Theory of Mind Beyond Beliefs: Testing Attention-Based Social Micro-Processes in LLMs

John Muchovej (john.muchovej@yale.edu)[1], Paula Rubio-Fernandez[2] & Julián Jara-Ettinger[1]
[1] Department of Psychology, Yale University
[2] Multimodal Language Department, Max Planck Institute for Psycholinguistics

Abstract

Vision-Language Models (VLMs) can now produce fluent, socially appropriate dialogue, leading to interest in whether they have Theory of Mind (ToM). While recent work suggests that VLMs still lack coherent mental-state reasoning, this work has focused on classical propositional belief representations. Here we test a complementary, communication-relevant form of ToM: attention-based social micro-processes that support referential communication in the here and now (i.e., selecting efficient descriptions to identify a particular object for someone else). In face-to-face communication, people strategically add redundant color adjectives to facilitate the listener's visual search, and omit them when they provide no benefit. We evaluate whether VLMs show the same strategy in a referential communication paradigm where the usefulness of redundant color words varies based on the set size and color distribution of objects. We find that VLMs can produce successful referential expressions but lack the attention-guiding strategies that make human communication so efficient. This suggests that VLMs lack the more implicit representations of attention that people use in everyday communication.

Keywords: Reference; Overspecification; Visual search; Pragmatics; VLMs; Color Adjectives

Introduction

Large Language Models (LLMs) and Vision-Language Models (VLMs) (jointly, V/LLMs) can now, with some proficiency, tutor (Scarlatos et al., 2025), read between the lines of what people say (Hu et al., 2023), and even help people question their incorrect beliefs (Costello et al., 2024). This, along with many other findings, raises the question of whether these systems understand how human minds work, a capacity known as Theory of Mind (ToM).

Research into this question has largely focused on classical ToM tasks like the false-belief task, where the goal is to test whether LLMs can represent propositional mental states like beliefs or desires. This work revealed that LLMs pass standard ToM tasks (Kosinski, 2024), but this is likely the result of superficial heuristics because simple changes to the stimuli result in striking failures (Shapira et al., 2024; Ullman, 2023). Since then, research has revealed that the reasoning signatures expected from ToM-like reasoning are largely absent from LLMs and VLMs (Muchovej et al., 2025).

However, ToM does not only include a capacity to represent other people's knowledge and preferences, but also a capacity to represent dynamic cognitive processes in other minds, like people's attention, visual search, and distraction (Berke et al., 2023). These forms of ToM do not constitute a classical form of belief-desire representation, but representations of people's real-time reasoning, which give us expectations about how people's attention is directed in real time. These types of representations have not been tested in V/LLMs, but they are particularly important for communication (Rubio-Fernandez et al., 2025b), and hence might be likely to spontaneously emerge in V/LLMs.

Representations of others' cognitive processes shape how people communicate. The canonical case comes from referential communication (i.e., identifying something among a set of alternatives). Here, the way people describe an object is designed to help guide the listener's visual search so that they can quickly identify what is being talked about, thereby reducing the collaborative effort of referring (Clark & Wilkes-Gibbs, 1986). For example, if there's a single cup resting on a table between the speaker and the listener, most people would simply call it "the cup". However, if the table were crowded with other objects, people start to mention the color (e.g., "the blue cup"), even if this is technically unnecessary because there is only one cup (i.e., the color word is redundant; e.g., Rubio-Fernandez et al., 2021). Although people are more likely to use more redundant color words as the number of objects increases, this behavior disappears when the target's color is shared with many other objects. In the same example, if everything in the crowded table were blue, speakers revert to using a bare noun ("the cup"; Rubio-Fernandez, 2019) or use a different adjective that might be helpful to the listener, like mentioning the cup's material or spatial location.

Together, these behaviors efficiently guide the listener's attention to the target object. To do this, speakers rely on an implicit model of other people's visual search (Jara-Ettinger & Rubio-Fernandez, 2021). Unlike belief and desire representations in ToM, which require stable and structured representations, tracking and anticipating others' attention requires only considering the visual features of a scene to decide what might be salient or not. These rapid forms of ToM, called *social micro-processes*, hence might be easier to compute (Rubio-Fernandez et al., 2025b).

