How infants' understanding of goal-directed actions differs from a large language model

Alyson Wong (alyson_wong@berkeley.edu), Bill D. Thompson* (wdt@berkeley.edu), Fei Xu* (fei_xu@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

Abstract

Theory of mind (ToM) is a hallmark feature of human cognition that emerges very early in development. Much work has explored human infants' implicit social reasoning abilities. Recent work has examined whether LLMs reliably make ToM inferences in explicit social reasoning tasks. However, it remains unclear how reliably LLMs generate human-like social reasoning when this capacity is invoked implicitly. We systematically examined GPT-4's ability to implicitly reason about goal-directed actions by adapting well-studied infant paradigms. Our results suggest that, unlike infants who can understand goal-directed actions from a very young age, GPT-4 fails to correctly attribute goal-directed actions to agents. These findings suggest that LLMs may lack key aspects of implicit social reasoning and provide insight into the emergence of these abilities in infants.

Keywords: social reasoning**;** theory of mind; human infants; large language models

Introduction

The capacity to make consistent inferences about another agent's goals and intentions is critical to the viability of conversational AI systems. AI systems that do not understand the complex relations between the actions people take and their underlying beliefs and goals would be severely limited in their usefulness and create potential risks when applied to real-world problems such as education or healthcare. For this reason, there has been a surge of recent research in cognitive science examining Theory of Mind (ToM) abilities, or the ability to reason about others' mental states, in Large Language Models (LLMs). Existing studies exploring *explicit* ToM show mixed results, finding that LLMs could succeed at simple Sally-Anne tasks and predict other agents' mental states, but failed at more complex variations of the same task and to use this knowledge to reason about and predict agents' future behavior (Gandhi et al., 2023; Gu et al., 2024). Other work has additionally explored ToM in more applied settings, finding that an LLM struggled to update its beliefs upon being provided new information about the user's preferences (Qiu et al., 2024). However, less work has explored *implicit* ToM in LLMs, which is potentially relevant to a wider range of contexts: LLMs may be building implicit models of a user's mental states during any multi-turn interaction. Additionally, relations between perceptions, beliefs, and intentions are often not verbalized explicitly during discourse and may therefore be under-represented in the data that LLMs are trained on.

Much work has explored implicit ToM in infants, finding that these abilities are early emerging in humans. By 12 months of age, infants show an understanding of goal-directed actions, efficiency, preferences, value of goals based on costs, and can even make moral evaluations based on social behavior (Gergely & Csibra, 1995; Hamlin, Wynn & Bloom, 2007; Liu et al., 2007; Onishi & Baillargeon, 2005; Wellman et al., 2016; Woodward, 1998). Although some work has explored implicit social reasoning abilities in LLMs using infant-inspired methods (Kosoy et al., 2023; Ruis et al., 2023), to our knowledge, no research has systematically investigated whether LLMs perform similarly on these tasks as infants.

In the current study, we explore LLMs' implicit social reasoning abilities by converting a well-established infant paradigm to verbal prompts and systematically manipulating variables to investigate which factors LLMs may be using to reason about social scenarios. Given that these social reasoning abilities in humans are early emerging well before language is developed and may also be present in non-human animals (see Krupenye & Call, 2019 for a review), implicit ToM may be independent of language, thus suggesting LLMs may not acquire the same social reasoning abilities given their training is entirely language-based.

Relatedly, infants have a set of domain-specific mechanisms to reason about the world, e.g., an intuitive psychology that reasons specifically about agents versus an intuitive physics that reasons specifically about inanimate objects. In particular, infants expect a human hand to exhibit goal-directed behavior but not other inanimate, perceptually hand-like objects (Spelke & Kinzler, 2006; Spelke, 2022; Woodward, 1998; see also Carey, 2009). The agent system consists of its own set of core concepts such as goals, preferences, desires, and beliefs, whereas the object system is governed by principles of solidity, cohesion, continuity, and contact (Anguiar & Baillargeon, 1999; Leslie & Keeble, 1987; Spelke et al., 1992). The training data that the LLMs receive include many, many domains of knowledge, and LLMs are not instructed to categorize their knowledge into specific domains (see Palmarini & Mitchell, 2024, for evidence that a multimodal LLM failed to learn aspects of the object concept). It is possible that LLMs may be lacking this domain-specificity that is foundational to infant conceptual development. Specifically, the distinction between agents and non-agents may be critical for developing these implicit social reasoning abilities. In our

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

study, we explore whether LLMs perform comparably to infants in interpreting goal-directed actions, and whether goal-directed actions are specific to understanding agents.

