# Comparing human and machine communication patterns through a Tangram game

**Haoran Zhao**
University of Washington
hjzhao@uw.edu

**Colin Conwell**
MIT CSAIL
conwell@mit.edu

## Abstract

When humans communicate about visual objects, they develop shared linguistic conventions that progressively reduce referential ambiguity through collaborative dialogue. To better understand the representational patterns underlying human communication and test whether vision-capable large language models (VLLMs) exhibit similar communicative behaviors, we compare human-human and agent-agent interactions in the tangram communication game. In this task, two players establish shared references for abstract shapes through dialogue across six repeated rounds. We analyzed existing human-human data and conducted agent-agent experiments with five VLLMs, measuring performance and using representational probes to explore the potential structure underlying performance. Humans demonstrate clear convention formation, with representations becoming increasingly distinguishable across rounds as task accuracy improves from 78% to 96%. In contrast, AI agents fail to exhibit similar collaborative patterns in our experiments, achieving consistently low performance (10-30%) with minimal improvement and no evidence of convention development, despite access to interim accuracy reports, full conversation history, and (most curiously) what appear to be largely accurate initial descriptions by "director" agents. Taken together, these preliminary results suggest that current VLLMs—without task-specific training—may still struggle with kinds of grounded, evolving, coreferential communication of that defines human language in collaborative contexts.

## 1 Introduction

People develop shared linguistic conventions over repeated interactions when referring to the same objects [7; 15; 1]. For example, two friends in a reference game might initially describe an abstract figure as "the one that looks like a person raising their arm and leaning left", but over repeated rounds come to refer to the same shape simply as "the dancer". This collaborative (near-universal) process clearly enables increasingly efficient communication [6]. Far less clear are the underlying computational, algorithmic, and representational structures that enable it [25]. The ascendance in recent years of multi-modal generative AI systems and vision-capable large language models (VLLMs, e.g. `GPT4o, GPT5, Gemini2.5`, etc.), coupled with the growing sophistication of natural language processing techniques (e.g. vector-semantic analysis) that link human-comparable linguistic behavior to numerical latents, now offers renewed opportunity to address the gaps in our understanding of efficient coreferential communication [12; 16; 18; 24]. This also led us to ask: do otherwise competent language-generating AI systems exhibit human-like communicative patterns when referring to grounded visual content?

In this work, we study these questions through the tangram communication game, where 2 paired players establish shared references for abstract shapes through dialogue across repeated rounds. We first analyze human-human interactions from existing data [13], examining how linguistic dis-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: DBM.

tinguishability evolves across rounds by measuring embedding-based (dis)-similarity scores and visualizing representational trajectories through Principal Component Analysis. We then conduct agent-agent versions of the game with 5 state-of-the-art VLLM models to investigate whether artificial agents exhibit similar communicative patterns.

Our comparison reveals significant differences between human and AI players. Humans show clear convention formation, with embeddings-based linguistic distinguishability scores increasing monotonically from 0.65 to 0.75 across rounds as correspondent behavioral performance improves from 78% to 96% accuracy. In our preliminary experiments, AI agents fail to demonstrate similar patterns, achieving consistently low performance between 10% and 30% with little-to-no discernible improvement, suggesting that current VLLMs may not generalize to collaborative coreference resolution in this (or similar) tasks—despite sophisticated individual vision and language capabilities and opportunity for in-context learning by way of interim accuracy reports and full conversation history. We conclude with a discussion of possible culprits for this performance gap and possible paths to improvement.

## 2 Related Works

**Referential communication game in humans** Referential communication games investigate how humans establish shared understanding through interactive dialogue. Foundational work by Clark and Wilkes-Gibbs [7] demonstrated collaborative titration of specific expressions through repeated interactions, and showed how common ground (successful coreference) emerges through communicative grounding. Subsequent research examined speaker adaptation to partner knowledge and familiarity [15], conceptual pacts for difficult-to-describe objects [1], and how speakers balance informativeness and efficiency [13]. Recent work has investigated pragmatic reasoning in reference resolution [14] and linked successful performance to deeper theory of mind [29].