On the one hand, one might expect V/LLMs to develop these strategies spontaneously because their goal is to communicate in a human-like way, trained from large text corpora and fine tuned via reinforcement learning from human feedback (RLHF; Kaufmann et al., 2025). On the other hand, this might be a particularly challenging task for V/LLMs because

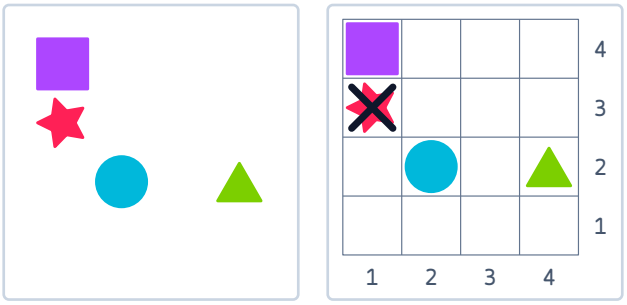

Figure 1: Example set of inputs provided to VLMs in our experiment. The left display shows the target image, and the right display is the reference sheet that indicates which object the VLM must communicate about.

most of their training reflects *displaced* communication: talking about things removed from the here and now, like past events, hypothetical situations, and distant places (Ma et al., 2023). Developing a capacity to use language to efficiently guide the listener's attention in a visual scene or display might then require true interactive training data.

Our goal in this paper is to test whether VLMs can strategically use adjectives in referential communication about the here-and-now (i.e., in *non-displaced* events), to see whether they have developed strategies to guide listener attention. This helps us further our understanding of whether VLMs communicate like humans do, and whether they are developing human-like social reasoning through attention tracking (a social micro-process) rather than explicit belief representations characteristic of full fledged ToM.

While the question of whether VLMs are good at referential communication has been a topic of recent interest, most work has focused on whether VLMs can produce referential expressions that successfully identify the target (e.g., Ma et al., 2025; Tang et al., 2024), rather than asking whether they use the same communicative strategies a human would.

Experiments

Our general experimental approach is based on referential communication games used to explore how humans strategically build useful referential descriptions for listeners (e.g., Rubio-Fernandez et al., 2021 see Rubio-Fernandez et al., 2025 for review).

In this setup, participants see a set of shapes with a target referent, which always has a unique shape, such that mentioning the shape alone is guaranteed to lead to communicative success. However, by varying the number of objects and the relative distribution of colors, the usefulness of adding a redundant color adjective varies. Specifically, in polychrome displays (i.e., when all objects are different colors), people are more likely to use redundant color adjectives as the number of objects increases. This use is then even sharper when all non-targets are the same color, creating a popout effect for the target (popout condition). By contrast, use of referential

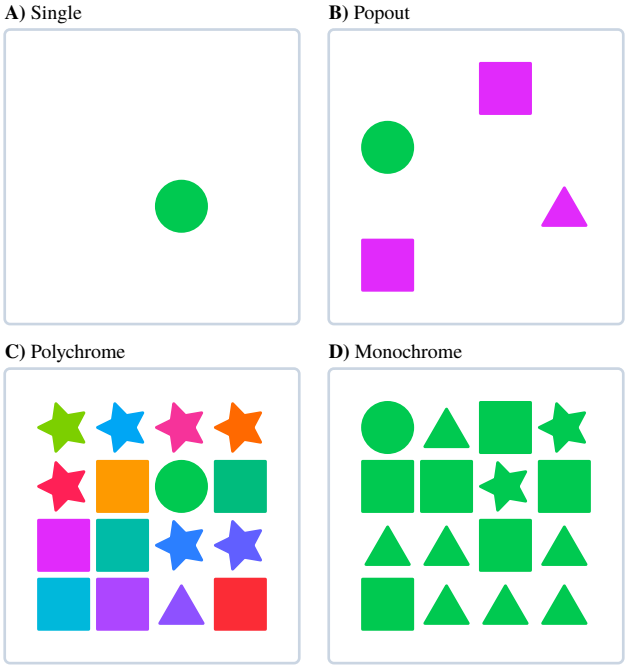

Figure 2: Example trial types used in the study, that use the green circle (●) as the target.

overspecification decreases as all objects become the same color as the target (monochrome condition) because the color adjective is no longer useful information to filter the set of possible referents. Thus, our experimental approach focuses on testing referential expression varying set size, and color salience (monochrome, polychrome, popout).

