Using Perspectival Words Is Harder Than Vocabulary Words for Humans —and Even More So for Multimodal Language Models

Anonymous EACL submission

Abstract

Multimodal language models (MLMs) increasingly demonstrate human-like communication, yet their use of everyday perspectival words remains poorly understood. To address this gap, we compare humans and MLMs in their use of three word types, which we predict impose increasing cognitive demands: vocabulary (e.g., 'boat' or 'cup'), possessives (e.g., 'mine' vs. 'yours'), and demonstratives (e.g., 'this one' vs. 'that one'). Testing seven MLMs against human participants, we find that perspectival words are harder than vocabulary words for both groups. The gap is even larger for MLMs: while models approach human-level performance on using vocabulary, they exhibit clear deficits with possessives and even greater difficulties with demonstratives. Ablation analyses point to limitations in perspective-taking and spatial reasoning as key sources of these gaps in MLMs. Instruction-based prompting helps close the gap for possessives but still leaves demonstratives far below human performance. These results show that, unlike vocabulary, perspectival words pose a greater challenge in human communication—and this difficulty is further amplified in MLMs, revealing a crucial shortfall in their pragmatic and social-cognitive abilities.

1 Introduction

A central goal in NLP has long been to develop systems that can communicate about the world as humans do. Recent advances in multimodal language models (MLMs) have brought us closer to this goal, enabling systems to use text and image so naturally that users often perceive them as real conversational partners (Elkins and Chun, 2020). Existing evaluations of MLMs have largely focused on vocabulary—words with relatively stable, context-independent meanings in conversation, such as object names, colors, and verbs (Yuksekgonul et al., 2022; Chaoyou et al., 2023). This vocabulary-centric approach, however, overlooks a fundamentally different yet ubiquitous word class: perspectival words such as 'I', 'you', 'mine', and 'this one'.[1]

Unlike vocabulary words, perspectival words are part of grammar, and their meaning is **grounded in the context of use**. For example, 'I' and 'you' refer to the speaker and listener in a conversation, so they denote a different interlocutor at every turn. Perspectival words are also sensitive to **speaker-listener perspectives**: whether a given spoon is 'mine' or 'yours' depends on who is speaking, and whether it counts as 'this one' or 'that one' depends on the object's spatial position relative to the speaker's (Rubio-Fernandez, 2021, 2022). Therefore, using perspectival words requires **pragmatics**—the ability to ground language use in its physical and social context, and **social cognition** (or 'Theory of Mind')—the ability to understand others' minds and take their perspective.

Despite their frequent and central role in everyday communication (Clark, 1996; Carston, 2002; Levinson, 2006, 2024; Rubio-Fernandez, 2021, 2024; Jara-Ettinger and Rubio-Fernandez, 2024), perspectival words have received surprisingly little attention in MLM evaluation. This gap is consequential because correctly using these words requires pragmatic reasoning and perspective-taking. Investigating how MLMs use perspectival words therefore provides a direct test of whether these systems possess the pragmatic capabilities (Jeretic et al., 2020; Tong et al., 2021; Liu et al., 2022; Ruis et al., 2022; Stowe et al., 2022; Barattieri di San Pietro et al., 2023; Hu et al., 2023) and Theory of Mind (for reviews, see Ma et al., 2023; Shapira et al., 2023; Hu et al., 2025) that are necessary for genuinely human-like communication.

In cognitive science, perspectival words are also less well understood than sentence-level pragmatic phenomena (e.g., metaphor, irony, or scalar implicature). Recent behavioral and computational

[1] 'Perspectival words' is a non-technical label for indexicals, specifically personal, possessive, and demonstrative pronouns.

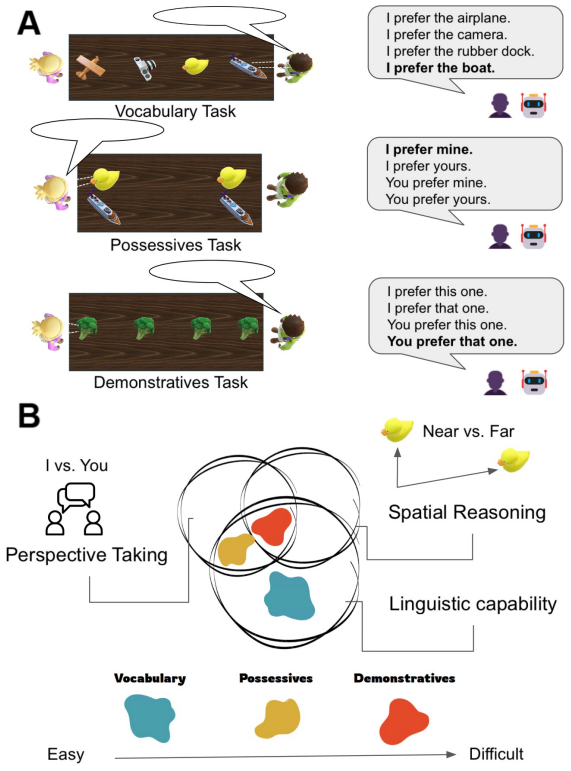

Figure 1: **A:** We test human participants and MLMs across three tasks: (1) Vocabulary task: selecting the correct label for a referent (e.g., 'boat'); (2) Possessive task: selecting 'I' vs 'you' and 'mine' vs 'yours', and (3) Demonstrative task: selecting 'I' vs 'you' and 'this one' vs 'that one'. **B:** We predict a hierarchy of task difficulty.

work suggests that perspectival words rely on **social micro-processes** (Rubio-Fernandez et al., 2024): temporally-bound mentalistic representations that allow us to reason about others during real-time communication, without fully modeling their minds (see also Holler, 2025). Comparing how humans and models use perspectival words can therefore offer valuable insights into the socio-cognitive basis of conversation—for instance, whether pragmatic behaviors can emerge without explicit mentalistic representations (Andreas, 2022), or in the absence of communicative interaction in the physical world (Fried et al., 2023).

We introduce a multiple-choice task to test two novel hypotheses in both humans and MLMs. First, selecting the correct label for an object (e.g., 'spoon' vs. 'boat') should be easier than using possessives ('mine' vs. 'yours') or demonstratives ('this one' vs. 'that one') for the same objects (Figure 1). Second, we predict that demonstrative use will be more challenging than possessive use, as it requires both perspective-taking and spatial reasoning. To evaluate these hypotheses, we collected data from 360 adult participants using materials carefully designed to avoid training contamination in MLMs (Akata et al., 2023), and compared human performance with that of seven state-of-the-art MLMs.

We observe the predicted hierarchy of word usage difficulty in both humans and most MLMs: Vocabulary < Possessives < Demonstratives. For Vocabulary, models achieve near-human ceiling performance, but their accuracy drops below human levels for possessives and declines even further for demonstratives. Ablation analyses further suggest these difficulties arise from limitations in perspective-taking and spatial reasoning in the models. Finally, instruction-based prompting improves model performance with possessives, bringing it close to human-level accuracy, while gains on demonstratives remain far below human performance, and primarily reflect enhanced perspective-taking rather than improved demonstrative use. Overall, these findings provide novel insights into the cognitive demands of perspectival words relative to vocabulary words for both humans and MLMs, and highlight the challenge of producing grammatical forms that require pragmatic reasoning and social cognition—capacities that remain limited in current NLP systems.

2 Related work

Prior research on NLP systems' pragmatic abilities has mostly focused on non-literal language use. Hu et al. (2023), for example, tested whether humans and models select pragmatic interpretations of utterances representing seven pragmatic phenomena (i.e., deceits, indirect speech, irony, conversational maxims, metaphor, humor and coherence). Sravanthi et al. (2024) released a Pragmatics Understanding Benchmark (PUB) including fourteen tests of four pragmatic phenomena (implicature, presupposition, metonymy and co-reference) using multiple-choice question answers. Findings from these studies suggest that language models demonstrate pragmatic language comprehension comparable to humans across multiple tasks. However, with the exception of co-reference (see also Zheng et al., 2021), all evaluated pragmatic phenomena are non-literal language uses, which limits pragmatics to the message level (cf. Carston, 2002). As a re-

sult, these model evaluations rest on a skewed view of pragmatics, casting doubts on models' communicative abilities once word-level pragmatics and face-to-face interaction are considered (Bisk et al., 2020; Rubio-Fernandez et al., 2024). Our work differs from these approaches in two key respects: by testing physical reference, we study word-level pragmatics and ground language use in a simulation of face-to-face interaction.