**Visual reasoning through communication games in LLMs** Recent research shows agents can communicate about visual inputs through referential signaling games [17; 11; 23], with modern approaches incorporating structured representations for interpretable emergent communication [2]. However, it has also been shown that current vision-language models (including modern VLLMs) face limitations in visual reasoning due to modular architectural designs in which vision and language components remain largely separate [4; 21]. Researchers are addressing these challenges in multiple ways, including enhanced multimodal cross-talk [22], reward-optimized, context-dependent captioning [8], and deeper integration of perceptual processing with language-based reasoning [19; 27].

## 3 Data Collection

**What is the Tangram game?** The tangram game is a collaborative puzzle where a **director** and **matcher** work together to arrange twelve abstract shapes (see classic examples from [7] in Figure 1). Both players see a 6×2 grid of tangrams, but while the director's pieces are in a fixed target order, the matcher's are randomly scrambled. Through multi-turn dialogue, the director guides the matcher to rearrange their pieces to match the target order. After the matcher submits their arrangement and receives a score (out of 12), the same twelve tangrams are scrambled again in a new random order for the next round. This process repeats for 6 rounds total.

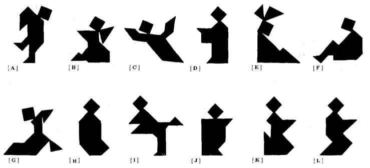


Figure 1: 12 tangrams stimulus set (taken from Clark and Wilkes-Gibbs [7])

**Human-human game** : We utilized the data collected from Hawkins et al. [13], which recruited 200 Amazon Mechanical Turk participants paired into dyads for real-time tangram games. After excluding incomplete sessions and non-native speakers, the dataset contains 9,967 utterances from 67 complete games. Participants were randomly assigned as directors or matchers and interacted with numerically-labeled 6×2 grids containing twelve tangram shapes from Clark and Wilkes-Gibbs [7].

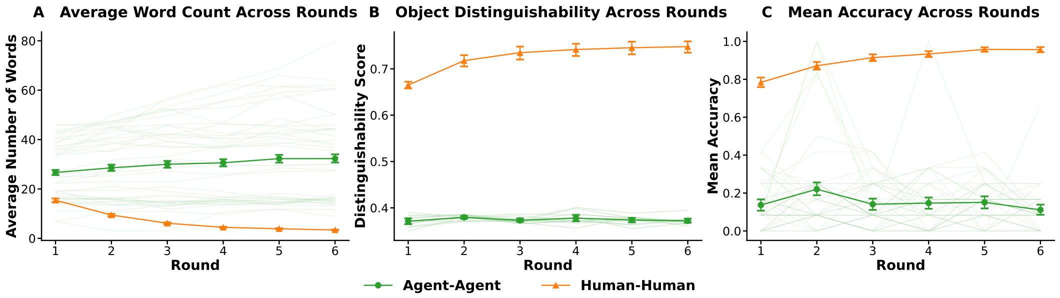


Figure 2: Comparison between human-human and `GPT4o-GPT4o` games on averaged distinguishability of 12 tangrams, average word count of directors' description, and mean accuracy across rounds. Light green lines show the results of individual Agent-Agent game sessions. Error bars are bootstrapped 95% confidence intervals across vignettes.

**Agent-agent game** : We replicated the tangram game with five VLLMs: `GPT-4o`, `GPT-5-mini`, `Claude-3.7-sonnet`, `Claude-sonnet-4`, and `Gemini-2.5-flash`. These models were selected as leading representatives of major VLLM families (OpenAI, Anthropic, Google), each known for strong performance across multimodal and reasoning benchmarks. We paired models of the same type as director-matcher dyads (e.g., `GPT-4o` director with `GPT-4o` matcher, etc.). We then ran multiple game sessions with every model pair and collected 2,880 utterances from 40 complete games after removing incomplete sessions [1] (see Table 1 and 2 for sample descriptions). A verbatim reproduction of the prompts we provided to the director and matcher agents is available in subsection A.1.

