov et al., 2024) require MLMs to communicate and compare visual input, which assess MLMs’s abilities in effective communication and image reasoning.

Similarly, we adopt MLMs in playing games, but uniquely focuses on the game Dixit, which has yet to be explored in the context of MLMs. Although Dixit has been studied in psychology, education, and cognitive science, its use as a benchmark for multimodal reasoning remains underexplored (Kunda and Rabkina, 2020b). Finally, a few works have explored MLM captioning capabilities in Dixit (Hsu et al., 2019), but we are the first to have MLM agents compete in Dixit end-to-end in a full game to holistically evaluate their capabilities.

2.2 Evaluation of MLMs

Evaluation of vision-language tasks typically employs two main approaches: 1) pointwise (Liu et al., 2024; Yu et al., 2023; Sun et al., 2023; Zhang et al., 2025) and 2) pairwise settings. Pointwise evalua-

tions focus on isolated tasks such as image captioning (Nguyen et al., 2023), classification (Abdelhamed et al., 2024), or visual understanding (Wang et al., 2024a). While effective for assessing specific skills, this approach is limited in its ability to jointly test and compare models’ broader multimodal capabilities and often introduces subjectivity in interpreting results. On the other hand, pairwise evaluations, where outputs from two models are sampled and compared directly (Lu et al., 2024; Yu et al., 2024), address some of these limitations but are resource-intensive, requiring significant time, computational power, and human effort. In response to these, we propose a game-based evaluation framework that simultaneously tests multiple capabilities within a structured, rule-governed environment.

3 Method: The Rules of Dixit

Each game of Dixit uses a deck of m cards $\mathcal{C} = \{c_1, c_2, \dots, c_m\}$ and consists of $3 \leq n \leq 6$ players $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$. Each card $c \in \mathcal{C}$ contains a fantasy picture (Figure 1) to allow players to exercise creativity in caption generation. Each player $p_i \in \mathcal{P}$ draws n cards to form their hand $\mathcal{C}_i \subset \mathcal{C}$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset \quad \forall p_i, p_j \in \mathcal{P} \text{ for } p_i \neq p_j$ (i.e. each card is unique). In a single round of Dixit, one player $p_j \in \mathcal{P}$ is the *story-teller*; the story-teller role cycles between players in each round. After every round, each player draws an unseen card from $\mathcal{C}$ to replace the card they played in the round. The game ends when any player scores 30 or more points, or if the deck $\mathcal{C}$ is out of unseen cards. Below, we describe each step of Dixit and the scoring system (§3).

Step 1: The Story-Teller Generates a Caption.

In the first step of each turn in Dixit, the story-teller $p_j \in \mathcal{P}$ must generate a caption t for one of the cards in their hand $c \in \mathcal{C}_j$. We decompose this into two steps. First, using their hand $\mathcal{C}_j$ as input, p_j must sample a card $c \sim p_j(\mathcal{C}_j) \in \mathcal{C}_j$ to generate a caption for. Next, using the story-teller’s selected card c as input, p_j must generate a caption $t \sim p_j(c)$ for this card. For the story-teller to score points, the caption t should allow some of the other players $p_i \in \mathcal{P} \setminus \{p_j\}$ to identify the correct card c , but not all of them. As a result, when humans play Dixit, captions are often vague. This poses a unique challenge beyond traditional image captioning tasks, where the goal is to produce a highly-accurate caption for an input image.



Figure 2: Visualization of our graphical user interface for humans to play Dixit. Human players can generate captions for cards and select cards from their hand, just like in the original game of Dixit.

Step 2: Players Select Cards from their Own Hand for the Caption.

After the story-teller generates a caption t , it is read aloud or shown to all players $\mathcal{P}$. After understanding the caption, each player except the story-teller $p_i \in \mathcal{P} \setminus \{p_j\}$ must select a card from their hand $c_i \sim p_i(t, \mathcal{C}_i) \in \mathcal{C}_i$ that they believe corresponds to the caption t . Players score points in the third step of Dixit (§3) if other players incorrectly pick their card when they guess which is the story-teller's, so each player is incentivized to select the card from their hand that they believe best corresponds to the caption t_j .