Using perspectival words requires perspective-taking, which is part of human social cognition. Language models have been extensively evaluated for their Theory of Mind skills, with mixed results (for review and discussion, see Hu et al., 2025). In human communication, the role of Theory of Mind has traditionally been limited to the sentence level (e.g., interpreting 'What a great friend' as a sarcastic remark). However, it has recently been argued that human communication also involves social micro-processes (Rubio-Fernandez et al., 2024). As shown by behavioral and computational work (Rubio-Fernandez et al., 2021; Jara-Ettinger and Rubio-Fernandez, 2022, 2024; Woensdregt et al., 2022), physical reference relies on social micro-processes. Our study is the first to compare humans and MLMs on this form of social cognition.

Our work is also relevant to a substantial body of research that develops MLMs to align with human dialogue in reference tasks, where participants share access to a set of referents and the speaker describes a target for the listener to identify (De Vries et al., 2017; Das et al., 2017). Prior studies show that models often rely on fewer or less effective referential expressions, but that incorporating human-like referential strategies in these contexts improves both task success and conversational quality (Shekhar et al., 2018; Dong et al., 2021; Testoni et al., 2024). We extend this line of research by introducing a controlled and interpretable framework for investigating how humans and models use perspectival words in simulated face-to-face interaction, providing a cognitive foundation for understanding such referential expressions in naturalistic communication.

3 Method

3.1 Tested Phenomena

Our experiment investigates word usage in humans and MLMs across three tasks: Vocabulary, Possessive, and Demonstrative tasks, which we hypothesize increase in production difficulty.

Vocabulary We predict that both humans and models will be able to identify the referent object and select the correct label. This is the easiest task because it does not require pragmatic grounding or perspective taking (Clerkin and Smith, 2022).

Possessive pronouns We predict humans will perform better than models, and both groups will perform worse than in the Vocabulary task. This task is harder because it requires pragmatic grounding (i.e., anchoring on the speaker) and perspective taking (i.e., monitoring whose preference is being expressed) (Levinson, 2006).

Demonstrative pronouns We predict humans will perform better than models, and both groups will perform worse than in the Possessive task. This is the hardest task because it requires not only pragmatic grounding and perspective taking, but also spatial reasoning (i.e., which objects are near vs far from the speaker) (Rubio-Fernandez, 2022).

3.2 Task Designs

Each experiment uses a multiple-choice format with four options, only one of which is correct (as shown in Figure 1). The experimental design systematically manipulates two independent variables: spatial arrangement (i.e., speaker-target distance) and perspective ('I prefer' vs 'You prefer'). We selected 128 objects that are reliably recognizable across models and designed two parallel task setups—one optimized for humans and one for MLMs—that present identical communicative scenarios but differ in how information is provided. This design choice is deliberate: by allowing each participant group to process information in their native modality, we preserve their ability to produce the correct referential expression while avoiding confounds (e.g., failure in object identification). We validated this dual design through systematic sanity checks that confirmed models perform better in their own tasks than in those for humans (see Appendix A).

Task setups for humans We built three separate tasks, with the same instructions: participants would see a series of visual scenes depicting two characters (a boy and a girl) standing on opposite sides of a table with four objects. Participants were asked to play the role of the speaker (who had a speech bubble) and select the correct message from among four options. The speaker was expressing either their own or the other character's preference

for one of the four objects (as indicated by their line of gaze).

In the Vocabulary task, four different objects were placed along the table. In the Possessive task, two identical pairs of objects were placed at either end of the table, with object-character proximity indicating ownership. In the Demonstrative task, four identical objects were placed along the table. Four versions of each visual display were generated, in a fully-crossed Speaker (Boy vs Girl) by Perspective ('I prefer' vs 'You prefer) design (see Fig.1). Object ownership (in the Possessive task) and distance to the speaker (closest vs furthest, in the Demonstrative task) were fully counterbalanced. In the Vocabulary task, all four possible responses revealed the same perspective (either 'I prefer...' or 'You prefer...') and each contained the name of one of the objects on the table (e.g., 'the plane,' 'the camera,' 'the rubber duck,' 'the boat'). In the Possessive and Demonstrative tasks, two responses expressed the speaker's perspective and another two the other agent's, each followed by either 'mine' vs 'yours,' or 'this one' vs 'that one,' in a fully-crossed design.

Task setups for MLMs We designed three tasks for MLMs, each comprising 6144 trials combining an image and a text prompt. Task images mirrored human stimuli but were adapted for model input: speech bubbles and gaze lines were removed, a green square was added to mark the target object, and speaker and gazer identity were included in the prompt. To ensure comparability with human tasks, we applied controlled augmentations to task-irrelevant features (e.g., object order; see Appendix D). We designed prompts that closely match human instructions (Appendix E.1). All models were tested in zero-shot settings with temperature set to zero to evaluate the knowledge learned from training. To ensure that performance truly reflects referential abilities rather than prompt-specific artifacts, we tested multiple prompt variants and confirmed that results remained consistent across formulations (see Appendix E.2 for variant comparisons).

3.3 Evaluating humans

Production tasks The Vocabulary task included 16 trials, and the Possessive and Demonstrative tasks included 32 trials each. The tasks were built using Qualtrics and administered to three separate groups of monolingual native English speakers recruited through Prolific (N=120 per task). To ensure that the original results replicated when the experiment was run within participants, a single task was built using 16 trials from each condition distributed in two lists of trial blocks (Vocabulary-Possessives-Demonstratives and Vocabulary-Demonstratives-Possessives) and administered to a different group of native English speakers (N=100) in a follow-up task (for details and results, see Appendix B).

3.4 Evaluating MLMs

MLMs We evaluated seven models (more details in Appendix C): (1) GPT-4o, a closed-source model with real-time multimodal reasoning capabilities (OpenAI, 2024); (2) Llama-4-Scout-17B-16E-Instruct (hereafter Llama-4-Scout), an open-source vision-language model integrating a visual encoder and language model (Meta AI, 2025); (3) Gemma-3-27B-it (Gemma-3), which tokenizes images for multimodal input (GemmaTeam, 2025); (4–5) InternVL 2.5-4B-MPO (InternVL 2.5-4B) and 8B-MPO (InternVL 2.5-8B), differing in scale, using a visual encoder and LLM bridged by an MLP projector (Chen et al., 2024); and (6–7) Ovis 2-4B and 8B, aligning visual and textual embeddings for enhanced reasoning (Lu et al., 2024). We used the LLM-as-a-Judge method (Zheng et al., 2023), with GPT-4o as the evaluator, and our thorough manual validation of response subsets confirmed complete reliability.

Ablation analysis To further examine the contribution of spatial and perspectival cues to reference production, we conducted an ablation analysis on the MLMs. Spatial cues were manipulated under three conditions: (i) *Spatial Cue Removal (Linguistic)*: ownership markers (for possessives) and proximity terms (for demonstratives) were omitted from the prompts; (ii) *Spatial Cue Removal (Visual)*: all objects were displayed using identical green squares; and (iii) *Spatial Cue Removal (Visual + Linguistic)*: both linguistic and visual cues were removed, such that ownership markers, proximity terms, and green squares were all removed. For perspectival cues, we implemented (iiii) *Perspectival Cue Removal (Linguistic)*: the speaker's identity and the gazer's identity were specified, while removing any explicit indication of the viewpoint to be adopted by the model. Further implementation details are provided in Appendix F.

In-context learning and Instruction-based prompting We here evaluated two widely used

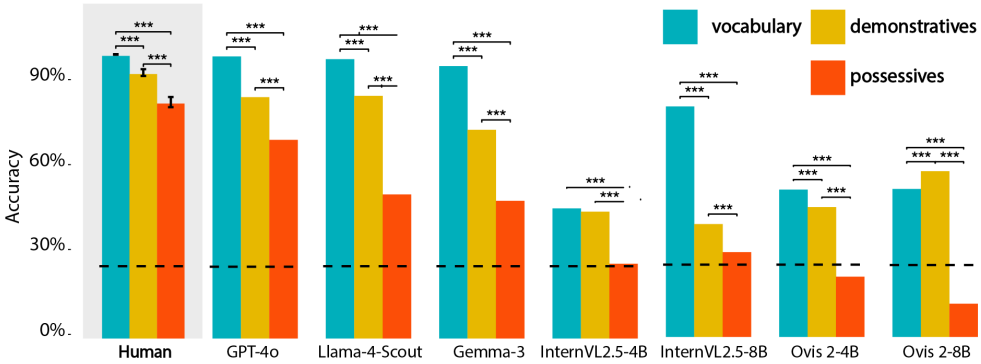

Figure 2: **Human and MLMs production task accuracy**. Error bars represent ±1 standard error across human participants. Since models were evaluated deterministically, no error bars are shown for MLMs. Significance (*) was established via logistic regression with Task as predictor, followed by post-hoc pairwise comparisons using estimated marginal means. The dashed line indicates the chance level.

prompting strategies to enhance MLMs' performance in tasks involving perspectival words (Wei et al., 2022; Qin and Eisner, 2021). First, in in-context learning, we presented four annotated exemplars—one for each answer type—prior to the task. Second, in instruction-based prompting, we crafted prompts that emphasized relevant linguistic cues (e.g., speaker identity, preference, ownership, spatial proximity) and encouraged the models to produce step-by-step reasoning along with their responses. All prompts were customized for the specific task. We further examined the error patterns that emerged under these strategies and compared them with those observed in the main experiment. More details are provided in Appendix G.

