Context informs pragmatic interpretation in vision–language models

Alvin W.M. Tan*, Ben Prystawski*, Veronica Boyce, Michael C. Frank
Department of Psychology
Stanford University
Stanford, CA 94305
tanawm@stanford.edu

Abstract

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 trials from iterated reference games, varying the given context in terms of amount, order, and relevance. Without relevant context, models were above chance but substantially worse than humans. However, with relevant context, model performance increased dramatically over trials. Few-shot reference games with abstract referents remain a difficult task for machine learning models.

1 Introduction

Recent advances in machine learning have produced multi-turn conversational agents, which are a public face of artificial intelligence [1]. The utility of conversational agents depends on capacities including natural language understanding, world knowledge, and instruction following [2]. 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 language 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 [3, 4]. *Iterated reference games* are a common experimental paradigm for studying this type of communication. In these paradigms, a describer produces a description of a referent such that a matcher can correctly select the referent from a set of options. Games repeat over multiple rounds, typically resulting in the emergence of conventionalised referring expressions over repeated descriptions of the same referent [5]. These experiments demonstrate that humans dynamically adapt to their conversational partners, creating ad hoc context- and/or partner-specific meanings and conventions [5–7].

Conversations—whether human–human or human–AI—require their participants to use contextual information to understand the meanings of utterances. A system with the ability to do this would show two behavioural signatures: (1) the ability to interpret linguistic meaning pragmatically and (2) sensitivity to contextual information [8–14]. In this work, we employ iterated reference games as a minimal test case to investigate whether state-of-the-art open 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. Iterated reference games have previously been used to study AI systems for both generation and comprehension of descriptions, but this domain continues to prove challenging for AI systems, especially for generation with abstract images [11, 15, 16].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First Workshop on Interpreting Cognition in Deep Learning Models (CogInterp).

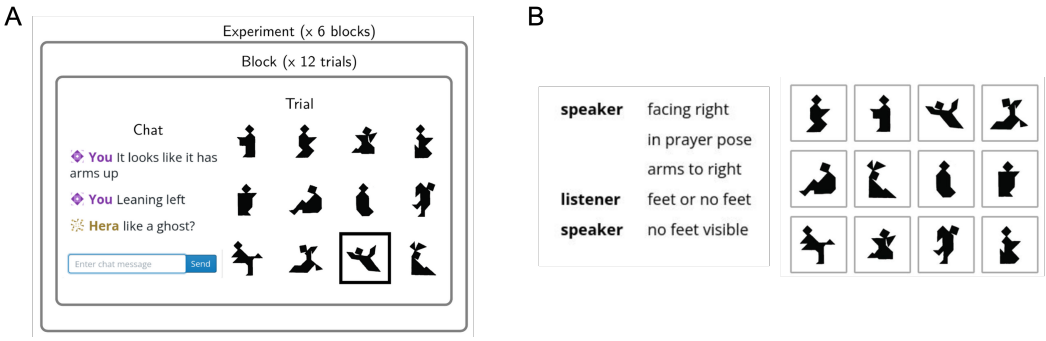

Figure 1: A. Overview of the experimental structure for the original interactive games in Boyce et al. [7]. B. User interface for each trial in the experiments with naïve participants.

2 Methods

All model data and code are available on Github at `https://github.com/benpry/vlm-tg-context/tree/coginterp`.

2.1 Dataset

We used an iterated reference game dataset from Boyce et al. [7]. In each game, players saw a grid of 12 tangram images (Figure 1A). On each trial, one player (the describer), saw one image highlighted. The describer described a target image to the other players via chatbox so that the matchers could each select the target image 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. [9] to test for context sensitivity in naïve human matchers. Naïve matchers read transcripts of the conversations between the players in the original games and then made a selection of what image was being describer (Figure 1B). Naïve matchers received feedback on whether they were correct, but were not told the correct answer if their response was incorrect. Each participant saw all trials from one game, in an order that varied based on condition. Boyce et al. [9] collected participants in the yoked condition (N = 99) and the shuffled condition (N = 97). We collected additional human data on Prolific in the backward (N = 89) and random (N = 107) conditions (described in Section 2.3) using the same procedure.

2.2 Experiment setup

Our experiments with VLMs aimed to match the naïve human setup. We evaluated the instruction-tuned versions of four leading open-weights VLMs of different sizes: Qwen 2.5 VL 32B [17], Gemma 3 27B [18], Llama 3.2 11B [19], and Kimi VL A3B [20]. Models were given a general system prompt describing the task setup, followed by the set of 12 tangram image options labeled A through L presented as a single image. The exact prompt and image are included in Appendix A. 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 through L; note that models were given the correct target (unlike humans). Finally, the description for the test trial was provided, and the model log probabilities for the tokens A through 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.

