Which one is banana man? Evaluating vision–language models in multi-turn pragmatic interpretation

Anonymous ACL submission

Abstract

Flexible adaptation to context and shared pragmatic intuitions contribute to smooth human conversation. Iterated reference games—in which players repeatedly pick out novel referents using language—present a test case for agents' ability to perform context-sensitive pragmatic reasoning in multi-turn linguistic environments. We tested humans and vision–language models on their ability to identify the intended meaning of descriptions produced in iterated reference games, varying the provided context in terms of amount, order, and relevance. While humans performed well consistently, the models we evaluated could make use of prior context to interpret humans' referring expressions, but they struggled to build up the relevant context to interpret those expressions effectively. Our results suggest that models lack core skills needed for efficient linguistic collaboration.

1 Introduction

Recent advances in machine learning have produced multi-turn conversational agents, which are a public face of artificial intelligence (OpenAI, 2024). The utility of conversational agents depends on capacities including natural language understanding, world knowledge, and instruction following (Guan et al., 2025). Furthermore, success in multi-turn conversation requires the underlying language model to respond to the user's message appropriately given the preceding context, which in turn requires the model to retain relevant contextual information and use it to interpret new messages.

Such multi-turn interactions are a core feature of human communication: shared conversational history supports a shared system of semantic meaning (Clark, 1996; Geurts, 2024). *Iterated reference games* are a common experimental paradigm for studying this type of communication. In iterated reference games, a describer produces a description of a referent so that a matcher can select the referent from a set of options. Over multiple rounds, conventionalised referring expressions usually emerge from repeated descriptions of the same referent (Clark and Wilkes-Gibbs, 1986). These experiments demonstrate that humans dynamically adapt to their conversational partners, creating ad hoc context- and/or partner-specific meanings and conventions (Clark and Wilkes-Gibbs, 1986; Hawkins et al., 2021; Boyce et al., 2024). Humans can also comprehend conventionalized descriptions from games they were not part of, especially when provided with the original context (Boyce et al., 2025; Schober and Clark, 1989; Hawkins et al., 2023).

Conversations—whether human–human or human–AI—require their participants to use contextual information to understand the meanings of utterances. A system that uses context to understand conversation would show two behavioural signatures: (1) the ability to interpret linguistic meaning pragmatically and (2) sensitivity to contextual information (Hawkins et al., 2023; Boyce et al., 2025; Junker and Zarrieß, 2025; Gul and Artzi, 2024; Hua and Artzi, 2024; Chen et al., 2025b; Sravanthi et al., 2024).

In this work, we employ iterated reference games as a minimal test case to investigate whether state-of-the-art vision–language models (VLMs) reason pragmatically about referring expressions given varying types of prior context, and how their sensitivity to context compares to that of humans.

2 Related work

Model performance in interpreting descriptions from iterated reference games rely on several capacities, including the use of visual information for grounding, tracking and deploying information over multi-turn conversation, pragmatic reasoning, and conversational adaptation.

2.1 Integrating visual information in VLMs

Reference games use images to provide grounded context, but the vision abilities of VLMs are much weaker than their language understanding abilities. Some image information encoded by the vision encoder is not effectively decoded and used by VLMs, leading to poor performance at image classification for lower frequency classes (Zhang et al., 2024), poor performance at identifying shape overlaps or counting shapes (Rahmanzadehgervi et al., 2025), and poor spatial reasoning (Kamath et al., 2023; Liu et al., 2023; Chen et al., 2025a). VLMs also tend to overlook atypical features of images and rely on general world knowledge instead (Vo et al., 2025). Thus, VLMs seem to remain limited in their ability to flexibly integrate relevant visual information (Tong et al., 2026).

2.2 Evaluating multi-turn conversational abilities in language models

Several recent benchmarks, such as MT-Eval (Kwan et al., 2024), MultiChallenge (Sirdeshmukh et al., 2025), Multi-Turn Puzzles (Badola et al., 2025), and MT-PingEval (Eisenstein et al., 2026), have evaluated language models in multi-turn contexts to assess their ability to maintain coherent responses across a conversation and use earlier-provided information to inform later responses. Even contemporary commercial models do poorly on these benchmarks, exhibiting incoherence across responses, and failure to use earlier information when it is relevant but not explicitly noted as such. Models also struggle to efficiently collaborate to establish common ground (Eisenstein et al., 2026). Model behaviour in multi-turn contexts is weaker than in single-shot contexts in part because errors can compound across the interaction; models perform better given correct context than their own previously generated responses (Kwan et al., 2024).

2.3 Assessing pragmatics in VLMs

Pragmatics is a broad term that encompasses language that is sensitive to different types of context, including cultural knowledge, grounded visual context, the knowledge state of the interlocutor, and the recent conversational context (Floyd et al., 2025). Different aspects of pragmatics have been tested on language models (Fried et al., 2023), including single-shot reference games to assess how models understand and generate contextually appropriate descriptions, and to evaluate whether they match human preferences (Ma et al., 2025). Referring tasks can also test interlocutor-specific adjustments when a specific listener has a different visual perspective or a different set of background knowledge and vocabulary (Tang et al., 2024; Takmaz et al., 2023; Bao et al., 2022).

2.4 VLMs and iterated reference games

Iterated reference games represent a difficult test of VLM pragmatics, sometimes incorporating difficult visual context and requiring online adaptation to the conversational partner. Prior work suggests VLMs can do a reasonable job at comprehending human-generated descriptions of naturalistic images that increase in efficiency (Hawkins et al., 2020; Hua and Artzi, 2024; Greco et al., 2023).

Several recent studies have compared human–human dyads to human–AI and AI–AI dyads using more difficult contexts such as tangrams or similar-looking visual stimuli (Jones et al., 2026; Zeng et al., 2026). While frontier models can achieve reasonable accuracy when playing with each other, their dynamics and descriptions do not match the patterns of humans (Imai et al., 2025; Zhao and Conwell, 2025). Human–AI dyads perform poorly, with low accuracy, compared to either AI–AI or human-human dyads.

Training models on the task and giving explicit prompts both lead to somewhat higher performance and superficially more human-like descriptions than base models (Hua and Artzi, 2024; Takmaz et al., 2020; Gul and Artzi, 2024; Chen et al., 2025b; Vaduguru et al., 2025), but no model shows the partner-sensitive convention-formation characteristic of human interaction.

While they did not compare directly to human performance, Wang et al. (2025) finds that VLMs are generally weak at understanding descriptions for iterated reference games, regardless of the amount of earlier context.

3 Methods

All model code is available on Github, and data are available on OSF. AI assistants were used for some coding tasks, but not for any other purpose in this project.