VLMs Used

Throughout, we focused on two models: `gpt-4o-2024-11-20` (GPT-4o) and `gpt-5.2-2025-12-11` (GPT-5.2). We chose these models to evaluate how standard VLMs perform (GPT-4o) and how this performance changes with reasoning models (GPT-5.2). We tested these models through OpenAI's API. In each trial we provide a system prompt describing the task (detailed in each experiment's procedure) along with a JSON schema so it can generate a structured-output. The JSON schema requires the VLM to provide the coordinates of the target object and the description used to guide "the User." Responses which failed to accurately extract the target's coordinates were re-sampled online and excluded from analyses. In lieu of token-level logprobs (as these are not available for GPT-5.2), we collected 200 samples per trial to estimate the distribution of production patterns.

Experiment 1

Experiment 1 began with a test querying VLMs to simply produce valid referential expressions so as to successfully communicate.

Materials

Stimuli consisted of images like those shown in Figure 1. Each trial consisted of a test image including a set of objects randomly placed on a 4x4 grid (with the grid not observable), and a reference image, which showed the same test image, but with the 4x4 grid overlaid and an X marking the target shape to be described. This approach was based on how people are tested in these games, where there is a public display of shapes, and a private one that participants receive to know which object they should talk about (e.g., Rubio-Fernandez et al., 2021).

We specifically began considering three set sizes of objects: one object, four objects, and sixteen objects. For the four-object and sixteen-object set sizes, we designed three sub-conditions. In the monochrome condition all objects were the same color, including the target; in the polychrome condition all objects were of different colors; and in the popout condition all objects except the target were the same color. This resulted in a total of seven trial types (see Figure 2 for examples). Each trial was represented through a sampler that randomly selected a target shape, and then populated the grid with the target and the distractors in random locations (shape options: star, square, triangle, circle). While the target always had a unique shape, the distractors could overlap in shape. We then randomly sampled colors from a set of seventeen colors according to the condition specification. We drew from a fixed set of 17 colors commonly used in web development tools that VLMs are familiar with. The first color was uniformly sampled from the fixed set. Afterwards, each additional color was selected by maximizing the perceptual difference (in RGB terms) between itself and the current set of colors. Thus, each VLM run on a trial consisted of a different visual display of the target category.

In each trial, VLMs were provided the trial (see Figure 1 for an example) and given the prompt described in Prompt 1.

Procedure

VLMs were provided with Prompt 1 as their system prompt. The test and reference images were provided as "User" input. Similar to tasks with humans (e.g., Rubio-Fernandez, 2019; Rubio-Fernandez et al., Long et al.; 2021, 2020), the input clarified that the objects would be in different locations in the listener's grid, removing the possibility of identifying the referent through simple spatial strategies (something that VLMs are already known to struggle with; e.g., Chen et al., 2025).

The first part of the task (identification) was included to help confirm that the VLMs correctly identified the target in each trial. Thus, when the VLM fails to refer to an object correctly, we can distinguish cases where it failed because it did not identify the target referent (a failure in the vision module), from cases where it failed because it did not produce an appropriate referential description (a failure in communication). The second component (communication) was then our target of analysis. Events where the VLMs failed to correctly identify the target were discarded as that was not the focus of our study.

Prompt 1

You will be given two images of a 4×4 grid. The first image contains colored shapes. Both you and the User can see this image. The second image shows a labeled grid with a "×" marking the target. Only you can see which object is marked.

Your task has two parts:

1. IDENTIFICATION (for verification): Report the (x, y) coordinates of the crossmark.
2. COMMUNICATION (for the User): Produce a referential description to help the User click on the target object. This description should be grammatical, but does not need to include articles or instructions like "click on" or "the".

CRITICAL: The User will see the same objects but in COMPLETELY DIFFERENT positions on their grid. Think carefully about what information will actually help them identify the correct object when spatial positions don't match.