2.3 Context conditions

Do VLMs make use of context to the same extent, and in the same way, as humans? We considered eight conditions varying the amount and type of in-context trials seen by the model (Table 1).

The conditions varied in their fidelity to the context experienced by actual human participants in the original iterated reference games. Four conditions contained trials drawn from a single original

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
Yoked	720	Y	Y	Original	Y
Shuffled	7056	Y	Y	Permuted	Y
Backward	720	Y	Y	Reversed	Y
Ablated	720	Y	Y	Original	N
Other-within	6480	N	Y	Original	Y
Other-across	6480	N	N	Original	Y
Random	7704	N	N	Permuted	Y
No context	720	N	N	None	N

game. The *yoked*, *shuffled*, and *backward* conditions presented all the trials in the original game, in the original order, a permuted order, or the reversed order respectively. The *ablated* condition presented the trials in the original order, but in-context trials containing the same target tangram 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. We also included a *random* condition that fully shuffled across games and trial orders. Finally, we had a baseline condition of *no context*, in which no in-context trials were presented.

3 Results

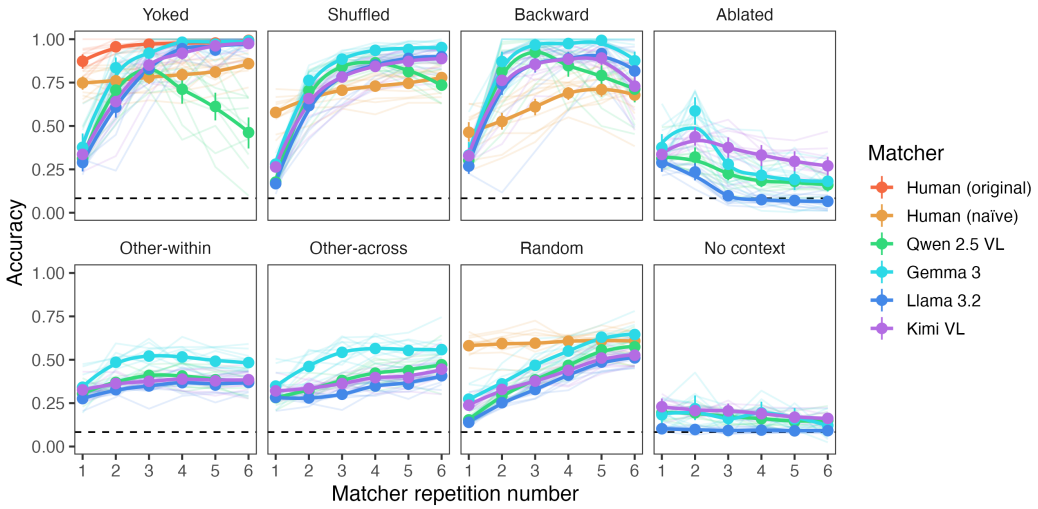

Figure 2: Matcher accuracy across all conditions and matcher types (both human and model), with best-fit LOESS curves, shown by repetition number as seen by the matcher, except for the no context condition where repetition number is from the original game. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

Our first question was whether models are able to perform ad hoc pragmatic reference resolution in this abstract, out-of-domain context. Results from all models and conditions, along with comparative human data from the original reference games [7] and from naïve humans [9], are shown in Figure 2. All models demonstrated above-chance accuracy on the task across all conditions, but without context, models were only slightly above chance. In the random condition, model performance improved as the number of in-context examples increased, suggesting some adaptation to

the task. In contrast, naïve humans' performance was relatively high even in the first repetition and remained steady over time. In the absence of conversational history, models appear to have much worse "intuition" for interpreting pragmatic references to tangram images than humans.

Our second question was how models would perform with relevant conversational history. The yoked, shuffled, and backward conditions show how model performance was affected by the *amount* and *order* of contextual information. Performance in the shuffled condition was generally poorer than in the yoked condition, matching the trend of naïve humans. Models were more accurate in the backward condition than in the shuffled condition, unlike humans, suggesting that working backwards from conventions to earlier and less conventionalised expressions may be easier for models than for humans. While humans outperformed models on the first trials seen, models were able to rapidly exploit the conversational history to achieve $\sim$0.8 accuracy. These disparities suggest that VLMs may be worse at zero- or few-shot novel task performance than humans, but are able to learn relatively quickly with sufficient examples.