3.1 Dataset

We used an iterated reference game dataset from Boyce et al. (2024). In each game, players saw a

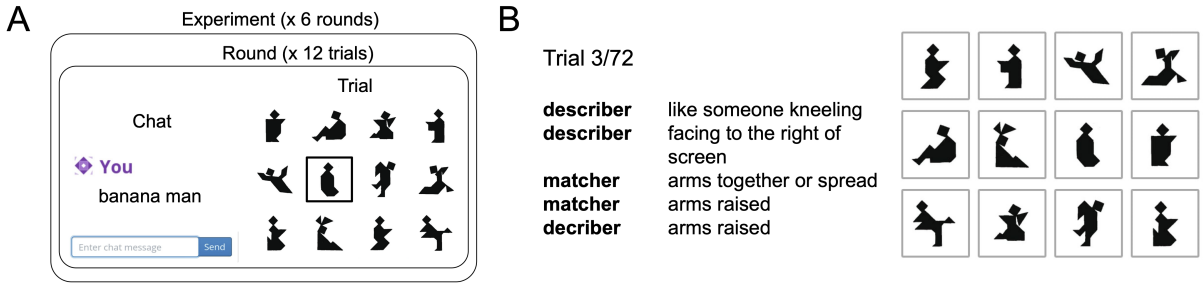

Figure 1: (A) Overview of the experimental structure for the original interactive games in Boyce et al. (2024). (B) User interface for each trial in the experiments with naïve participants.

grid of 12 tangram images (Figure 1A). On each trial, the describer saw one image highlighted and described it to the other players via chatbox so that matchers could each select the target from the 12 options. Each game had 2–6 players, who played 6 rounds of 12 trials each (one trial for each image).

Here we use ten of these games, which were subsequently used by Boyce et al. (2025) to test for context sensitivity in naïve human matchers. Naïve matchers read transcripts of the conversations between the players in the original games and chose what image was being described (Figure 1B). They received feedback on whether they were correct, but were not told the correct answer if they were incorrect. Boyce et al. (2025) collected participants who saw the trials in the original (yoked) order (N = 99) or shuffled (N = 97). We collected additional human data on Prolific in the backward (N = 89) and random (N = 107) conditions (described in Section 3.3) using the same procedure. See Appendix A.1 for more details.

3.2 Experiment setup

Our experiments with VLMs aimed to match the naïve human setup. We focused on evaluating the instruction-tuned versions of five leading open-weights VLMs of different sizes: Qwen 3 VL 32B (Bai et al., 2025), Gemma 3 27B (Gemma Team et al., 2025), Llama 3.2 11B (Llama Team, 2024), Molmo 2 8B (Clark et al., 2026), and Kimi VL A3B (Kimi Team et al., 2025). Models were given a general system prompt describing the task setup, followed by the set of 12 tangram image options labeled A–L presented as a single image. All open models were run with the vLLM inference framework (Kwon et al., 2023). See Appendix A for additional details. Models were then given the context of all preceding trials formatted as a chat history, with the user supplying the text written by participants and the model serving as the assistant providing the target tangram, denoted A–L. Finally, the description for the test trial was provided, and the model log probabilities for the tokens A–L were obtained. We renormalised the log probabilities using a softmax, and treated the probability assigned to the correct target as the model's accuracy.

3.3 Context conditions

To investigate the role of prior contextual information, we considered eight conditions varying the amount and type of in-context trials (Table 1). The conditions varied in their fidelity to the context experienced by human participants in the original iterated reference games.

Our main set of conditions contained trials that were cumulative (i.e., in-context trials had been seen exactly in order), for which we had comparative naïve human data. Three conditions contained trials drawn from a single original game: the *yoked*, *backward*, and *shuffled* conditions presented all the trials in the original game, in the original, reversed, or permuted order respectively. We also included a *random* condition that fully shuffled across games and trial orders.

To further explore how prior context informed VLM interpretation, we considered additional conditions for which the trials were not cumulative. The *ablated* condition presented trials in the original order, but in-context trials containing the same target as the test trial were removed. Two other conditions involved in-context trials sampled from different original games than the test trial. The *other-within* condition drew all in-context trials from a single different original game than the test trial, whereas the *other-across* condition drew in-context trials randomly from all other original games. Finally, we had a baseline condition of *no context*, in which no in-context trials were presented.

Table 1: Conditions of in-context trials seen by models. Columns indicate whether the in-context trials are from the same original game as the test trial, whether the in-context trials are all from the same original game, what the order of the in-context trials is, and whether the tangram in the test trial occurs in the in-context trials.

	Condition	N_{trials}	Original game	Same game	Trial order	Same tangram seen
Cumulative	Yoked	720	Y	Y	Original	Y
	Backward	720	Y	Y	Reversed	Y
	Shuffled	7056	Y	Y	Permuted	Y
	Random	7704	N	N	Permuted	Y
Non-cumulative	Ablated	720	Y	Y	Original	N
	Other-within	6480	N	Y	Original	Y
	Other-across	6480	N	N	Original	Y
	No context	720	N	N	None	N

3.4 Feedback availability

How much do models rely on information provided in feedback? We varied the type of feedback provided in the in-context trials to distinguish retrieval from learning.[1] Our primary type of feedback was *interactive limited feedback*: in-context trials contained models' own prior guesses, and whether those guesses were correct or not. This setup was the most similar to what naïve humans experienced in our experiments. As additional comparisons, we also evaluated models given *human-yoked limited feedback*: in-context trials contained human participants' prior guesses, and whether those guesses were correct or not; and given *full feedback*: in-context trials contained the correct answers, always marked as correct. In full feedback, many trials can be solved by retrieving the most semantically similar in-context trial and repeating the answer (whereas this heuristic may not have been available in the other feedback types since the previous guess may not have been correct).

4 Results

We first compare model and human performance across different trial orders and different amounts of prior context. We then use additional control conditions to more precisely evaluate models.

4.1 Models struggle to reason correctly in iterated reference games

Our first question was whether VLMs were able to perform ad hoc pragmatic reference resolution in our abstract, out-of-domain context. Figure 2A shows the results from all open-weights models in the cumulative conditions under interactive limited feedback, along with comparative human data from the original reference games and from naïve humans. All models demonstrated above-chance accuracy on the task across all conditions. Additionally, model performance began low and improved very slightly across rounds, especially from rounds 1 to 3, suggesting some adaptation to the task. In contrast, naïve humans' performance was relatively high even in the first repetition and showed greater improvement over rounds in the yoked, backward, and shuffled conditions. Performance in the shuffled and random conditions were generally poorer than in the yoked condition, matching naïve humans. Broadly, interpreting pragmatic references to tangram images remains a challenging task for models in comparison to humans.

4.2 Providing additional contextual information improves model performance

We next examined how feedback availability affected model performance. Figure 2B and 2C show the results from all models in the cumulative conditions under human-yoked limited feedback and full feedback respectively, again with comparative human data (although it is important to note that the humans themselves only received interactive limited feedback). Models showed substantial improvement with human-yoked limited feedback, and further improvements still with full feedback. Notably, in the yoked, backward, and shuffled conditions, models asymptoted close to human levels in the human-yoked limited feedback condition, and close to 100% accuracy in the full feedback condition. In the random condition, models mostly remained worse than humans; drawing in-context trials from different games may have limited the amount of relevant contextual information available. These trends suggest that models were able to exploit the correct responses found in the in-context trials to attain better subsequent performance.

[1] Note that the no context condition did not include any in-context trials, so there was no effect of feedback condition.

