Artificial Language Learning Paradigm Reveals Pragmatic Blind Spots in Vision-Language Models

Yan Cong[1], **Julia Rayz**[2]
[1]School of Languages and Cultures, Purdue University
[2]School of Applied and Creative Computing, Purdue University
{cong4, jtaylor1}@purdue.edu

Abstract

Humans are pragmatic language users who naturally and effortlessly reason about the choice of utterances that help collaborate and engage in social interactions. In this paper, we examine whether vision-language models (VLMs) exhibit similar pragmatic reasoning effects through a validated artificial language learning paradigm. Across four experiments, we evaluate five VLMs' sensitivity to production cost, ambiguity-driven competition effects, and the influences of visual features and model properties. We find evidence of cost effects in some VLMs. However, no model consistently exhibits competition effects driven by ambiguity risk, a hallmark of Gricean pragmatic reasoning. We also find that model scale alone does not predict pragmatic alignment; architectural choices play a larger role. Moreover, probability-based methods reveal clearer effects than prompting. Overall, current VLMs capture only a restricted subset of pragmatic effects central to Gricean reasoning, suggesting gaps in multimodal pragmatic reasoning.

1 Introduction

In humans' daily conversations, utterances and their *alternatives* (expressions that have the same literal semantics and that a speaker could have used but chose not to) compete. Rational speakers choose expressions that are true, informative, relevant, and unambiguous, while listeners infer the speaker's intended meaning from those choices (Grice, 1989; Geurts, 2010). For example, the speech act of using ambiguous expressions violates Grice's *Manner* Maxim (avoid ambiguity) and leads the listener to draw a manner implicature inference: the speaker *intends* to convey a non-literal, *disambiguation*-based inference (Grice, 1989; Rett, 2015, 2019). This speaker-listener iterative process of pragmatic reasoning rests on evaluating the competition of *alternatives*. Recent Gricean pragmatics theories recast the notions of cost, including utterance length, syntactic and phonological complexity, or dependency structure, as refined dimensions shaping speakers' reasoning (Katzir, 2007; Buccola et al., 2021).

Our study concerns vision-language models (VLMs) as pragmatic language users. We build on the Artificial Language Learning (ALL) framework studied in Buccola et al. (2018), which shows that humans engage in pragmatic reasoning when even using nonce expressions in minimal communicative systems, suggesting that humans instinctively and spontaneously apply strategic reasoning. ALL provides a controlled way to study language use by isolating specific inferential strategies and pressures such as cost and competition. We adopt this paradigm to examine whether similar pragmatic reasoning patterns emerge in VLMs, exploring whether VLMs display sensitivity to the same pressures and strategies that influence human pragmatic reasoning. Evidence of such emergence can shed light on the development and optimization of pragmatically aware VLMs (Ma et al., 2025).

Our results align with and add to previous research by contributing three diagnostic aspects in VLM assessment: (a) scale invariance about model size (7B to 13B); (b) the generation-representation gap (probability vs. prompting); and (c) architectural factors (e.g., Janus showed cost effects that other models lack).[1]

2 Background

2.1 Pragmatic reasoning in humans

Humans are pragmatic language users who reason about alternatives to derive enriched, implied meaning. Under Gricean probabilistic frameworks (Grice, 1989, 1975; Bergen et al., 2016; Franke and Bergen, 2020; Frank and Goodman, 2012; Franke and Jager, 2016; Qing and Frank, 2014; Franke,

[1]Scripts are available in this public repository: https://doi.org/10.17605/OSF.IO/E6TKM.

Third Workshop on the Bridges and Gaps between Formal and Computational Linguistics, pages 50–62

2011), listeners consider not only what a speaker says but also what they could have said yet chose not to, drawing inferences derived from competition with alternatives. Although most work focuses on alternatives and competition, the role of cost as an independent dimension in pragmatic inference remains relatively underexplored. Moreover, the intersection of cost and ambiguity-based (Manner) inference is less studied due to the relative lack of robustness, making it difficult to empirically study ambiguity-based implicature in humans (Wilson, 2017) and even fewer studies are reported in language models (Cong, 2024). We hope to bridge the gap and pragmatically capture ambiguity in VLMs.

2.2 The artificial language learning paradigm

Artificial Language Learning (ALL) paradigms are used to investigate linguistic universals and domain-general cognitive preferences by isolating minimal, explicitly defined alternatives (Martin et al., 2019; Buccola et al., 2018; Culbertson, 2012; Culbertson and Schuler, 2019; Motamedi et al., 2019). Using a minimal artificial lexicon, Buccola et al. (2018) show that human learners exhibit scalar-implicature-like behavior even in the absence of prior natural language experience, demonstrating that pragmatic competition can arise spontaneously from general reasoning dynamics, going beyond language-specific conventions. Cong (2021) used ALL to study *cost* and found that humans behave pragmatically but not optimally, showing a trend to use short, ambiguous expressions when production cost is high, and significantly adjusting their choices to ambiguity risk.

This line of work suggests that humans treat cost as a principled factor in pragmatic inference, which is manifested in (i) their preference to ambiguous phrases when the competing unambiguous alternative is costly to produce or when the competition is weak and the ambiguous phrase becomes effectively unambiguous, and (ii) the disappearance of such preference when both phrases have approximately equal cost. Overall, because of ALL's minimal ingredients and exclusion of prior linguistic knowledge, it is especially useful for us to diagnose whether VLMs exhibit emergent pragmatic behavior rather than reproducing surface patterns from training data.

2.3 Pragmatic capacity of VLMs

Previous studies shed light on referring expression generation and comprehension, visual encoders, and alignment of visual and textual representations (Shu et al., 2025; Ma et al., 2025; Fu et al., 2025; Cao et al., 2020; Salin et al., 2022; Shirnin et al., 2024; Liu et al., 2021; Hua and Artzi, 2024; Vaduguru et al., 2025). Pertaining to our work, Pitta et al. (2025) evaluated VLMs in relation to the visual entailment task, where models infer whether a given image supports a specific textual caption. They note that although some VLMs perform well on visual relation detection and linguistic alignment, they still show biases rooted in text processing. Kouwenhoven et al. (2024) used ALL to examine how VLMs can develop structured communication protocols, a crucial aspect for enhancing effective collaboration among agents in multi-agent systems (Lazaridou et al., 2018). No studies have been done to apply a theory-informed ALL paradigm to identify Gricean reasoning effects in VLMs, a critical step toward rigorous multimodal pragmatics benchmark.

