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 model's bias. 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 one model shows a selective bias for agents acting purposefully, but is biased differently than humans. 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]

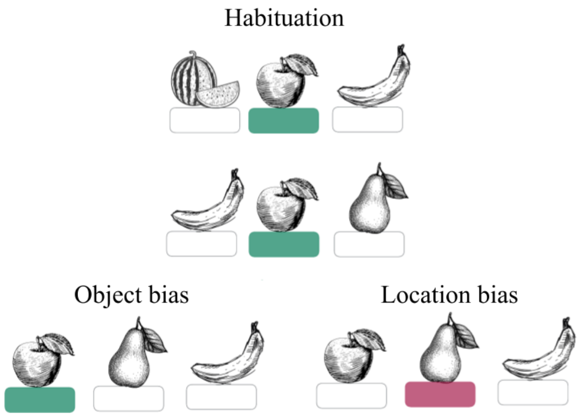

Figure 1. A visual depiction of our test inspired by Woodward (1998). We prompt an LLM with k ambiguous linguistic *habituations* that can be explained either by the goal being the object or the location ($k = 2$ in the image). We then test the bias the model shows for assuming the goal was the object (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, Kosinski (2023) shows certain LLMs can pass classic false-belief tests, but Ullman (2023) shows those same models fail on minimal alterations to these tasks that change the expected answer.[2] This 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 repeat training patterns for adversarial examples is precisely because these examples

[*] 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. Copyright 2023 by the author(s).

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

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

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 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, if machine ToM is like human ToM 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 GPT-4 passes the criterion for saying that it selectively encodes the goal of an actor's reach. However, whilst humans show no bias in the inanimate test case (Woodward, 1998), GPT-4 shows a strong location bias in both the inanimate and control test cases. Additionally, the selective encoding of the goal is drastically less strong for a semantically equivalent prompt variation. Another model (GPT-3.5-turbo), seemingly shows object bias in the animate case, and location bias in the inanimate case. However, it also shows object bias in the control task, which means we cannot say that it selectively encodes the goal of an agent's reach. From these results we conclude that although we can say that GPT-4 selectively encodes the goal of an actor's reach sometimes, it does not do so robustly, and moreover is 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.

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 fine-tuned 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.

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

(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. The question we want to investigate is 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 $p(\cdot \mid \text{prompt})$.

We do not have access to the probabilities assigned to each target by the GPT-3.5-turbo and GPT-4 models due to their 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. 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 (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 cases. 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

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

Template variation	Test case	Template text
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

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* … 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.

4. Experiments

We evaluate three different models on our test cases, all of which are OpenAI's 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 doesn't have strong bias for a target a priori. These numbers should ideally show no bias, but in reality there is a slight bias for each model. This is to be expected for a sample of 20 data points, and in Appendix C we increase the sample to 100 for one prompt template and show that this bias goes away, but the bias for $k = 6$ (which is the result we care about) remains. Below, we discuss the results for $k = 6$ habituations.

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 agent's reach. From Table 1 we know that for the fruit target templates, simply repeating the target from the habituations would result in an object bias, and for the pillar target templates repeating the pillar from the habituations would result in a location bias. Hence, a priori for language models it is unsurprising that a stronger object bias shows up in the top right of Figure 2, and a location bias in the bottom right. Any deviation from this is meaningful, as it goes against a repeated pattern in the habituations. All models show a higher object bias in the animate case than the inanimate case, which points at the result that Woodward showed (object bias in the animate case, no bias in the inanimate case). However, 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 and GPT-3.5-turbo do *not* appear to selectively encode the goal of an agent's reach. Although text-davinci-003 and GPT-3.5-turbo show an object bias in the animate case and no object bias in the inanimate case, they also show an object bias in the control test case. This means we cannot say the models selectively encode the goal of an agent's reach, because they also do so 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 show drastically more location bias when the target is the pillar than when the

