# Do LLMs selectively encode the goal of an agent's reach?

Laura Ruis[1] Arduin Findeis[2] Herbie Bradley[3 4 5 6] Hossein A. Rahmani[7] Kyoung Whan Choe[3]
Edward Grefenstette[* 1] Tim Rocktäschel[* 1]

## Abstract

In this work, we investigate whether large language models (LLMs) exhibit one of the earliest Theory of Mind-like behaviors: selectively encoding the goal object of an actor's reach (Woodward, 1998). We prompt state-of-the-art LLMs with ambiguous examples that can be explained both by an object or a location being the goal of an actor's reach, and evaluate the models' biases. We compare the magnitude of the bias in three situations: i) an agent is acting purposefully, ii) an inanimate object is acted upon, and iii) an agent is acting accidentally. We find that two models show a selective bias for agents acting purposefully, but are biased differently than humans in comparable studies. Additionally, the encoding is not robust to semantically equivalent prompt variations. We discuss how this bias compares to the bias infants show and provide a cautionary tale of evaluating machine Theory of Mind (ToM). We release our dataset and code.[1]

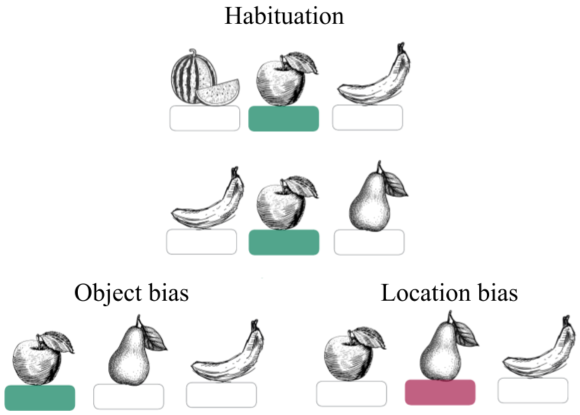

*Figure 1.* A visual depiction of our test inspired by Woodward (1998). We prompt an LLM with $k$ ambiguous linguistic *habituations* that can be explained either by the goal being the object or the location ($k = 2$ in the image). We then test the bias the model shows for assuming the goal was the object (left-bottom) or the location (right-bottom). We say a model *selectively* encodes the goal if it shows a distinct bias when an agent appears to be acting purposefully.

## 1. Introduction

Theory of Mind (ToM) is the socio-cognitive ability to reason about unobserved mental states of other agents. It is considered central to many aspects of human cognition, like linguistic communication (Milligan et al., 2007). In light of rapidly advancing linguistic capabilities of large language models (LLMs), recent studies have explored the emergence of ToM in these models. The results are as of yet inconclusive; some works suggest it has emerged (Kosinski, 2023; Moghaddam & Honey, 2023), and others suggest it has not (Ullman, 2023) or at least not at a level comparable to humans (Trott et al., 2022; Sap et al., 2022; Shapira et al., 2023).

A reason for these conflicting results is that we cannot simply apply the tests we use to study ToM in humans to LLMs. Many of these tests appear in the training data, meaning that models can pass the tests without reasoning about other agent's mental states. For example, whilst Kosinski (2023) shows certain LLMs can pass classic false-belief tests, Ullman (2023) demonstrates that those same models fail on minimal alterations to these tasks that change the expected answer.[2] This evidence suggests models memorize training patterns without actually mentalizing. Although Ullman (2023) shows that the model fails on adversarial alterations to the task, highlighting that their capabilities are far from robust, we cannot conclude that the model cannot reason about the mental states of others. Perhaps the reason models

[*] Equal contribution [1] DARK Lab, University College London, UK [2] Department of Computer Science and Technology, University of Cambridge, UK [3] CarperAI [4] Stability AI [5] EleutherAI [6] CAML Lab, University of Cambridge, UK [7] Web Intelligence Group, University College London, UK. Correspondence to: Laura Ruis <laura.ruis.21@acl.ac.uk>.

*Proceedings of the First Workshop on Theory of Mind in Communicating Agents*, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1] https://github.com/LauraRuis/tom

[2] Ullman tests LLMs on unexpected contents tasks where the contents are in see-through containers, altering the answer to the false-belief tests.

repeat training patterns for adversarial examples is precisely because these examples follow patterns from training so closely. LLMs are trained to mimic the training distribution and are known to repeat training patterns regardless of truthfulness (Lin et al., 2022). In many cases they need in-context examples of the task — not to in-context learn (Brown et al., 2020), but simply to adhere to the right output format (Min et al., 2022). To more fairly evaluate these models' mentalizing capabilities, we need to properly set them up for the task and provide examples (Lampinen, 2023).