4 Results

4.1 Human vs. MLM Performance on Vocabulary and Perspectival Words

Figure 2 reports task accuracy for human participants and seven MLMs across the three tasks. We first established the human baseline and then evaluated MLM performance relative to it.

Despite overall high accuracy, human participants showed systematic performance differences across tasks: Vocabulary elicited the highest accuracy, followed by Possessives, with Demonstratives exhibiting the lowest performance. This reduction in task accuracy confirms that vocabulary and perspectival words pose different cognitive demands on language users. Within perspectival words, we interpret the performance difference as an effect of cognitive complexity, reflecting the semantics of these words (Levinson, 2006): possessive use relies on a binary object categorization (as either self-owned or other-owned), whereas demonstrative use requires distinguishing objects along a continuous spatial dimension (based on their relative proximity to the speaker). Supporting this interpretation, a comprehension version of our task revealed that possessives ('mine' vs. 'yours') show binary rating patterns—high for target objects, low for non-targets—while demonstratives ('this' vs. 'that') exhibit proximity-based gradients, with 'this' showing sharper distance sensitivity than 'that' (See Appendix H for full human comprehension experiments and results).

A similar hierarchy of task performance was observed across most models. Five of the seven tested models—GPT-4o, Llama-4-Scout, Gemma-3, InternVL2.5-8B-MPO, and Ovis2-4B—achieve the highest accuracy on Vocabulary, intermediate accuracy on Possessives, and the lowest accuracy on Demonstratives. The remaining two models (InternVL2.5-4B-MPO and Ovis2-8B) show broadly comparable trends, except for deviations in the Vocabulary task due to notably poor performance, primarily stemming from difficulties in using personal pronouns (see Appendix I for a detailed error analysis).

Next, we compared human and MLM performance across tasks. On the Vocabulary task, most models approach ceiling performance comparable to humans, indicating that vocabulary word selection poses minimal difficulty in current MLMs. In contrast, a substantial performance gap emerges in the Possessive and the Demonstrative tasks compared with humans. Even the top-performing commercial models (GPT-4o, Llama-4-Scout, and Gemma-3) achieve accuracy scores that remain

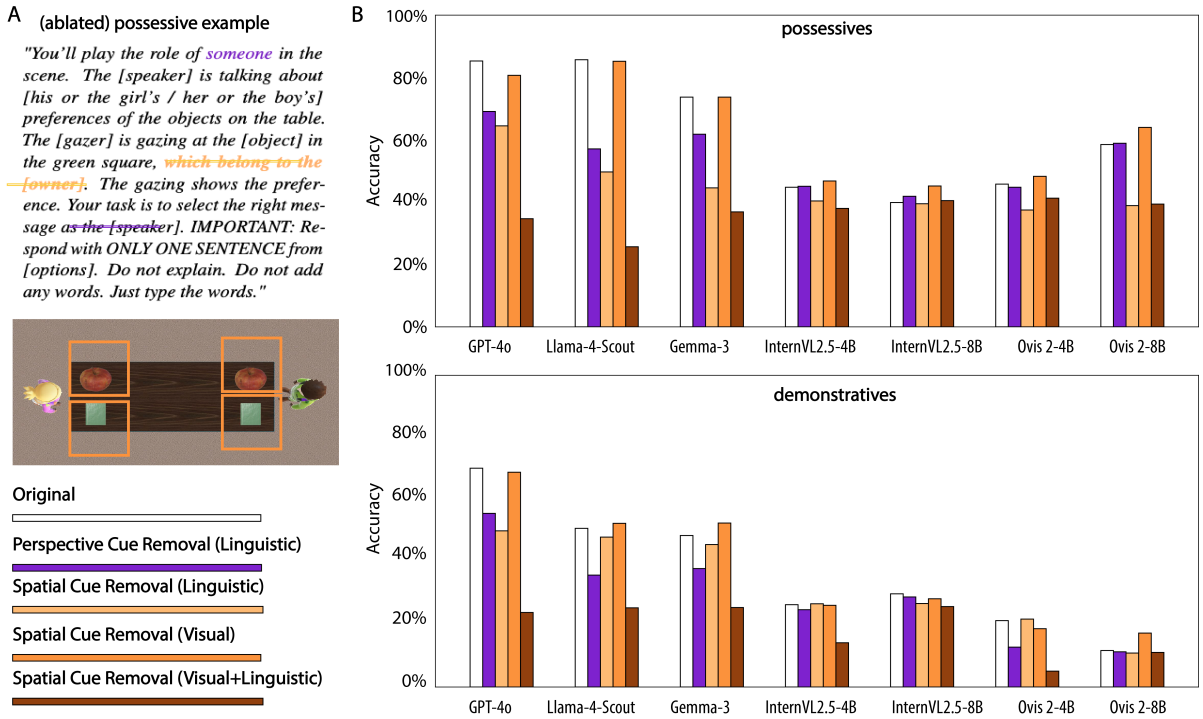

Figure 3: **Ablation analysis of perspectival and spatial cues in possessive and demonstrative reference production across MLMs.** Performance is shown for the original task and four ablation conditions: *Perspective Cue Removal (Linguistic)*, where model viewpoint (boy vs. girl) is removed from the prompt; *Spatial Cue Removal (Linguistic)*, where ownership markers (possessives) or proximity terms (demonstratives) are omitted; *Spatial Cue Removal (Visual)*, where all objects appear inside identical green squares; and *Spatial Cue Removal (Visual+Linguistic)*, combining both spatial manipulations. **(A)** Example of an ablated possessive trial (see Appendix F for complete ablation materials). **(B)** Top panel: Possessive ablated task accuracy; Bottom panel: Demonstrative ablated task accuracy.

well below human performance, despite humans themselves not attaining perfect accuracy on these tasks. For the Possessive task, humans achieve 93% accuracy compared to the best model (Llama-4-Scout) at 85.7%, while the gap is more pronounced for the Demonstrative task, with humans at 83% versus the best model (GPT-4o) at only 70.2%. The remaining models exhibit even greater deficits, with markedly lower accuracy on the Possessive task and near-chance performance on Demonstratives. These results reveal a clear divergence between MLMs and humans in perspectival word use, contrasting with their comparable performance on vocabulary word selection.

4.2 Where Models Fall Short: Ablations on Spatial and Perspective Cues

The results above show that MLMs struggle with perspectival words more than humans—but where exactly do models fall short? For humans, perspectival words are grammatical forms whose correct use depends on integrating two types of information: perspective cues (identifying the speaker and listener) and spatial cues (determining the objects' relative positions). We hypothesize that successfully combining these cues requires perspective-taking and spatial reasoning—cognitive capacities that likely account for the difficulties MLMs encounter relative to humans.

To identify the source of these deficits, we conducted an ablation study examining how models use and integrate perspective and spatial cues when producing possessive and demonstrative pronouns. We systematically manipulated the availability of perspective and spatial information to determine the origin of the observed difficulties (See an ablated example in Figure 3**A**).

We first examined the effect of removing perspective cues for both tasks (Figure 3**B**), comparing the original condition with the *Perspective Cue Removal (Linguistic)* condition. Models show a clear split in their ability to use perspective information. The top-performing models (GPT-4o, Llama-4-Scout, Gemma-3) demonstrate clear accu-

racy drops when perspective cues (indicating which role they play in the scene) were removed, suggesting their reliance on viewpoint information to select the correct perspectival term. In contrast, the remaining models showed minimal change—or even slight improvements—after the removal of perspective cues, indicating that they either fail to leverage these cues or are misled by them. These results reveal a divergence in perspective-taking capabilities: advanced models demonstrate sensitivity to shifts in speaker–listener perspectives comparable to humans, whereas other models appear insensitive to—or unable to properly integrate—perspectival information. This difference may relate to training data, as commercial models are likely exposed to more conversational contexts requiring perspective tracking across turns—a question that deserves further investigation.