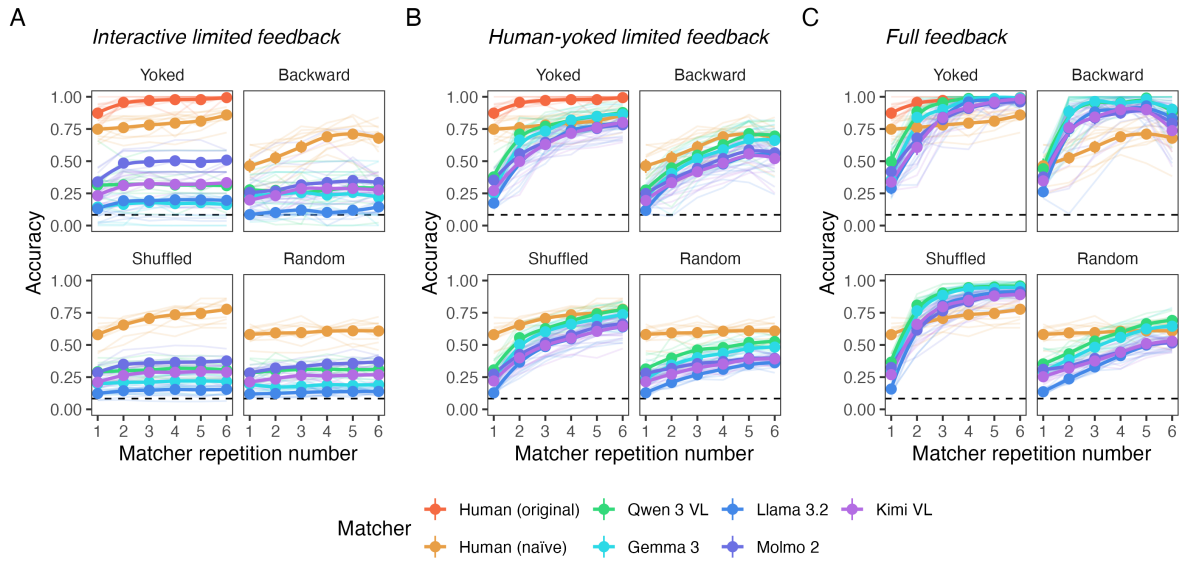

Figure 2: Matcher accuracy across all cumulative conditions and matcher types (both human and model), with best-fit LOESS curves, shown by repetition number as seen by the matcher. Subplots reflect the type of feedback received by model matchers: (A) interactive limited feedback, (B) human-yoked limited feedback, or (C) full feedback. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

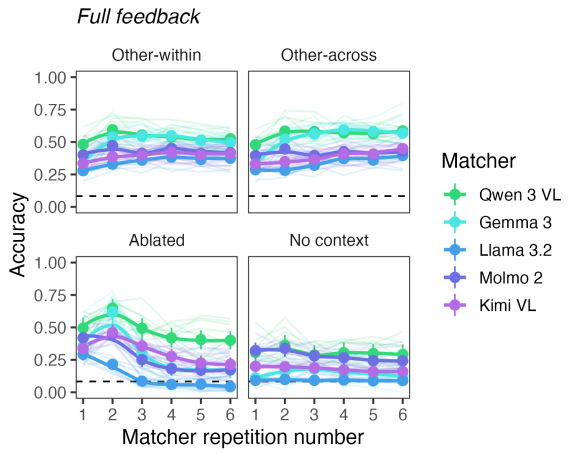

Figure 3: Matcher accuracy in full feedback across all non-cumulative conditions, with best-fit LOESS curves, shown by repetition number as seen by the matcher except for the no context condition, which is shown by the original repetition number. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

4.3 Only relevant contextual information improves model performance

We investigated what types of contextual information help model performance by turning to the non-cumulative conditions under full feedback, as shown in Figure 3. In these conditions, information relevant to the current trial was unlikely to be present in the in-context trials, since previous instances of the specific target–game combination of the test trial were not present. When the conversational history came from a different game or games (other-within and other-across), models had much lower accuracy of 0.3 – 0.6. The lower performance compared to the yoked, shuffled, and backward conditions suggests that model improvement with relevant context was not simply adaptation to the task. Rather, the boost depended on the in-context and test trials coming from the same original game, as conventions reached by one game are not necessarily predicted by the context from other games. Performance was also low in the ablated condition, implying that experience with the test trial tangram itself was crucial to understand a convention, which was not systematically inferrable from other tangram conventions, even from the same game. These results further support the hypothesis that models are mostly retrieving from relevant in-context trials rather than learning over the course of the task. Nonetheless, model performance in the other-within, other-across, and ablated conditions tended to be better than the no context condition, suggesting that at least some learning was occurring.

4.4 Models fail to learn from past mistakes

As an alternative window into the effect of feedback on performance, we visualised matcher accuracy

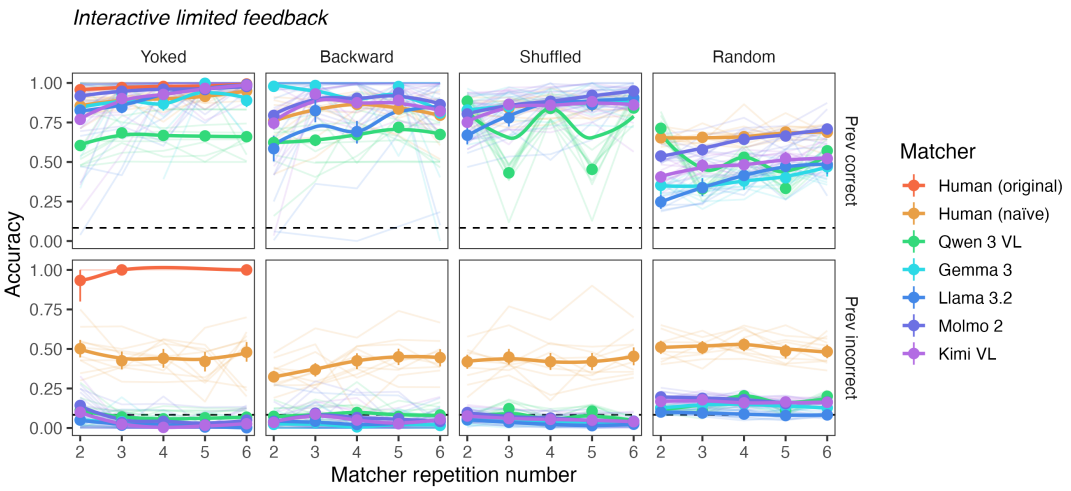

Figure 4: Trial accuracy as a function of previous-repetition accuracy for the same tangram across matchers for the yoked, backward, shuffled, and random conditions under interactive limited feedback. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

from the second repetition onward as a function of whether the matcher selected correctly in the previous repetition of the same tangram, as shown in Figure 4.[2]

In the yoked, shuffled, and backward conditions, when both humans and models answered a previous repetition correctly, they were also very likely to answer the next repetition correctly. However, results were different when matchers answered a previous repetition incorrectly: naïve humans were able to correct their previous mistakes approximately half the time, whereas models were at or worse than chance. Thus, models' overall performance may have been largely driven by the correct answers they happened to attain and retain, rather than learning over the course of the task. Follow-up analyses where models saw the same items repeatedly demonstrate that models did not learn over repetitions, and instead perseverated, selecting the same wrong answer (see Appendix B.6).

4.5 VLMs make limited use of visual information

To determine the extent to which VLMs were using the tangram images to perform pragmatic reasoning, we also ran all models using a prompt that did not include the tangram images. The difference in accuracy between the image–text and the text-only paradigms is shown in Figure 5.

[2]We quantised previous repetition accuracy by determining whether the argmax of the probabilities was the target tangram for models, and by determining whether any matcher selected correctly for the original-game humans; naïve human performance was already binarised.