How can we investigate machine theory of mind in models that have seen all the classic tasks from developmental psychology and regurgitate their patterns even when these tasks are worded differently? In this work, we take a step back and avoid pre-existing text-based tasks. Instead, we investigate whether LLMs encode situations differently when a goal-directed agent is involved. Specifically, we look at one of the earliest ToM-like human biases: selectively encoding the goal object of an actor's reach (Woodward, 1998). In her seminal study, Woodward shows that infants as young as six months old exhibit a bias for encoding an agent's goal object over a goal location. Similarly, we ask the question: *do large language models selectively encode the goal object of an actor's reach?* We prompt a set of LLMs with *habituations* that can be explained both by the goal of an actor's reach being an object, as well as a location. We then look at whether LLMs exhibit a bias for assuming the goal is the object or the location (see Figure 1). We investigate the bias the model shows in three situations: an agent is purposefully reaching for an object, an inanimate object moves and touches an object, and an agent is acting accidentally and touches an object. We say a model *selectively encodes the goal of an agent's reach* if it shows a distinct bias between the agent acting purposefully and otherwise. For a behavior to be considered theory of mind, the same behavior should not show up when the task does not involve a goal-directed agent (Frith & Frith, 2012; Devaine et al., 2014).

Our protocol has several benefits over other approaches of investigating machine ToM from literature. Firstly, the underlying task logic is visually presented to pre-linguistic human infants in literature, making it less likely that the exact task appears in the training data of pre-trained language models. Nonetheless, the reasoning pattern might be numerously described. In similar spirit to Ullman (2023), we extend Woodward (1998) by adding a control task where the agent acts accidentally, nullifying the assumption that the agent is acting in a goal-directed way. Like the inanimate case, the object bias should not show up in this control task. Another benefit is the habituations that are reminiscent of few-shot prompting in LLMs (Brown et al., 2020), but unlike true few-shot examples these do not leak any information about the expected output. These examples both serve to habituate a model in order to probe for a bias, as well as to guide the model to the task. Importantly however, even though we can use our protocol to make empirically backed claims about whether or not LLMs selectively encode the goals of agents, we can make no statements about how the model does it and whether there is reasoning involved. Similarly, Woodward makes no assumptions about what kind of knowledge infants use to encode the goal object of an actor's reach; she just shows that they do.

Our results show that both GPT-3.5-turbo and GPT-4 pass the criterion for saying that they selectively encode the goal of an actor's reach. However, whilst infants show no bias in the inanimate test case in the visual task by Woodward (1998), both models show biases in the inanimate and control test cases. We hypothesise that this is due to the sensitivity of models to irrelevant surface-level patterns in text, which we elaborate upon below when discussing the results. Additionally, the selective encoding of the goal does not show up for two semantically equivalent prompt variations. For those variations, although models show more object bias in the animate case than the inanimate case, they also show more object bias in the control case, which means we cannot say that they selectively encode the goal of an agent's reach. From these results we conclude that although we can say that GPT-3.5-turbo and GPT-4 selectively encode the goal of an actor's reach sometimes, they do not do so robustly, and moreover are biased differently from humans. Our results contribute to the picture from existing work on ToM in LLMs, concluding that even the developmentally earliest ToM-like human behavior does not robustly show up in current SotA LLMs. Our findings further highlight the importance of designing multiple prompt variations for each task: depending on how the task is framed, conclusions can be opposite.

## 2. Related Work

Recently, classic ToM tests from developmental psychology have been extensively applied to LLMs. However, these studies have conflicting results. Kosinski (2023) claims theory of mind has emerged in a subset of OpenAI's API models, but the evaluation protocol has been pointed out as flawed by Ullman (2023). Similarly, Sap et al. (2022) show that GPT-3 achieves well below human performance on a range of different ToM tasks. The methodology used in that study is however critiqued by Moghaddam & Honey (2023), who apply similar tests but use SotA prompting techniques and show that OpenAI's models that are finetuned with RLHF achieve human-level performance on the ToM tasks. By contrast, Shapira et al. (2023) show that LLMs can robustly solve some ToM tasks, but not others, and conclude that models have some ToM capabilities, but that these are not robust.

Woodward (1998) conducts her study with the aim of exploring how infants perceive and comprehend others' actions[3]. The study focuses on investigating infants' ability to selectively encode the goal object of an actor's reach. Drawing inspiration from Woodward (1998)'s work, Gandhi et al. (2021) apply a similar task to neural networks, aiming to determine whether machines can represent an agent's preferred goal object. However, to our knowledge, there is currently no study that applies the task from Woodward specifically to pre-trained LLMs.

## 3. Method

In this section we outline the method we use to answer the research question: *do language models selectively encode the goal object of an actor's reach?*

**Defining object and location bias.** We want to investigate the question whether models store knowledge that leads them to encode the goal-related properties of an agent's reaching event, and that this knowledge does not get encoded in similar events involving inanimate objects. To this end, we design the following test cases: an animate test case where the prompt contains $k$ habituations in which an agent reaches for the same object in the same location. A test case is appended to this prompt where the goal object is placed in a different location. We then obtain the likelihoods the model assigns to continuing the full prompt as if the same location with a novel object is reached for by the agent (*location bias*), or the same object at a different location (*object bias*, see Figure 1). Below is an example for an agent, Wendy, who has a preference for kiwis, with $k = 2$ habituations:

> There is a kiwi on the first pillar, an orange on the second pillar, and a fig on the third pillar. Wendy grasps the item on the first pillar.
> There is a kiwi on the first pillar, a fig on the second pillar, and an orange on the third pillar. Wendy grasps the item on the first pillar.
> There is an orange on the first pillar, a kiwi on the second pillar, and a fig on the third pillar. Wendy grasps the item on the *first/second*