We next evaluated how models use and integrate spatial cues from linguistic and visual inputs (Figure 3, comparing the original condition with three ablation conditions: *Spatial Cue Removal (Linguistic)*, *Spatial Cue Removal (Visual)*, and *Spatial Cue Removal (Visual+Linguistic)*). Removing spatial cues from the linguistic prompt (*Spatial Cue Removal (Linguistic)*) decreases task performance overall but produces markedly different effects across the two tasks. For Possessives, all models exhibit substantial accuracy drops, indicating their critical reliance on ownership markers in the prompt. For Demonstratives, however, only GPT-4o shows a clear decline. This trend aligns with the differing spatial semantics of these word classes: possessives rely on binary categorization (self-owned vs. other-owned) that can be explicitly specified linguistically, whereas demonstratives encode proximity-based gradients (near vs. far) that are more difficult to capture through linguistic description alone.

In contrast, removing visual spatial cues alone (*Spatial Cue Removal (Visual)*) has minimal impact on task accuracy: all models maintain—or even slightly improve—their performance, indicating that visual information provides little benefit on its own and can sometimes mislead when processed without linguistic support. Taken together, these results suggest that MLMs rely more heavily on linguistic than visual input, a trend consistent with previous findings on modality imbalance in these models (Deng et al., 2025; Wu et al., 2025).

In the combined removal condition (*Spatial Cue Removal (Visual+Linguistic)*), performance drops substantially more for both possessive and demonstrative tasks than when only linguistic cues are removed, demonstrating that visual information contributes to model performance—but only when paired with linguistic context. This reveals a critical bias: MLMs rely on linguistic input as the primary scaffold for spatial reasoning, with visual information refining (rather than independently establishing) spatial relationships. In contrast, humans integrate visual and linguistic spatial information in parallel during conversation (Kita and Özyürek, 2003; Rubio-Fernandez, 2022), highlighting a fundamental difference in multimodal processing that warrants further investigation in future model development.

4.3 Bridging the Gap: Toward More Human-like Referential Expressions

In the previous section, we showed a clear performance gap between humans and models in perspectival word use, even for the advanced MLMs (GPT-4o, Llama-4-Scout, and Gemma-3). To improve model alignment with human-level competence on the Possessive and Demonstrative tasks, we explored two widely used approaches with these models: (1) instruction-based prompting, where models reason through explicit pragmatic rules, and (2) in-context learning, where models learn from solved examples (see implementation details in Appendix G). Figure 4**A** presents accuracy and error patterns for possessives and demonstratives, comparing original performance with in-context and instruction-based prompting conditions.

For the Possesive task, instruction-based prompting substantially enhance performance for all three models, with explicit prompts directing models to consider who is speaking, who expresses a preference, and who owns the object. This approach achieves improvements over the original condition and approaches human-level accuracy. In contrast, in-context learning—providing solved examples without explicit reasoning—proves markedly less effective: only GPT-4o shows a significant improvement, while the other two models exhibit notable performance decreases. Error analysis (Figure 4**B**) shows that gains from instruction-based prompting primarily stem from reductions in possessive pronoun errors, reflecting improved understanding of ownership distinctions when explicitly guided.

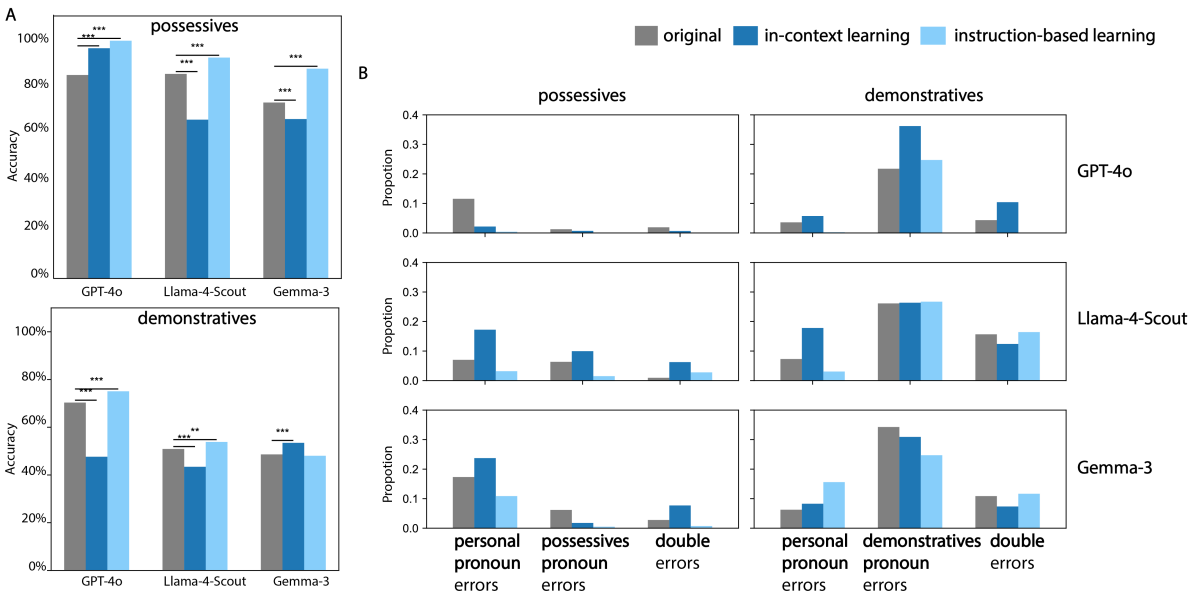

Figure 4: **Effect of prompt approaches on MLM performance and error patterns. A:** Task accuracy for top-performing models (GPT-4o, Llama-4-Scout, Gemma-3) across three prompt conditions: original zero-shot, in-context learning with examples, and instruction-based prompting. Significance (*) was established via logistic regression analysis with response correctness as outcome and Prompt Approach as predictor, followed by estimated marginal means pairwise comparisons. **B:** Error type distributions across three conditions

However, for the Demonstrative task, even with similar prompts plus additional guidance on the referent's distance to the speaker, models continue to struggle. Only GPT-4o and Llama-4-Scout show marginal improvements, and their overall performance remains well below human levels. In-context learning generally yields worse results, except for Gemma-3, which exhibits a modest improvement over the original condition. Error analysis (Figure 4) indicates that performance gains mainly reflect improvements in perspective-taking (reductions in personal pronoun errors) rather than enhanced spatial reasoning (reductions in demonstrative pronoun errors). Demonstrative pronoun errors persist, highlighting that while perspective-taking can be improved through explicit instructional prompts, the spatial reasoning required for demonstrative use remains a major challenge for current MLMs.

5 Discussion

In this work, we propose a unified framework for directly comparing the usage of three word classes in referential communication. Our results reveal a clear difficulty hierarchy: Vocabulary < Possessives < Demonstratives. Vocabulary words rely on stable object-label associations, possessive pronouns require perspective-taking to anchor reference to the speaker's viewpoint, and demonstratives additionally demand fine-grained spatial reasoning for proximity-based distinctions.

Critically, this hierarchy is far more pronounced for MLMs than humans, revealing fundamental limitations in perspective-taking and spatial reasoning in these models. Instruction-based prompting substantially improves possessive production, bringing model accuracy near human levels through explicit pragmatic scaffolding. However, demonstratives show no comparable gains—even with explicit spatial guidance, performance remains well below the human benchmark. This asymmetry suggests that while perspective-taking can be enhanced through instruction, spatial reasoning deficits represent a persistent challenge for current MLMs.

We hypothesize that successful demonstrative use may require embodied, interactive communication in physical environments (Peeters and Özyürek, 2016)—experiences through which humans, but not current MLMs, acquire spatial language. Future work should investigate whether models trained with richer spatial interaction data or equipped with explicit spatial reasoning modules can bridge this persistent gap in referential communication. In doing so, future studies could provide insights into MLMs' social micro-processes (Rubio-Fernandez et al., 2024).

6 Limitations

Our study has several methodological limitations. First, we employed a simple multiple-choice task to facilitate direct comparison between models and humans. Although this approach offers clear metrics, we acknowledge it does not capture the full complexity of perspectival word usage in face-to-face communication (Holler, 2025). Second, we only tested two strategies for improving models' production of human-like perspectival words. Other methods remain unexplored, such as persona-based prompting (Tan et al., 2024), hyperparameter optimization (Oliver and Wang, 2024), and advanced reasoning frameworks (Wei et al., 2022; Yao et al., 2023), which future work can further investigate. Third, due to computational resource limitations, we tested only four small models (Interns and Ovis <10B) in addition to models accessed via APIs. Finally, we evaluated models exclusively on English language materials, with our task design being based on English perspectival word usage. We acknowledge that perspectival word usage varies across languages and cultures (Rubio-Fernandez, 2022; Jara-Ettinger and Rubio-Fernandez, 2024). Future studies should examine these phenomena across diverse linguistic and cultural contexts.