Models generally performed worse without images than with images in the first repetition, and performance collapsed to chance without context; however, by the third repetition, models were not different whether or not they had images in most conditions. Models remained slightly worse without images in conditions with limited relevant contextual information (random, other-within, other-across, and ablated). Notably Llama 3.2 was better without the image in many conditions, especially in the second repetition. Some models seem to primarily using text information to select the correct match, which corroborates results from Hua and Artzi (2024), showing that when images are shuffled from trial to trial, VLMs exhibit little to no learning over repetitions.

4.6 Models and humans diverge in their responses

To examine whether models generally showed the same response patterns as humans, we compared model and naïve human performance at the trial level under interactive limited feedback (Figure 6). Models were relatively poorly calibrated, with weak correlations between model and human performance (yoked model r = -.07 – .46, human split-half r = .42 [.32, .50]; backward model r = -.20 – .24, human split-half r = .48 [.40, .55]; see Appendix B.2 for more details). Models may not have the same factors affecting their performance as humans, resulting in a lack of trial-wise similarity.

We investigated the divergent human and model results by exploring features of the tangram de-

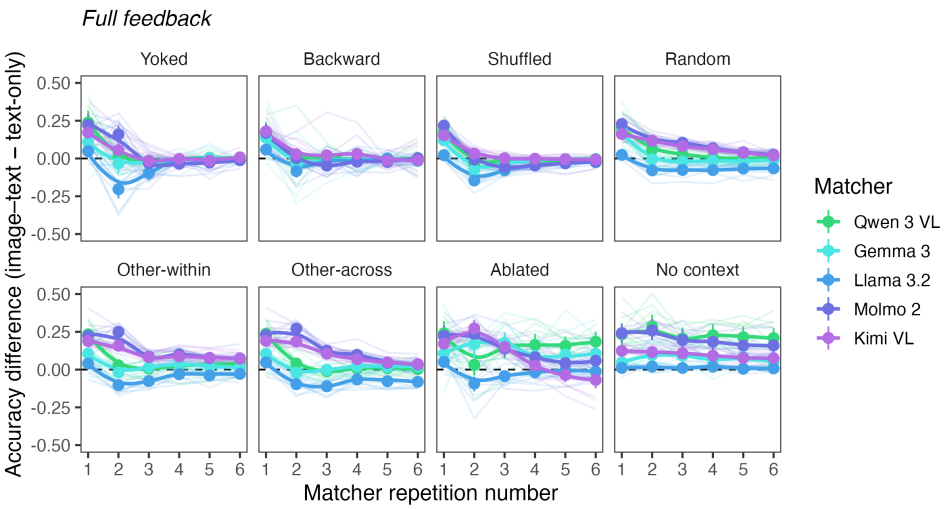

Figure 5: Difference in accuracy between the text-only and image–text paradigms across all conditions and models under full feedback. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate no difference.

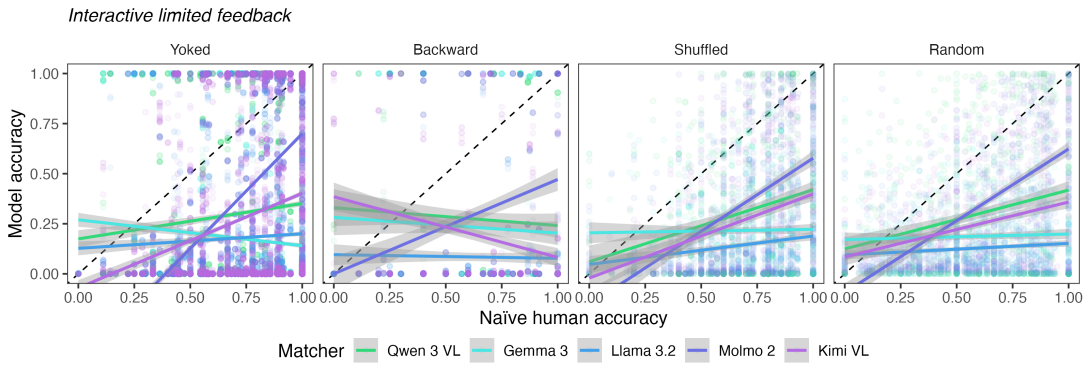

Figure 6: Comparison between naïve human accuracy and model accuracy under interative limited feedback, with best-fit linear regressions. Shaded regions indicate bootstrapped 95% confidence intervals. Dashed lines indicate perfect calibration ($y = x$).

scriptions driving human and model performance. Specifically, we investigated the possible roles of metaphorical and holistic descriptions. (See Appendix B.5 for a complementary approach focusing on lemmas.)

We used LLM-assisted annotation to annotate spans of referential descriptions in each tangram description; see Appendix A.4 for more details. We then ran a quasibinomial regression predicting matcher accuracy from matcher type (human / model, with human as the reference level), the proportion of metaphorical spans, the proportion of holistic spans, and all interactions among these predictors. As shown in Figure 7, there was a significant effect of the proportion of metaphorical spans (b = 1.38 [1.08, 1.68], $p < .001$) as well as model type (b = -0.93 [-1.23, -0.63], $p < .001$). Most interestingly, there was also a significant interaction between the proportion of metaphorical spans and model type (b = -1.04 [-1.38, -0.69], $p < .001$); estimated marginal trends showed that humans showed a strong positive effect of metaphorical spans (b = 1.08 [0.84, 1.31]) whereas models showed a much weaker effect (b = 0.22 [0.10, 0.35]). No effects including the proportion of holistic spans were significant.

Additional analyses on item-level effects and condition-wise correlations are shown in Appendix B.

4.7 Frontier models still fail to reach human-level performance

As an extension, we evaluated three frontier VLMs on the yoked and no context conditions under interactive limited feedback to determine whether better and larger models trained on more data could approach human-level performance. The frontier

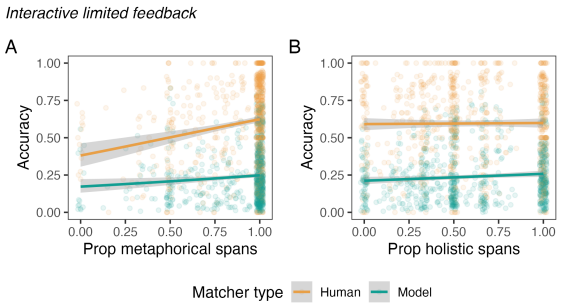

Figure 7: Matcher accuracy as a function of the proportion of metaphorical and holistic spans in descriptions.

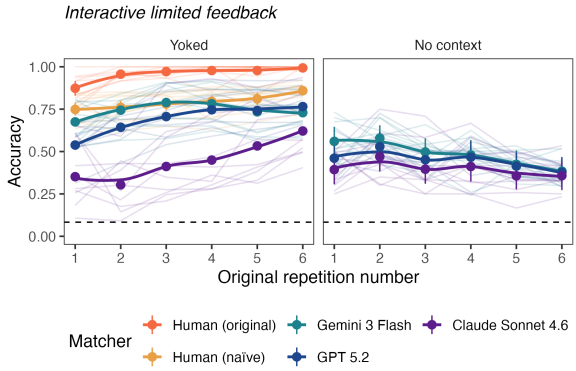

Figure 8: Matcher accuracy for frontier models under interactive limited feedback.

models we evaluated were Gemini 3 Flash (Team, 2025), GPT 5.2 (OpenAI, 2025), and Claude Sonnet 4.6 (Claude Team, 2026). The log probabilities were not fully obtainable for all of these models, so we resampled generations 10× for each trial (see Appendix A.3 for more details). We used the lowest reasoning level to approximate humans' relatively fast response times on this task. Models sometimes failed to return valid responses; we ignored such errors and renormalised excluding them (see Appendix B.7).