In this example, a model that assigns a higher probability to *first* is said to exhibit a location bias, whereas a model that assigns a higher probability to *second* exhibits object bias. Independently, we test the model on the same example with an inanimate object:

> There is a kiwi on the first pillar, an orange on the second pillar, and a fig on the third pillar. A pole moves to and touches the item on the first pillar.
> There is a kiwi on the first pillar, a fig on the second pillar, and an orange on the third pillar. A pole moves to and touches the item on the first pillar.
> There is an orange on the first pillar, a kiwi on the second pillar, and a fig on the third pillar. A pole moves to and touches the item on the *first/second*

We generate 100 examples with a roughly equal distribution over object and location targets (in this example template, the targets can be one of "first", "second", and "third"). We define the *object bias* $o_b$ as the conditional probability that the object bias target is chosen by a model given that the model has to either choose the object or location bias target, as in

$$o_b = \frac{p(\text{object bias target})}{p(\text{object bias target}) + p(\text{location bias target})}, \quad (1)$$

where each probability $p(\cdot)$ is conditioned on the prompt like $p(\cdot \mid \text{prompt})$.

In some cases we do not have access to the probabilities assigned to each target by a model (i.e. GPT-3.5-turbo and GPT-4 have restrictive APIs). Instead, we sample those models ten times for each prompt with a temperature of 1, recording how often they output the object bias target $c_o$ (*second* in the previous example) or the location bias target $c_l$ (*first* in the previous example). Using these counts, we estimate the object bias $o_b$ of a model for each example as the fraction of times it chooses the object bias target:

$$\hat{o}_b = \frac{c_o}{c_o + c_l} \quad (2)$$

We discard all samples where a model does not choose the object or location bias target and record them separately as unclassified in the $c_u$ count. We report summary statistics for the obtained probabilities and the counts $c_o$, $c_l$, and $c_u$ for each model and prompt template in Appendix A.

**The criterion for selective encoding.** As mentioned in the introduction, we add a control task where the agent accidentally reaches for the item, meaning that the object is no longer the agent's goal. We do this by slightly changing the animate prompts. For example in one template we change *Wendy grasps the item . . .* to *Wendy falls and accidentally grasps the item . . . .* Note that although this is similar in spirit to Ullman (2023), the difference is that we show the model multiple habituations with the same change.

The criterion for saying that a model selectively encodes the goal of an actor's reach is if it exhibits a distinct bias in the animate case compared with the bias shown in the inanimate and control case. In other words, the bias in the animate case should be different from the bias in the

[3] More detailed background on this study can be found in Appendix B.

*Table 1.* The prompt variations we use in our evaluations. For each template text, the target word is **bolded**.

| Template variation | Test case | Example of differing template part |
|---|---|---|
| Fruit targets | Animate | Wendy grasps the **kiwi** |
| | Inanimate | A rod moves to and touches the **kiwi** |
| | Control | Wendy accidentally touches the **kiwi** |
| Fruit targets (anim) | Animate | A person named Wendy grasps the **kiwi** |
| | Inanimate | An inanimate rod moves to and touches the **kiwi** |
| | Control | A person named Wendy accidentally touches the **kiwi** |
| Pillar targets | Animate | Wendy grasps the item on the **first** pillar |
| | Inanimate | A rod moves to and touches the item on the **first** pillar |
| | Control | Wendy accidentally grasps the item on the **first** pillar |
| Pillar targets (anim) | Animate | A person named Wendy grasps the item on the **first** pillar |
| | Inanimate | An inanimate rod moves to and touches the item on the **first** pillar |
| | Control | A person named Wendy accidentally grasps the item on the **first** pillar |

inanimate and control case, and the latter two should be similar. If this criterion is passed, it means the model has a different bias when there is a goal-directed agent involved than when there is an inanimate or non-goal-directed agent involved. Besides the selective encoding of the goal, we can also contrast the specific bias the model demonstrates with human infants, who show an object bias in the animate case, and no bias in the inanimate case in Woodward's visual test (infants are not tested with a control task).

**Prompt variations.** We vary the agent names, pillar fruits, and inanimate objects to get a larger set of test examples (namely 100 per test case). Additionally, for each test case we design a set of four different prompts, to test for things like irrelevant alterations of the text. The first prompt has already been presented in this section. This prompt is of the type *pillar target*, because the target on which the model is evaluated is a pillar choice (first, second, or third). In the second prompt the target is not the pillar location, but the fruit itself (e.g. replace *Wendy grasps the item on the first pillar* with *Wendy grasps the kiwi*), and so the prompt is of the type *fruit target*. For both of these prompts, we also construct a variation in which we explicitly denote that the agent is animate and the inanimate object is not (e.g. replace *A pole moves to . . .* with *An inanimate pole moves to . . .* and replace *Wendy grasps . . .* with *A person named Wendy grasps . . .*). This leaves us with four prompt variations in total, which are fully presented in Table 1. Templates only differ in the sentences describing the agent's reach, otherwise they share the pattern previously shown.