After all non-story-teller players select cards from their hands, these cards are combined with the story-teller's card to create a pool of cards $\mathcal{C}_{pool} = \{c\} \cup \{c_i \mid \forall p_i \in \mathcal{P} \setminus \{p_j\}\}$. Pooling cards also ensures that the story-teller does not generate a fully random caption t_j with the goal of having just one other player selecting this card in the third step (§3) by chance; if the caption is random and another player has a card that better matches it, most players will pick this relevant card and the story-teller will not score any points (§3).

Step 3: Players Select the Card they Believe is the Story-Teller's.

After the pool of cards $\mathcal{C}_{pool}$ is created, all non-story-teller players $p_x \in \mathcal{P} \setminus \{p_i\}$ must select the card $\hat{c}_x \sim p_x(t, \mathcal{C}_{pool} \setminus \{c_x\}) \in \mathcal{C}_{pool}$ they believe was the story-teller's card c_j used to create the caption t_j . Each player p_i knows that their own card c_x is not the story-teller's card c , so for simplicity, we remove this card from $\mathcal{C}_{pool}$ when giving each

player the pool of cards they can select from.

Scoring System

After each turn, the player's votes are tallied, and scoring is calculated as follows. The story-teller receives: 0 points if *no* player picked the story-teller's card, i.e., $\hat{c}_i \neq c; \forall p_i \in \mathcal{P} \setminus \{p_j\}$; 0 points if *all* players picked the story-teller's card, i.e., $\hat{c}_i = c; \forall p_i \in \mathcal{P} \setminus \{p_j\}$; 3 points otherwise. Each non-story-teller receives: 2 points if the story-teller received 0 points; 3 points if the story-teller did not get 0 points and the player picked the story-teller's card; 0 points otherwise. Non-story-tellers also receive an additional +1 point for every player who incorrectly believed their card was the story-teller's (i.e. player p_i gets additional points equal to $\sum_x \mathbb{I}(\hat{c}_x = c_i)$).

4 Experimental Setup

4.1 Dataset

To have MLMS play Dixit, we need images of Dixit cards that we can use as inputs. The Dixit game is proprietary and subject to copyright laws, so we cannot use images from the actual game. Thus, we use cards from an open-source version of Dixit,² where cards are obtained from Pinterest, ensuring that we use our data within its intended use. In total, we obtain $m = 100$ cards to form the deck $\mathcal{D}$.

²<https://github.com/jminuscula/dixit-online>

Model	Avg Points ($\uparrow$)	Avg Position ($\downarrow$)	Dixit Rank ($\downarrow$)	OpenVLM Rank ($\downarrow$)	ChatArena Rank ($\downarrow$)
GPT-4o	29.25	1.725	1	1 (4)	1 (3)
Claude-3.5	28.55	2.050	2	2 (7)	2 (7)
Qwen-2-VL	25.25	3.075	3	3 (18)	3 (19)
InternVL2	22.70	3.725	4	4 (30)	4 (21)
Molmo	18.80	4.525	5	5 (73)	5 (25)
Random	8.85	5.900	6	-	-

Table 1: Benchmarking of MLM Dixit capabilities (left) and comparing MLM Dixit rankings versus rankings on popular MLM benchmarks (right). The *best* performing model is in **bold**.

4.2 Models

We evaluate five strong MLMs in Dixit consisting of two proprietary models (GPT-4o (Achiam et al., 2023) and Claude 3.5 Sonnet (Anthropic, 2023)), and three open-weights models (InternVL2-8B (Chen et al., 2024b), Molmo-7B-D (Deitke et al., 2024), Qwen2-VL-7B (Bai et al., 2023)).³

We also implement a trivial Random player to ensure these models have above-random Dixit capabilities; when the Random player is the story-teller, it selects a caption randomly from a list of 30 captions generated by GPT-4o, and when selecting cards, it randomly samples a card using a Uniform distribution. All models are zero-shot prompted and asked to provide the rationales for their decisions. Exact prompts are shown in Appendix A.

4.3 Human Interface

To allow humans to intervene in our Dixit-based evaluation and compare human and model abilities in the game, we also implement a graphical user interface (GUI) for users to play Dixit (Figure 2). Our GUI is implemented with the PyQt5 library.⁴

5 Results

5.1 Which MLM is the Strongest Dixit Player?

We first understand the relationship between typically measured MLM capabilities and those measured in Dixit by having our six MLMs play 20 games of Dixit. To measure overall MLM performance on Dixit, we use two metrics: **1) Average Points:** the total number of points scored by each MLM divided by the number of games played (20); and **2) Average Position:** the game position of each MLM (e.g. 1 for first, 2 for second) averaged over all games played (20). The player who starts as the story-teller in each game is randomized.