As shown in Figure 8, frontier models demonstrated better performance than open-weights models in general, but were still worse than naïve humans on the whole. The no context condition had only slightly worse accuracies than the yoked condition, also suggesting limited learning over the course of the task. Trials from later rounds were also more difficult for frontier models.

5 Discussion

Multi-turn conversation is a key aspect of human communication and a challenge for AI agents. We demonstrated that state-of-the-art open-weights VLMs were able to perform ad hoc pragmatic reasoning across multiple turns of a reference game, but were still substantially worse than humans. Models were sensitive to the presence of relevant contextual information, showing greater improvement when correct answers were present in their context, and limited learning when the correct answers were unavailable. This pattern appeared due to models' failure to learn from past mistakes, whereas humans could. Our results align with previous work on in-context learning, where increasing the number of valid in-context examples improved model performance, especially on difficult tasks (Chen et al., 2023; Agarwal et al., 2024; Jiang et al., 2024).

Furthermore, models displayed different patterns of performance than humans and showed weak trial-wise calibration to human accuracies. These discrepancies were partially explained by the fact that humans were much more accurate with more metaphorical spans, whereas models were only weakly so. The differential sensitivity to quantity and type of context between humans and models could suggest different approaches to the task. For example, models can (theoretically) exactly retrieve previous trial descriptions, whereas humans have more limited working memories and may need to use compression or other strategies to recall appropriate description–referent mappings.

Overall, reference games with abstract referents remain a difficult task for machine learning models, particularly in the few-shot setting (Gul and Artzi, 2024; Chen et al., 2025b). Our results point to the important role of contextual information, but also highlight ways in which current VLMs are limited in bootstrapping their own learning. Further work is needed to more comprehensively characterise the role of prompt engineering, causal and attentional factors underlying context use, and humans' own sensitivity to context. Furthermore, our study focused on interpretation, but generating appropriate descriptions is a yet harder task that is worth exploring. These directions will help us to build language models that are flexible but robust, and that can adapt to a range of tasks and settings as humans do.

Limitations

While we attempted to match the task for humans and models as much as possible, some differences still remained; for example, we ignored model con-

tinuations which were not one of A–L, whereas human participants were explicitly limited in the range of response options (i.e., they were only able to select one of A–L, and could not offer any other response type). Additionally, we tested a limited set of abstract stimuli as a representative example, and focused on English language responses; generalisability to other (e.g., naturalistic) stimuli sets as well as other languages (especially low-resource languages) is left to future work.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-Shot In-Context Learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 76930–76966.

Kartikeya Badola, Jonathan Simon, Arian Hosseini, Sara Marie Mc Carthy, Tsendsuren Munkhdalai, Abhimanyu Goyal, Tomáš Kočiský, Shyam Upadhyay, Bahare Fatemi, and Mehran Kazemi. 2025. Multi-Turn Puzzles: Evaluating Interactive Reasoning and Strategic Dialogue in LLMs. *Preprint*, arXiv:2508.10142.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to Mediate Disparities Towards Pragmatic Communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842, Dublin, Ireland. Association for Computational Linguistics.

Veronica Boyce, Robert Hawkins, Noah D. Goodman, and Michael C. Frank. 2024. Interaction structure constrains the emergence of conventions in group communication.

Veronica Boyce, Ben Prystawski, Alvin Wei Ming Tan, and Michael C. Frank. 2025. Idiosyncratic but not opaque: Linguistic conventions formed in reference games are interpretable by naïve humans and vision–language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How Many Demonstrations Do You Need for In-context Learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore. Association for Computational Linguistics.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025a. Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas. *Preprint*, arXiv:2503.01773.

Zizhao Chen, Mustafa Omer Gul, Yiwei Chen, Gloria Geng, Anne Wu, and Yoav Artzi. 2025b. Retrospective Learning from Interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24580–24606, Vienna, Austria. Association for Computational Linguistics.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, and 2 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *Preprint*, arXiv:2601.10611.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

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

Claude Team. 2026. Introducing Sonnet 4.6.

Jacob Eisenstein, Fantine Huot, Adam Fisch, Jonathan Berant, and Mirella Lapata. 2026. Mt-pingeval: Evaluating multi-turn collaboration with private information games. *arXiv preprint arXiv:2602.24188*.

Sammy Floyd, Olessia Jouravlev, Moshe Poliak, Zachary Mineroff, Edward Gibson, and Evelina Fedorenko. 2025. Three distinct components of pragmatic language use: Social conventions, intonation, and world knowledge–based causal reasoning. *Proceedings of the National Academy of Sciences*, 122(50):e2424400122.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Bart Geurts. 2024. Common Ground in Pragmatics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2024 edition. Metaphysics Research Lab, Stanford University.

Claudio Greco, Diksha Bagade, Dieu-Thu Le, and Raffaella Bernardi. 2023. She adapts to her student: An expert pragmatic speaker tailoring her referring expressions to the Layman listener. *Frontiers in Artificial Intelligence*, 6.

Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian-guang Lou. 2025. Evaluating LLM-based Agents for Multi-Turn Conversations: A Survey. *Preprint*, arXiv:2503.22458.

Mustafa Omer Gul and Yoav Artzi. 2024. CoGen: Learning from Feedback with Coupled Comprehension and Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12966–12982, Miami, Florida, USA. Association for Computational Linguistics.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2021. From partners to populations: A hierarchical Bayesian account of coordination and convention. *Preprint*, arXiv:2104.05857.

Robert D. Hawkins, Minae Kwon, Dorsa Sadigh, and Noah D. Goodman. 2020. Continual adaptation for efficient machine communication. *Preprint*, arXiv:1911.09896.

Robert D. Hawkins, Megumi Sano, Noah D. Goodman, and Judith E. Fan. 2023. Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, 14(1):2199.

Yilun Hua and Yoav Artzi. 2024. Talk Less, Interact Better: Evaluating In-context Conversational Adaptation in Multimodal LLMs. In *First Conference on Language Modeling*.

Saki Imai, Mert Inan, Anthony B Sicilia, and Malihe Alikhani. 2025. Measuring how (not just whether) vlms build common ground. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 441–451.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H. Chen, and Andrew Y. Ng. 2024. Many-Shot In-Context Learning in Multimodal Foundation Models. *Preprint*, arXiv:2405.09798.

Cameron R. Jones, Agnese Lombardi, Kyle Mahowald, and Benjamin K. Bergen. 2026. LLMs and people both learn to form conventions – just not with each other. *Preprint*, arXiv:2602.08208.

Simeon Junker and Sina Zarrieß. 2025. SceneGram: Conceptualizing and Describing Tangrams in Scene Context. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23976–23992, Vienna, Austria. Association for Computational Linguistics.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 76 others. 2025. Kimi-VL Technical Report. *Preprint*, arXiv:2504.07491.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Llama Team. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Ziqiao Ma, Jing Ding, Xuejun Zhang, Dezhi Luo, Jiahe Ding, Sihan Xu, Yuchen Huang, Run Peng, and Joyce Chai. 2025. Vision-Language Models Are Not Pragmatically Competent in Referring Expression Generation. *Preprint*, arXiv:2504.16060.