## 4 Results Analysis

Our analysis focuses on utterances produced by *directors*, as they provide the primary communicative signal for referent identification. For all sentence embedding extractions, we used the sentence-transformers model `all-mpnet-base-v2`.[2]

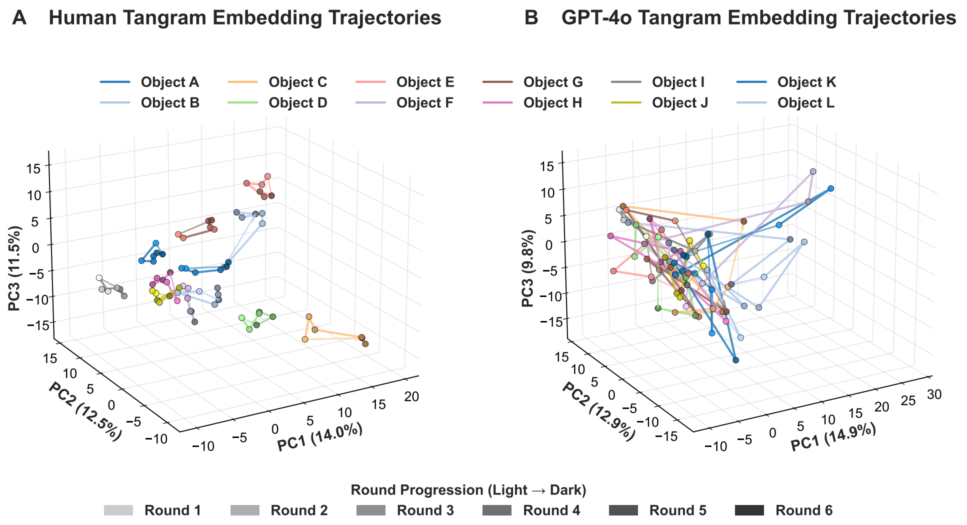
**Human communication patterns and performance** We observed systematic changes in linguistic patterns across rounds, with average distinguishability scores increasing from 0.65 in Round 1 to 0.75 in Round 6 (see Figure 2B), showing strong correlation with round number ($R^2 = 0.731, p = 0.030$). The distinguishability score was calculated as $\frac{1}{n(n-1)} \sum_{i \neq j}(1 - \text{cosine\_similarity}(\mathbf{e}_i, \mathbf{e}_j))$, where $n = 12$ in our case, and $\mathbf{e}_i$ and $\mathbf{e}_j$ are embedding vectors for objects $i$ and $j$, averaged across all pairwise combinations. Complementarily, between-object similarity decreased correspondingly, demonstrating that coreference resolution led to increasingly differentiated linguistic representations. As objects became more linguistically distinguishable, humans demonstrated corresponding improvements in task accuracy, with performance increasing from 78.4% in Round 1 to 95.7% by Round 6 (see Figure 2C), confirming that reduced referential ambiguity drives successful collaborative communication.

**Agent game analysis** We identified notable contrasts between agent-agent and human-human communication patterns. Humans showed a clear decrease in word count per tangram description across rounds, whereas VLLMs became increasingly verbose, sometimes exceeding 80 words per description (see Figure 2A). Human games also exhibited steadily rising object distinguishability, while VLLMs remained largely stable with little variation. Notably, VLLMs showed no systematic gains in task accuracy, maintaining performance between 10% and 30%, significantly above chance (8.33%), but far below human levels (see Table 3 for details). Although they occasionally achieved perfect accuracy in isolated rounds (almost always in Round 2), these successes were inconsistent and did not indicate reliable learning or adaptation.