7 Ethical considerations

Our research followed ethical guidelines throughout. We recruited monolingual native English speakers through Prolific, obtained approval from our institutional review board, and kept all participant data anonymous and secure. Understanding what multimodal language models can and cannot do matters for researchers and the public alike. Our work offers new ways to test how different these models are from humans in handling perspectival words that require pragmatics and social cognition. We hope this approach will increase researchers' interest in evaluating human-centric capabilities of these models and leads to improvements that make them more useful for everyday conversations. .

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.

Jacob Andreas. 2022. Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2):379–400.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Robyn Carston. 2002. *Thoughts and utterances: The pragmatics of explicit communication*. John Wiley & Sons.

Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Elizabeth M Clerkin and Linda B Smith. 2022. Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences*, 119(18):e2123239119.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876.

Tianai Dong, Alberto Testoni, Luciana Benotti, Raffaella Bernardi, et al. 2021. Visually grounded follow-up questions: a dataset of spatial questions which require dialogue history. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 22–31. ACL.

Katherine Elkins and Jon Chun. 2020. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 5(2).

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

GemmaTeam. 2025. Gemma 3 technical report. https://arxiv.org/abs/2503.19786. ArXiv preprint arXiv:2503.19786.

Judith Holler. 2025. Facial clues to conversational intentions. *Trends in Cognitive Sciences*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Reevaluating theory of mind evaluation in large language models. *arXiv preprint arXiv:2502.21098*.

Julian Jara-Ettinger and Paula Rubio-Fernandez. 2022. The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological review*, 129(6):1394.

Julian Jara-Ettinger and Paula Rubio-Fernandez. 2024. Demonstratives as attention tools: Evidence of mentalistic representations within language. *Proceedings of the National Academy of Sciences*, 121(32):e2402068121.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. *arXiv preprint arXiv:2004.03066*.

Sotaro Kita and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language*, 48(1):16–32.

Stephen C Levinson. 2006. Deixis. *The handbook of pragmatics*, pages 97–121.

Stephen C Levinson. 2024. *The dark matter of pragmatics: Known unknowns*. Cambridge University Press.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. *arXiv preprint arXiv:2310.19619*.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: 2025-05-12.

Michael Oliver and Guan Wang. 2024. Crafting efficient fine-tuning strategies for large language models. *arXiv preprint arXiv:2407.13906*.

OpenAI. 2024. GPT-4o Technical Report. Technical report, OpenAI.

David Peeters and Aslı Özyürek. 2016. This and that revisited: A social and multimodal approach to spatial demonstratives.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Paula Rubio-Fernandez. 2021. Pragmatic markers: The missing link between language and theory of mind. *Synthese*, 199(1):1125–1158.

Paula Rubio-Fernandez. 2022. Demonstrative systems: From linguistic typology to social cognition. *Cognitive Psychology*, 139:101519.

Paula Rubio-Fernandez. 2024. Cultural evolutionary pragmatics: Investigating the codevelopment and coevolution of language and social cognition. *Psychological Review*, 131(1):18.

Paula Rubio-Fernandez, Marlene D Berke, and Julian Jara-Ettinger. 2024. Tracking minds in communication. *Trends in Cognitive Sciences*.

Paula Rubio-Fernandez, Francis Mollica, and Julian Jara-Ettinger. 2021. Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*, 150(3):583.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Ravi Shekhar, Tim Baumgärtner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernandez. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1218–1233, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *arXiv preprint arXiv:2401.07078*.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. Impli: Investigating nli models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388.

Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. 2024. Phantom: Persona-based prompting has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

Alberto Testoni, Barbara Plank, and Raquel Fernández. 2024. Racquet: Unveiling the dangers of overlooked referential ambiguity in visual llms. *arXiv preprint arXiv:2412.13835*.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4673–4686.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Marieke S Woensdregt, Julian Jara-Ettinger, and Paula Rubio-Fernandez. 2022. Language universals rely on social cognition: Computational models of the use of this and that to redirect the receiver's attention. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. 2025. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. Demystifying clip data. In *12th International Conference on Learning Representations (ICLR)*, Vienna, Austria. Hybrid.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085.

Appendix

A Sanity-check: MLMs' Performance on Object Recognition and Speaker/Gazer Identification

In the human experiments, the speaker was indicated with a speech bubble and their preference with an intermittent gaze line connecting the eyes to the target object. In the MLM experiments, these cues were instead conveyed in text prompts, with the only visual cue being a green box around the target object.

We did not use the human-adapted materials for MLMs because models may not interpret speech bubbles or gaze lines as humans do (See more details in the **Human Experiment Setting Check** below). To avoid confounds, we provided information primarily in text, supplemented with a simple visual cue. Conversely, humans did not receive written descriptions of speaker or gazer identity, as this would burden working memory compared to perceiving the cues directly in the display (e.g., in some trials, participants would have to remember that they boy is the speaker and the girl is the gazer while looking at a bare display, rather than receiving that information from the speech bubble and line of gaze in the actual display).

In short, we treated visual information as the native modality for humans and written text as the native modality for models. To verify comparability, we first ran a sanity check by testing MLMs on the original human stimuli.

Human Experiment Setting Check We tested MLMs on the human experiment stimuli by probing each image three times: object in the green square, speaker, and gazer. As shown in Table 1, MLMs struggled—especially with object recognition, indicating that the human-optimized stimuli were not well-suited to the models' capacities.

MLM Experiment Setting Check We next tested MLMs on a subsample of the MLM stimuli, ensuring that each of the 128 objects was queried once. Using the same three recognition tasks (object, speaker, gazer), we obtained the results in Table 2. MLMs robustly recognized all scene elements, confirming that the adapted stimuli support parallel task designs differing only in modality (visual for humans, text+visual for models). This ensures that divergences in behavior reflect word use rather than confounds in scene understanding.

Test Type	Model	Accuracy
Object Recognition	gpt-4o	0.375
	Llama-4-Scout	0.438
	Gemma-3	0.375
	InternVL2_5-4B-MPO	0.438
	InternVL2_5-8B-MPO	0.375
	Ovis2-8B	0.562
Gazer Identification	gpt-4o	0.875
	Llama-4-Scout	0.938
	Gemma-3	0.938
	InternVL2_5-4B-MPO	0.750
	InternVL2_5-8B-MPO	0.750
	Ovis2-4B	0.562
	Ovis2-8B	0.750
Speaker Identification	gpt-4o	1.000
	Llama-4-Scout	0.938
	Gemma-3	0.812
	InternVL2_5-4B-MPO	0.750
	InternVL2_5-8B-MPO	0.750
	Ovis2-4B	0.562
	Ovis2-8B	0.875

Table 1: Overall accuracies of MLMs in the Human Experiment Setting.

Test Type	Model	Accuracy
Object Recognition	gpt-4o	1.000
	Llama-4-Scout	0.984
	Gemma-3	0.977
	InternVL2_5-4B-MPO	0.969
	InternVL2_5-8B-MPO	0.961
	Ovis2-4B	1.000
	Ovis2-8B	0.992
Gazer Identification	gpt-4o	0.906
	Llama-4-Scout	0.992
	Gemma-3	1.000
	InternVL2_5-4B-MPO	0.992
	InternVL2_5-8B-MPO	0.953
	Ovis2-4B	1.000
	Ovis2-8B	0.961
Speaker Identification	gpt-4o	1.000
	Llama-4-Scout	1.000
	Gemma-3	1.000
	InternVL2_5-4B-MPO	0.984
	InternVL2_5-8B-MPO	1.000
	Ovis2-4B	1.000
	Ovis2-8B	1.000

Table 2: Overall accuracies of MLMs in the MLM Experiment Setting.

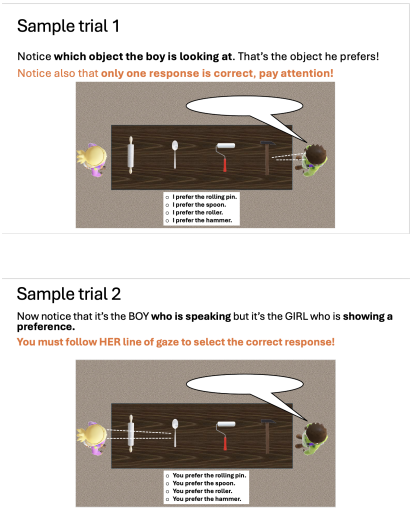

Figure 5: Sample trials in the Vocabulary task for humans.