3 Rationale and hypothesis

Inspired by Bergen et al. (2016); Buccola et al. (2018); Cong (2021), we hypothesize that a rational agent reasons about ambiguity-based inference by weighing the communicative benefits of disambiguation against the production costs of longer expressions. In any context where both a short but ambiguous form ϕ and a longer, more costly but unambiguous alternative ψ are logically true, the rational choice depends on a context-sensitive trade-off between the cost of producing ψ and the risk of miscommunication triggered by using ϕ. This leads to four specific predictions pinpointing pragmatic reasoning effects in VLMs:

1. `Cost effect`: When production costs are high (a long expression denotes the referent), a rational agent is predicted to choose the short, ambiguous form ϕ more often than when production costs are low.
2. `Competition effect`: When ambiguity risks are high (strong competition and no contextually salient reading), a rational agent is predicted to use the short, ambiguous form ϕ less often than when ambiguity risks are low.
3. `Lack of Effect`: Image features (color intensity and objects position) should not modulate pragmatic reasoning: VLMs responses should remain stable regardless of image features.
4. `Model Effect`: Depending on model's properties (scale and language backbone), the sign

and magnitude of the above effects are predicted to differ across VLMs.

We designed four experiments under the ALL paradigm to operationalize the measurement of the above pragmatic reasoning effects in VLMs. First, we examine the cost and competition effects by comparing VLMs to humans (as rational agents). Second, we examine the cost and competition effects by comparing VLMs to the *ideal* rational agents, ones that consistently favor the short, ambiguous phrase whenever it can truthfully denote the intended referent, always relying on listeners to reason about the speaker's choice and recover the intended meaning (Bergen et al., 2016; Franke and Bergen, 2020). Third, we examine how image features such as color intensity and object position influence VLMs' choices. Fourth, we examine how model properties such as parameter sizes and architectural choices influence VLMs' responses.

4 Experiments

4.1 Experiments for humans

The original ALL experiment in Buccola et al. (2018) and Cong (2021), on which our VLM experiments design is based, examined how human speakers (n=58) balance utterance cost and ambiguity when choosing among competing alternatives. Participants learn an artificial Martian language whose nonce phrases denote geometric objects. The lexicon includes three expression types: (i) short, ambiguous phrases that can refer to two objects, (ii) long (phonologically complex), unambiguous phrases, and (iii) longest, unambiguous phrases (the longest among the three types). The task is framed as a friendly interaction with Martians who value efficiency, preferring the shortest phrase that still ensures true, correct reference.

The experiment has a learning phase and a testing phase (Figure 1). In the learning phase, participants infer meanings from trial-by-trial feedback: each trial presents a nonce phrase as in Table 1 and three candidate objects, and participants guess the referent. Utterance cost is made salient by requiring more key-presses to reveal longest phrases.

In the testing phase, participants act as speakers: they view three objects horizontally in one display, one circled to denote that it is the intended target referent, and must type one of the learned phrases to communicate the target. No feedback is given, and instructions stress that Martians always choose the phrase that can concisely and uniquely identify the object, prompting participants to reason about cost and ambiguity. Participants complete both the *experimental* trials, where two geometric objects make the short phrase contextually ambiguous, and the *control* trials, where ambiguity risk is low, since only one geometric object is present and the other is replaced by a non-geometric distractor.

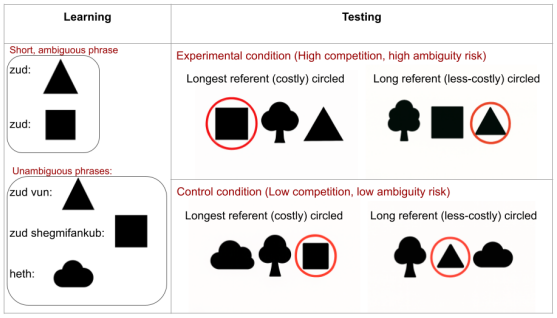


Figure 1: Artificial language learning design. Long(est) referent: a referent that the long(est) phrase denotes.

Referents	Phrases
Geometric shapes	**Short:** *kad, zud*; **Long:** *kad mof, zud mof, kad vun, zud vun*; **Longest:** *kad shegmifanzub, zud shegmifankub*
Organic shapes	*sot, heth*

Table 1: Nonce phrases used in our experiments.

We used the nonce phrases (Table 1) from Cong (2021), created following Buccola et al. (2021) and Buccola et al. (2018), which have phonetically and syntactically validated that these phrases do not exist in any natural languages. The design of these nonce phrases used sequence *length* as a proxy to *cost*. Controlled experiments in psychology show that the length of phrases signifies their conceptual complexity and cost, with longest phrases corresponding to more complex concepts (Pimentel et al., 2023; Lewis and Frank, 2016). Lewis and Frank (2016) suggested that human participants inherently link word length to conceptual complexity during both comprehension and production, even when factors like frequency, familiarity, imageability, and concreteness are considered.

4.2 Experiments for VLMs

Our VLM experiments are the same as the human ALL experiment in Buccola et al. (2018) and Cong (2021) (Figure 1), except for the following three differences, ensuring valid comparison and feasible translation of experiments between humans

and VLMs. First, we additionally examined image properties, including objects position and color intensity. As a consequence, we expanded the original dataset of 96 sets (unique combinations of image and phrase) into 768 sets. There were 36 base images, organized into 6 spatial configurations of three geometric objects (*triangle*, *square*, *circle*), in addition to 2 distractor organic objects (*cloud* and *tree*). Object positions were permuted to prevent spatial cues from biasing disambiguation. Each object was centered within its respective quadrant. Color intensity was manipulated using *solid* or *transparent* renderings. All the images were 1536x1024 pixels in size (objects did not fully occupy these pixels), with no text on the images. There were 16 unique phrase-object mappings. Each phrase denoted at least one object (Table 1).