Kenneth O. McGraw and S. P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

OpenAI. 2024. Introducing ChatGPT. https://openai.com/index/chatgpt/.

OpenAI. 2025. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2025. Vision language models are blind: Failing to translate detailed visual features into words. *Preprint*, arXiv:2407.06581.

Michael F Schober and Herbert H Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *Preprint*, arXiv:2501.17399.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernández. 2023. Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind. *Preprint*, arXiv:2305.19933.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Zineng Tang, Lingjun Mao, and Alane Suhr. 2024. Grounding Language in Multi-Perspective Referential Communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19727–19741, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team. 2025. A new era of intelligence with Gemini 3.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, and 2 others. 2026. Beyond Language Modeling: An Exploration of Multimodal Pretraining. *Preprint*, arXiv:2603.03276.

Saujas Vaduguru, Yilun Hua, Yoav Artzi, and Daniel Fried. 2025. Success and Cost Elicit Convention Formation for Efficient Communication. *Preprint*, arXiv:2510.24023.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision Language Models are Biased. *Preprint*, arXiv:2505.23941.

Zhengxiang Wang, Weiling Li, Panagiotis Kaliosis, Owen Rambow, and Susan E. Brennan. 2025. LVLMs are Bad at Overhearing Human Referential Communication. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou. Association for Computational Linguistics.

Peter Zeng, Weiling Li, Amie Paige, Zhengxiang Wang, Panagiotis Kaliosis, Dimitris Samaras, Gregory Zelinsky, Susan Brennan, and Owen Rambow. 2026. LVLMs and Humans Ground Differently in Referential Communication. *Preprint*, arXiv:2601.19792.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are Visually-Grounded Language Models Bad at Image Classification? *Preprint*, arXiv:2405.18415.

Haoran Zhao and Colin Conwell. 2025. Comparing human and machine communication patterns through a tangram game. In *Proceedings of the Workshop on Data on the Brain & Mind at NeurIPS 2025*.

A Additional methodological details

A.1 Human experiments

Human experiments were approved by [REDACTED] University IRB under protocol number [REDACTED]. Participants were recruited via Prolific and compensated $3.50 for 20 minutes of their time (hourly wage $10.50, which is above the minimum wage in the United States), as well as $0.05 for each correct answer as a bonus (up to $3.60). Participants had to be fluent in English, and not have participated in prior related studies. Before beginning the experiment, participants gave informed consent about participating in the experiment, that they were assured anonymity, and that they could decline participation at any time. All responses were anonymised.

The following instructions were presented to participants at the start of the experiment:

Please read these instructions carefully!

In this task, you will see a **transcript of a conversation** some previous participants had where a **speaker**

identified one of the images to some listeners.

Your goal is to **read the transcript** and **figure out which image** is being described!

You will **read the transcript word-by-word, pressing SPACEBAR to reveal each word.** When you have finished reading, you will **click the image** you think is being described. Then you will **find out if your selection is correct**.

You will see **72 trials**. The image choices will be the same on each trial, but the descriptions may come from different groups of people.

Some of the transcripts may be much shorter or longer than others. Some of the descriptions may be hard to understand – just take a guess if you are not sure.

You will get a bonus of **5 cents** for each image you get right!

Note: These descriptions came from real people and may contain inappropriate or offensive language. If you find a description inappropriate, please let us know via Prolific message or in the exit survey, so we can filter it out in the future.

On each trial, participants saw instructions "Press space to reveal the next word." until they had revealed all the text, when the instructions switched to "Click on the image that was described."

A.2 Prompt and in-context trials

The system prompt for all language model generations included the image in Figure 9 as well as the following text:

You will be presented with a list of messages between people playing a reference game, where the describer has to get the matcher to choose a shape from a set of shapes. Your goal is to guess which of the shapes the describer is trying to get the matcher to choose. The shapes, with their labels, are shown in the image. Please answer with just the letter corresponding to the image you think the describer is trying to get the matcher to choose, and no other text. You will receive feedback telling you whether your choice was correct or incorrect.

In-context trials were presented as if the Assistant had responded with a single letter guess, and the User then responds with whether the guess was correct or incorrect, as shown in Figure 10.

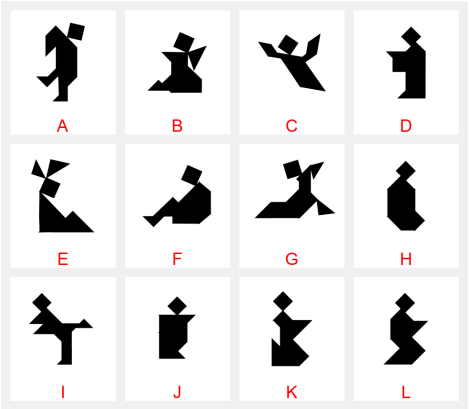

Figure 9: The image that was presented to the vision–language models, containing all 12 tangram shapes with their letter labels.

A.3 Inference details

We ran open-weights models with the vLLM inference framework (Kwon et al., 2023). Models were run in bfloat16 precision on two NVIDIA A40 GPUs. We generated the top 1000 logprobs for the first completion token from these models. In the Llama 3.2 implementation, the model returned the top logprobs for the 10 most likely tokens, then returned -9999 for all other tokens. We called frontier models using the corresponding APIs: the OpenAI Responses API, the Anthropic API, and the Google GenAI API for GPT 5.2, Claude 4.6 Sonnet, and Gemini 3 Flash, respectively. For each of these models, we turned the reasoning level to the lowest possible setting. We sampled ten completions of at most 256 tokens, then took the last capital letter in the completion as the model's choice.

A.4 Span annotation

We prompted Olmo 3 7B Instruct (Team Olmo et al., 2025) to identify spans and tag them along two dimensions: literal / metaphorical, and part-based / holistic. One author manually corrected the spans and tags. Another author independently corrected the spans and tags for 20% of the trials to determine inter-rater reliability. For each trial, we calculated the proportion of spans tagged as metaphorical, and the proportion tagged as holistic. We calculated the single score two-way agreement intraclass correlation coefficient for each dimension (McGraw and Wong, 1996). Inter-rater reliability was excellent for both dimensions, ICC(A, 1) = .954 [.937, .967] for metaphorical, ICC(A, 1) =

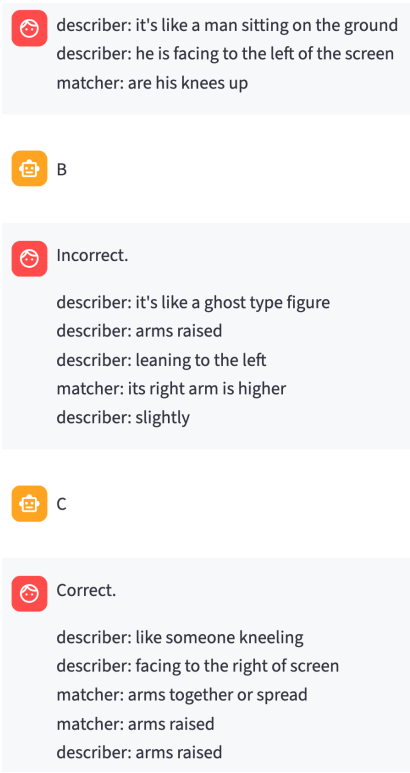

Figure 10: Example of the in-context trials presented to a model, with the User providing the trial text, the Assistant providing a guess, and the User responding with the correctness of the guess.