Note that for the pillar target prompt variations, a prompt with multiple habituations repeats the action of reaching for the same pillar multiple times (e.g. *grasps the item on the first pillar*). Hence, a language model that is sensitive to surface-level patterns in text might put a high probability on the same pillar from the habituations to complete the test case phrase *grasps the item on the _*, which would result in a recorded location bias for this model. This is why we construct the prompt variations where the target is the fruit instead (*Fruit targets* in Table 1). In those variations, a prompt with multiple habituations repeats the action of grasping the fruit (e.g. *grasps the kiwi*). This might cause a language model to put high probability on the same object from the habituations to complete the test case phrase *grasps the _*, in which case an object bias would be encoded. Therefore, if the model is not encoding the semantics of the prompt and simply repeats surface-level patterns, we expect an object bias for the fruit target variations, and a location bias for the pillar target variations, regardless of whether the test case is animate, inanimate, or control.

## 4. Experiments

We evaluate three different models on our test cases, all of which are OpenAI API models (text-davinci-003, GPT-3.5-turbo, and GPT-4). For the latter two, we do not have access to their likelihoods — to obtain an estimate despite this we apply a sampling strategy as described in Section 3. The results are presented in Figure 2, and the numbers underlying this figure are presented in Appendix A. The left column in Figure 2 shows the results for $k = 0$ habituations, which is a sanity check that the model does not have a strong bias for a target a priori. These numbers should ideally show no bias (0.5 object bias and location bias), which is roughly the case. Below, we discuss the results for $k = 6$ habituations, which is the number of habituations Woodward (1998) uses with infants.

**Insight 1: All models show a stronger object bias in the animate case than in the inanimate case, but only GPT-3.5-turbo and GPT-4 selectively encode the goal of an**

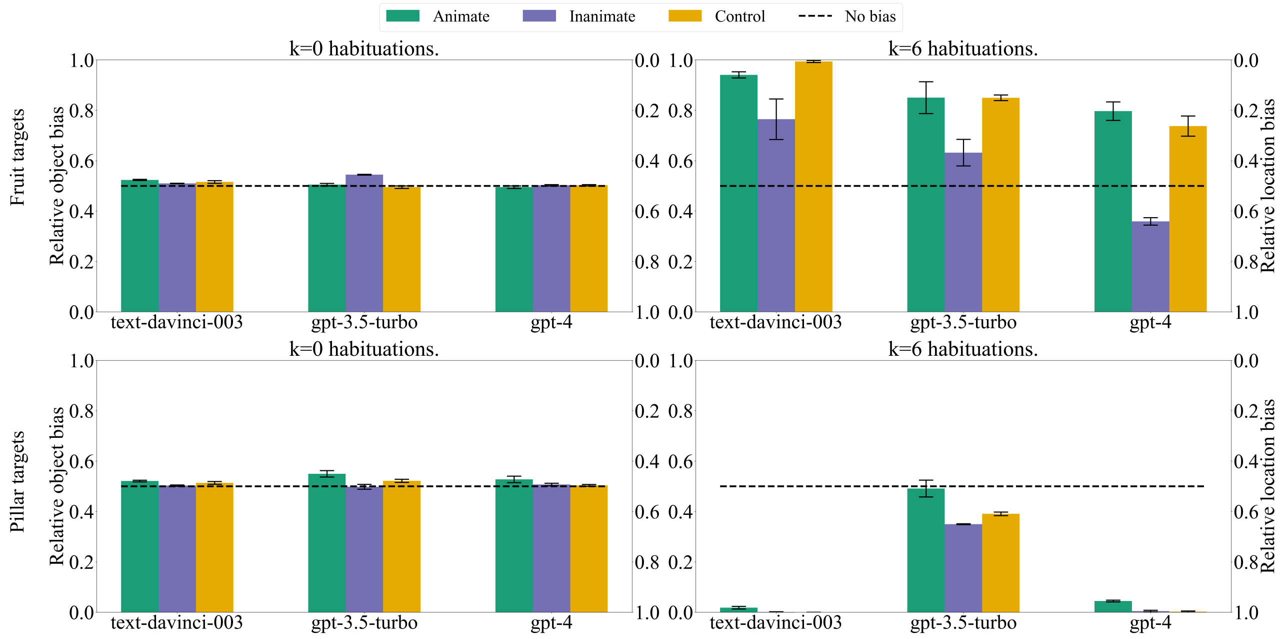

*Figure 2.* The results for text-davinci-003, GPT-3.5-turbo, and GPT-4 for $k = 0$ (left) and $k = 6$ (right) habituations. For $k = 0$, we expect the object bias to be roughly 0.5 (equal selection of object bias target and location bias target). For $k = 6$ in the right column of the figure, recall that if the model is encoding irrelevant surface-level patterns of the prompt, we expect a strong object bias for the fruit target prompt variations (top-right) and a strong location bias for the pillar target prompt variations (bottom-right), regardless of whether the test case is animate, inanimate, or control. Indeed, we observe a general stronger object bias for the top row than the bottom row when $k = 6$. We further see that all models have a higher object bias for the animate test cases than for the inanimate, but show a similar bias for the control test case as the animate case for the fruit target variations (top-right plot). GPT-3.5-turbo and GPT-4 are the only models that also show a similar bias for the control test case as the inanimate case, which means they selectively encode the goal of an agent's reach (i.e. the biases for inanimate and control are similar and distinct from animate). However, only when the target is the pillar (bottom-right plot), and GPT-4 does so only very weakly. The error bars represent the standard deviation over the two prompt templates in each group (fruit targets and pillar targets).