Second, instead of learning via iterative correctness feedback, inspired by Verhoef et al. (2024); Pitta et al. (2025), we designed experiments in such a way that VLMs learn the mapping of phrase and object through the following structured declarative prompt: "*The label for this object is <label>*", pairing each nonce phrase with a visual reference in a single exposure per display, effectively a one-shot declarative learning setup. Namely, we give each shape in a single image (c.f., Figure 1 – Learning). If a phrase is ambiguous and can refer to two objects, it is learned in two separate displays.

We additionally conducted a supplementary, qualitative analysis on a single trial rather than the full dataset, examining the impact of prompting on model performance. Specifically, to address concerns that the models might lack sufficient information or context to exhibit pragmatic reasoning, a follow-up experiment was conducted using few-shot declarative learning together with an interactive Martian game context. For example, we showed three images where the object "zud" is a different color and in a different position each time. This setup explicitly provided the models with feature-invariance rules stating that color, size, and position are irrelevant to the object names, explicit cost instructions that shorter phrases are preferred if both candidate descriptions are true, and communicative framing that placed the model in a game with a listener in order to establish a shared conversational context. All the prompt templates were presented in the Appendix A.

Third, instead of typing phrases in the testing phase, we used two methods to probe VLMs. In forced-choice format, we prompted VLMs to generate and choose a phrase from their learned phrases. Model input included images of three objects and an instruction prompt in (1) `Default` plain form: "*Identify the label for the circled object in the image from the following options: <zud>, <zud vun>, <zud shegmifankub>, <heth>. Respond with only the label without explaining*"; and (2) in `Explicit` instruction form with the following instruction appended to the Default form "*Choose the option that concisely and uniquely identifies the circled object*" (Ma et al., 2025; Park et al., 2024). An Image was appended to the prompt accordingly (c.f., Figure 1 – Testing). VLMs output an option out of four. The ordering of the label was randomized at each trial to alleviate positional bias. The prompting variants (1,2) allowed us to examine if explicit efficiency instruction (as given to humans) would improve VLMs reasoning.

Besides prompting, we used forced-choice logprob scoring (Misra et al., 2020, 2021; Misra, 2022; Misra et al., 2023; Wolf et al., 2020) to compute phrase surprisals given an image and instruction prompts (as in the prompting method), normalized by phrase token length. Surprisal is the token-level negative log probability of the exact target token sequence under each VLM's native tokenizer, conditioned on the same multimodal input across VLMs. For each trial, we computed surprisal for all four phrases and took the highest-probability phrase as the model's choice, then compared these choices across models and conditions. We did not directly compare raw surprisal values across models or conditions, but instead derived a discrete choice from VLM surprisals first, to ensure valid comparisons.

We included five VLMs with differences in scale, visual encoding, and language backbone: Llava-1.5-7b and Llava-1.5-13b (Liu et al., 2023), Llava-1.6-7b (Liu et al., 2024), Gemma-2-3b (Steiner et al., 2024), and Janus-pro-1.3b (Wu et al., 2024).

4.3 Statistical tests

For Experiments 1 and 2 on cost and competition effects, to assess VLMs' alignment with human responses as reported in Cong (2021), we conducted chi-square tests of independence to determine whether the proportion of short, ambiguous responses differed significantly as a function of either *CircledObject* (longest≈costly vs. long≈less-costly; cost effect) or *Condition* (control vs. experimental; competition effect). To evaluate VLMs' alignment with ideal rational agents, we fit logistic

regression models in which match rates between each VLM's predicted phrases and the true labels were regressed on cost (*CircledObject*) and competition (*Condition*).

For Experiment 3, which examined the absence of image feature effects, we fit additional logistic regression models to test how color intensity (solid vs. transparent) affected these match rates. We applied Levene's tests to assess the homogeneity of variance in VLMs responses across object position. These tests evaluated whether VLMs response variability differed across trials with identical content and conditions but varied only in object spatial arrangement. p-values were adjusted for multiple comparisons (Benjamini–Hochberg false discovery rate (FDR)), with FDR-adjusted p-values below 0.05 indicating significant variance heterogeneity across images with different object positions.

In Experiment 4, we fitted linear mixed-effects models (LMEs) to investigate how different Llava model variants predict the accuracy of matches with an ideal rational agent. In these LMEs, match served as the dependent variable and model type as the fixed-effect predictor, and we specified a random intercept for the item identifier to account for the repeated-measures structure of the data. We constructed two LMEs: on parameter size by comparing Llava-1.5-13b with Llava-1.5-7b; on language model backbone by comparing Llava-1.6-7b with Llava-1.5-7b.

5 Results

5.1 Experiment 1: Cost and competition effects - VLMs vs. humans

5.1.1 Prompting-based metric

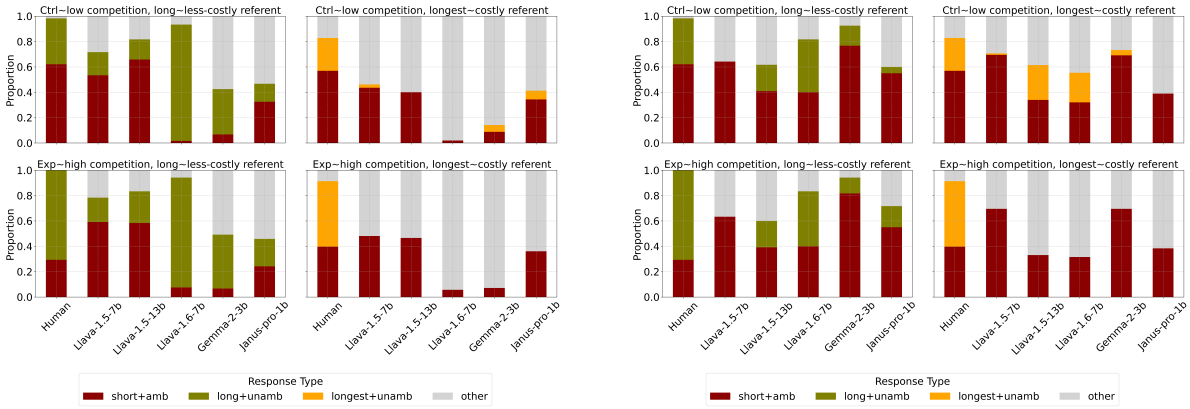
Figure 2a shows the proportion of each response type in VLMs' prompting-based responses, with "other" representing irrelevant or false responses, for instance, *zud vun* would be false in terms of literal semantics when the circled object can only be referred to using *zud* or *zud shegmifankub*. First, we observed a cost effect in Janus, which chose short, ambiguous expressions more often when cost was high (longest referent) than when cost was low (long referent), consistent with how humans perform in Cong (2021). Second, inconsistent with how humans perform in Cong (2021), we did not observe a competition effect in any of the VLMs: relative to control conditions with low ambiguity risk, the experimental conditions did not lead the models to choose short, ambiguous expressions less often. Third, relative to humans' responses in Cong (2021), we found a substantial proportion of "other" responses in Llava-1.6-7b and Gemma.