Experiment 1

Methods

We assessed GPT-4 using five prompts designed to capture the structure of the Woodward (1998) infant paradigm. Responses were sampled from the model using a moderately high sampling temperature of 0.7, determined from pilot results indicating low variability in responses despite the high sampling temperature.

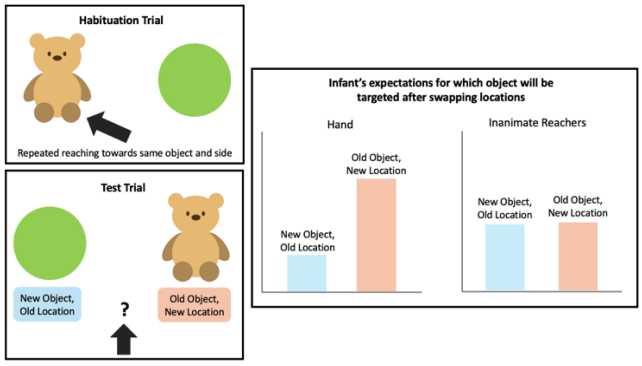

Figure 1. Depiction of Woodward (1998) paradigm and results

In the original study, infants viewed different types of agents (e.g. a hand) or inanimate agent-like objects (e.g. a metal claw, a hand-shaped sponge) repeatedly reaching towards one of two objects during a series of habituation trials (Figure 1). Most importantly, the two objects were always presented on their same respective sides (e.g. teddy bear always on left and ball always on right), thus confounding whether the reacher was displaying a preference for the specific object (e.g. the teddy bear) or the specific location (e.g. any object on the left). During the critical test trials, the location of the previously targeted object was switched (e.g. teddy bear now on right and ball on left). Using a violation of expectation paradigm, Woodward (1998) measured infants' looking times to two test events: (1) the reacher reaching for the previously targeted object (e.g. teddy bear now on right) or (2) the reacher reaching for a new object in the previously targeted location (e.g. ball on left despite object being different).

To generate prompts depicting this scenario verbally, we first set the context with a description of the scene (*"Here are two toys…"*). We systematically manipulated the following variables to create a structured set of scenarios that assess whether GPT-4 provides answers similar to those of human infants on a comparable task or whether it uses lower-level features to reason about goal-directed actions.

Reacher. As in Woodward (1998), we manipulated the type of reacher performing the reaching action. Although Woodward (1998) tested three different variations of inanimate objects as reachers in addition to the human hand,

Table 1: General prompt template (left) and prompt variation differences in habituation trials (right)

Section	Template
Habituation Intro	Here are two toys, a [TARGET TOY] and a [OTHER TOY]. The [TARGET TOY] is on the [TARGET DIRECTION] and the [OTHER TOY] is on the [OTHER DIRECTION].
*Habituation Trial *(Omitted from Habituations Collapsed)*	[A/The] [REACHER] [again] moves towards the [TARGET DIRECTION] and [grasps/touches] the [TARGET TOY] for a few moments. Then, the [REACHER] [releases/moves away from] the [TARGET TOY] and retracts from view.
Test Intro	The toys now swap positions, with the [TARGET/OTHER TOY] on the [OTHER/TARGET DIRECTION] and the [OTHER/TARGET TOY] on the [TARGET/OTHER DIRECTION]. The same [REACHER] moves in to [grasp/touch] one of the toys.
Question	Which toy will the [REACHER] [grasp/move towards]? Here are two options
Answers	A: The [REACHER] will [grasp/move towards] the [TARGET/OTHER TOY] B: The [REACHER] will [grasp/move towards] the [OTHER/TARGET TOY]