**Representation comparison** We further visualized the representational structure by applying Principal Component Analysis (PCA) to the aggregated descriptions of each tangram across rounds.

[1] We used temperature $\tau = 1.0$ for GPT and Claude models, and $\tau = 0.7$ for Gemini.

[2] All analyses were performed on local machines with 32–64GB RAM and up to 16 CPU cores.


A Human Tangram Embedding Trajectories

B GPT-4o Tangram Embedding Trajectories

Figure 3: 3D visualization of embedding trajectories for descriptions of all 12 Tangrams

Although the first three principal components captured less than 40% of the total variance, the 3D visualization nevertheless reveals distinct patterns between human and AI communication. For human interactions (Figure 3A), the twelve tangram descriptions form discrete clusters with relatively stable trajectories across the six rounds, indicating consistent representational structure with gradual refinement. In contrast, the `GPT-4o` representations (Figure 3B) exhibit highly scattered and erratic trajectories with no clear clustering or systematic evolution across rounds, demonstrating the absence of coherent communicative conventions comparable to human patterns.

## 5 Discussion

Our preliminary results reveal notable differences between human and AI communicative patterns in an iterative referential task. While previous work on tangram games has primarily focused on studying AI models' visual and abstract reasoning capabilities [5; 10], our study takes a different angle by using the game to examine the evolution of communication patterns across repeated interactions. We found that VLLMs mostly failed to exhibit human-like linguistic adaptation: they neither developed increasingly concise descriptions nor established distinguishable conventions for different objects, and showed no systematic performance improvement despite access to accuracy reports and conversation history that could provide the basis for (in-context) learning.

One key finding is the asymmetric performance between AI directors and matchers. Manual inspection reveals that AI directors can often generate accurate, distinctive descriptions that differentiate between tangram shapes (see Table 1 and Table 2). AI matchers, on the other hand, consistently struggle with the ordinal arrangement required by the game, failing to correctly map descriptions to their corresponding positions. This disconnect suggests that one potential bottleneck to performance in the tangram game may simply be the ordingal arrangement process. This may in turn reflect known limitations of VLLMs in numerical, spatial, and relational reasoning [9; 26; 31; 3; 20; 28; 30], as the tangram matching task fundamentally requires understanding positional relationships.

Several experimental design choices we made offer avenues for future investigation. Our setup required AI matchers to process all twelve tangrams simultaneously and provide a single complete arrangement at the last, mirroring the human experimental protocol [13]. This may have overwhelmed the models' ordinal reasoning abilities. Future work could explore a sequential matching version, where tangrams are presented one at a time for individual matching. Additionally, mixed human-AI experiments (pairing human directors with AI matchers or vice versa) could further isolate and distill the specific impacts of description quality versus the reasoning and task-specific behaviors required for success human-like collaborative communication. New prompt variations that unlock further performance are always possible, as well; though at the same time, of course, the engineering of the perfect prompt is a human-AI collaborative communication endeavor still early in its evolution.

## 6 Code + Data Availability

Code and data for this project will be made available (or linked) via the Project GitHub repo:
- github.com/ColinConwell/DBM-Tangrammer

## References

[1] Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 6 (1996), 1482–1493.

[2] Ruxiao Chen, Dezheng Han, Wenjie Han, and Shuaishuai Guo. 2025. Cognitively-Inspired Emergent Communication via Knowledge Graphs for Assisting the Visually Impaired. *arXiv.org* (2025).

[3] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773* (2025).

[4] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024. Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning. *AAAI Conference on Artificial Intelligence* (2024).