Model	Cost effect		Competition effect	
	OR	p	OR	p
Prompting-based response				
Llava-1.5-7b	1.52	0.009	1.22	0.193
Llava-1.5-13b	2.137	<0.0001	1.087	0.612
Llava-1.6-7b	1.22	0.748	3.552	0.004
Gemma-2-3b	0.827	0.632	0.851	0.652
Janus-1.3b	0.73	0.072	0.939	0.74
Probability-based response				
Llava-1.5-7b	0.772	0.134	0.993	1.000
Llava-1.5-13b	1.322	0.097	0.951	0.796
Llava-1.6-7b	1.429	0.033	0.984	0.977
Gemma-2-3b	1.682	0.006	1.097	0.622
Janus-1.3b	1.941	<0.0001	0.982	0.957
Humans	0.901	0.793	2.789	0.0002

Table 2: VLMs alignment with humans. OR (Odds Ratio) < 1 indicates lower odds of the short, ambiguous response when production costs are high (*cost effect*) or when ambiguity risks are low (*competition effect*), Odds Ratio > 1 indicates higher odds in such conditions; aligned (sign and magnitude); misaligned; improved with explicit instruction.

Statistical tests results were summarized in Table 2. Humans showed a strong competition effect but a weak cost effect (Cong, 2021): circling costly objects showed a small increase in short, ambiguous choices (48.28% for costly and 45.69% for less-costly), corresponding to just a 2.59% difference, and this pattern was not statistically significant ($\chi^2 = 0.069$, $p = 0.793$). Because humans themselves exhibit only a weak directional cost effect, we considered VLM as "aligned with humans" as long as it shows at least a 2.59% higher rate of short, ambiguous responses for costly relative to less-costly referents, matching the magnitude of the human pattern and ensuring that VLM-human comparisons reflect the empirical behavior of human participants.

Table 2 shows that statistically, relative to the behavioral threshold humans display, only Janus exhibit cost effect, and no VLMs show competition effect, suggesting that with the prompting method, VLMs struggle with choosing the true and relevant responses, especially in scenarios when a longest≈costly referent is circled (i.e., intended) and production cost is high (Figure 2a). We did not find solid evidence of competition effects, suggesting lack of sensitivity to ambiguity risks. With explicit instruction, only Llava-1.6-7b prompting showed cost effect (χ^2=6.054, p=0.013); and no VLMs showed competition effects (details in Ap-



(a) Prompting-based metric

(b) Probability-based metric

Figure 2: VLM responses in the testing phase, grouped by condition and circled object.

pendix Figure 4, 5).

As to the supplementary, qualitative experiment with more elaborate, interactive prompts, results suggested that all tested models (PaliGemma-2-3B, Janus-pro-1.3B, and Llava) failed to perform the task, and in many cases behaved worse than in the minimal one-shot setup. PaliGemma-2-3B repeatedly disregarded the negative constraint to respond with only the label and instead produced descriptive sentences listing the features it had been instructed to ignore, for example, statements such as *The object is square* and *The object is red*. In addition, the model frequently hallucinated objects and shapes that were not present in the dataset, including hearts, stars, and diamonds. Under the increased complexity introduced by the few-shot interactive prompt, Janus and Llava showed similar behaviors, with many "other" outputs.

5.1.2 Probability-based metric

Figure 2b shows proportions of each response type in VLMs' probability-based responses, with generally different patterns than the prompting responses. First, we observed cost effect in Llava-1.5-7b: when the intended referent is costly, it chose the short, ambiguous phrases more often than when the referent is less costly. Second, similar to prompting results in Figure 2a, we did not observe competition effects in any VLMs. Third, we observed generally fewer "other" in probability than in prompting responses. Statistically, Table 2 confirmed our observation that Llava-1.5-7b exhibited cost effect and that no VLMs aligned with humans in competition effects, suggesting that regardless of prompting or probability methods, VLMs did not show consistent sensitivity to ambiguity risks.

5.2 Experiment 2: Cost and competition effects - VLMs vs. ideal rational agent

5.2.1 Prompting-based metric

Figure 3a shows the percentage of match between VLM's prompting-based responses and true labels, which are always short, ambiguous phrases. Since there were four options in the testing phase, we drew the dotted line at 25% as the chance level match accuracy. First, we observed cost effect in Janus and Gemma, where there were more matches in "longest referent" (a scenario that the costly object is circled and intended) than in "long referent" (less-costly intended). Second, we observed no competition effects that more matches should be in control than in experimental condition, resonating with Experiment 1 (Figure 2a). Third, we observed that Llava-1.5-7b and Llava-1.5-13b consistently showed above chance match accuracies, across conditions and circled objects, suggesting high tolerance of ambiguity risks since short, ambiguous phrases were their frequent choices.

Table 3 shows that statistically, circling a long≈less-costly object decreased match odds for Gemma and Janus, confirming our observation about cost effects that when the intended is a costly referent, there is an increase in choosing short, ambiguous phrases. This also resonates with Experiment 1 where we found cost effect in Janus. Across VLMs, we did not find statistical evidence for competition effects, echoing Experiment 1. With explicit instruction, no VLMs showed cost effects; Llava-1.6-7b prompting showed competition effect (β=-0.7, p=0.024) (Appendix Figure 6, 7).