³Model version or HuggingFace names are outlined in Appendix Table 4.

⁴<https://pypi.org/project/PyQt5/>

Our games of Dixit with MLMs reveal two notable findings (Table 1). First, all MLMs greatly outperform the Random player, showcasing above-random Dixit capabilities. MLMs are unlikely to have been trained on Dixit logs, meaning that an MLMs acquired skills in image comprehension, reasoning, and classification from other tasks are transferrable to Dixit—a likely out-of-domain task. Second, Dixit game rankings are perfectly correlated with rankings on OpenVLM and ChatbotArena. As a result, we believe game-based evaluations have the potential to evaluate MLMs as holistically as multi-task benchmarks. Dixit assesses several MLM capabilities—such as image comprehension, reasoning, classification, and captioning—within a single task, eliminating the need for practitioners to use several diverse task formats, and datasets to comprehensively assess MLM capabilities.

5.2 Are Humans Dominated in Dixit?

To study if MLMs have the potential to surpass humans in playing Dixit, our three authors play Dixit with an MLM. As we did not have the resources to host our human Dixit interface online, we run it locally, precluding the use of open-weight MLMs. Due to API cost constraints, we selected GPT-4o Mini for this analysis. Since GPT-4o Mini is the weakest MLM out of the ones we studied, if the model is able to beat any human, it provides evidence that MLMs can play Dixit on par with humans. In future iterations of this work, we plan to extend our human versus MLM analysis to more capable and a wider variety of models.

We run three practice rounds of Dixit so the human players are confident on how to play the game. We play three games of Dixit, each lasting around 30 minutes. All players verbally expressed that the games were enjoyable (average Likert rating of 4.433), showing that Dixit and game-based evaluations could be a more engaging method to allow

	Game 1		Game 2		Game 3		Average	
Player	Points ($\uparrow$)	Rank ($\downarrow$)	Points ($\uparrow$)	Rank ($\downarrow$)	Points ($\uparrow$)	Rank ($\downarrow$)	Points ($\uparrow$)	Rank ($\downarrow$)
Player 1	31.00	1	31.00	1	29.00	2	30.33	1.33
Player 2	28.00	2	19.00	4	30.00	1	25.67	2.33
Player 3	24.00	3	23.00	2	14.00	4	20.33	3.00
GPT-4o Mini	23.00	4	22.00	3	21.00	3	22.00	3.33

Table 2: GPT-4o mini vs. three authors in Dixit over three games. The *worst* player is in **bold**.

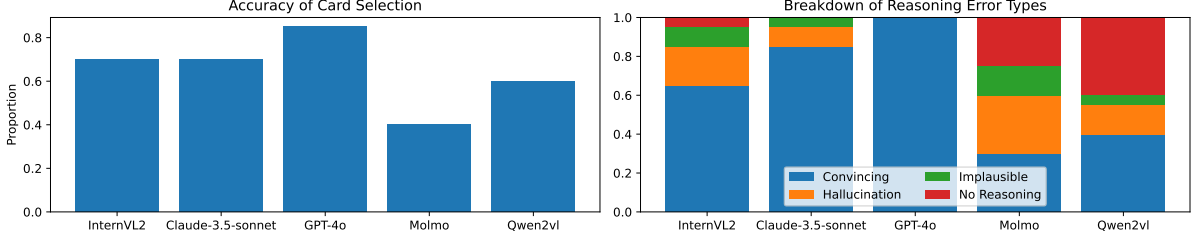


Figure 3: Accuracy of MLM selections for story-teller captions (left) along with an error analysis of the rationales provided for the selections (right) in each model's generated chain-of-thought.

Player	Player 1	Player 2	Player 3	GPT-4o Mini
Player 1	1.00	-	-	-
Player 2	0.61	1.00	-	-
Player 3	0.35	0.53	1.00	-
GPT-4o Mini	0.53	0.47	0.56	1.00

Table 3: Raw agreement between human and Dixit players. Highest non-perfect agreement is bold.

practitioners to directly intervene in their evaluations. As engagement is often linked to education, it would be interesting for future work to see if game-based evaluations allow model designers to learn more about their models (Oblinger, 2004).