**agent's reach.** In general for all three test cases, we see a stronger object bias for the fruit target variations (top-right in Figure 2), and a stronger location bias for the pillar target variations (bottom-right in Figure 2). As mentioned at the end of Section 3, this is unsurprising given the repeated patterns in the habituations, and any deviation from this pattern is notable and points to encoding of semantics over surface-level properties. All models show a higher object bias in the animate case than the inanimate case, which is similar to the effect that Woodward finds for infants (object bias in the animate case, no bias in the inanimate case). The only models which pass the criterion for saying they selectively encode the goal of an agent's reach on our test set are GPT-3.5-turbo and GPT-4 (recall that the criterion is a strong difference in bias for the animate test case compared to the inanimate and control cases). However, the criterion is primarily passed for the two prompts where the targets are the pillars instead of the fruits (bottom-right), and only very weakly in GPT-4's case.

**Insight 2: text-davinci-003 does *not* appear to selectively encode the goal of an agent's reach.** Although text-davinci-003 shows a stronger object bias in the animate case than the inanimate case, it shows the same bias as the animate case in the control test case (overlapping error bars for the pillar targets in the bottom-right plot). This means we cannot say the model selectively encodes the goal of an agent's reach, because it encodes text similarly when an agent is acting purposefully as when the agent is not acting in a goal-directed fashion. Looking at the magnitude of the biases again, we see that text-davinci-003 shows a strong object bias for the fruit target templates, whereas it shows a full location bias for the pillar target templates. For the latter, it might simply be using the heuristic of repeating the pillar from habituations.

**Insight 3: All three models are heavily influenced by semantically irrelevant alterations of the prompt, but are clearly not only encoding surface-level statistics of the text.** Comparing the top-right and bottom-right plots in Figure 2, we find that all three models show much more location bias when the target is the pillar instead of the fruit. However, it is not the case that the models simply have an object bias when the target is the fruit and a location bias when the target is the pillar. Although this shows that the models' internal reasoning can be heavily influenced by superficial differences in output requirements, the strong biases that go against the surface-level repetitions do indicate encoding of the semantics of the text.

## 5. Discussion

Our results show that the tested LLMs do not robustly encode the goal-related properties of an agent's reaching action. GPT-3.5-turbo and GPT-4 do treat text differently when there is a goal-directed agent involved, but do not do this equally for semantically equivalent prompt variations. Additionally, the biases they show are very different from the bias human infants show in Woodward (1998). The specific bias we investigate is very basic, appearing in infants as young as six months old. Our results indicate that ToM-like human biases might not emerge from large-scale pre-training on text or instruction fine-tuning, at least not in the way we might expect them to. This suggests that studies investigating the emergence of ToM in LLMs should not expect a machine ToM that is comparable to human ToM, but should instead focus on identifying in what way machines reason about the mental states of others, if they do so at all. Additionally, our results show that studies need to take into account the sensitivity of models to semantically irrelevant surface-level patterns in text, which might be very different from humans' responses to such patterns. In our study we deal with this by designing prompt variations that would result in an opposite effect if only surface-level patterns are encoded. Any deviation from this pattern indicates encoding of semantics over irrelevant patterns. Our results serve as a first step towards comparing human theory of mind and machine theory of mind without preconceived notions of the kind of mentalizing the machine should do.

We take the approach of linguistically presenting a ToM test to LLMs that is traditionally only tested *visually* in prelinguistic infants. Although we view this as a strength of the protocol because it makes it less likely that the test appears in the training data, it also means that a lack of human-like bias in LLMs may simply indicate that this bias does not show up linguistically. To say LLMs show a different bias than humans in this task, we need to administer the same tests to human adults. In future work, we want to conduct human evaluations on our linguistic test to identify the biases humans show.

One hypothesis for why selectively encoding the goal object of an actor's reach has not yet emerged is that learning such a bias might simply not be consistently useful for next-token prediction in pre-training on text. Another hypothesis is that pre-training on large-scale internet data representing too many agents with noisy beliefs hindered the ToM-like ability (Andreas, 2022). In a future version of this study, we want to test if fine-tuning a pre-trained Pythia model (Biderman et al., 2023) on data reflecting agent preferences for objects, and random reaching events for inanimate objects can lead to the emergence of ToM-like ability. Successful next-token prediction on this dataset requires inferring the underlying agent preferences of the agents that occur in the data, as well as learning that inanimate objects have no preferences. Using this protocol, we can control how consistently useful the object bias is for next-token prediction by adding noise to the data, and seeing how this affects the resulting biases in the model for novel agents and objects.