Within-game context was critical to models. When we varied the *relevance* of the context by drawing the conversational history from a different game or games (other-within and other-across), models had much lower accuracy of 0.3 – 0.5. This lower performance compared to yoked, shuffled, and backwards suggests that model improvement with relevant context is not simply adaptation to the task. Rather, the boost depends on the context and the test trial coming from the same original game. The convention reached by one game is 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 is crucial to understand a convention, which is not systematically inferrable from the other tangram conventions, even from the same game.

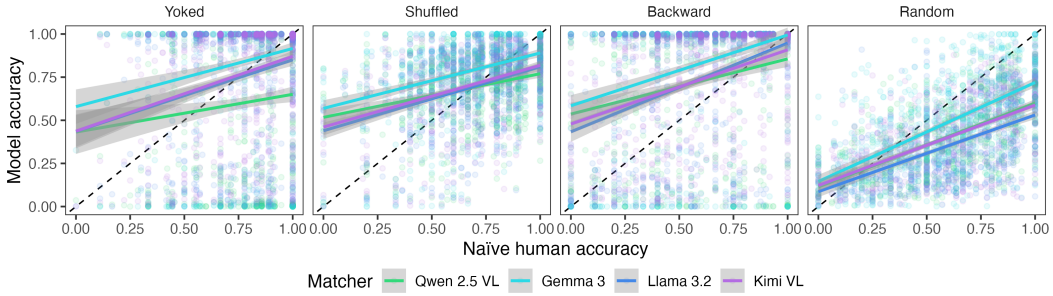

Figure 3: Comparison between naïve human accuracy and model accuracy, with best-fit linear regressions. Shaded regions indicate bootstrapped 95% confidence intervals. Dashed lines indicate perfect calibration ($y = x$).

To determine whether models generally show the same pattern of difficulties as humans, we also compared model and naïve human performance at the trial level (Figure 3). Models were relatively poorly calibrated, with weak correlations between model and human performance (yoked model r = .10 – .27, human split-half r = .42 [.32, .50]; backward model r = .23 – .40, human split-half r = .48 [.40, .56]; 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.

Across models, Gemma 3 was the best performing model across most conditions. Interestingly, Qwen 2.5's performance diverged from the other VLMs in the yoked, shuffled, and backward conditions, performing much worse in later rounds; we speculate this finding is due to the longer context lengths in later rounds, which Qwen 2.5 may be worse at handling.

Additional analyses on item-level effects, condition-wise correlations, and errors are shown in Appendix B, and additional results from a no-image, text-only paradigm are shown in Appendix C.

4 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. We found that models were sensitive to the amount,

order, and relevance of contextual information, showing rapid improvement when in-context examples came from the same source game and included the target image. This performance reflects a capacity for flexible meaning construction in a range of different contexts.

Nonetheless, models displayed different patterns of performance than humans and showed weak trial-wise calibration to human accuracies. Our results align with previous work on in-context learning, where increasing the number of in-context examples improved model performance, especially on difficult tasks [21–23]. While humans do show practice effects, these effects are typically much smaller. 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.

While we tested humans and models on the same stimuli and some of the same stimulus orders, the models received more informative feedback than humans. Models received the correct answers for the in-context examples, while humans only learnt if their selections were correct; this disparity may have contributed to differences in performance (see Section B.4). Additionally, we tested a limited set of abstract stimuli as a representative example; generalisability to other (e.g., naturalistic) stimuli sets is left to future work.

Overall, reference games with abstract referents remain a difficult task for machine learning models, particularly in the few-shot setting [11, 13]. Our results point to the important role of contextual information, but also highlight ways in which current VLMs are overly sensitive to context. 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.

References

[1] OpenAI, *Introducing ChatGPT*, https://openai.com/index/chatgpt/, Mar. 2024.

[2] S. Guan, H. Xiong, J. Wang, J. Bian, B. Zhu, and J.-g. Lou, *Evaluating LLM-based Agents for Multi-Turn Conversations: A Survey*, Mar. 2025. DOI: 10.48550/arXiv.2503.22458. arXiv: 2503.22458 [cs].

[3] H. H. Clark, *Using Language*. Cambridge: Cambridge University Press, 1996, ISBN: 978-0-511-62053-9.

[4] B. Geurts, "Common Ground in Pragmatics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Winter 2024, Metaphysics Research Lab, Stanford University, 2024.

[5] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, no. 1, pp. 1–39, Feb. 1986. DOI: 10.1016/0010-0277(86)90010-7.

[6] R. D. Hawkins *et al.*, *From partners to populations: A hierarchical Bayesian account of coordination and convention*, Dec. 2021. DOI: 10.48550/arXiv.2104.05857. arXiv: 2104.05857 [cs].

[7] V. Boyce, R. Hawkins, N. D. Goodman, and M. C. Frank, *Interaction structure constrains the emergence of conventions in group communication*, 2024.