[5] Christopher Clark, Jordi Salvador, Dustin Schwenk, Derrick Bonafilia, Mark Yatskar, Eric Kolve, Alvaro Herrasti, Jonghyun Choi, Sachin Mehta, Sam Skjonsberg, Carissa Schoenick, Aaron Sarnat, Hannaneh Hajishirzi, Aniruddha Kembhavi, Oren Etzioni, and Ali Farhadi. 2021. Iconary: A Pictionary-Based Game for Testing Multimodal Communication with Drawings and Text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1864–1886. `https://doi.org/10.18653/v1/2021.emnlp-main.141`

[6] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*, Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (Eds.). American Psychological Association, 127–149. `https://doi.org/10.1037/10096-006`

[7] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.

[8] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. 2025. Perceptual Decoupling for Scalable Multi-modal Reasoning via Reward-Optimized Captioning. *arXiv.org* (2025).

[9] Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language Models Hallucinate, but May Excel at Fact Verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1090–1111. `https://doi.org/10.18653/v1/2024.naacl-long.62`

[10] Mustafa Omer Gul and Yoav Artzi. 2024. CoGen: Learning from Feedback with Coupled Comprehension and Generation. arXiv:2408.15992 [cs.CL] `https://arxiv.org/abs/2408.15992`

[11] Serhii Havrylov and Ivan Titov. 2017. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. *Neural Information Processing Systems* (2017).

[12] Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual Adaptation for Efficient Machine Communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Raquel Fernández and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 408–419. `https://doi.org/10.18653/v1/2020.conll-1.33`

[13] Robert D. Hawkins, Mike Frank, and Noah D. Goodman. 2017. Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

[14] Robert X.D. Hawkins, Noah D. Goodman, and Robert L. Goldstone. 2019. The Emergence of Social Norms and Conventions. *Trends in Cognitive Sciences* 23, 2 (2019), 158–169. https://doi.org/10.1016/j.tics.2018.11.003

[15] Ellen A Isaacs and Herbert H Clark. 1987. References in conversation between experts and novices. *Journal of Experimental Psychology: General* 116, 1 (1987), 26–37.

[16] Angeliki Lazaridou and Marco Baroni. 2020. Emergent Multi-Agent Communication in the Deep Learning Era. arXiv:2006.02419 [cs.CL] https://arxiv.org/abs/2006.02419

[17] Angeliki Lazaridou, A. Peysakhovich, and Marco Baroni. 2016. Multi-Agent Cooperation and the Emergence of (Natural) Language. *International Conference on Learning Representations* (2016).

[18] Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7663–7674. https://doi.org/10.18653/v1/2020.acl-main.685

[19] Heekyung Lee, Jiaxin Ge, Tsung-Han Wu, Minwoo Kang, Trevor Darrell, and David M. Chan. 2025. Puzzled by Puzzles: When Vision-Language Models Can't Take a Hint. *arXiv.org* (2025).

[20] Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics* 11 (2023), 635–651.

[21] Jingming Liu, Yumeng Li, Boyuan Xiao, Yichang Jian, Ziang Qin, Tianjia Shao, Yao-Xiang Ding, and Kun Zhou. 2024. Autonomous Imagination: Closed-Loop Decomposition of Visual-to-Textual Conversion in Visual Reasoning for Multimodal Large Language Models. (2024).

[22] Shuhang Liu, Zhenrong Zhang, Pengfei Hu, Jie Ma, Jun Du, Qing Wang, Jianshu Zhang, Quan Liu, Jianqing Gao, and Feng Ma. 2025. MMC: Iterative Refinement of VLM Reasoning via MCTS-based Multimodal Critique. *arXiv.org* (2025).

[23] Daniela Mihai and Jonathon S. Hare. 2021. The emergence of visual semantics through communication games. *arXiv.org* (2021).

[24] Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics* 5 (2017), 325–338. https://doi.org/10.1162/tacl_a_00064

[25] Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences* 27, 2 (2004), 212–225. https://doi.org/10.1017/S0140525X04450055

[26] Roussel Rahman. 2025. Large Language Models in Numberland: A Quick Test of Their Numerical Reasoning Abilities. arXiv:2504.00226 [cs.AI] https://arxiv.org/abs/2504.00226

[27] Yeonsang Shin, Jihwan Kim, Yumin Song, Kyungseung Lee, Hyunhee Chung, and Taeyoung Na. 2025. Generating Animated Layouts as Structured Text Representations. *arXiv.org* (2025).