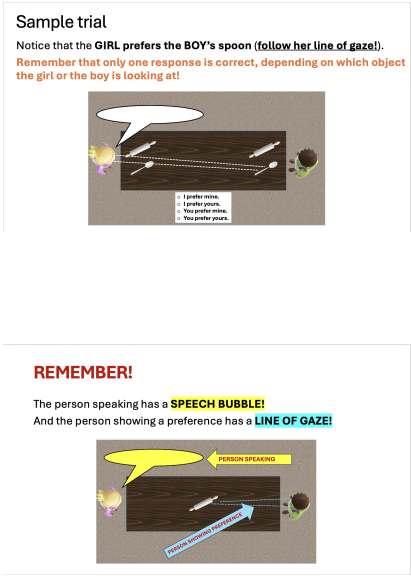

Figure 6: Sample trials in the Possessive task for humans.

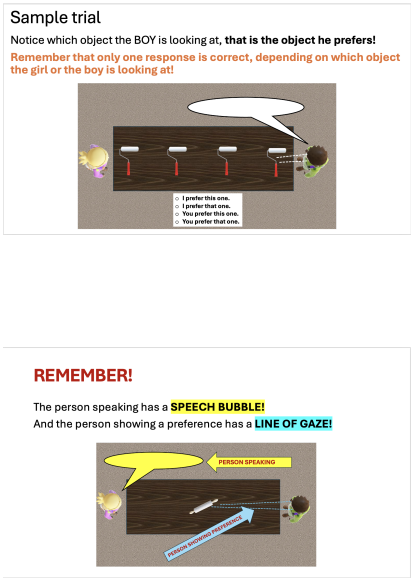

Figure 7: Sample trials in the Demonstrative task for humans.

B Tasks for Humans

Informed consent was obtained from all participants prior to the start of the task. All tasks had an average duration of 6 minutes for more accurate across-task comparisons, and to avoid participant fatigue. Participants were paid 9 GBP per hour, which was evaluated as a "Good rate" by Prolific.

In the following sections we provide further details of the tasks described in the main text.

B.1 Production Task

Participants were instructed to play the role of the speaker and complete the speech bubble by selecting the correct message (from among 4 options) given the object preference shown by one of the characters through their line of gaze. Participants were shown two sample trials in each task.

B.2 Follow-up Production Task

We aimed to replicate the results of the Production tasks using the same instructions and materials in a within-participants design.

B.3 Production and Follow-up Production Task Results

The production task and follow-up production task accuracy are shown in Table 3. We excluded one participant from the production vocabulary task and one from the follow-up production vocabulary task due to random performance. No significant difference in accuracy was detected across the first and follow-up runs of the three tasks.

Task Type	Production	Follow-up	*p*-value
Vocabulary	99.84%	99.43%	0.0764
Possessive	93.44%	92.44%	0.2037
Demonstrative	83.00%	82.19%	0.4908

Table 3: Comparison of production and follow-up production task accuracy.

C Model Specifications

We conducted all experiments using GPT-4o, Llama-4-Scout, and Gemma-3 through their respective APIs, locally hosted the remaining models on a standard GPU setup, and completed the full evaluation suite in approximately 40 hours.

C.1 GPT-4o

OpenAI's omni transformer jointly encodes text, vision, and audio in a single network (OpenAI, 2024). The public API exposes a 128 k token context window.

C.2 Llama-4-Scout 17B (16E)

A 17B active-param mixture-of-experts (MoE) auto-regressive LM with early-fusion MetaCLIP vision branch (Meta AI, 2025; Xu et al., 2024). It is advertised as a 10M-token window.

C.3 Gemma 3 (27B)

Decoder-only LM distilled from a larger teacher model and paired with a frozen 400 M-parameter SigLIP encoder at 896×896 px, adopting a Pan and Scan (P&S) method for flexible resolutions (GemmaTeam, 2025), providing a 128k token window.

C.4 InternVL 2.5-4B / 8B (MPO)

InternViT-300M-448px encoder is coupled with different pretrained LLMs (Qwen2.5-3B-Instruct for InternVL 2.5 4B and internlm2_5-7b-chat for InternVL 2.5 8B) via an MLP projector (Chen et al., 2024). Adopting Mixed Preference Optimization (MPO) on the Multi-Modal Preference (MMPR) dataset, InternVL 2.5 is the first open model to pass 70 % MMMU.

C.5 Ovis 2-4B / 8B

Ovis 2-4B and 8B are separately based on Qwen2.5-3B-Instruct/ Qwen2.5-7B-Instruct LLM, with implementing structural embedding alignment by inserting a learnable visual-embedding table inside an aimv2-huge-patch14-448 vision encoder, mirroring word-token structure (Lu et al., 2024). All sizes support videos and multi-image processing.

D Trial Augmentation and Matching with Human Evaluation

In each model experiment, we augmented the trials to 6144 while retaining alignment to the human evaluations by varying task-irrelevant factors. To compare with human experiment results, we obtained the MLM experiment accuracy as the main metric. The response time of MLMs can be influenced by network conditions or API provider, therefore it is not selected as the primary measurement of MLM performance.

Vocabulary Task In both human and model evaluations, we grouped the 128 test objects into 32 groups. In the model evaluation, we showed models each object group in half of the possible orders on the table, randomly sampled, which rendered the trials = 6144 = 32 (groups) $\times$ 4 (possible targets) $\times$ 12 (possible orders) $\times$ 2 (possible speakers) $\times$ 2 (possible perspectives). In the human evaluation, the trial number is 16 = 4 (groups) $\times$ 1 (possible targets) $\times$ 2 (possible speakers) $\times$ 2 (possible perspectives).

Possessive Task In model evaluation, for each object group, we had 6 possible two-object combinations, and we placed the objects in two spatial arrangements, rendering the trials = 6144 = 32 (groups) $\times$ 4 (possible targets) $\times$ 6 (two-object combinations) $\times$ 2 (spatial arrangements) $\times$ 2 (possible speakers) $\times$ 2 (possible gazers). In the human evaluation, the trial number is 32 = 4 (groups) $\times$ 2 (two-object combinations) $\times$ 1 (possible targets) $\times$ 2 (possible speakers) $\times$ 2 (possible perspectives).

Demonstrative Task In the model demonstrative task, we made 6 scenes in which the two characters stood on either side of the table. The model task trial number is 6144 = 128 (objects) $\times$ 2 (possible targets) $\times$ 6 (scenes) $\times$ 2 (possible speakers) $\times$ 2 (possible gazers). In the human evaluation, the trial number is 32 = 8 (objects) $\times$ 2 (possible speakers) $\times$ 2 (possible perspectives).

Object-Group Effects on Experiment Accuracy We used chi-squared test (table 4) to see if models performance in the experiments varied with the objects appeared in the trial images. We do not find significant effect in Possessive or Demonstrative experiments. However, we observed significant results in Vocabulary experiment, which is likely due to the task demanding the model to recognize the object and include the object name in the final answer.

Experiment	Model	χ^2 (df)	p-value
Vocabulary	GPT-4o	173.116 (31)	<0.001
	Llama-4-Scout	126.189 (31)	<0.001
	Gemma-3	239.487 (31)	<0.001
	InternVL2_5-4B-MPO	93.080 (31)	<0.001
	InternVL2_5-8B-MPO	600.099 (31)	<0.001
	Ovis2-4B	71.249 (31)	<0.001
	Ovis2-8B	56.744 (31)	0.003
Possessive	GPT-4o	26.220 (31)	0.711
	Llama-4-Scout	35.969 (31)	0.247
	Gemma-3	25.112 (31)	0.763
	InternVL2_5-4B-MPO	21.418 (31)	0.900
	InternVL2_5-8B-MPO	15.860 (31)	0.989
	Ovis2-4B	15.021 (31)	0.993
	Ovis2-8B	6.984 (31)	1.000
Demonstrative	GPT-4o	14.442 (31)	0.995
	Llama-4-Scout	32.695 (31)	0.384
	Gemma-3	27.129 (31)	0.666
	InternVL2_5-4B-MPO	20.342 (31)	0.928
	InternVL2_5-8B-MPO	8.989 (31)	1.000
	Ovis2-4B	40.622 (31)	0.116
	Ovis2-8B	36.492 (31)	0.229

Table 4: Chi-square test results for object-group effects on model accuracy across vocabulary, possessive, and demonstrative experiments.

E Images and Prompts for MLM Experiment

In the following sections we provide our sample images prompts used in the MLM experiments. The square brackets in the prompts signify dynamic variables that change by trial.

E.1 Main Experiment

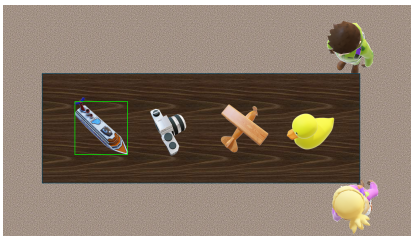

Figure 8: Sample image in the Vocabulary task for models.