[8] R. D. Hawkins, M. Sano, N. D. Goodman, and J. E. Fan, "Visual resemblance and interaction history jointly constrain pictorial meaning," *Nature Communications*, vol. 14, no. 1, p. 2199, Apr. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-37737-w.

[9] V. Boyce, B. Prystawski, A. W. M. Tan, and M. C. Frank, "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*, vol. 47, 2025.

[10] S. Junker and S. ZarrieSS, "SceneGram: Conceptualizing and Describing Tangrams in Scene Context," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 23 976–23 992, ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.1229.

[11] M. O. Gul and Y. Artzi, "CoGen: Learning from Feedback with Coupled Comprehension and Generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 966–12 982. DOI: 10.18653/v1/2024.emnlp-main.721.

[12] Y. Hua and Y. Artzi, "Talk Less, Interact Better: Evaluating In-context Conversational Adaptation in Multimodal LLMs," in *First Conference on Language Modeling*, Aug. 2024.

[13] Z. Chen, M. O. Gul, Y. Chen, G. Geng, A. Wu, and Y. Artzi, "Retrospective Learning from Interactions," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 24 580–24 606, ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.1200.

[14] S. Sravanthi, M. Doshi, P. Tankala, R. Murthy, R. Dabre, and P. Bhattacharyya, "PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12 075–12 097. DOI: 10.18653/v1/2024.findings-acl.719.

[15] A. Ji *et al.*, "Abstract Visual Reasoning with Tangram Shapes," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 582–601. DOI: 10.18653/v1/2022.emnlp-main.38.

[16] Z. Wang, W. Li, P. Kaliosis, O. Rambow, and S. E. Brennan, "LVLMs are Bad at Overhearing Human Referential Communication," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou: Association for Computational Linguistics, Sep. 2025. DOI: 10.48550/arXiv.2509.11514.

[17] S. Bai *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[18] Gemma Team *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[19] Llama Team, *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/, 2024.

[20] Kimi Team *et al.*, *Kimi-VL Technical Report*, Jun. 2025. DOI: 10.48550/arXiv.2504.07491. arXiv: 2504.07491 [cs].

[21] J. Chen, L. Chen, C. Zhu, and T. Zhou, "How Many Demonstrations Do You Need for In-context Learning?" In *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11 149–11 159. DOI: 10.18653/v1/2023.findings-emnlp.745.

[22] R. Agarwal *et al.*, "Many-Shot In-Context Learning," in *Advances in Neural Information Processing Systems*, vol. 37, Dec. 2024, pp. 76 930–76 966.

[23] Y. Jiang, J. Irvin, J. H. Wang, M. A. Chaudhry, J. H. Chen, and A. Y. Ng, *Many-Shot In-Context Learning in Multimodal Foundation Models*, Oct. 2024. DOI: 10.48550/arXiv.2405.09798. arXiv: 2405.09798 [cs].

[24] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis - connecting the branches of systems neuroscience," *Frontiers in Systems Neuroscience*, vol. 2, 2008, ISSN: 1662-5137.

[25] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen, *Vision language models are blind: Failing to translate detailed visual features into words*, Mar. 2025. DOI: 10.48550/arXiv.2407.06581. arXiv: 2407.06581 [cs].

[26] S. Chen *et al.*, *Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas*, Mar. 2025. DOI: 10.48550/arXiv.2503.01773. arXiv: 2503.01773 [cs].

[27] A. Vo, K.-N. Nguyen, M. R. Taesiri, V. T. Dang, A. T. Nguyen, and D. Kim, *Vision Language Models are Biased*, May 2025. DOI: 10.48550/arXiv.2505.23941. arXiv: 2505.23941 [cs].

A Additional methodological details

A.1 Prompt

The following text was used as the system prompt for all language model generations.

```
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.
```

The image shown in Figure 4 was included with this prompt.

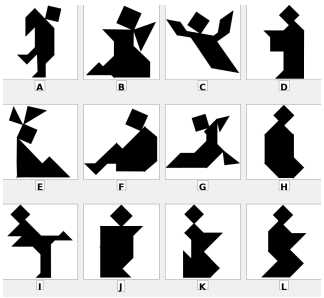

Figure 4: The image that was presented to the vision-language models, containing all 12 tangram shapes with their letter labels.

B Additional analyses

B.1 Role of 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 5; note that the results for the no context condition are identical to those in Figure 2 as they were shown by original repetition number in that plot. The general worsening performance over rounds across these conditions suggests that increasing conventionalisation in human referring expressions leads to more idiosyncratic expressions (as in [9]) that are harder to resolve. Interestingly, humans (and some models) showed highest performance in the second repetition (rather than the first). 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 exhibiting conventionalisation of expressions.