Response coding

In some cases, VLMs produced expressions that used an unexpected but valid color or shape. For example, it might refer to a square as a "rectangle" or call blue "turquoise", in other cases it might use adjectival forms of color or shape like "golden" instead of gold or "triangular" instead of triangle. To identify unexpected color and shape words, we used the 2025 release of Open English WordNet (McCrae et al., 2019). If the noun form shared hypernymy with `color.n.01` or `plane_figure.n.01`, we considered it a color or shape, respectively. We then used Google's EmbeddingGemma (Vera et al., 2025) to find potential matches through embeddings. Specifically, we computed the cosine similarity between the unknown color produced by the VLM and all colors used in a given trial. We then selected all colors which were at least 1 standard deviation above the mean similarity to produce a set of candidates. If this candidate set contained the target color, we categorized the VLM as producing a valid match. We used a parallel logic with shapes.

We applied this coding scheme to 17.40% and 19.83% of trials for GPT-4o and GPT-5.2 (i.e., the cases where the color or noun was not immediately identifiable). Of these, 61% and 71.48% of the decoded responses correctly referred to the target shape and its color (if a color was specified). The remainder of decoded responses were categorized as incorrect descriptions of the target.

Results

Overall, the procedure resulted in valid referential expressions (only 1.07% and 0.21% of all responses were excluded, for GPT-4o and GPT-5.2 respectively). On average, GPT-4o and GPT-5.2 incorrectly identified the target on 51.99% and

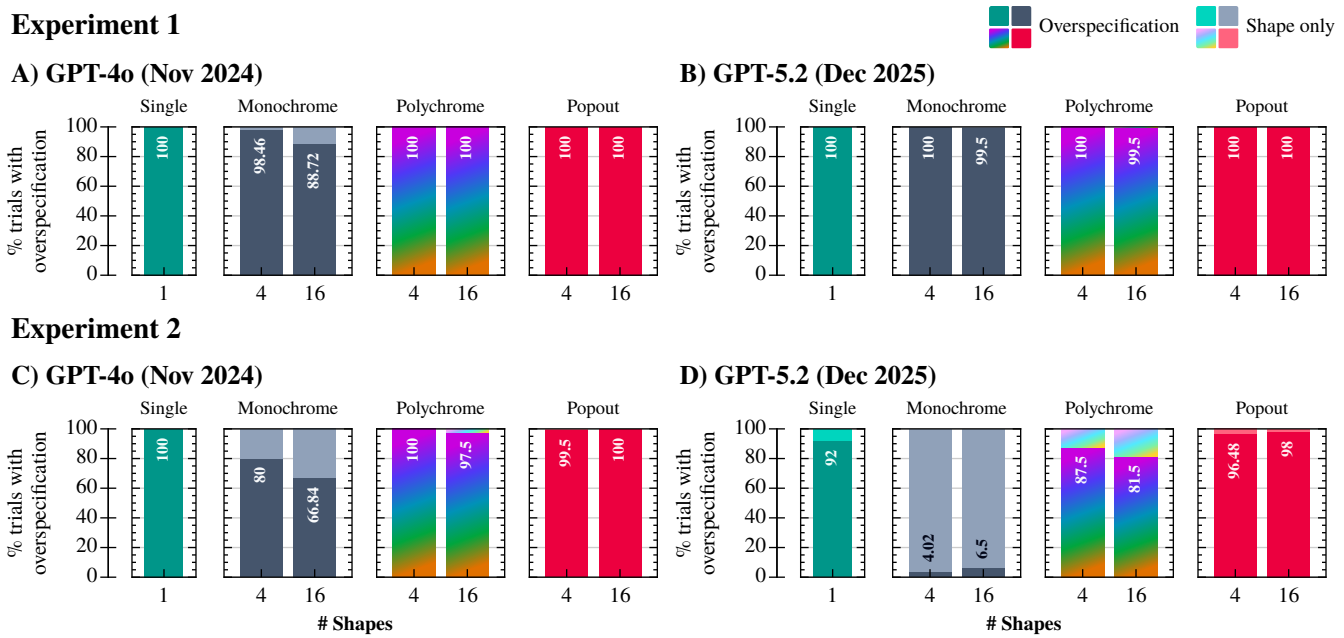

Figure 3: Results for Experiments 1 and 2 with GPT-4o and GPT-5.2. Percentage of descriptions across conditions using redundant color adjectives (across all trials regardless of referent correctness). (A-B) Results from Experiment 1, when simply asked to communicate the referent to an interlocutor for GPT-4o and GPT-5.2, respectively. (C-D) Results from Experiment 2, when asked to consider visual search for GPT-4o and GPT-5.2, respectively.