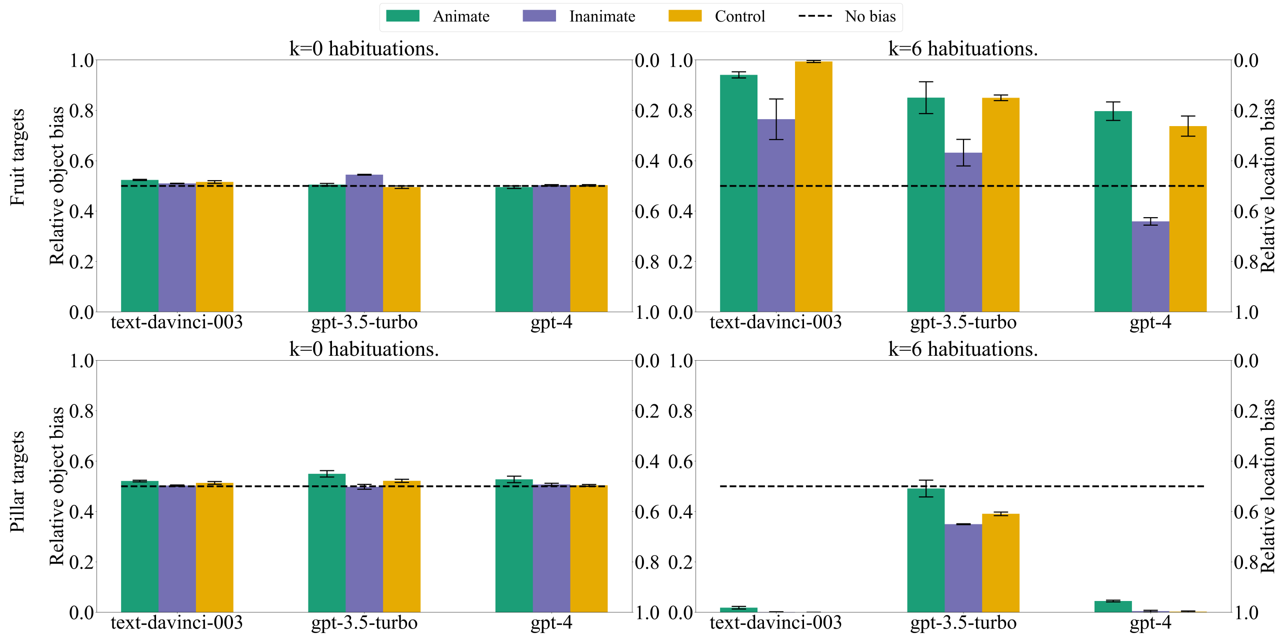

Figure 2. The results for text-davinci-003, GPT-3.5-turbo, and GPT-4 for $k = 0$ (left) and $k = 6$ (right) habituations. For $k = 0$, we expect the object bias to be roughly 0.5 (equal selection of object bias target and location bias target). For $k = 6$, we see that all models have a higher object bias for the animate test cases than for the inanimate, but the only models that pass the criterion for saying they selectively encode the goal of an agent's reach are GPT-3.5-turbo and GPT-4 (i.e. the biases for inanimate and control are similar and distinct from animate). However, only when the target is the pillar (bottom row), 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).

target is the fruit. 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 model simply has an object bias when the target is the fruit and a location bias when the target is the pillar, which can be seen from the location biases that GPT-3.5-turbo and GPT-4 show when the target is the fruit. Intuitively, this appears to indicate that the models' internal reasoning can be heavily influenced by superficial differences in output requirements that in humans might not lead to the same biases.

5. Discussion

Our results show that the tested LLMs do not robustly encode the goal-related properties of an agent's reaching action. GPT-4 does treat text differently when there is a goal-directed agent involved, but it does not do this strongly for all prompt variations. Additionally, the bias it shows is very different from the bias human infants show. 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. 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 pre-linguistic 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. This could be tested in future work by conducting human evaluations on our linguistic test.

Another 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. In a future version of this study, we want to test this hypothesis by 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. Successful next-word 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-word 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 (**?**). Since our evaluation protocol allows varying the number of habituations, future work might take a similar approach, and investigate how repeated observations change the model's predictions of an agent's behavior. In studies investigating human ToM, experimenters capture how repeated observation of others' behavior affect mentalizing (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-4 selectively encodes an agent's goal in such a way that it does not fail on our control 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

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.