We show the results from our games in Table 2. Humans tend to have the upper-hand on MLMS, measured through the points in each game and the rank of the players. However, GPT-4o Mini was able to beat one player (Player 3) in points on average. Since this model is our weakest closed-source model, we speculate that MLMS could demonstrate Dixit capabilities that are similar to or maybe even surpass humans.

5.3 What Strategies do MLMS Use in Dixit?

For a fine-grained analysis into the types of strategies MLMS employ in Dixit, we study the logs and interactions between MLMS when they play Dixit. We decompose our analysis into three questions.

5.3.1 How Accurate are MLM Story-Teller Selections?

MLM Card Selection Accuracy: To assess MLM image comprehension and reasoning capabilities in Dixit, we study the accuracy of MLM decisions when they are prompted to select the story-teller's card. We first have one Ph.D. student annotate if the card selected by the MLM plausibly matches the caption of the story-teller, which we call **accuracy**.

In Figure 3 (left), we find that most MLMS have similar, moderate accuracy, around 0.70. The notable exception in GPT-4o, which has an accuracy over 0.80; we believe this improvement over other MLMS can explain the model's high win rate when models compete against each other.

MLM Reasoning Errors: Since all MLMS are prompted to select cards via chain-of-thought, we analyze the rationales provided by the models to identify any potential reasoning gaps that could be improved upon in future work. We classify rationales into four types: **1) Convincing:** No issues; **2) Implausible:** The rationale behind the selected card is unrealistic or far-fetched; **3) Hallucination:** The rationale references something not present in the card; and **4) No Reasoning:** The model does not engage in reasoning (e.g. simply repeats the caption).

In Figure 3 (right), we find that some models, in particular closed-source models, are able to provide convincing justifications for their decisions via chain-of-thought; open-source models lag behind, revealing that model practitioners can aim

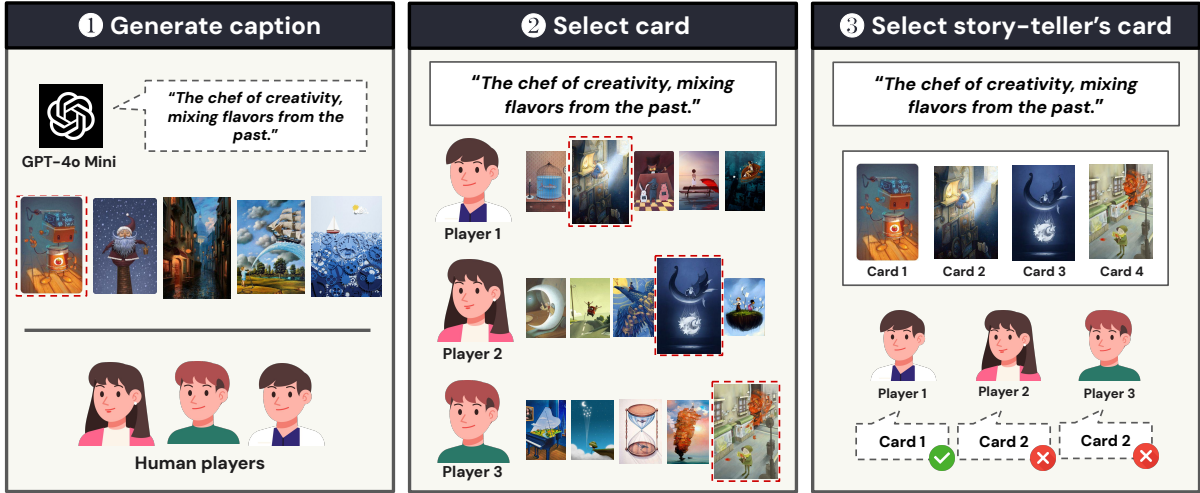


Figure 4: Illustration of example when GPT-4o Mini generates a caption that fooled 2/3 human players.

to design models with improved chain-of-thought reasoning. Further, we find that the majority of errors are from models either refusing to engage in reasoning or hallucinating items that are not present in the image. Thus, Dixit could serve as a valuable testbed to study MLM hallucinations and reasoning (Wu et al., 2024).