Model	Cost effect		Competition effect	
	β	p	β	p
Prompting-based response				
Llava-1.5-7b	0.429	0.013	0.191	0.189
Llava-1.5-13b	0.635	<0.0001	0.087	0.553
Llava-1.6-7b	0.0093	0.982	1.2675	0.004
Gemma-2-3b	-0.735	0.022	-0.178	0.523
Janus-1.3b	-0.441	0.017	-0.091	0.555
Probability-based response				
Llava-1.5-7b	-0.958	<0.0001	-0.019	0.907
Llava-1.5-13b	0.2569	0.145	-0.05	0.741
Llava-1.6-7b	0.588	0.001	-0.0118	0.939
Gemma-2-3b	0.331	0.100	0.093	0.569
Janus-1.3b	0.529	0.002	-0.019	0.898

Table 3: VLMs alignment with ideal rational agent. Values: β and p; aligned (sign and magnitude); misaligned; improved with explicit instruction.

Model	Default prompt		Explicit instruction	
	Acc	F1	Acc	F1
Prompting-based response				
Llava-1.5-7b	0.510	0.674	0.241	0.370
Llava-1.5-13b	0.527	0.685	0.241	0.370
Llava-1.6-7b	0.042	0.079	0.076	0.139
Gemma-2-3b	0.073	0.136	0.000	0.000
Janus-1.3b	0.318	0.480	0.290	0.448
Probability-based response				
Llava-1.5-7b	0.666	0.799	0.675	0.805
Llava-1.5-13b	0.368	0.537	0.291	0.447
Llava-1.6-7b	0.359	0.527	0.202	0.336
Gemma-2-3b	0.742	0.851	0.730	0.841
Janus-1.3b	0.468	0.634	0.425	0.593
Human			0.470	0.628

Table 4: Human and VLMs overall performance in matching with ideal rational agents. Acc: Accuracy.

5.2.2 Probability-based metric

Figure 3b demonstrates the percentage of match between VLM's probability-based responses and true labels. First, we observed cost effect in Llava-1.5-7b, resonating with our Experiment 1 finding about this same VLM (Table 2). Second, we did not observe any competition effects, reproducing prompting-based findings as well as Experiment 1 results. Third, all VLMs showed above-chance match accuracies across conditions and circled objects, surpassing prompting-based results in Figure 3a. Statically, Table 3 confirmed cost effects in Llava-1.5-7b probability-based responses. Overall, Experiment 2 findings were mostly in line with Experiment 1.

5.2.3 Overall match

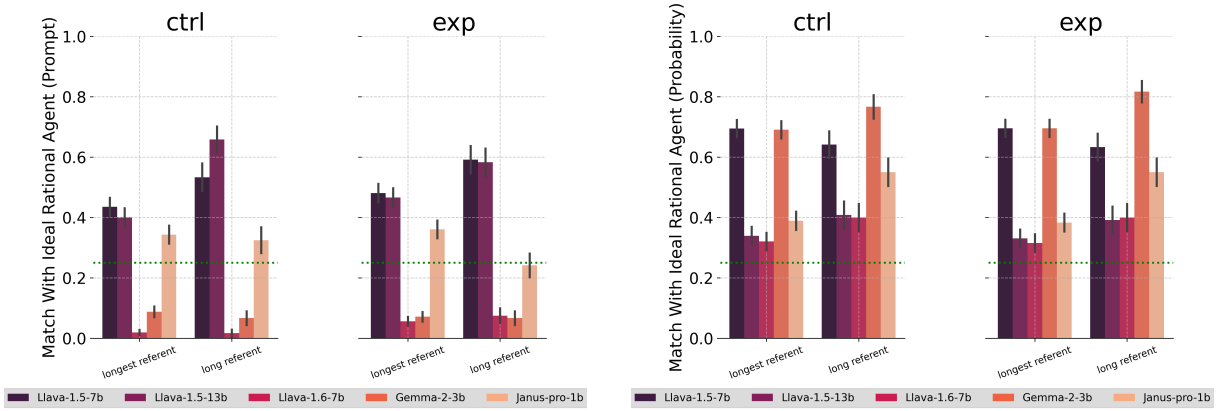
We additionally calculated humans' and VLMs' overall match accuracies and weighted F1 scores aggregating all the trials (Table 4). First, Llava-1.5-7b is the sole VLM that showed good alignment with ideal rational agents in both prompting and probability methods, surpassing human participants' match rates metrics. Second, Gemma is the most sensitive to methods, which showed much higher accuracies and F1 scores in probability method than in prompting method. Third, we found generally higher accuracies and F1 values in probability than in prompting method. Overall, Table 4 echoed the observation in Figures 3a and 3b, suggesting that depending on VLMs, probabilistic measurements can outperform prompting. With explicit instruction, we observed improvements for Llava-1.6-7b prompting and Llava-1.5-7b probability, whereas Gemma (prompting) was biased by the instruction and never chose the ambiguous phrase.

5.3 Experiment 3: Lack of effect

Table 5 demonstrates statistical results of how image features influence VLMs' responses. First, color intensity significantly influenced match odds for Gemma, Janus, Llava-1.5-7b, and Llava-1.6-7b, in prompting or probability methods, suggesting that these models' responses varied depending on whether the image was solid or transparent. Second, we found that for Llava-1.5-13b and Llava-1.6-7b, there were significant variance of VLMs probability responses in the position of objects, suggesting that these VLMs did not remain stable when image features varied. Third, we observed generally stable responses in prompting method than in probability. Fourth, we found color intensity, but not position of objects, more often triggered VLMs' responses variance. With explicit instruction, Gemma and Janus no longer showing significant variance in color, and Llava-1.5-13b and Llava-1.6-7b no longer exhibiting significant variance in position. Meanwhile, adding explicit instruction triggered significant variance in color for Llava-1.6-7b prompting (p=0.015) and Llava-1.5-13b probability (p <0.0001).