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. 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	7.4 +/- 3.7	1.9 +/- 3.3	0.7 +/- 2.3	7.1 +/- 4.0	2.3 +/- 3.5	0.6 +/- 2.3	6.0 +/- 3.9	1.9 +/- 2.3	2.1 +/- 3.0
gpt-4	6	N	0.8 +/- 1.6	9.2 +/- 1.6	0.0 +/- 0.0	0.0 +/- 0.0	10.0 +/- 0.0	0.0 +/- 0.0	0.1 +/- 0.2	9.9 +/- 0.2	0.0 +/- 0.0
gpt-4	0	Y	5.7 +/- 4.5	2.5 +/- 3.3	1.8 +/- 3.1	7.7 +/- 3.9	2.3 +/- 3.9	0.0 +/- 0.0	4.8 +/- 3.3	2.9 +/- 2.1	2.3 +/- 2.8
gpt-4	6	Y	0.5 +/- 1.1	9.5 +/- 1.1	0.0 +/- 0.0	0.1 +/- 0.2	9.9 +/- 0.2	0.0 +/- 0.0	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	4.6 +/- 3.9	3.1 +/- 3.4	2.3 +/- 3.2	5.1 +/- 3.5	1.8 +/- 2.0	3.1 +/- 3.7	3.7 +/- 3.5	2.2 +/- 3.3	4.1 +/- 3.6
gpt-4	6	N	4.0 +/- 4.1	6.0 +/- 4.1	0.0 +/- 0.0	1.8 +/- 3.2	8.2 +/- 3.2	0.0 +/- 0.0	2.8 +/- 3.7	7.2 +/- 3.7	0.0 +/- 0.0
gpt-4	0	Y	4.9 +/- 3.8	3.7 +/- 3.5	1.4 +/- 2.8	6.2 +/- 4.1	2.5 +/- 4.0	1.3 +/- 2.4	1.7 +/- 1.6	1.4 +/- 2.6	6.9 +/- 2.4
gpt-4	6	Y	4.8 +/- 4.0	5.2 +/- 4.0	0.0 +/- 0.0	1.0 +/- 2.5	9.0 +/- 2.5	0.0 +/- 0.0	1.8 +/- 3.1	8.2 +/- 3.1	0.0 +/- 0.0

Table 4. Animate, inanimate, and control object and location bias for GPT-3.5-turbo on 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	2.0 +/- 3.6	5.0 +/- 4.8	3.0 +/- 3.9	2.2 +/- 2.8	3.9 +/- 4.0	3.9 +/- 3.5	2.4 +/- 4.0	5.8 +/- 4.9	1.9 +/- 3.6
gpt-3.5-turbo	6	N	3.4 +/- 4.2	6.3 +/- 4.2	0.3 +/- 0.9	2.5 +/- 3.4	6.8 +/- 3.5	0.6 +/- 1.2	2.5 +/- 4.0	6.7 +/- 4.3	0.8 +/- 2.1
gpt-3.5-turbo	0	Y	1.9 +/- 3.3	6.4 +/- 4.4	1.7 +/- 2.9	2.1 +/- 3.5	4.4 +/- 4.2	3.4 +/- 3.7	2.5 +/- 4.0	5.4 +/- 4.5	2.1 +/- 3.3
gpt-3.5-turbo	6	Y	4.0 +/- 4.4	6.0 +/- 4.3	0.1 +/- 0.4	1.9 +/- 2.9	7.5 +/- 3.1	0.7 +/- 1.5	3.4 +/- 4.3	6.5 +/- 4.2	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	4.2 +/- 3.1	2.6 +/- 2.6	3.2 +/- 2.8	5.0 +/- 3.6	2.2 +/- 2.9	2.8 +/- 3.1	3.3 +/- 2.8	3.5 +/- 2.8	3.2 +/- 2.4
gpt-3.5-turbo	6	N	5.2 +/- 4.4	4.7 +/- 4.3	0.2 +/- 0.5	2.8 +/- 3.8	6.8 +/- 3.8	0.4 +/- 1.0	6.5 +/- 4.3	3.5 +/- 4.4	0.1 +/- 0.2
gpt-3.5-turbo	0	Y	3.9 +/- 3.1	2.6 +/- 2.7	3.5 +/- 3.1	3.8 +/- 3.2	2.4 +/- 2.1	3.8 +/- 3.1	2.9 +/- 2.7	2.8 +/- 2.6	4.3 +/- 2.6
gpt-3.5-turbo	6	Y	8.0 +/- 3.7	1.9 +/- 3.7	0.1 +/- 0.4	2.8 +/- 3.9	7.2 +/- 4.0	0.1 +/- 0.3	5.8 +/- 4.4	3.8 +/- 4.5	0.4 +/- 1.1