46.31% of trials, respectively. Notably, all failures were cases where both VLMs incorrectly identified the color of the target shape. This contrasts with related work that has found much higher failure rates in referential communication in VLMs (Ma et al., 2025; Tang et al., 2024), suggesting that our paradigm was simple enough for the GPTs to succeed, allowing us to focus on its communicative strategies. The remainder of the analyses focused exclusively on the cases where the referential description correctly identified the target.

Figure 3 A-B shows the results from Experiment 1. Overall, both VLMs produced an overwhelming rate of color adjectives. Across all conditions, the VLMs only produced the bare noun description (e.g., "the star") on 1.81% and 0.01% of trials (for GPT-4o and GPT-5.2, respectively). These bare descriptions correctly appeared in the monochrome trials for GPT-4o (1.54% / 11.28% for the four- and sixteen-object trials; $p < 0.001$ across set sizes by test of equal proportions), but not with GPT-5.2, which produced the most bare descriptions in the sixteen-object displays on both the monochrome and polychrome trials (0.0% / 0.5% in the four- and sixteen-object cases; $p = 1.0$ across set sizes within conditions).

This pattern contrasts sharply with human choices (Figure 4). To begin, no study has documented any case where humans reach near close to 100% rates of overspecification in displays with the number of shapes we tested. The highest rates have been found in popout conditions (■; Figure 4) and in polychrome conditions (■) with sixteen objects. The largest divergences come from how people behave with four objects, where overspecification does not go above 50% (contrasted with VLMs at 100%), and in monochrome trials (■), where any overspecification virtually disappears with as little as four objects.

Experiment 2

Experiment 1 showed that neither GPT-4o or GPT-5.2 produce human-like referential expressions spontaneously. This contrasts with humans, where evidence shows that they naturally prefer expressions that guide listener visual search without need to be explicitly prompted (Jara-Ettinger & Rubio-Fernandez, 2022). We next tested whether we could produce this behavior when explicitly prompted to consider listener visual search and the usefulness of color adjectives.

Materials

This experiment used the same materials from Experiment 1.

Procedure

The procedure was identical to Experiment 1 but using a modified prompt that emphasized the importance of visual search (Prompt 2). We additionally included an extra task for the VLM which explicitly requested that it analyze the usefulness of color, and emphasized the tradeoff between efficiency and communicative support at the end. For clarity, the key changes from Experiment 1 are underlined.

Prompt 2

You will be given two images of a 4×4 grid. The first image contains colored shapes. Both you and the User can see this image. The second image shows a labeled grid with a "×" marking the target. Only you can see which object is marked.

Your task has three parts:

1. IDENTIFICATION (for verification): Report the (x, y) coordinates of the crossmark.
2. ANALYSIS: Assess whether color information would meaningfully help the User find the target faster in this display.
3. COMMUNICATION (for the User): Based on your analysis, produce a referential description. This description should be grammatical, but does not need to include articles or instructions like "click on" or "the".

CRITICAL: The User will see the same objects but in COMPLETELY DIFFERENT positions on their grid. Think carefully about what information will actually help them identify the correct object when spatial positions don't match. Your thinking should aim to balance description length with search efficiency. On the one hand, shorter descriptions are more efficient as a speaker but can be slower for the listener to find the object. On the other hand, longer descriptions are less efficient as a speaker but can help the listener find the object faster.

We additionally modified the output format in the JSON schema to include an additional judgment as to whether color information would help the user find the object efficiently.

Results

Like Experiment 1, both models largely produced valid referential expressions (only 0.43% and 0.14% of responses were excluded, for GPT-4o and GPT-5.2 respectively). Similarly, both models had high rates of expressions which failed to correctly refer to the target object (49.07% and 31.76% for GPT-4o and GPT-5.2, respectively). As in Experiment 1, these failures were strictly in color-resolution.