Our evaluation protocol opens up further interesting avenues for future work. Although prior work in machine ToM mostly views it as a static ability that you can either have or not, current approaches to ToM in humans and other animals recognize that mentalizing inferences are dynamic (Baker et al., 2017) and graded in performance (Devaine et al., 2014). These insights have recently been applied to make progress on the Baby Intuitions Benchmark (Gandhi et al., 2021) by applying a Bayesian hierarchical framework (Langley, 2000). Since our evaluation protocol allows varying the number of habituations, future work might take a similar approach, and investigate how varying degree of observations change the model's predictions of an agent's behavior, as the studies investigating human ToM did (Devaine et al., 2014; Baker et al., 2009; 2017; Shafto et al., 2014; Yoshida et al., 2008). For example, repeated trials of hide and seek (Devaine et al., 2014) can differentiate ToM abilities in different clinical populations (d'Arc et al., 2020) and even across primate species (Devaine et al., 2017). Models taking this approach successfully generate precise quantitative predictions of how people infer preferences and beliefs of other agents over a range of parametrically controlled stimuli (Baker et al., 2017).

## 6. Conclusion

In this paper, we introduce a new evaluation protocol to test large language models' (LLMs) capabilities in the context of Theory of Mind (ToM). Inspired by Woodward (1998), we prompt LLMs with ambiguous examples of agents interacting with objects. We let the models predict the agent's next interaction, which can be either explained as an explicit agent goal in terms of location or object choice, or by random chance—allowing us to assess if *a model selectively encodes the goal of an agent's reach*. Extending the original study, we do not only test against inanimate interactions but also use a control task with accidental interactions. We apply our evaluation to a number of recent LLMs, namely text-davinci-003, GPT-3.5-turbo, and GPT-4. Our results indicate that all tested models appear to make some form of distinction between animate and inanimate actors, but only GPT-3.5-turbo and GPT-4 demonstrate the ability to selectively encode an agent's goal in such a way that it does not fail on our control task. We show that all models are highly susceptible to be influenced by minor prompt variations that do not semantically change the task.

## Acknowledgements

Hossein A. Rahmani's effort was supported by the Engineering and Physical Sciences Research Council (EP/S021566/1). Herbie Bradley's and Arduin Findeis's contributions were supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (EP/S022961/1). Arduin Findeis was also supported by a University of Cambridge School of Physical Sciences Award. Laura Ruis' contributions were supported by the EPSRC Grant EP/S021566/1 and UCL International Scholar Award for Doctoral Training Centres. We thank Stability AI for providing OpenAI API credits for our experiments.

## References

Andreas, J. Language models as agent models. *ArXiv*, abs/2212.01681, 2022.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. Action understanding as inverse planning. *Cognition*, 113:329–349, 2009.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, 2017.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

d'Arc, B. F., Devaine, M., and Daunizeau, J. Social behavioural adaptation in autism. *PLoS Computational Biology*, 16, 2020.

Devaine, M., Hollard, G., and Daunizeau, J. The social bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, 10, 2014.

Devaine, M., San-Galli, A., Trapanese, C., Bardino, G., Hano, C., Jalme, M. S., Bouret, S. G., Masi, S., and Daunizeau, J. Reading wild minds: A computational assay of theory of mind sophistication across seven primate species. *PLoS Computational Biology*, 13, 2017.

Frith, C. D. and Frith, U. Mechanisms of social cognition. *Annual review of psychology*, 63:287–313, 2012.

Gandhi, K., Stojnic, G., Lake, B. M., and Dillon, M. Baby intuitions benchmark (BIB): Discerning the goals, preferences, and actions of others. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=TFEFvU0ZV6Q.

Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *ArXiv*, abs/2302.02083, 2023.

Lampinen, A. K. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Milligan, K., Astington, J. W., and Dack, L. A. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2):622–646, 2007. doi: https://doi.org/10.1111/j.1467-8624.2007.01018.x. URL https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8624.2007.01018.x.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

Moghaddam, S. and Honey, C. J. Boosting theory-of-mind performance in large language models via prompting. *ArXiv*, abs/2304.11490, 2023.

Sap, M., Bras, R. L., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

Shafto, P., Goodman, N. D., and Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89, 2014.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. Clever hans or neural theory of mind? stress testing social reasoning in large language models, 2023.

Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. Do large language models know what humans know? *arXiv preprint arXiv:2209.01515*, 2022.

Ullman, T. D. Large language models fail on trivial alterations to theory-of-mind tasks. *ArXiv*, abs/2302.08399, 2023.

Woodward, A. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. doi: 10.1016/s0010-0277(98)00058-4.

Yoshida, W., Dolan, R. J., and Friston, K. J. Game theory of mind. *PLoS Computational Biology*, 4, 2008.

## A. Detailed results

Table 2 and 3 show the results for GPT-4 on the pillar target templates and fruit target templates respectively. Table 4 and 5 show the results for GPT-3.5-turbo on the pillar target templates and fruit target templates respectively. Table 6 and 7 show the results for text-davinci-003 on the pillar target templates and fruit target templates respectively.