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	0.4 +/- 0.2	0.2 +/- 0.1	0.6 +/- 0.2	0.2 +/- 0.1	0.1 +/- 0.1	0.6 +/- 0.2	0.4 +/- 0.2	0.2 +/- 0.1	0.6 +/- 0.2
text-davinci-003	6	N	0.0 +/- 0.0	1.0 +/- 0.0	0.0 +/- 0.0	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	0.4 +/- 0.2	0.2 +/- 0.1	0.6 +/- 0.2	0.2 +/- 0.2	0.2 +/- 0.1	0.6 +/- 0.2	0.3 +/- 0.2	0.2 +/- 0.1	0.6 +/- 0.2
text-davinci-003	6	Y	0.0 +/- 0.0	1.0 +/- 0.0	0.0 +/- 0.0	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	0.3 +/- 0.2	0.3 +/- 0.1	0.5 +/- 0.2	0.2 +/- 0.2	0.2 +/- 0.1	0.6 +/- 0.2	0.4 +/- 0.2	0.3 +/- 0.2	0.6 +/- 0.2
text-davinci-003	6	N	0.9 +/- 0.2	0.1 +/- 0.2	0.9 +/- 0.2	0.6 +/- 0.4	0.4 +/- 0.4	0.6 +/- 0.4	1.0 +/- 0.0	0.0 +/- 0.0	1.0 +/- 0.0
text-davinci-003	0	Y	0.4 +/- 0.2	0.3 +/- 0.2	0.6 +/- 0.2	0.2 +/- 0.1	0.1 +/- 0.1	0.6 +/- 0.3	0.3 +/- 0.1	0.2 +/- 0.1	0.6 +/- 0.2
text-davinci-003	6	Y	0.9 +/- 0.2	0.1 +/- 0.3	0.9 +/- 0.3	0.7 +/- 0.4	0.3 +/- 0.4	0.7 +/- 0.4	1.0 +/- 0.1	0.0 +/- 0.1	1.0 +/- 0.1

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.

C. Results on a larger dataset

We ran the study for one of the templates (fruit targets with animate denotation) on GPT-3.5-turbo, GPT-4, and text-davinci-003 on 80 additional examples, to see whether that would change the results. In Figure 3 the results are shown. We observe that the bias almost fully disappears for $k = 0$ (which is the expected result), and the result for $k = 6$ is similar to the result on the 20 examples from the main paper. We therefore refrain from running the evaluation on more than 20 examples in the rest of the experiments.

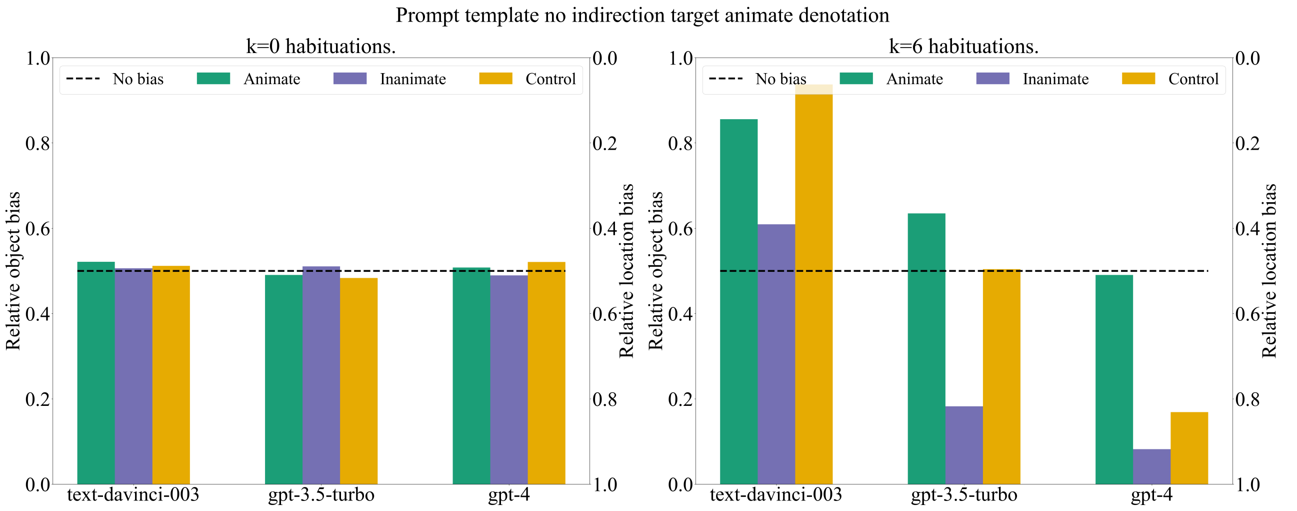

Figure 3. The results for 100 examples with one of the prompt variations (fruit targets with animate denotation). We find that the biases with $k = 0$ habituations disappear as expected, and the biases for $k = 6$ remain similar to those found for 20 examples (see Appendix A for the numbers broken down per prompt variation underlying this plot).