Vocabulary Task

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

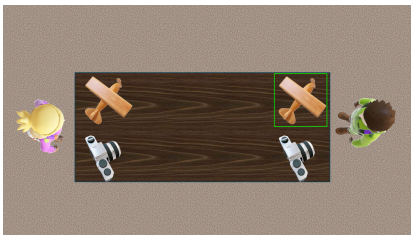

Figure 9: Sample image in the Possessive task for models.

Possessive Task

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square,* ***which belong to the [owner]****. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

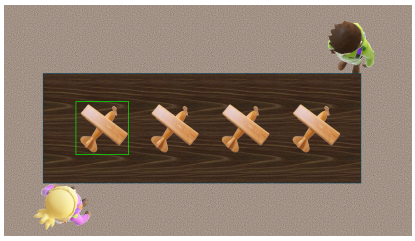

Figure 10: Sample image in the Demonstrative task for models.

Demonstrative Task

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square,* ***which is close to the [person near target]****. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

E.2 Prompt Variation Experiment

We provided three prompt variations to assess the robustness of observed effect. The assessment results of the context-first, condensed, and simple prompts are shown in table 5, 6, 7. In general, we found models' hierarchical performance across the three experiments under those prompt variations remains stable. No prompt variation shows substantial advantage over the original prompt over three experiment. The only exception was the condensed prompt improved InternVL2.5-4B-MPO and Ovis-4B's performance on Vocabulary task, while it also reduced several models' accuracy on Possessive and Demonstrative task.

Context-First

> *"In this scene, the [speaker] is talking about their own or the other person's preferences. The [gazer] is looking at the*

Experiment	Model	Accuracy
Vocabulary	GPT-4o	0.999
	Llama-4-Scout	0.997
	Gemma-3	0.990
	InternVL2_5-4B-MPO	0.483
	InternVL2_5-8B-MPO	0.871
	Ovis2-4B	0.520
	Ovis2-8B	0.516
Possessive	GPT-4o	0.826
	Llama-4-Scout	0.742
	Gemma-3	0.721
	InternVL2_5-4B-MPO	0.402
	InternVL2_5-8B-MPO	0.432
	Ovis2-4B	0.436
	Ovis2-8B	0.428
Demonstrative	GPT-4o	0.633
	Llama-4-Scout	0.444
	Gemma-3	0.480
	InternVL2_5-4B-MPO	0.237
	InternVL2_5-8B-MPO	0.319
	Ovis2-4B	0.236
	Ovis2-8B	0.242

Table 5: Accuracy of models under the **Context-First** prompt variation.

object in the green square, which shows the preference. Now take the role of the [speaker]. From [options], choose ONE sentence. Write only that sentence."

Condensed

"Act as the [speaker]. The [speaker] discusses object preferences. The [gazer] looks at the object in the green square, showing the preference. Choose ONE sentence from [options]. Respond with only that sentence."

Simple

"You are the [speaker]. The [gazer] looks at the object in the green box, which means they like it. Choose ONE sentence from [options]. Write only that sentence."

Experiment	Model	Accuracy
Vocabulary	GPT-4o	0.996
	Llama-4-Scout	0.973
	Gemma-3	0.965
	InternVL2_5-4B-MPO	0.952
	InternVL2_5-8B-MPO	0.974
	Ovis2-4B	0.831
	Ovis2-8B	0.514
Possessive	GPT-4o	0.721
	Llama-4-Scout	0.728
	Gemma-3	0.702
	InternVL2_5-4B-MPO	0.409
	InternVL2_5-8B-MPO	0.443
	Ovis2-4B	0.435
	Ovis2-8B	0.622
Demonstrative	GPT-4o	0.475
	Llama-4-Scout	0.547
	Gemma-3	0.490
	InternVL2_5-4B-MPO	0.220
	InternVL2_5-8B-MPO	0.243
	Ovis2-4B	0.262
	Ovis2-8B	0.246

Table 6: Accuracy of models under the **Condensed** prompt variation.

Experiment	Model	Accuracy
Vocabulary	GPT-4o	0.996
	Llama-4-Scout	0.961
	Gemma-3	0.954
	InternVL2_5-4B-MPO	0.630
	InternVL2_5-8B-MPO	0.971
	Ovis2-4B	0.516
	Ovis2-8B	0.513
Possessive	GPT-4o	0.672
	Llama-4-Scout	0.568
	Gemma-3	0.551
	InternVL2_5-4B-MPO	0.430
	InternVL2_5-8B-MPO	0.475
	Ovis2-4B	0.452
	Ovis2-8B	0.553
Demonstrative	GPT-4o	0.530
	Llama-4-Scout	0.376
	Gemma-3	0.312
	InternVL2_5-4B-MPO	0.247
	InternVL2_5-8B-MPO	0.277
	Ovis2-4B	0.288
	Ovis2-8B	0.233

Table 7: Accuracy of models under the **Simple** prompt variation.

F Ablation Experiment

In this section, we showcase our images and prompts used for the ablation experiment. The bold text highlights the changed text in the ablated prompt and the struck-through text indicates the text was deleted in the ablated prompt.

Possessive Task Ablation Spatial cue removal (linguistic only, removal of ownership information): the image used was the same as in the main Possessive experiment. The prompt used shows below:

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square~~, which belongs to the [owner]~~. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

Spatial cue removal (visual only): the prompt used was the same as in the main Possessive experiment. See figure 11 for the image used in this ablation.

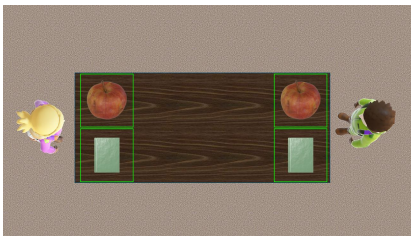

Figure 11: Sample image in the Possessive ablation experiment, spatial cue removal (visual only). We removed the visual cue by adding green squares to all objects

Spatial cue removal (linguistic and visual): See figure 12 for the image used in this ablation. The prompt used shows below:

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. ~~The [gazer] is gazing at the [object] in the green square, which belongs to the [owner]~~. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

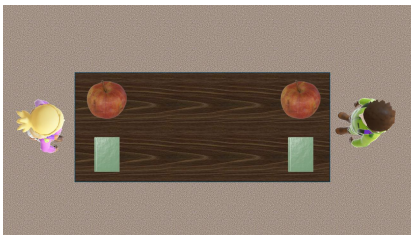

Figure 12: Sample image in the Possessive ablation experiment, spatial cue removal (both linguistic and visual). We removed the visual cue by removing all green squares of all objects

Perspective cue removal The image used was the same as in the main Possessive experiment. The prompt used shows below:

> *"You'll play the role of the **someone** in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square, which belong to the [owner]. The gazing shows the preference. Your task is to select the right message~~ as the [speaker]~~. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

Demonstrative Task Ablation Spatial cue removal (linguistic only, removal of object proximity information): the image used was the same as in the main Demonstrative experiment. The prompt used shows below:

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square~~, which is close to the [person near the target]~~. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain.*

Do not add any words. Just type the words."

Spatial cue removal (visual only): the prompt used was the same as in the main Demonstrative experiment. See figure 13 for the image used in this ablation.

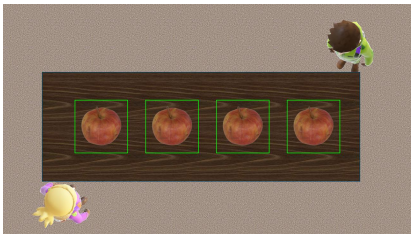

Figure 13: Sample image in the Demonstrative ablation experiment, spatial cue removal (visual only). We removed the visual cue by adding green squares to all objects

Spatial cue removal (linguistic and visual): See figure 14 for the image used in this ablation. The prompt used shows below:

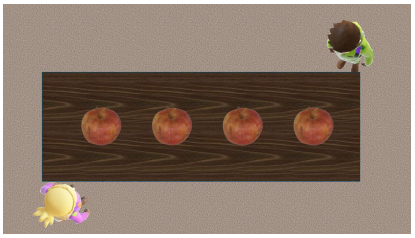

Figure 14: Sample image in the Demonstrative ablation experiment, spatial cue removal (both linguistic and visual). We removed the visual cue by removing all green squares of all objects

"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. ~~The [gazer] is gazing at the [object] in the green square, which is close to the [person near the target]~~. The gazing shows the preference. Your task is to select the right message as the [speaker]. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."