Figure 3 C-D shows the results from this experiment. Because the model performance now diverged, we discuss each model one at a time.

Overall, GPT-4o continued to over-use redundant color adjectives (92.04% of trials), and its rate of overspecification was above 97.5% across all conditions except the two monochrome ones. Here, GPT-4o reduced over-specification to 80% of trials in the four-object monochrome condition (a significant drop compared to Experiment 1; $p < .001$), and down to 66.84% of trials in the sixteen monochrome condition (a significant drop compared to the four-object set; $p = .004$). GPT-4o was notably consistent in its expression generation, 98.21% of the time it concluded that color would be helpful and indeed referenced a color (disagreeing with itself 1.79% of the time); however, it was less consistent with itself when

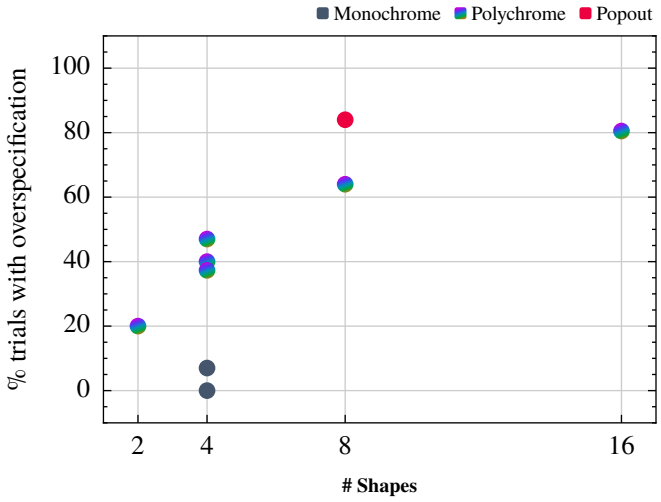

Figure 4: Overspecification results in human studies from Long et al. (2020), Rubio-Fernandez (2019), and Rubio-Fernandez et al. (2021a). In polychrome displays (■), people increase their propensity to overspecify as a function of display size reaching 80% overspecification on 16-item displays. By contrast, at just four objects of the same color (■), participants suppress overspecification almost entirely (0% overspecifying in one sample and 7% in another). Finally, popout displays (■) with eight objects trigger about 84% of overspecification.

color was deemed unhelpful by only omitting color references 82.88% of the time (thus producing a color when considered unhelpful 17.12% of the time). This pattern reveals a correct identification of the kinds of trials where overspecification ceases to be useful, although in both cases it continues to over-specify more than half of time, producing a very strong divergence to how humans communicate in these situations.

By contrast, GPT-5.2 overall showed a major alignment with human judgments. GPT-5.2 was remarkably consistent in cases where it judged that color would not be helpful, never referencing a color in these trials. However, when GPT-5.2 judged a color would be helpful, it referenced a color 82.81% of the time and failed to reference a color 17.19% of the time. Though GPT-5.2 included redundant adjectives on 66.6% of trials, these are all clustered on the non-monochrome trials. This model overspecified 92% of the time on the single object trial. However, in the monochrome trials, GPT-5.2 only overspecified on 4.02% and 6.5% of trials for the four- and sixteen-object conditions (no significant difference across set size conditions). This reveals that reasoning models are able to recognize that overspecification is no longer useful in monochrome trials. Next, we look at GPT-5.2′s internal judgement about whether color would be helpful. At the same time, the fine-grained structure continued to be inconsistent with humans. In both the monochrome and polychrome conditions, its rate of overspecification moved in the incorrect direction: increasing as a function of objects in the monochrome condition, and decreasing as a function of objects in the polychrome condition.

General Discussion

Effective human communication requires not only producing descriptions that successfully reveal what the speaker is talking about, but also producing descriptions that help interlocutors understand the message efficiently. In referential communication in the here-and-now, past work has shown that people strategically use redundant color adjectives to guide listener attention. Speakers add them when they speed up the listener's visual search and avoid them when they would hinder the search (Rubio-Fernandez et al., 2025b). This behavior reflects a form of ToM grounded in rapid social micro-processes, where speakers model the listener's visual attention within an event (Jara-Ettinger & Rubio-Fernandez, 2021). This contrasts with more propositional forms of ToM like belief reasoning, which have historically dominated research on social cognition in LLMs and VLMs.