*Table 2.* Animate, inanimate, and control object and location bias for gpt-4 on prompts from the group Pillar targets. H stands for habituations, and Anim for whether (Y) or not (N) the prompt template has animate denotation.

| | | | | Animate | | | Inanimate | | | Control | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | H | Anim | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified |
| gpt-4 | 0 | N | 3.8 +/- 4.5 | 3.0 +/- 4.3 | 3.2 +/- 4.4 | 3.6 +/- 4.5 | 3.4 +/- 4.4 | 3.1 +/- 4.3 | 3.6 +/- 3.9 | 3.2 +/- 3.5 | 3.2 +/- 3.6 |
| gpt-4 | 6 | N | 0.4 +/- 1.4 | 9.6 +/- 1.4 | 0.0 +/- 0.1 | 0.1 +/- 0.7 | 9.9 +/- 0.7 | 0.0 +/- 0.0 | 0.1 +/- 0.3 | 9.9 +/- 0.3 | 0.0 +/- 0.0 |
| gpt-4 | 0 | Y | 3.7 +/- 4.3 | 3.2 +/- 4.0 | 3.0 +/- 3.9 | 3.6 +/- 4.6 | 3.2 +/- 4.5 | 3.3 +/- 4.5 | 3.7 +/- 3.5 | 3.3 +/- 3.4 | 3.0 +/- 3.4 |
| gpt-4 | 6 | Y | 0.5 +/- 1.6 | 9.5 +/- 1.6 | 0.0 +/- 0.0 | 0.0 +/- 0.1 | 10.0 +/- 0.1 | 0.0 +/- 0.1 | 0.0 +/- 0.0 | 10.0 +/- 0.0 | 0.0 +/- 0.0 |

*Table 3.* Animate, inanimate, and control object and location bias for gpt-4 on prompts from the group Fruit targets. H stands for habituations, and Anim for whether (Y) or not (N) the prompt template has animate denotation.

| | | | | Animate | | | Inanimate | | | Control | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | H | Anim | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified |
| gpt-4 | 0 | N | 0.0 +/- 0.0 | 0.0 +/- 0.1 | 10.0 +/- 0.1 | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 10.0 +/- 0.0 | 0.0 +/- 0.2 | 0.0 +/- 0.2 | 9.9 +/- 0.3 |
| gpt-4 | 6 | N | 6.1 +/- 3.7 | 2.0 +/- 3.2 | 1.8 +/- 3.0 | 3.0 +/- 3.5 | 5.7 +/- 4.1 | 1.3 +/- 2.5 | 5.5 +/- 3.5 | 1.6 +/- 2.9 | 2.9 +/- 3.1 |
| gpt-4 | 0 | Y | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 10.0 +/- 0.0 | 0.0 +/- 0.1 | 0.0 +/- 0.1 | 10.0 +/- 0.2 | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 10.0 +/- 0.0 |
| gpt-4 | 6 | Y | 5.4 +/- 3.8 | 1.0 +/- 2.4 | 3.6 +/- 3.7 | 2.3 +/- 3.1 | 5.8 +/- 4.0 | 1.9 +/- 3.1 | 5.0 +/- 3.6 | 2.6 +/- 3.7 | 2.4 +/- 3.0 |

*Table 4.* Animate, inanimate, and control object and location bias for gpt-3.5-turbo on prompts from the group Pillar targets. H stands for habituations, and Anim for whether (Y) or not (N) the prompt template has animate denotation.

| | | | | Animate | | | Inanimate | | | Control | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | H | Anim | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified |
| gpt-3.5-turbo | 0 | N | 3.7 +/- 4.4 | 2.7 +/- 4.0 | 3.6 +/- 4.4 | 3.2 +/- 3.7 | 3.0 +/- 3.6 | 3.9 +/- 3.8 | 3.1 +/- 4.4 | 2.9 +/- 4.4 | 4.0 +/- 4.7 |
| gpt-3.5-turbo | 6 | N | 4.5 +/- 4.3 | 5.3 +/- 4.2 | 0.2 +/- 0.6 | 3.3 +/- 3.7 | 6.0 +/- 3.8 | 0.7 +/- 1.5 | 3.7 +/- 4.4 | 5.3 +/- 4.3 | 0.9 +/- 2.2 |
| gpt-3.5-turbo | 0 | Y | 3.2 +/- 4.4 | 2.7 +/- 4.2 | 4.1 +/- 4.6 | 3.1 +/- 4.2 | 2.9 +/- 4.0 | 4.0 +/- 4.3 | 3.2 +/- 4.4 | 2.8 +/- 4.1 | 4.1 +/- 4.6 |
| gpt-3.5-turbo | 6 | Y | 5.2 +/- 4.5 | 4.7 +/- 4.5 | 0.1 +/- 0.5 | 3.3 +/- 3.9 | 6.2 +/- 3.9 | 0.5 +/- 1.2 | 3.8 +/- 4.4 | 6.0 +/- 4.4 | 0.2 +/- 0.7 |

*Table 5.* Animate, inanimate, and control object and location bias for gpt-3.5-turbo on prompts from the group Fruit targets. H stands for habituations, and Anim for whether (Y) or not (N) the prompt template has animate denotation.