Perspective cue removal The image used was the same as in the main Possessive experiment. The prompt used shows below:

*"You'll play the role of **someone** in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square, which is close to the [person near target]. The gazing shows the preference. Your task is to select the right message ~~as the [speaker]~~. IMPORTANT: Respond with ONLY ONE SENTENCE from [options]. Do not explain. Do not add any words. Just type the words."*

G Instruction-based Prompting and In-context Learning Experiments

G.1 In-context Learning

We first present the models four examples capturing all personal/ possessive/ demonstrative pronoun combinations with correct reference solutions. The prompt for the examples is:

> *"In this picture, the [speaker] is speaking and the [gazer] is gazing at the object in the green square, and the [speaker] says '[correct reference]'"*

After the example presentations, the models would receive the same image and prompt combination as in the zero-shot experiments.

G.2 Instruction-based prompting

We sent the prompt including task image plus hints as input to the models and obtained their answer. We use bold text to highlight the added hints to the original prompts.

Possessive Task

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square, which belongs to the [owner]. The gazing shows the preference.* ***Your task is to produce a reference in this format: "{personal pronoun} prefer {possessive pronoun}" as the [speaker]. Choose the correct option based on: Who is speaking (you as the [speaker]); Who is expressing preference (the gazer: the [gazer]); Who owns the object being preferred (the owner: the [owner]). Use the instructions form before to select the right message as the [speaker].*** *IMPORTANT: Respond with ONLY ONE SENTENCE from [options]."*

Demonstrative Task

> *"You'll play the role of the [speaker] in the scene. The [speaker] is talking about [his or the girl's / her or the boy's] preferences of the objects on the table. The [gazer] is gazing at the [object] in the green square, which belong to the [owner]. The gazing shows the preference.* ***Your task is to produce a reference in this format: "{personal pronoun} prefer {demonstrative pronoun}" as the [speaker]. Choose the correct option based on: Who is speaking (you as the [speaker]); Who is expressing preference (the gazer: the [gazer]); Who is close to the object being preferred (the person being close to: the [person near target]). Use the instructions form before to select the right message as the [speaker].*** *IMPORTANT: Respond with ONLY ONE SENTENCE from [options]."*

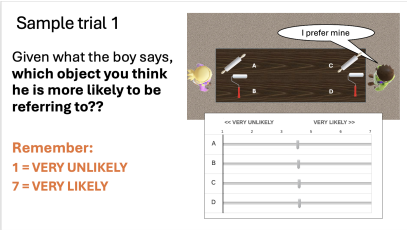

Figure 15: Sample trial in the Possessive Comprehension task.

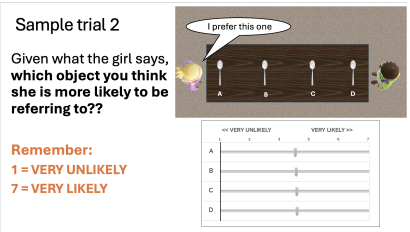

Figure 16: Sample trial in the Demonstrative Comprehension task.

H Human Comprehension Task

To better understand the human error patterns, a comprehension version of the Possessive and Demonstrative tasks was built where one of the two characters expressed either their own or the other's preference for an object (e.g., 'I prefer yours' or 'You prefer this one') and participants had to indicate on a 1-7 scale the probability that each object on the table (labeled A-D) was the referent. The Comprehension task included 8 Possessives trials and 8 Demonstratives trials, individually randomized, and was administered to a new group of native English speakers (N=100). Figure 15 and 16 shows two samples from Possessive comprehension and Demonstrative comprehension tasks, respectively.

As the Figure 17 shows, possessive pronouns 'mine and 'yours' display a clear binary distribution, with target objects receiving high likelihood ratings and non-target objects consistently being rated low. In contrast, demonstrative interpretation follows a graded pattern based on the target's distance from the speaker. For 'this,' the objects closer to the speaker receive the highest likelihood ratings, while for 'that,' the objects farther from the speaker are rated highest. Interestingly, we also observe subtle distinctions within demonstratives: 'this' exhibits a sharper proximity-based gradient (indicating a more specific meaning), whereas 'that' shows more similar ratings across intermediate and distant positions (indicating a vaguer meaning).

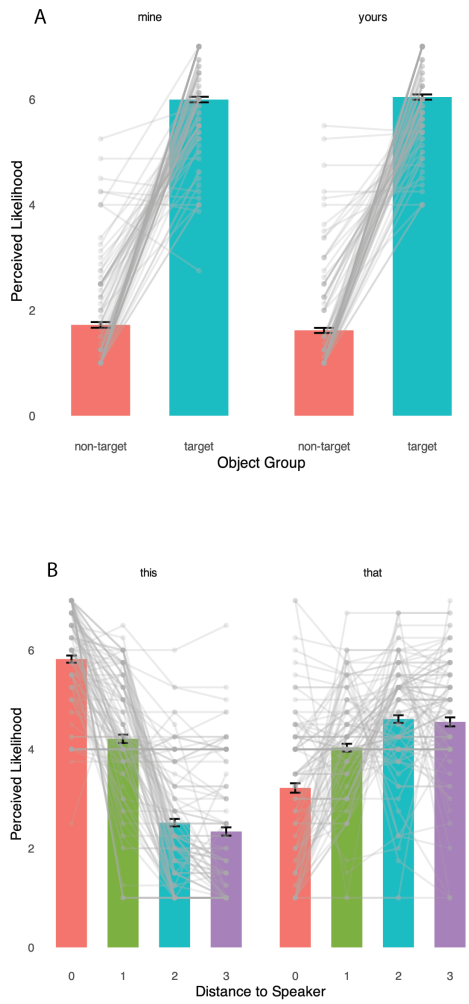

Figure 17: **(A)** Perceived probability ratings in the Possessive comprehension task, comparing target object pair versus non-target object pair. **(B)** Perceived probability ratings in the Demonstrative comprehension task as a function of object distance from speaker, ranging from 0 (nearest) to 3 (farthest).

I Error Pattern of Human and MLMs

Figure 18 presents response distributions for humans and all models across all three tasks. For the Vocabulary task, humans and the three advanced models (GPT-4o, Llama-4-Scout, and Gemma-3) demonstrate near-perfect accuracy, while other models show errors primarily in personal pronoun selection. In the Possessive task, humans make relatively few errors, whereas even advanced models occasionally make incorrect personal pronoun choices, suggesting limitations in perspective-taking abilities that may account for their performance gap relative to humans.

Recall that the Demonstrative task presents the greatest challenge for both humans and models. Interestingly, despite quantitative differences in performance, we observe a convergence in error types between humans and the three advanced models, with the most frequent errors occurring in demonstrative pronoun use, followed by errors in personal pronoun selection or combined errors. We take this finding as evidence of the intrinsic demands that the Demonstrative task poses compared to Vocabulary and Possessive tasks to both humans and models. We do not analyze the remaining models in detail, as they perform at near-chance level and exhibit error patterns that are less interpretable.

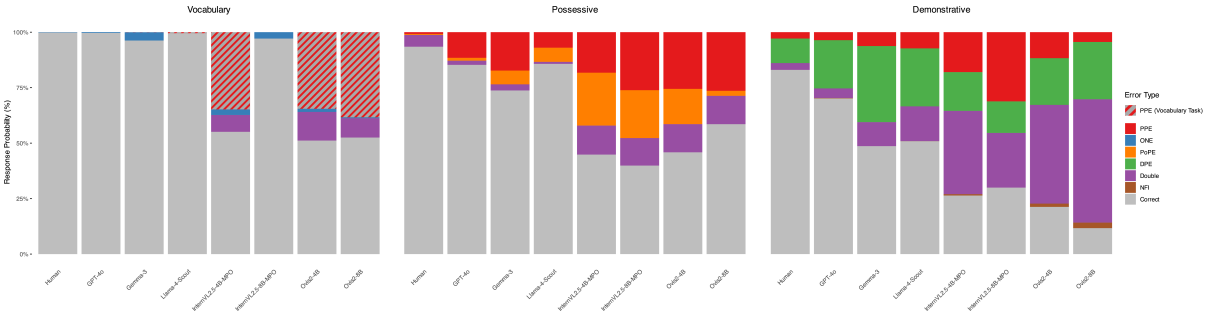

Figure 18: Response probability by task type with respect to **PPE**: personal pronoun error; **ONE**: object noun error; **PoPE**: possessive pronoun error; **DPE**: demonstrative pronoun error; **Double**: both personal and possessive/demonstrative pronoun error; **NFI**: not following instructions (only in MLM evaluations), and **Correct**: correct responses. Gray lines mark PPE in the Vocabulary task, where personal pronouns were not among the response options, in contrast to the other tasks.