5.4 Experiment 4: Model effect

In the first linear mixed effects model (LME), we compared Llava-1.5-13b against Llava-1.5-7b. The results show no difference between the two (Estimate = 0.001, p = 0.958). In contrast, the second LME compared Llava-1.6-7b to Llava-1.5-7b and revealed that Llava-1.6-7b shows lower match scores (Estimate = -0.451, $p < 0.0001$). Together, these results demonstrate that language backbone, rather than parameter size, played a role in matching with ideal rational agent.



(a) Prompting-based metric

(b) Probability-based metric

Figure 3: VLM response match with ideal rational agents. Error bars indicate standard error of the mean. Dotted lines show chance-level match accuracy.

Model	**Color** (p)	**Position** (FDR-p)
Prompting-based response		
Llava-1.5-7b	0.863	0.253
Llava-1.5-13b	0.066	0.453
Llava-1.6-7b	0.246	0.920
Gemma-2-3b	<0.0001	0.306
Janus-1.3b	0.008	0.584
Probability-based response		
Llava-1.5-7b	<0.0001	0.998
Llava-1.5-13b	0.759	<0.0001
Llava-1.6-7b	0.003	<0.0001
Gemma-2-3b	0.015	0.530
Janus-1.3b	0.049	0.770

Table 5: As predicted (sign and magnitude); not as predicted; improved with explicit instruction.

6 Discussion

Some evidence for cost effects Our first two experiments provide evidence that cost effects can emerge in VLMs, although inconsistently across models and measurement methods. In Experiment 1, Janus-pro-1.3b (prompting) and Llava-1.5-7b (probability) showed a higher proportion of short, ambiguous expressions when the costly referents were circled and intended, mirroring the weak directional cost effect observed in humans. Experiment 2 reinforced this pattern: the same models exhibited higher match rates with the ideal rational agent when the intended referent was costly. These results illustrate a first application of ALL as a diagnostic tool for probing cost-related pragmatic behavior in VLMs, suggesting that some models can exhibit strategic cost-sensitive reasoning without relying solely on natural language frequency or lexical priors. One possible explanation for why only Janus and Llava-1.5-7b show this pattern is that both architectures preserve a degree of separation between visual and textual processing, with explicit cross-modal alignment mechanisms (Shu et al., 2025), which may better augment reasoning. Overall, while certain VLMs approximate cost-based pragmatic reasoning in minimal settings, further evidence is needed to establish that this is not only method- or architecture-specific but reflects a robust, model-internal capacity.

Absence of competition effects We found no compelling evidence that VLMs exhibit competition effects driven by ambiguity risk. This contrasts with human sensitivity to ambiguity risk. The absence of competition effects suggests that VLMs do not reliably reason over alternatives in a Gricean way. This finding challenges our hypothesis and aligns with prior work showing that VLMs struggle with pragmatic reasoning despite strong surface-level performance (Verhoef et al., 2024; Cao et al., 2020; Bihani and Rayz, 2025; Misra et al., 2023; Cong and Rayz, 2025; Hua and Artzi, 2024; Vaduguru et al., 2025). The weak pragmatics and lack of robust alternatives-based reasoning also align with broader findings in multimodal evaluations, reporting failures in referential disambiguation, ambiguity handling, among other pragmatic tasks (Qiu et al., 2025; Ma et al., 2025; Testoni et al., 2025; Fu et al., 2025; Shu et al., 2025).

Image features affect some VLMs Contrary to our hypothesis that rational choices should be robust to irrelevant visual cues, color intensity significantly modulated responses in several models, and object position induced variance in larger Llava variants. These effects were more pronounced in probability-based analyses, indicating

that visual salience and low-level perceptual differences likely affect probabilistic pragmatic decision-making. This finding reinforces concerns that VLMs conflate perceptual salience with communicative relevance, a limitation also noted in Liu et al. (2021); Shirnin et al. (2024); Fu et al. (2025).

Model properties As predicted, the sign and magnitude of pragmatic reasoning effects differ across VLMs. Model scale does not influence pragmatic alignment. Increasing parameter size from Llava-1.5-7b to 13b yielded no improvement, whereas replacing the Llama language backbone with Mistral (Llava-1.5 vs. 1.6) substantially reduced alignment with the ideal rational agent. This resonates with previous work advocating for moving beyond sheer scale (Kouwenhoven et al., 2024). Both Llava-1.5 and Llava-1.6 are Vicuna- and CLIP-based, our preliminary findings are also suggestive of limitations in such multimodal architectures. Moreover, consistent improvement in Llava-1.6-7b thanks to explicit instruction in prompting suggests its sensitivity to instructions, but we did not find generalization of such improvement for other VLMs, in fact, explicit instructions triggered VLMs to attend the irrelevant color intensity features, suggesting misalignment with human pragmatic preference (Ma et al., 2025).

The influence of more elaborate prompts For the supplementary prompting experiment, the failure to follow explicit, elaborate instructions regarding feature-invariance suggests that the models conflate perceptual salience with communicative relevance. Rather than ignoring color and position when instructed to do so, the models became fixated on describing these visually salient properties. The results provide another piece of preliminary evidence that the absence of competition effects in VLMs is likely not an artifact of an underspecified prompt, but instead reflects a fundamental reasoning deficit that persists even when the rules of pragmatic behavior are explicitly provided in the input.

Proxy for cost: token length versus phrase length Gemma-based models (e.g., Google's PaliGemma) use a SentencePiece tokenizer that tends to preserve short nonce strings as single tokens (e.g., *zud* and *heth* each map to 1 token, while *zud shegmifankub* maps to 2 tokens). In contrast, Llama- and Mistral-based Llava models apply BPE tokenization, segmenting the same strings into multiple subword units (e.g., *zud* becomes 2 tokens, and *zud shegmifankub* becomes 5 – 6 tokens). Janus, which uses the DeepSeek tokenizer, tokenizes the same phrase *zud shegmifankub* into 3 tokens.

This variation is relevant because *length* as a cost proxy can refer either to human-perceived phrase length or to token count, which more directly reflects a model's representational and processing load. Under this view, a 6-token sequence (e.g., *zud shegmifankub*) is effectively longer for a model than a 2- or 3-token one, even when humans perceive the strings as equally long. Model preferences may therefore track token length rather than phrase length, and this could differ systematically across architectures. A natural follow-up would construct nonce stimuli with matched token lengths across VLMs, enabling cleaner comparison against a common human baseline. Since prior work links greater length to higher cost, aligning token length across models would help clarify whether observed effects are consistent with that literature and whether they hold within VLMs whose behavior more closely mirrors human judgments.