.824 [.764, .860] for holistic.

B Additional results

B.1 Performance varies by original trial number

Four conditions displayed trials in a different order than the original games: shuffled, backward, random, and no context. To understand the effect of the original trial number, we replotted matcher performance for these conditions by original repetition number in Figure 11.[3] The general worsening performance over rounds across these conditions suggests that increasing conventionalisation in human referring expressions leads to more idiosyncratic expressions (as in Boyce et al., 2025) that are harder to resolve. Interestingly, humans (and some models) showed highest performance in the second repetition (rather than the first); this trend was also observed in the ablated condition results in Figure 3. This phenomenon suggests that there might be a trace of practice effects in human describers, whereby they exhibit increasing systematisation from the first to the second repetition, before showing conventionalisation of expressions.

[3]Note that the results for the no context condition are identical to those in Figure 3 as they were shown by original repetition number in that plot.

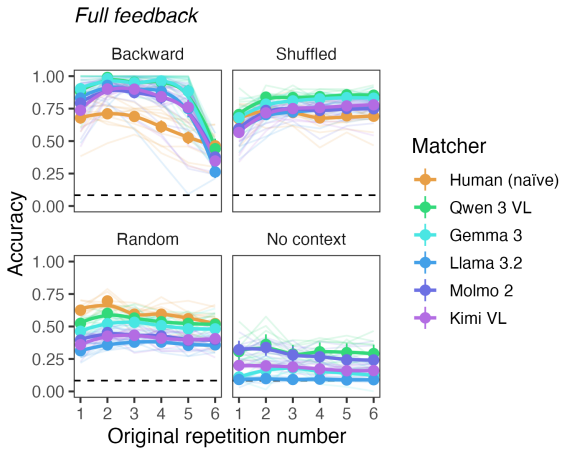

Figure 11: Matcher accuracy for the shuffled, backward, random, and no context conditions across all matcher types (both human and model), with best-fit LOESS curves, shown by repetition number from the original game. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

B.2 Humans are more correlated with each other than with models

Correlation values between models and naïve humans, as well as human split-half correlations, are shown in Table 2. For the shuffled and random conditions, we averaged accuracies for each original trial prior to calculating correlations. Human split-half correlations were computable only for yoked and backward conditions, in which multiple participants saw the same trials in the exact same order. We also note that the human split-half correlations are likely underestimated as some splits did not have the whole set of original trials in each split, and thus had some dropped trials. Overall, models were less similar to humans than humans were to other humans in the yoked and backward conditions; we hypothesise that the same would be true for the shuffled and random conditions if appropriate data were available.

B.3 Items show large variation in performance

Figure 12 shows the item-wise performance for all matchers and conditions. Both naïve humans and models display large item-wise variation, with the magnitude of this variation sometimes exceeding that due to repetition number, matcher, or condition,

Table 2: Human split-half correlations and model–human correlations for the yoked, shuffled, backward, and random conditions. Human split-half correlations show the mean and 95% confidence interval over 1000 random splits.

	r					
Condition	Human split-half	Qwen 2.5 VL	Gemma 3	Llama 3.2	Molmo 2	Kimi VL
Yoked	.42 [.32, .50]	.08	-.07	.03	.46	.19
Backward	.48 [.40, .55]	-.08	.02	-.00	.24	-.20
Shuffled	—	.23	.04	.18	.47	.31
Random	—	.23	.03	.12	.60	.29

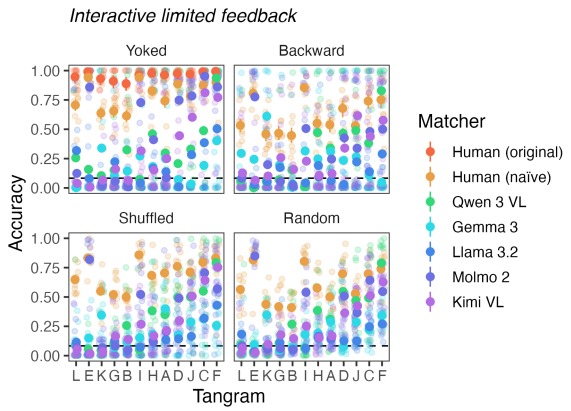

Figure 12: Item-wise variation in accuracy across matchers and conditions. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

replicating the item-wise variation found by Boyce et al. (2025). There appears to be some shared variance across models, although each model also has particular idiosyncrasies and biases (e.g., Molmo 2 appears to be especially accurate for tangram E compared to other models).

B.4 Conditions pattern with each other systematically

Drawing inspiration from representational similarity analysis (Kriegeskorte et al., 2008), we conducted a correlation analysis over the confusion matrices for each model across conditions under full feedback. We constructed confusion matrices for each model–condition combination by taking the mean of the probability distributions across trials for each target tangram. We then calculated the Pearson's correlation among the confusion matrices for all conditions. The resultant correlograms are shown in Figure 13. These correlograms demonstrate relatively high correlations across conditions for all models (all $r > .85$), except for the ablated condition (r = .25 – .90) and the no context condition (r = -.01 – .85). Furthermore, all models show a similar pattern of correlations, with the yoked, shuffled, and backward conditions clustering together, the other-within, other-across, and random conditions clustering together, and the ablated and no context conditions being the most distinct from the other conditions. This pattern of results partly reflects the differences in overall performance across the different conditions, but also supports the idea that relevance is a crucial organising dimension of the content of contextual information.

B.5 Lemmas are related to model–human discrepancies

To understand the discrepancies between models and humans, we used model and naïve human responses in the random condition under interactive limited feedback, in which there was limited relevant context, although we note that our method could be applied to other conditions as well. We estimated the discrepancy between model and human performance by calculating the difference between model and human accuracy for each unique trial, averaging across all models and runs. Then, we lemmatised the message texts in each trial using UDPipe (Straka, 2018), and estimated the means and 95% confidence intervals of the model–human discrepancy for each unique lemma across all trials in which that lemma occurred. We recentred these discrepancy scores by subtracting the grand mean in discrepancy across all trials (-0.36), and determined the significance of the centred discrepancy (CD) estimates by calculating whether the 95% confidence intervals contained zero. These CD values reflected whether models were better or worse than average, in comparison to humans. We also filtered down to lemmas that occurred in at least 4 games and in at least 20 trials (i.e., their discrepancies were not due to a small number of idiosyncratic games). The CD values for selected lemmas are shown in Figure 14.