[28] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems* 37 (2024), 75392–75421.

[29] Lane Wardlow. 2013. Individual differences in speakers' perspective taking: The roles of executive control and working memory. *Psychonomic Bulletin & Review* 20, 4 (2013), 766–772.

[30] Zihan Weng, Lucas Gomez, Taylor Whittington Webb, and Pouya Bashivan. 2025. Caption This, Reason That: VLMs Caught in the Middle. *arXiv preprint arXiv:2505.21538* (2025).

[31] Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2025. Number Cookbook: Number Understanding of Language Models and How to Improve It. In *International Conference on Learning Representations*. `https://openreview.net/forum?id=BWS5gVjgeY`

## A Appendix + Supplementary Information

### A.1 Prompts provided to AI player agents (director and matcher)

Below, we reproduce the prompts provided to the director and matcher agents in our AI version of the tangrams games, with some variable names left as-is for clarity:

- **Director (System Prompt)**: “In this conversation, you will be playing a 2-player communication game. Your role in this game that of the DIRECTOR. The role of the other player is the MATCHER. Your goal is to describe the shapes in your board order so the MATCHER can reorder their images to match yours. Do not mention filepaths or URLs to images in your descriptions. If playing for multiple rounds, try to update your descriptions based on the MATCHER’s previous responses to help the MATCHER find the correct order."
- **Director (User Prompt), Round = 1**: “Please describe each of these 12 shapes in <description></description> tags ordered correctly and numbered 1 through 12, e.g. as follows:
  - <description>1. [Your description of the 1st shape...] </description>
  - <description>2. [Your description of the 2nd shape...] </description>
  - ...
  - <description>12. [Your description of the 12th shape...]</description>

  Each description should be sufficient for the MATCHER to identify and place the shapes in the correct order. You MUST provide exactly 12 <description></description> tags, numbered 1 through 12.
- **Interim System Prompt, Director + Matcher, Round > 1**: Accuracy of MATCHER’s order: ACCURACY
- **Director (User Prompt), Round > 1**: “Taking into account the included conversation history above and MATCHER’s accuracy (if any), please provide new or modified descriptions of the 12 shapes using the <description></description> tags specified above. Remember to use exactly 12 <description></description> tags."
- Matcher (System Prompt): “In this conversation, you will be playing a 2-player communication game. Your role in this game that of the MATCHER. The role of the other player is the DIRECTOR. Your goal is to place your shapes into the order intended by the DIRECTOR based on their description. Always provide your response using the specified formatting tags. You may include overall reasoning before along with these tags."
- **Matcher (User Prompt), Round = 1** “Here are the DIRECTOR’s descriptions: DIRECTOR DESCRIPTIONS Analyze the DIRECTOR’s descriptions and match them to your 12 shapes. The DIRECTOR has described the 12 shapes in order. Your task is to reorder your shapes to match the order described by the DIRECTOR. RESPONSE FORMAT:"
  1. You may include overall reasoning about your approach (optional).
  2. Provide reasoning for each position using <reasoning></reasoning> tags:
     - <reasoning>1. [Why you think position 1 matches the described shape] </reasoning>
     - <reasoning>2. [Why you think position 2 matches the described shape] </reasoning>
     - ...
     - <reasoning>12. [Why you think position 12 matches the described shape] </reasoning>
  3. Provide your final answer using the <order></order> tag, e.g. as follows: <order>1,5,6,7,4,11,3,9,12,10,2</order>

REQUIREMENTS:

1. If you include reasoning, you MUST provide exactly 12 <reasoning></reasoning> tags, numbered 1 through 12
2. You MUST provide exactly 1 <order></order> tag with 12 comma-separated integers
3. Each integer from 1 to 12 must appear exactly once in the <order></order> tag.
4. The <reasoning></reasoning> tags should explain your logic for matching each shape to each description.