7 Conclusion

VLMs exhibit some cost sensitivity and lack a core component of pragmatic competence: reasoning about ambiguity through competition with alternatives. Our study advocates for theory-driven evaluation frameworks and pragmatically aware VLMs.

8 Limitations

There are several limitations in our work. We operationalized the concept of cost using phrase length, as it has been studied in human behavior. This is an attempt to minimize divergence between human and VLM experiments design, while still allowing valid and feasible comparisons and experiments translation between the two. Future work could explore different approaches for cost manipulations and interactive, listener-based settings to further test whether richer pragmatic behavior can emerge in VLMs.

Tokenizations differ across the VLMs we evaluate. Our analyses rely exclusively on within-model comparisons, with phrase surprisal normalized by token length, and model choices are derived from relative surprisal rankings among competing phrases under a fixed tokenizer. We maintain that absolute differences in token granularity

across models do not necessarily affect the validity of our contrasts or the derived choice-based comparisons reported here. Future studies could systematically examine model-specific tokenization statistics, and show whether the observed cost effects persist when cost is defined by token count rather than phrase length.

Unlike human learners, VLMs in our study do not acquire mappings through iterative correctness feedback. While our structured prompting and one-shot learning design allow us to test VLMs' ability to represent and retain phrase-object associations, we acknowledge that learning mechanisms such as repetition, feedback, and long-term retention are underexplored here. While the primary experiments utilized 12 trials across various renderings, our supplementary qualitative analysis into few-shot interactive prompting was limited to one specific trial configuration. Unlike the main study, which systematically varied color intensity across 768 unique sets, this follow-up focuses on how the models (PaliGemma, Janus, and Llava) respond to a communicative game context within a single trial. Future work could expand our follow-up and systematically vary exposure regimes and incorporate more explicit post-learning comprehension checks within the ALL framework.

The present findings were derived from a highly controlled, narrowly defined task that employed a restricted set of materials. Subsequent research could consider more systematic comparisons between models and human behavior, thereby shedding light on the mechanisms of language learning and use in naturalistic, ecologically valid contexts.

9 Acknowledgments

We thank Brian Buccola for his insights in formal and experimental semantics and pragmatics, Trisha Godara for her help in computational linguistics, and BriGap reviewers for their constructive comments. This project is supported by the College of Liberal Arts, Purdue University.

References

Leon Bergen, Roger Levy, and Noah D. Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.

Geetanjali Bihani and Julia Rayz. 2025. Learning shortcuts: On the misleading promise of nlu in language models. In *Artificial Intelligence for Design and Process Science*, pages 147–158. Springer Nature Switzerland.

Brian Buccola, Isabelle Dautriche, and Emmanuel Chemla. 2018. Competition and symmetry in an artificial word learning task. *Frontiers in Psychology 9 (2176)*.

Brian Buccola, Manuel Križ, and Emmanuel Chemla. 2021. Conceptual alternatives. *Linguistics and Philosophy*.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020*, pages 565–580, Cham. Springer International Publishing.

Yan Cong. 2021. *Competition in Natural Language Meaning: The Case of Adjective Constructions in Mandarin Chinese and Beyond*. Michigan State University.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Yan Cong and Julia Rayz. 2025. Language models demonstrate the good-enough processing seen in humans. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Jennifer Culbertson. 2012. Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, 6(5):310–329.

Jennifer Culbertson and Kathryn Schuler. 2019. Artificial language learning in children. *Annual Review of Linguistics*, 5:353–373.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Michael Franke. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, 4(1):1–82.

Michael Franke and Leon Bergen. 2020. Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language*, 96(2).

Michael Franke and Gerhard Jager. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift fur Sprachwissenschaft*, 35(1):3–44.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

Bart Geurts. 2010. *Quantity Implicatures*. Cambridge University Press, Cambridge.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press, New York.

H. Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Yilun Hua and Yoav Artzi. 2024. Talk less, interact better: Evaluating in-context conversational adaptation in multimodal llms. *arXiv preprint arXiv:2408.01417*.

Roni Katzir. 2007. Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6):669–690.

Tom Kouwenhoven, Max Peeperkorn, and Tessa Verhoef. 2024. Searching for structure: Investigating emergent communication with large language models. *arXiv preprint arXiv:2412.07646*.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*.

Molly L Lewis and Michael C Frank. 2016. The length of words reflects their conceptual complexity. *Cognition*, 153:182–195.

Fenglin Liu, Xian Wu, Shen Ge, Xuancheng Ren, Wei Fan, Xu Sun, and Yuexian Zou. 2021. Dimbert: Learning vision-language grounded representations with disentangled multimodal-attention. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1):1–19.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr and world knowledge. *arXiv preprint arXiv:2401.17744*.

Haotian Liu, Yuntao Li, Yi Shen, Chunyuan Liu, and et al. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

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. *arXiv preprint arXiv:2504.16060*.

Alexander Martin, Theeraporn Ratitamkul, Klaus Abels, David Adger, and Jennifer Culbertson. 2019. Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, 5(1):20180072.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949.

Yasamin Motamedi, Marieke Schouwstra, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2019. Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, 192:103964.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024. Pragmatic competence evaluation of large language models for the Korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266, Tokyo, Japan. Tokyo University of Foreign Studies.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2240–2255.

Elena Pitta, Tom Kouwenhoven, and Tessa Verhoef. 2025. Probing vision-language understanding through the visual entailment task: promises and pitfalls. In *Proceedings of the 2nd LUHME Workshop*, pages 74–83.

Ciyang Qing and Michael C. Frank. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Proceedings of Semantics and Linguistic Theory (SALT) 24*, volume 24, pages 23–41.

Linlu Qiu, Cedegao E Zhang, Joshua B Tenenbaum, Yoon Kim, and Roger Levy. 2025. On the same wavelength? evaluating pragmatic reasoning in language models across broad concepts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19924–19946.

Jessica Rett. 2015. *The Semantics of Evaluativity*. Oxford University Press.