MLM and Human Card Selection Correlation:

We now explore the raw agreement between different human players and GPT-4o Mini from our human versus MLM games when players must select the story-teller’s card from the same set of cards. In Table 3, we find that the highest agreement are between two human players: Player 1 and Player 2. However, GPT-4o Mini also has fairly high agreement with some human players, always exceeding 0.50. This highlights that Dixit is a highly subjectively game, making it a valuable testbed for creativity and subjective decision-making. Further, it shows GPT-4o Mini aligns with human reasoning patterns reasonably well in the Dixit image classification tasks.

5.3.2 What Types of Captions do MLMS Generate?

In this section, we aim to understand the differences between the captioning strategies of humans and MLMS. We compute the average number of tokens in MLM and human-generated captions; MLM captions have 11.02 tokens on average, while human ones have 3.57 tokens, showing MLMS tend to generate lengthier, more descriptive captions compared to humans.

Further, upon manual inspection, we find MLMS tend to engage in *literal* Dixit strategies—

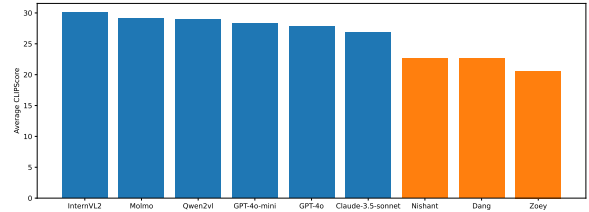


Figure 5: Average CLIPScore between images and generated captions among different players. Orange columns denote human players.

describing exactly what is occurring in the picture (e.g. “Child reaching for the moon”, “A forgotten journey through enchanted woods”). In contrast, humans tend to reference external knowledge and pop culture when generating captions (e.g. “Rapunzel”, “Marceline from Adventure Time”); we speculate such references are more fun for players and harder in the image selection stage, as it requires an extra hop of commonsense reasoning (Bauer et al., 2018). As a result, we believe future work in designing optimal MLM Dixit players can find ways to instill external commonsense knowledge while playing (Wang et al., 2020).

To confirm these differences in strategies quantitatively, we employ CLIPScore (Hessel et al., 2021), which measures the semantic similarity between each image and its corresponding generated caption. We compute the average CLIPScores of captions generated by MLMS and humans in Figure 5. Our results align with our qualitative assessment: humans tend to generate more ambiguous captions, allowing for various interpretations, while MLMS often produce captions that precisely de-

scribe the given images, making it easier for other players to identify the correct card.

5.3.3 When Do MLMs Generate Captions that Fool Most Human Players?

We analyze a case where GPT-4o Mini served as the storyteller and successfully generated a caption that fooled most human players in the human versus MLM game, as shown in Figure 4. The caption generated by GPT-4o Mini for its selected card was, “A chef of creativity, mixing flavors from the past.” The reasoning behind this caption was that the quirky character depicted in the card image appeared ready to “serve memories from a can.” Each non-storyteller human player then selected a card from their hand to contribute to the pool of options. Among the pool of cards, only one human player (Player 1) correctly identified the storyteller’s card, while the other two human players (Player 2 and Player 3) mistakenly selected Player 1’s card. GPT-4o Mini’s success in fooling the human players highlights its capability to craft a caption that is not *literal* but rather ambiguous, allowing for multiple interpretations. This made it challenging for the human players to identify the story-teller’s card.

6 Conclusion

In this paper, we explore the utility of Dixit as a testbed for holistically evaluating MLM capabilities, including: image comprehension, image captioning, and image classification. While MLMs excel in these capabilities, demonstrated by above-random Dixit performance in our offline analysis and competitive win-rates that rival human players, we also uncover several areas of improvement for future Dixit MLM models; notably, open-source MLM showcase significant reasoning weaknesses in image classification—prone to hallucinations and lack step-by-step breakdowns, while all models tend to generate detailed and literal captions for input images—unlike humans who often produce shorter caption that reference external knowledge and commonsense. We encourage future work to explore training strategies beyond basic prompting techniques to build stronger Dixit models. Specifically, we believe it will be an interesting challenge to design models that not only excel in Dixit win-rate, which can be achieved via strategies like self-play (Silver et al., 2017), but also make the game enjoyable for other human players. Overall, we argue that game-based evaluations like Dixit are a robust, challenging, and engaging testbed for MLM

capabilities, and we encourage future works to explore these setups further to evaluate their models within a single, unified task.