B.2 Model–human comparisons

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.

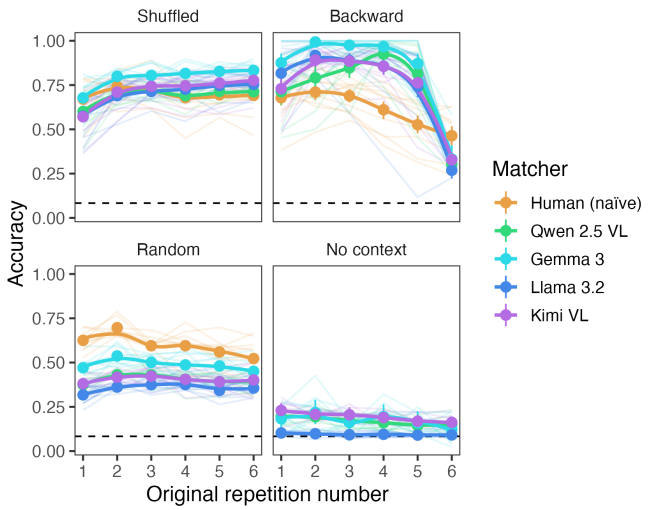

Figure 5: Matcher accuracy for the shuffled, backward, random, and no context conditions across all matcher types (both human and model), with best-fit LOESS curves, shown by repetition number from the original game. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

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	Kimi VL
Yoked	.42 [.32, .50]	.10	.20	.25	.27
Shuffled	—	.28	.34	.38	.37
Backward	.48 [.40, .56]	.23	.31	.40	.35
Random	—	.61	.58	.55	.57

B.3 Item-wise variation

Figure 6 shows the item-wise performance for all matchers and conditions. Figure 4 provides the grid of tangrams shown to the models, with correspondence between tangrams and letter labels.

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, replicating the item-wise variation found by Boyce et al. [9]. Item-wise variation appears to be largest for the other-within and other-across conditions, suggesting that some tangrams may be relatively more nameable or distinguishable than others, allowing for higher accuracy even without coherent context. There appears to be some shared variance across models, such that the mean accuracy for each item is somewhat consistent across models for each condition, although each model also has particular idiosyncrasies and biases (e.g., Gemma 3 appears to be especially accurate for tangrams A and E compared to other models).

B.4 Learning analyses

To investigate the potential role of feedback, we visualised matcher accuracy 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 7. 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.

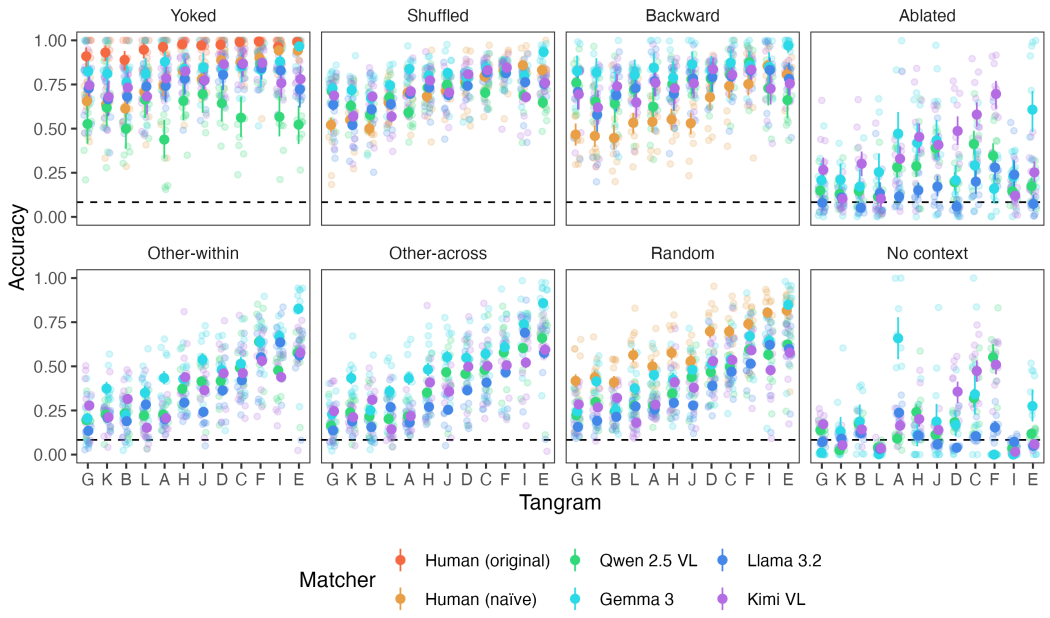

Figure 6: Item-wise variation in accuracy across matchers and conditions. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