Jessica Rett. 2019. Manner implicatures and how to spot them. *International Review of Pragmatics*, in press.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.

Alexander Shirnin, Nikita Andreev, Sofia Potapova, and Ekaterina Artemova. 2024. Analyzing the robustness of vision & language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2751–2763.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1713–1735, Suzhou, China. Association for Computational Linguistics.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, and 1 others. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Alberto Testoni, Barbara Plank, and Raquel Fernández. 2025. Racquet: Unveiling the dangers of overlooked referential ambiguity in visual llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23638–23658.

Saujas Vaduguru, Yilun Hua, Yoav Artzi, and Daniel Fried. 2025. Success and cost elicit convention formation for efficient communication. *arXiv preprint arXiv:2510.24023*.

Tessa Verhoef, Kiana Shahrasbi, and Tom Kouwenhoven. 2024. What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. *arXiv preprint arXiv:2407.17974*.

Elspeth Amabel Wilson. 2017. *Children's development of Quantity, Relevance and Manner implicature understanding and the role of the speaker's epistemic state*. Ph.D. thesis, University of Cambridge.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *Preprint*, arXiv:2410.13848.

A Prompt excerpts

Learning in one-shot format:
The label for this object is <label>.

Learning in few-shot format:

Note: the following prompt was presented three times, and each paired with an image where the object is in a different color and in a different position.

The label for the circled object is <label>. The color, size, and position of the object do not determine its name; only its shape matters.

Gaming without explicit instruction (default):
Identify the label for the circled object in the image from the following options: <label>zud</label>, <label>zud vun</label>, <label>zud shegmifankub</label>, <label>heth</label>.

Gaming with explicit instruction:
Identify the label for the circled object in the image from the following options: <label>zud</label>, <label>zud vun</label>, <label>zud shegmifankub</label>, <label>heth</label>. Respond with only the label without explaining. Choose the option that concisely and uniquely identifies the circled object.

Gaming with explicit instruction and interactive context:
You are playing a game with a Martian. The Martian sees the same three objects you see. You must send them a label so they can correctly pick the object you have circled. Identify the label for the circled object. Note that the size and position of the object do not determine its name; only its shape matters. In this Martian language, short phrases are preferred over long phrases if both are true. Your goal is to use the shortest possible label that uniquely identifies the target. Choose from the following options: <label>zud</label>, <label>zud vun</label>, <label>zud shegmifankub</label>, <label>heth</label>. Respond with only the label without explaining. Which label do you send?

B Supplementary figures

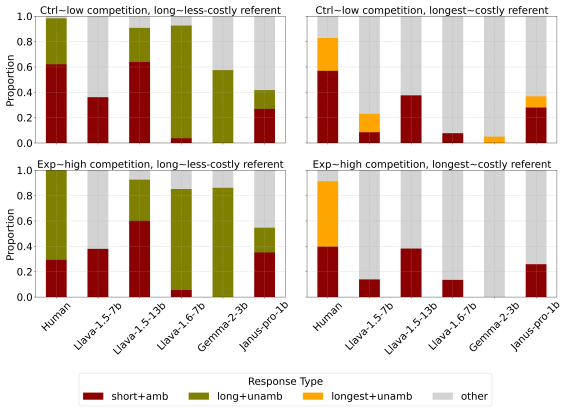


Figure 4: VLMs-prompt responses in the testing phase with explicit instruction prompt, grouped by condition and circled object.

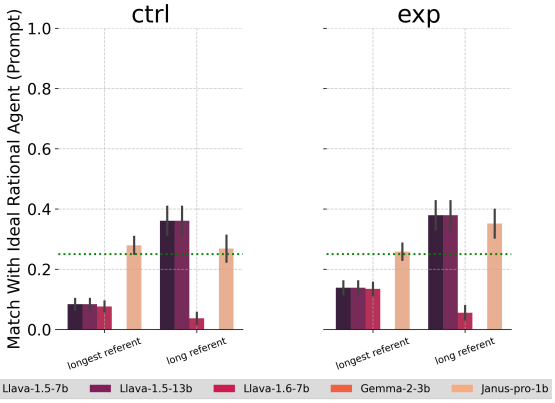


Figure 6: VLMs prompting responses match with ideal rational agents, with explicit instruction prompt. Error bars: standard error of the mean. Dotted line: chance-level match accuracy.

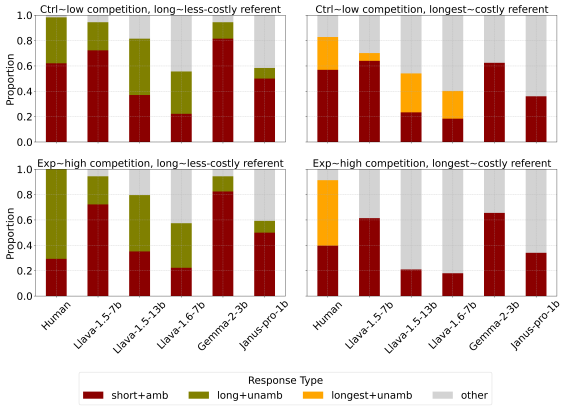


Figure 5: VLMs probability responses in the testing phase with explicit instruction prompt, grouped by condition and circled object.

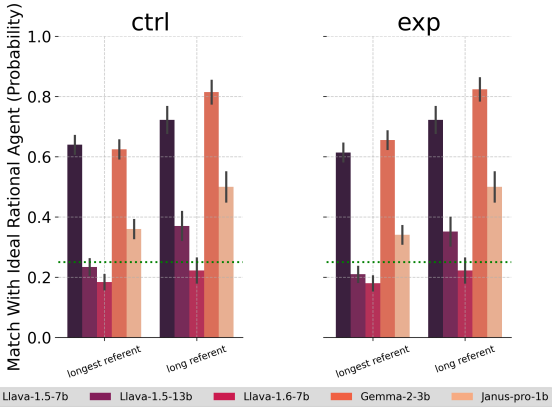


Figure 7: VLMs probability responses match with ideal rational agents, with explicit instruction prompt. Error bars: standard error of the mean. Dotted line: chance-level match accuracy.