7 Limitations

We selected Dixit as our testbed for the several reasons: (1) Dixit tests MLM’s creative captioning and calibration capability that only allow limited number of users to pick the correct card. (2) Dixit tests MLM’s theory-of-mind in image classification since they need to pick cards in consideration of the other players’ strategies. To this end, Dixit provides a unified evaluation test for testing multiple MLM capabilities. Despite these strengths, our findings should be interpreted with the limited scope of this task in mind. Nevertheless, we will make the card images, code, and framework publicly available so that new strategies, as well as tasks, can be seamlessly added.

Dixit is a game that inherently reflects the shared or non-shared knowledge of its players when generating captions. For instance, when playing Dixit with friends, one might strategically use non-shared knowledge to make it harder for others to select the correct card or rely on shared knowledge to make the game more entertaining. Consequently, a MLM’s performance in playing Dixit is likely influenced by its inherent training distribution and alignment process. Our current setup does not fully account for this, leaving open questions about how model family characteristics influence performance. Future research could explore this by comparing models within the same family to determine whether shared training distributions contribute to improved performance.

References

- Abdelrahman Abdelhamed, Mahmoud Afifi, and Alec Go. 2024. [What do you see? enhancing zero-shot image classification with multimodal large language models](#).
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

- Anthropic. 2023. Meet claude. <https://www.anthropic.com/product>. Accessed: 2024-09-10.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.
- Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*.
- Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. *arXiv preprint arXiv:2305.13455*.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.
- Rheta DeVries. 2015. Games with rules. In *Play from birth to twelve*, pages 151–158. Routledge.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. [Vlmevalkit: An open-source toolkit for evaluating large multi-modality models](#).
- Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. 2024. [Using game play to investigate multimodal and conversational grounding in large multimodal models](#).
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chao-Chun Hsu, Yu-Hua Chen, Zi-Yuan Chen, Hsin-Yu Lin, Ting-Hao Huang, and Lun-Wei Ku. 2019. Dixit: Interactive visual storytelling via term manipulation. In *The World Wide Web Conference*, pages 3531–3535.
- Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*.
- Maithilee Kunda and Irina Rabkina. 2020a. Creative captioning: An ai grand challenge based on the dixit board game. *arXiv preprint arXiv:2010.00048*.
- Maithilee Kunda and Irina Rabkina. 2020b. [Creative captioning: An ai grand challenge based on the dixit board game](#).
- Felix Lebed. 2006. System approach to games and competitive playing. *European Journal of Sport Science*, 6(01):33–42.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024a. [Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback](#).
- Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. 2024b. Vhelm: A holistic evaluation of vision language models. *arXiv preprint arXiv:2410.07112*.
- Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. 2024. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.
- Shalev Lifshitz, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila McIlraith. 2024. Steve-1: A generative model for text-to-behavior in minecraft. *Advances in Neural Information Processing Systems*, 36.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Jessica M Lopez, Kaitlin Woolley, and Ann L McGill. 2021. A preference for preference: Lack of subjective preference evokes dehumanization. *Organizational Behavior and Human Decision Processes*, 164:52–67.
- Yujie Lu, Dongfu Jiang, Wenhao Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in