Our results reveal three findings. First, VLMs can generally communicate successfully in simple referential games. They consistently produce expressions that uniquely identify the target, although our design was simple enough that just mentioning the target shape guaranteed success. Other work has shown that when ambiguity is likely, VLMs performance quickly breaks down (Ma et al., 2025; Tang et al., 2024). Second, VLMs show a strong propensity to include redundant color adjectives. When simply asked to communicate (Experiment 1), both models produced excessive redundant color adjectives, even when they were counterproductive, such as when every object shared the same color. In these cases, people recognize that redundant color adjectives add no information and delay communicative success, therefore omitting them entirely (Rubio-Fernandez, 2021).

Our last finding is that prompting VLMs to explicitly consider visual search (Experiment 2) improves their performance. Although the models still produced more redundant color words than humans, they reduced their use in the monochrome conditions and GPT-5.2 was able to almost entirely suppress them. These results suggest that VLMs, and particularly reasoning models, can improve how they guide listener attention when prompted to do so. Interestingly, however, these models only revealed sensitivity to the redundancy of color in monochrome displays, failing to modulate their use of color words as a function of the number of shapes on display (the way humans do; see Figure 4). Critically, even if VLMs performed in a human-like way in this experiment, this would still not capture human-like reasoning as people show this behavior without needing to be explicitly prompted to consider visual search (Jara-Ettinger & Rubio-Fernandez, 2022). This suggests a limited understanding of efficient referential communication.

One limitation of our work is that we did not manipulate color typicality, which affects efficiency (compare, e.g., "red strawberry" vs "red shoes") and therefore the use of redundant color words in humans (Long et al., 2020; Sedivy, 2003). We also only tested models in OpenAI's GPT family. We are currently evaluating VLMs from other model families. Note, however, that research on VLM and LLM social reasoning has generally found similar conclusions across model families (e.g., Muchovej et al., 2025; Ullman, 2023), suggesting that observed limitations likely reflect shared training regimes, training data, or both.

These communicative failures are consistent with the fact that these models are trained primarily on datasets in which people describe events rather than communicate about them. This suggests that models trained to describe are ill-suited as models of communication. One potential solution is for these models to be trained or fine-tuned on datasets specialized for communication, like RefCOCO (Yu et al., 2016). Unfortunately, datasets on interactive referential communication have been found to be highly faulty (Chen et al., 2025) and high rates of data overlap in these datasets reduces the available variance that current models can learn from when trained on these datasets (Chen et al., 2019; Ma et al., 2025).

One challenge we anticipate is that, even within referential communication, people use different strategies depending on what is the goal and context of the conversation. In face-to-face conversation, talking about the here-and-now, people strategically communicate by guiding listener attention. However, when talking about past events, people change their strategies to use different descriptions that are effective for guiding memory search (Suwal et al., 2025).

Altogether, our results show that current VLMs fail to spontaneously communicate in human-like ways, despite their fine-tuning specifically on human preferences. However, explicitly prompting them to consider visual search improves their performance. This is consistent with related work also finding that VLMs can improve performance through forms Chain of Thought reasoning (Dong et al., 2026). However, a critical aspect of human communication is that these complex communicative strategies are easy for humans, and don't require explicit attention or extended reasoning. Thus, even if VLMs can come to replicate these patterns through prompting and reasoning, in doing so, this capacity might still fail to be human-like, if not in performance, then in reasoning requirements.

Acknowledgments

We thank the VLMs for participating in this study. This work was supported by NSF award BCS-2045778.

Code and Data Availability

Code, model responses, and stimuli can be found on OSF.