Manual inspection of the resulting estimates reflects the role of metaphors in model–human dis-

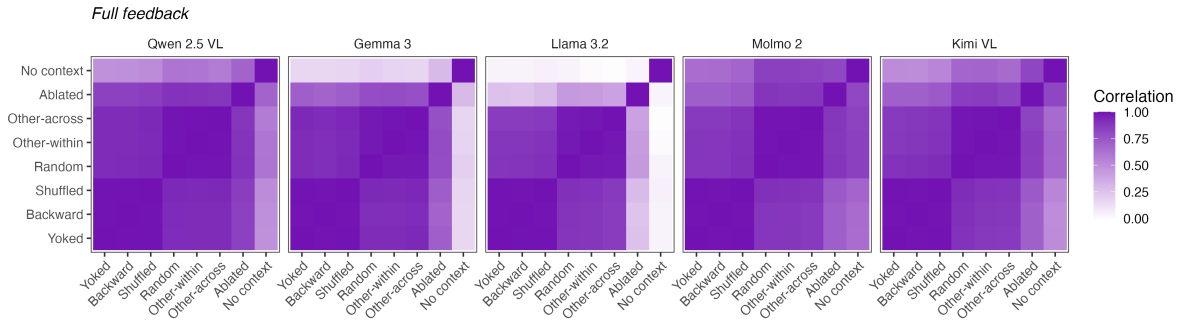

Figure 13: Correlation in confusion matrices across conditions for all models.

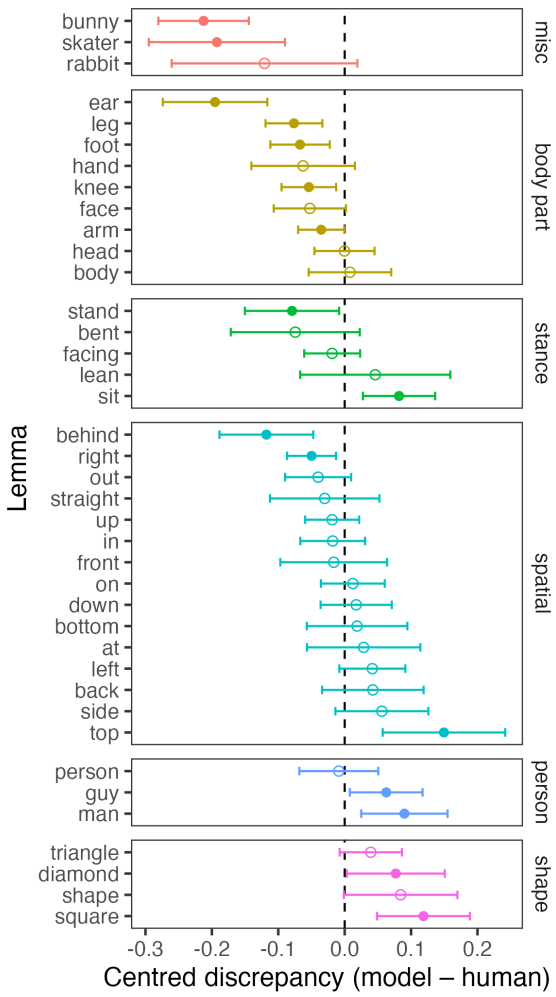

Figure 14: Centred discrepancy in model versus human accuracies in the random condition for selected lemmas, categorised by semantic categories. Error bars indicate 95% confidence intervals. Dashed line indicates zero (i.e., discrepancy equal to the grand mean). Empty/filled points indicate whether the confidence intervals do/do not include zero respectively.

crepancy. Lemmas relating to literal features of the tangram such as geometric shapes (e.g., "triangle", "diamond", "square") and spatial locations (e.g., "left", "side", "top") had CD values that were positive or not different than zero. Conversely, words which required more metaphorical interpretations, including "skater" and "bunny", as well as interpretations of tangram parts as body parts (e.g., "ear", "leg", "foot"), had negative CD values. Interestingly, some metaphorical words including "guy" and "man" had positive CD values; it is possible that these words were generic enough that they were highly frequent and could be found in many combinations with other lemmas, including with other literal descriptors. These results also converge with our span-based analysis of model and human performance in Section 4.6.

B.6 Given additional practice, models still fail to improve

The change in models' performance over repetitions are a function of both the change in the human descriptions of the tangrams, as well as models getting better at the task of interpreting descriptions. To isolate the effect of practice over trials, we tested models on additional conditions in which we repeated the descriptions from a single round of the reference game; thus, the descriptions were held constant, and models had multiple chances to learn from a specific set of descriptions. We tested two conditions drawing from descriptions that were originally in the first round and the last round, under interactive limited feedback.

As shown in Figure 15A, models performed extremely poorly in the practice conditions, revealing that they do not learn adequately with practice. After a small initial boost from the first to the second rounds, performance plateaus at <0.4. We then checked whether models' poor performance was

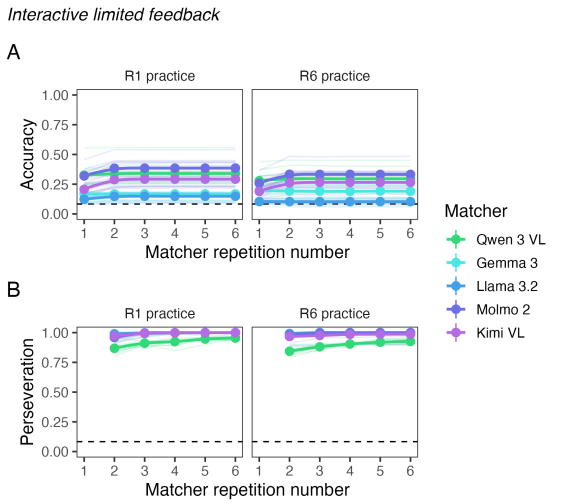

Figure 15: (A) Matcher accuracies and (B) likelihood of perseverating on their previous answer for practice conditions. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

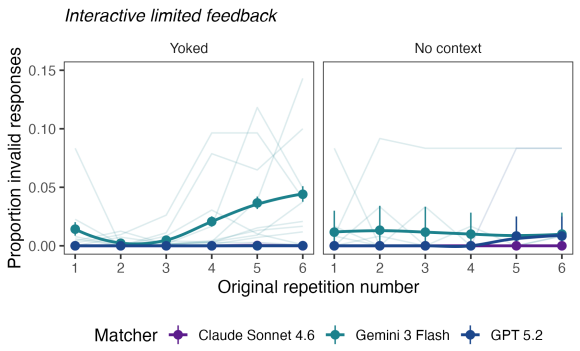

Figure 16: Proportion of invalid responses for frontier models under interactive limited feedback.

due to perseverating on their own previous incorrect answer; to do so, we examined whether the response with the highest log probability also had the highest log probability in the previous repetition (including ties). Figure 15B shows that this trend was indeed the case, with all models perseverating almost 100% of the time (except for Qwen 3, which still perseverated for >80% of trials). In fact, perseveration went up slightly over repetitions, suggesting that instead of learning, models became more entrenched in incorrect responses. Together with results from Section 4.4, these patterns show that invalid in-context information significantly worsens model performance.

B.7 Frontier models sometimes return invalid responses

As mentioned in A.3, for frontier models, we took the last capital letter in the completion as the model's choice. However, on occasion, models would fail to return a single capital letter within 256 tokens, leading to invalid responses. These were ignored when calculating response probabilities and accuracies. Figure 16 shows the rate of invalid responses across our frontier models. Gemini 3 Flash had the highest rate of invalid responses, which also increased over rounds 2 to 6 in the yoked condition. This pattern of responses may explain in part why we observed non-monotonic change in response accuracy in Figure 8. Claude Sonnet 4.6 did not return any invalid responses, and neither did GPT 5.2 in the yoked condition, although it did return a few invalid responses in the no context condition, specifically for trials originally in repetition 5 or 6. More careful qualitative analyses of these invalid responses may help to diagnose the observed failures.