- the wild with human preferences. *arXiv preprint arXiv:2406.11069*.
- Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. [Improving multimodal datasets with image captioning](#).
- Diana G Oblinger. 2004. The next generation of educational engagement. *Journal of interactive media in education*, 2004(1):10–10.
- Marc Prensky. 2001. Fun, play and games: What makes games engaging. *Digital game-based learning*, 5(1):5–31.
- Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. *arXiv preprint arXiv:2407.16711*.
- Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.
- Weihao Tan, Wentao Zhang, Shanqi Liu, Longtao Zheng, Xinrun Wang, and Bo An. 2024. [True knowledge comes from practice: Aligning llms with embodied environments via reinforcement learning](#).
- Jinyin Wang, Haijing Zhang, Yihao Zhong, Yingbin Liang, Rongwei Ji, and Yiru Cang. 2024a. [Advanced multimodal deep learning architecture for image-text matching](#).
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10760–10770.
- Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Shawn Ma, and Yitao Liang. 2024b. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. *Advances in Neural Information Processing Systems*, 36.
- Nicholas R. Waytowich, Devin White, MD Sunbeam, and Vinicius G. Goecks. 2024. [Atari-gpt: Benchmarking multimodal large language models as low-level policies in atari games](#).
- Steven Daniel Webb and Sieteng Soh. 2007. Cheating in networked computer games: a review. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts*, pages 105–112.
- Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddhartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.
- Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, et al. 2024. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *arXiv preprint arXiv:2406.10900*.
- Yuxiang Wu, Zhengyao Jiang, Akbir Khan, Yao Fu, Laura Ruis, Edward Grefenstette, and Tim Rocktäschel. 2023. Chatarena: Multi-agent language game environments for large language models. Version 0.1. GitHub repository: <https://github.com/chatarena/chatarena>.
- Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. 2024. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.
- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2025. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023a.

Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Sipeng Zheng, Jiazheng Liu, Yicheng Feng, and Zongqing Lu. 2023b. Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds. *arXiv preprint arXiv:2310.13255*.

Model	Name
GPT-4o	gpt-4o-2024-08-06
Claude 3.5 Sonnet	claude-3-5-sonnet-20241022
InternVL2-8B	InternVL2-8B
Molmo-7B-D	Molmo-7B-D-0924
Qwen2-VL-7B	Qwen2-VL-7B-Instruct

Table 4: Model version or HuggingFace model names of tested MLMs.

A Prompts

Prompt A.1: Dixit Rules Explanation

Dixit Rules
Dixit is a game of storytelling and imagination. Each round, one player takes on the role of the storyteller. The storyteller picks one of the six images in their hand and thinks of a sentence or phrase that describes that image. The other players then select an image from their hand that best matches the sentence given by the storyteller. The chosen images are shuffled and revealed, and players must guess which image belongs to the storyteller. Points are awarded to the storyteller if some, but not all, players guess the correct image. Points are also awarded to players whose images are selected. The game ends when the deck of cards is empty, and the player with the most points wins.

Scoring
- If all players guess the storyteller's image, the storyteller receives no points, and all other players receive 2 points.
- If no players guess the storyteller's image, the storyteller receives no points, and all other players receive 2 points (+1 bonus point for each vote their image receives).
- If some, but not all, players guess the storyteller's image, the storyteller receives 3 points, and each player who guessed correctly receives 3 points (+1 bonus point for each vote their image receives).

Prompt A.2: Selecting a Story-Teller Card

```
{{dixit_rules}}
```

Instructions
You are the storyteller for this round, and you have selected the following image from your hand. Think of a sentence or phrase that describes this image. Your output must follow the following JSON format:

```
```json
{
 "thought": "Your thought here.",
 "caption": "Your caption here."
}
```
```

Prompt A.3: Generating a Story-Teller Caption

```
{{dixit_rules}}
```

Instructions
You are the storyteller for this round, and the images below are the cards in your hand. Select an image to generate a caption for. Your output must follow the following JSON format:

```
```json
{
 "thought": "Your thought here.",
 "choice": "Write a single integer representing
the index of the card you selected. Valid
choices are {{valid_choices}}.",
}
```
```

Prompt A.4: Selecting a Card for the Pool

```
{{dixit_rules}}
```

Instructions
You are a non-storyteller player for this round. The storyteller provided the following caption: {{caption}}. The images below are the cards on the table. Select an image that you think belongs to the storyteller. Your output must follow the following JSON format:

```
```json
{
 "thought": "Your thought here.",
 "choice": "Write a single integer representing
the index of the card you selected. Valid
choices are {{valid_choices}}.",
}
```
```

Prompt A.5: Voting on the Story-Teller's Card

```
{{dixit_rules}}
```

Instructions
You are a non-storyteller player for this round. The storyteller provided the following caption: {{caption}}. The images below are the cards in your hand. Select an image that you think best matches the caption. Your output must follow the following JSON format:

```
```json
{
 "thought": "Your thought here.",
 "choice": "Write a single integer representing
the index of the card you selected. Valid
choices are {{valid_choices}}.",
}
```
```