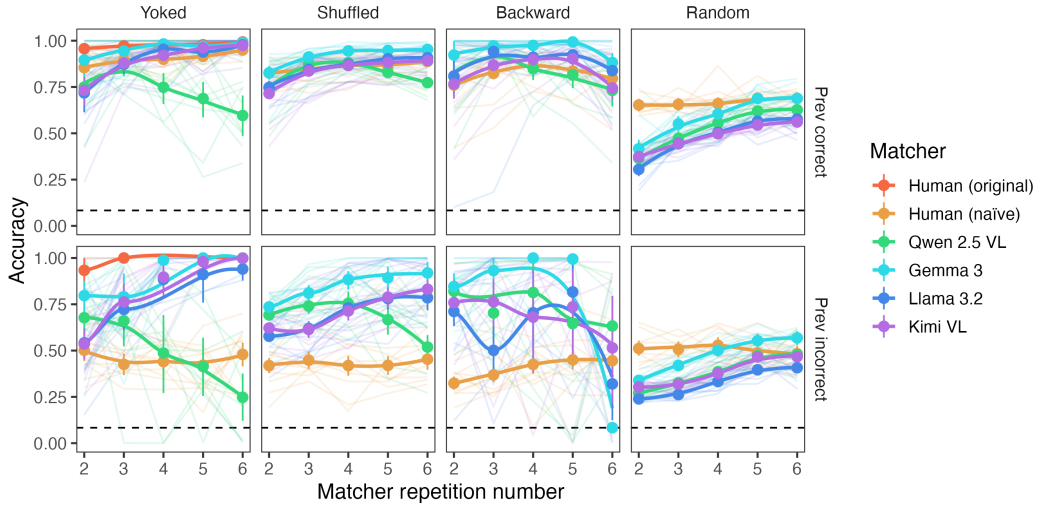

Figure 7: Trial accuracy as a function of previous-repetition accuracy for the same tangram across matchers for the yoked, shuffled, backward, and random conditions. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

The random condition replicated the overall accuracy curves in Figure 2; since the prior context did not provide any relevant cues as for subsequent test trials, both models and humans performed similarly regardless of whether the previous trial was correct or not.

In the yoked, shuffled, and backward conditions, for both humans and models, when matchers answered a previous repetition correctly, there were also very likely to answer the next repetition correctly, with accuracy > 0.7 (except for Qwen 2.5 in the yoked condition). However, the pattern of results was different when matchers answered a previous repetition incorrectly. In this case, models were generally able to learn from previous mistakes in the yoked and shuffled condition (again, ex-

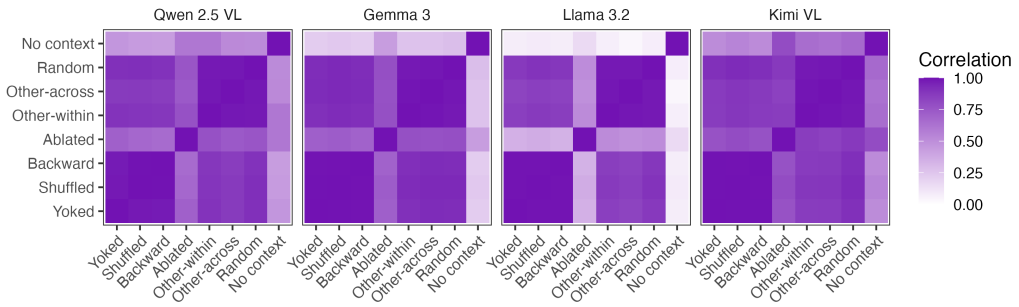

Figure 8: Correlation in confusion matrices across conditions for all models.

cept for Qwen 2.5), whereas naïve humans were at about 0.5 accuracy—in other words, they were able to correct their previous mistakes only approximately half the time.

Some of this discrepancy may have been due to the fact that models received the correct target for all previous trials, whereas naïve humans only received feedback about whether they had selected correctly (i.e., they knew the correct tangram only when they had selected it correctly themselves). Future work can more fully characterise the role of feedback by either allowing naïve humans to know the correct tangram if they answered incorrectly, or by presenting models with only selection correctness feedback. We note also that the construct of "learning" is slightly disanalogous between humans and models, since humans experienced trials sequentially whereas models did not. Evaluating models in a true multi-turn setting would help to more closely align the experiment setups.