- **Matcher (User Prompt), Round > 1**: (Injected conversation history, then same as Matcher (User Prompt) Round = 1).

### A.2 Sample descriptions from human and AI directors

Here, we show two tables of sample descriptions from humans and `GPT-4o`. Table 1 shows sample descriptions from humans and `GPT-4o` for all 12 objects from a random session in round 6. Table 2 shows sample descriptions of object D across rounds from humans and `GPT-4o`.

| Object | Human | GPT-4o |
|---|---|---|
| **A** | fat man leg up front head back | Square head, dynamic pose with one leg raised, like a mid-stride or stepping motion |
| **B** | angel praying | Square head, seated with one leg bent and the other facing forward, creating a relaxed position |
| **C** | Jesus hugging | Shape with a square head, arms raised widely in a flying motion, with a sharp, angled body |
| **D** | man takes up offering | Diamond head with a straight, tall body, left arm extended forward, resembling a composed stance |
| **E** | sitting under a plant, right | Diamond head tilted back, angular body and leg extended backwards in a running position |
| **F** | man sits on the floor, not plant | Square head leaning back slightly, body bent with legs forming a seated posture |
| **G** | not Jesus but hugging | Square head angled forward, crouched posture with one arm extended backward |
| **H** | bride with no feet no train | Diamond head, upright and solid body, appearing to sit with a curved leg base |
| **I** | twirl and dance | Diamond head, angular body, standing on one leg with the other extended horizontally like a ballet dancer |
| **J** | podium diamond | Diamond head, solid vertical posture with a pointed angular base, indicating stability |
| **K** | bride at desk with train upward | Diamond head, tall and straight body, with angled arms forming a zigzag pattern |
| **L** | bride with train down at desk | Diamond head in the middle of a zigzag body, with broad shoulders and poised stance |

Table 1: Sample descriptions of all objects in Round 6 from human and `GPT-4o` director, respectively

### A.3 Detailed performance (mean accuracy) table across agents

Table 3 shows the mean accuracy of different agent combinations.

| **Round** | **Human** | **GPT-4o** |
|---|---|---|
| **1** | Man standing holding a plate in front of him | Shape with a diamond on top and a right-angle triangle pointing left, forming an upright figure |
| **2** | Man holding an object facing left | Diamond head and angular arms, resembling a figure standing with one arm extended |
| **3** | Man holding an object to the left | Diamond head upright with broad shoulders, and a small triangular base, appearing to be standing attentively |
| **4** | Man with object facing left | Shape with a diamond head, long straight body, and small triangular foot, standing tall |
| **5** | Man with object facing left | Figure with a square head, arms widely extended upwards, resembling soaring or gliding |
| **6** | Man with object facing left | Diamond head with a straight, tall body, left arm extended forward, resembling a composed stance |

Table 2: Sample descriptions of object D across rounds from human and GPT-4o director, respectively

| **Agent Combination** | **Round** | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| GPT-4o vs. GPT-4o | 0.109 | 0.224 | 0.135 | 0.130 | 0.172 | 0.120 |
| GPT-5-mini vs. GPT-5-mini | 0.194 | 0.125 | 0.139 | 0.083 | 0.111 | 0.125 |
| Claude-3.7-sonnet vs. Claude-3.7-sonnet | 0.181 | 0.278 | 0.153 | 0.167 | 0.167 | 0.111 |
| Claude-sonnet-4 vs. Claude-sonnet-4 | 0.139 | 0.153 | 0.097 | 0.069 | 0.125 | 0.083 |
| Gemini-2.5-flash vs. Gemini-2.5-flash | 0.153 | 0.375 | 0.222 | 0.333 | 0.167 | 0.125 |

Table 3: Mean accuracy across rounds for each agent combination