| | | | | Animate | | | Inanimate | | | Control | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | H | Anim | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified | N obj bias | N loc bias | N unclassified |
| gpt-3.5-turbo | 0 | N | 0.1 +/- 0.3 | 0.1 +/- 0.4 | 9.8 +/- 0.4 | 0.3 +/- 0.6 | 0.2 +/- 0.7 | 9.5 +/- 0.9 | 0.0 +/- 0.2 | 0.0 +/- 0.2 | 9.9 +/- 0.2 |
| gpt-3.5-turbo | 6 | N | 6.0 +/- 3.4 | 1.8 +/- 3.2 | 2.2 +/- 2.3 | 6.3 +/- 3.8 | 2.8 +/- 3.6 | 0.8 +/- 1.5 | 7.8 +/- 3.1 | 1.6 +/- 3.2 | 0.6 +/- 1.0 |
| gpt-3.5-turbo | 0 | Y | 0.1 +/- 0.3 | 0.1 +/- 0.3 | 9.9 +/- 0.4 | 0.9 +/- 1.4 | 0.7 +/- 1.1 | 8.4 +/- 1.6 | 0.0 +/- 0.2 | 0.1 +/- 0.2 | 9.9 +/- 0.3 |
| gpt-3.5-turbo | 6 | Y | 7.4 +/- 2.6 | 0.8 +/- 2.2 | 1.8 +/- 2.0 | 5.3 +/- 4.2 | 4.0 +/- 4.3 | 0.7 +/- 1.4 | 8.2 +/- 2.8 | 1.3 +/- 2.8 | 0.5 +/- 0.8 |

*Table 6.* Animate, inanimate, and control object and location bias for text-davinci-003 on prompts from the group Pillar targets. H stands for habituations, and Anim for whether (Y) or not (N) the prompt template has animate denotation.

| | | | | Animate | | | Inanimate | | | Control | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | H | Anim | Obj p | Loc p | Obj bias | Obj p | Loc p | Obj bias | Obj p | Loc p | Obj bias |
| text-davinci-003 | 0 | N | 0.3 +/- 0.2 | 0.3 +/- 0.2 | 0.5 +/- 0.2 | 0.1 +/- 0.1 | 0.1 +/- 0.1 | 0.5 +/- 0.2 | 0.3 +/- 0.2 | 0.3 +/- 0.2 | 0.5 +/- 0.2 |
| text-davinci-003 | 6 | N | 0.0 +/- 0.1 | 1.0 +/- 0.1 | 0.0 +/- 0.1 | 0.0 +/- 0.0 | 1.0 +/- 0.0 | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 1.0 +/- 0.0 | 0.0 +/- 0.0 |
| text-davinci-003 | 0 | Y | 0.3 +/- 0.2 | 0.3 +/- 0.2 | 0.5 +/- 0.2 | 0.2 +/- 0.1 | 0.2 +/- 0.1 | 0.5 +/- 0.2 | 0.3 +/- 0.2 | 0.3 +/- 0.2 | 0.5 +/- 0.2 |
| text-davinci-003 | 6 | Y | 0.0 +/- 0.1 | 1.0 +/- 0.1 | 0.0 +/- 0.1 | 0.0 +/- 0.0 | 1.0 +/- 0.0 | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 1.0 +/- 0.0 | 0.0 +/- 0.0 |

*Table 7.* Animate, inanimate, and control object and location bias for text-davinci-003 on prompts from the group Fruit targets. H stands for habituations, and Anim for whether (Y) or not (N) the prompt template has animate denotation.

| | | | | Animate | | | Inanimate | | | Control | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | H | Anim | Obj p | Loc p | Obj bias | Obj p | Loc p | Obj bias | Obj p | Loc p | Obj bias |
| text-davinci-003 | 0 | N | 0.1 +/- 0.1 | 0.1 +/- 0.1 | 0.5 +/- 0.2 | 0.1 +/- 0.0 | 0.1 +/- 0.0 | 0.5 +/- 0.2 | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 0.5 +/- 0.3 |
| text-davinci-003 | 6 | N | 0.9 +/- 0.2 | 0.1 +/- 0.2 | 0.9 +/- 0.2 | 0.7 +/- 0.4 | 0.3 +/- 0.4 | 0.7 +/- 0.4 | 1.0 +/- 0.0 | 0.0 +/- 0.0 | 1.0 +/- 0.0 |
| text-davinci-003 | 0 | Y | 0.1 +/- 0.1 | 0.1 +/- 0.1 | 0.5 +/- 0.2 | 0.0 +/- 0.0 | 0.0 +/- 0.0 | 0.5 +/- 0.2 | 0.1 +/- 0.0 | 0.1 +/- 0.0 | 0.5 +/- 0.2 |
| text-davinci-003 | 6 | Y | 1.0 +/- 0.2 | 0.0 +/- 0.2 | 1.0 +/- 0.2 | 0.8 +/- 0.3 | 0.2 +/- 0.3 | 0.8 +/- 0.3 | 1.0 +/- 0.0 | 0.0 +/- 0.1 | 1.0 +/- 0.0 |

## B. Background

Woodward (1998) shows that infants of 6- and 9-months old selectively encode the aspects of a human action that are relevant to the actor's goals over other salient aspects of the event. She does this by habituating infants to reaching actions of a demonstrator that always reaches to the same object on the same location over another object in another location. The objects then switch positions, and infant looking times are then measured in two different test cases: the actor reaches to the same object from habituation that is now in a different location (object bias) or the actor reaches to another object in the same location from habituation (location bias). Infants look longer for the location bias case, suggesting that they selectively encode the goal object of the actor's reach and not the location. Moreover, they do not show this behavior when the actor is replaced by an inanimate rod that is moved to the object (the infants only see the rod and not whatever moves it). When they are habituated with a rod, the looking times in the object and location bias test cases are comparable.