B.5 Correlation analyses

Drawing inspiration from representational similarity analysis [24], we conducted a correlation analysis over the confusion matrices for each model across conditions. 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 8. These correlograms demonstrate relatively high correlations across conditions for all models (all $r > .83$), except for the ablated condition (r = .16 – .87) and the no context condition (r = .04 – .80). 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.6 Error analyses

We conducted an error analysis to investigate possible factors driving differences between model and human performance. We used model and naïve human responses in the random condition, 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, 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 (-0.186), 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 9.

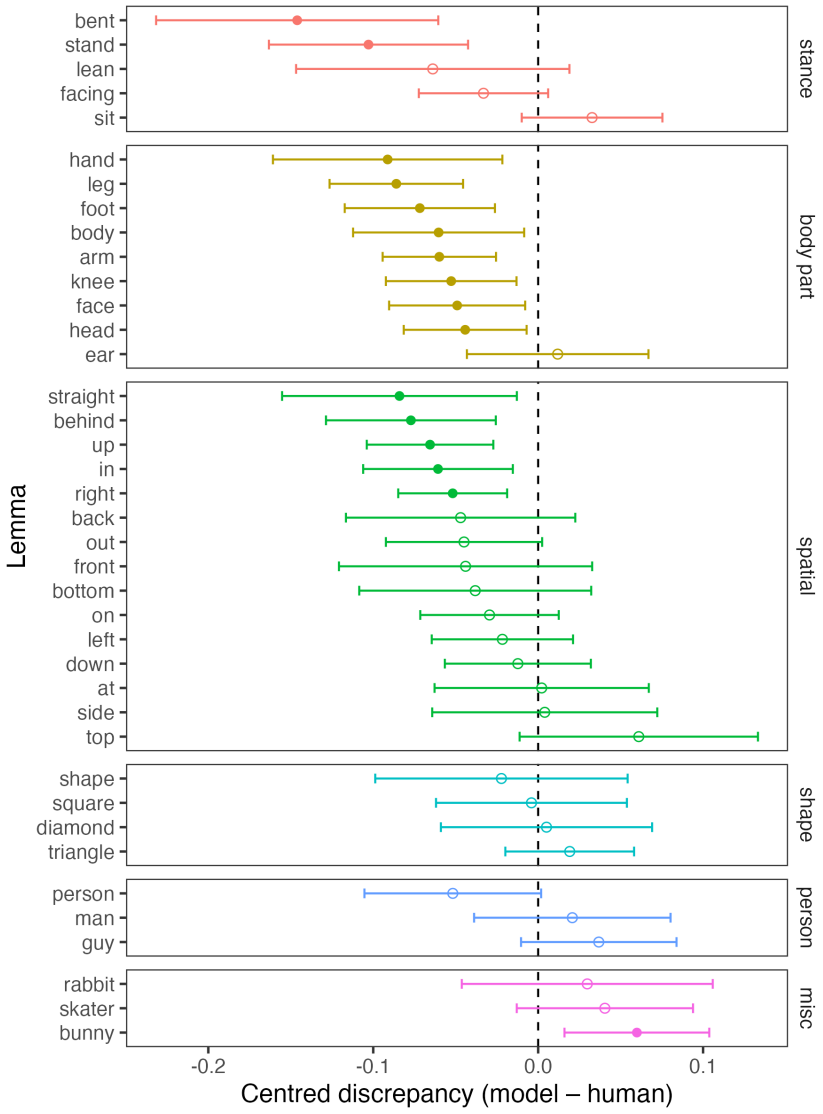

Figure 9: Centred discrepancy in model versus human accuracies in the random condition for selected lemmas, categorised by semantic categories. Error bars indicate 95% confidence intervals. Dashed line indicates zero (i.e., discrepancy equal to the grand mean). Empty/filled points indicate whether the confidence intervals do/do not include zero respectively.