References

Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023). *Thinking about Thinking as Rational Computation*. https://doi.org/10.31234/osf.io/e65p3

Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J. C., Geva, M., He, J., Wu, J., & Li, M. (2025). *Why is spatial reasoning hard for VLMs? An attention mechanism*

perspective on focus areas. https://doi.org/10.48550/arXiv.2503.01773

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2019). *UNITER: UNiversal Image-TExt Representation Learning*. https://doi.org/10.48550/arXiv.1909.11740

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science (New York, N.Y.)*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Dong, Q., Figueroa, L., Zhao, H., Kafle, K., Kuen, J., Ding, Z., Cohen, S., & Fu, Y. (2026). *CoT Referring: Improving referring expression tasks with grounded reasoning*. https://doi.org/10.48550/arXiv.2510.06243

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023,). A fine-grained comparison of pragmatic language understanding in humans and language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. https://doi.org/10.18653/v1/2023.acl-long.230

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science Advances*, *7*(47), eabj970. https://doi.org/10.1126/sciadv.abj0970

Jara-Ettinger, J., & Rubio-Fernandez, P. (2022). The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological Review*, *129*(6), 1394–1413. https://doi.org/10.1037/rev0000345

Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2025). *A survey of reinforcement learning from human feedback*. https://doi.org/10.48550/arXiv.2312.14925

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Long, M., Rohde, H., & Rubio-Fernandez, P. (2020). The pressure to communicate efficiently continues to shape language use later in life. *Scientific Reports*, *10*(1), 8214. https://doi.org/10.1038/s41598-020-64475-6

Ma, Z., Ding, J., Zhang, X., Luo, D., Ding, J., Xu, S., Huang, Y., Peng, R., & Chai, J. (2025). *Vision-language models are not pragmatically competent in referring Expression Generation*. https://doi.org/10.48550/arXiv.2504.16060

Ma, Z., Sansom, J., Peng, R., & Chai, J. (2023). *Towards A holistic landscape of situated theory of mind in large Language Models*. https://doi.org/10.48550/arXiv.2310.19619

McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., & Fellbaum, C. (2019). English WordNet 2019 – An Open-Source WordNet for English. *Proceedings of the 10th Global Wordnet Conference*, 245–252. https://doi.org/10.18653/v1/2019.gwc-1.31

Muchovej, J., Royka, A., Lee, S., & Jara-Ettinger, J. (2025). GPT-4o lacks core features of theory of mind. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. https://escholarship.org/uc/item/9h04686p

Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the Gricean maxim of quantity. *Cognitive Science*, *43*(11), e12797. https://doi.org/10.1111/cogs.12797

Rubio-Fernandez, P. (2021). Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. *Humanities & Social Sciences Communications*, *8*(1). https://doi.org/10.1057/s41599-021-00818-6

Rubio-Fernandez, P., Berke, M. D., & Jara-Ettinger, J. (2025b). Tracking minds in communication. *Trends in Cognitive Sciences*, *29*(3), 269–281. https://doi.org/10.1016/j.tics.2024.11.005

Rubio-Fernandez, P., Mollica, F., & Jara-Ettinger, J. (2021a). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology. General*, *150*(3), 583–594. https://doi.org/10.1037/xge0000963

Scarlatos, A., Liu, N., Lee, J., Baraniuk, R., & Lan, A. (2025). Training LLM-based tutors to improve student learning outcomes in dialogues. In *Lecture Notes in Computer Science* (pp. 251–266). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98414-3_18

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. https://doi.org/10.1023/a:1021928914454

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2024). Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2257–2273. https://doi.org/10.18653/v1/2024.eacl-long.138

Suwal, U., Morris, B., Lin, Q., Rubio-Fernandez, P., & Jara-Ettinger, J. (2025). *Speakers strategically adjust their descriptions based on perceived memorability*. https://doi.org/10.31234/osf.io/84kds_v1

Tang, Z., Mao, L., & Suhr, A. (2024). *Grounding language in multi-perspective referential communication*. https://doi.org/10.48550/arXiv.2410.03959

Ullman, T. (2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks*. http://arxiv.org/abs/2302.08399

Vera, H. S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S. R., Smoot, S., Naim, I., Zou, J., Chen, F., Cer, D., Lisak, A., Choi, M., Gonzalez, L., Sanseviero, O., Cameron, G., Ballantyne, I., Black, K., Chen, K., … Seyedhosseini, M. (2025). *EmbeddingGemma: Powerful and lightweight*

text representations. https://doi.org/10.48550/arXiv.2509.20354

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). *Modeling context in referring expressions*. https://doi.org/10.48550/arXiv.1608.00272