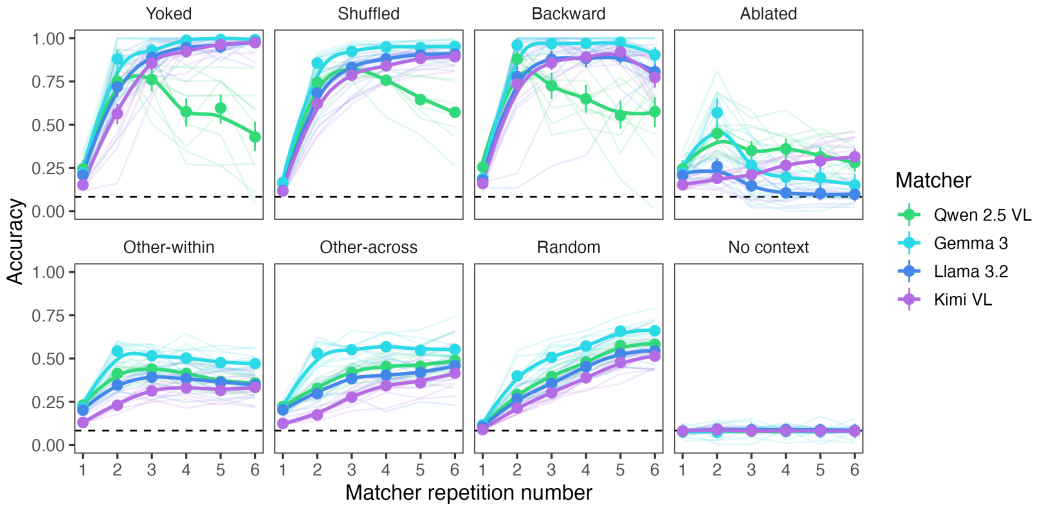

Figure 10: Matcher accuracy across all conditions and models using text input only, with best-fit LOESS curves, shown by repetition number as seen by the matcher, except for the no context condition where repetition number is from the original game. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate the chance level (0.083).

Manual inspection of the resulting estimates suggested that there were very few lemmas for which centred discrepancy was positive; the only such lemma that occurred in at least 4 games was "bunny". Conversely, there were 39 lemmas for which centred discrepancy was negative and which occurred in at least 4 games. Two notable groups of such lemmas were body parts (e.g., "hand", "leg", "foot", "arm"), and some prepositions and direction words (e.g., "behind", "up", "right"). Together, these discrepancies suggest that models were worse than humans at interpreting non-literal tangram parts and their relations. Intriguingly, models were not worse than average for several general person-related words (e.g., "person", "man", "guy"), shape words (e.g., "shape", "triangle", "square", "diamond"), and some other spatial terms (e.g., "left", "down", "side", "top"). This contrast suggests that models were average at interpreting more literal tangram parts and their relations, as well as more general or holistic humanoid descriptions.

C Additional results

C.1 Image–text vs. text-only performance

Previous studies with VLMs have suggested that they underutilise visual information due to attentional issues or bias from prior linguistic knowledge [25–27]. To investigate 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. Model accuracies are shown in Figure 10, and the difference in accuracy between the text-only and the image–text paradigms is shown in Figure 11.

All models showed near- or at-chance performance in the first repetition (across all conditions) and in the no context condition when no images were shown, suggesting that they were in fact using visual information to resolve pragmatic references. Models also generally performed worse without images than with images, although the change in performance varied across models—notably, Llama 3.2 performance in fact improved in most cases. This pattern of results suggests that some models (especially Llama 3.2) were primarily using the text information to select the correct match; this finding corroborates results from Hua et al. [12], showing that when images are shuffled from trial to trial, VLMs exhibit little to no learning over repetitions.

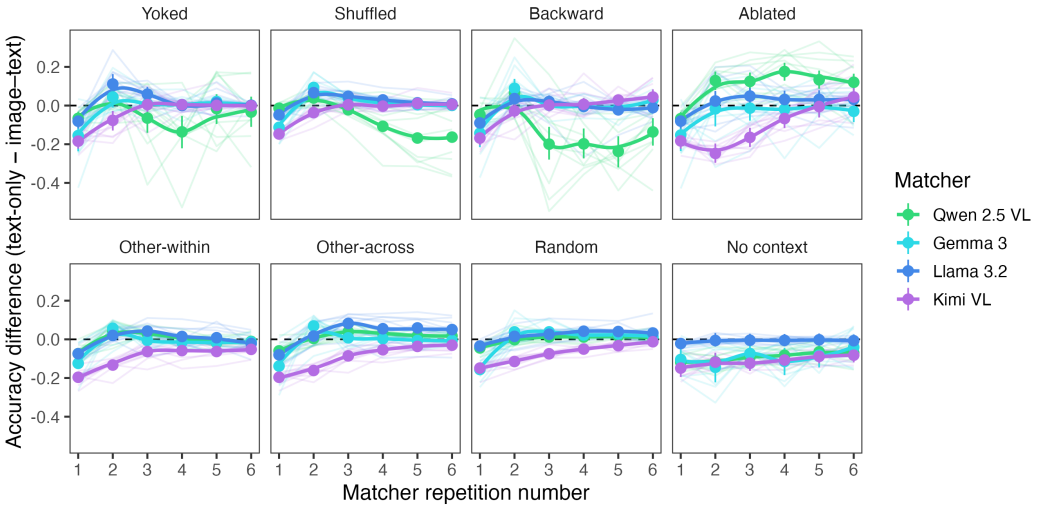

Figure 11: Difference in accuracy between the text-only and image–text paradigms across all conditions and models. Error bars indicate bootstrapped 95% confidence intervals. Dashed lines indicate no difference.