Prompt Variation	Prompt Differences
Target Direction First	A hand moves towards the left and grasps the white teddy bear…
Target Toy First	A hand moves towards the white teddy bear on the left and grasps the toy…
Target Direction Only	A hand moves towards the left and grasps the toy…
Target Toy Only	A hand moves towards the white teddy bear and grasps the toy…
Habituations Collapsed	Six times in a row, a hand moves towards the left and grasps the white teddy bear…

infants performed comparably on all three variations. We thus have limited our study to only explore the hand and a hand-resembling sponge. In addition to the difference in the reacher, a critical difference between the hand and sponge objects from Woodward (1998) is that the hand engages in a *grasping* action of the target object, whereas the sponge only *touches* the target object. We maintained these reacher-specific verbs when generating prompts.

Frequency of relevant terms. We manipulated the frequency of terms relating to either the target direction or target object of the reacher. Because we had hypothesized that GPT-4 may use low-level features to reason about the scenario, such as the frequency of terms indicating reacher direction, we examined three versions: (1) a prompt containing both the target direction and target toy during habituation trials, (2) only containing the target direction and removing the target toy, and (3) only containing the target toy and removing the target direction.

Ordering of relevant terms. We additionally manipulated the ordering of such relevant terms. Regardless of the frequency of these terms (target toy and target direction), we had reasoned that the model may be more sensitive to e.g. the first or most recent relevant term mentioned (i.e. report new goal/target direction if direction is mentioned first in the prompt).

Number of habituation trials. We manipulated the number of habituation trials within each of the above versions. LLMs are known to be able to learn from examples provided within the prompt and be sensitive to the number and structure of in-context examples. We made use of this connection between habituation trials and in-context learning to assess whether the model was systematically sensitive to the amount of evidence it was provided for goal-directed action. The number of habituation trials is infant-controlled in typical infant studies, such that the study proceeds to the critical test events once the infant's looking times to the habituation event substantially decreases. To adapt this for our purposes and because we are interested in exploring in-context learning, we varied the total number of habituation trials to be either one (to provide as little opportunity for in-context learning as possible), six (the minimum number of habituation trials for infants as per Woodward (1998)), nine (the average number of habituation trials for infants in Woodward (1998)), or 14 (the maximum number of habituation trials as per Woodward (1998)).

Additional Control. Previous work has not systematically explored habituation trials in this manner as far as we are aware. It is possible that repetitions within the prompt (reflecting the number of habituation trials) may interfere with GPT-4's reasoning abilities, as repetitions may seem unnatural and may be perceived as an error. To control for this, we additionally tested GPT-4 on a more naturalistic prompt with the repetitions collapsed into a single summary statement (e.g. *"four times in a row…"*).

Several other factors were randomized during prompt generation, such as the target toy (teddy bear, multi-colored ball), the target direction (left, right), the ordering of objects (target toy first, target toy last), as well as the correct response option (A, B). Each variation including randomizations was provided to GPT-4 a total of five times (independent calls to the GPT-4 API), resulting in a total of 3200 trials (640 for each of the five versions).

Data Coding and Analysis

To assess whether results from GPT-4 are comparable to the original infant results, we categorized GPT-4's responses into three categories: reporting that the reacher would go for (1) the old goal (i.e. new location), (2) the new goal (i.e. old location), or (3) reporting that there was not enough information in the prompt to guess which object will be selected. In one analysis, we analyzed the likelihood of responding *new goal* versus *old goal*. Infants implicitly expect agents (i.e. the hand) to engage in goal-directed action and therefore reach towards the old goal object despite it being in a new location. However, they do not expect this behavior from inanimate non-agents (i.e. the sponge). If the model responds in similar ways to infants, we should thus expect more old-goal responses for the hand and no difference in responses for hand-like sponge. In a separate analysis, we explored the likelihood of responding that there was not enough information provided.

Results

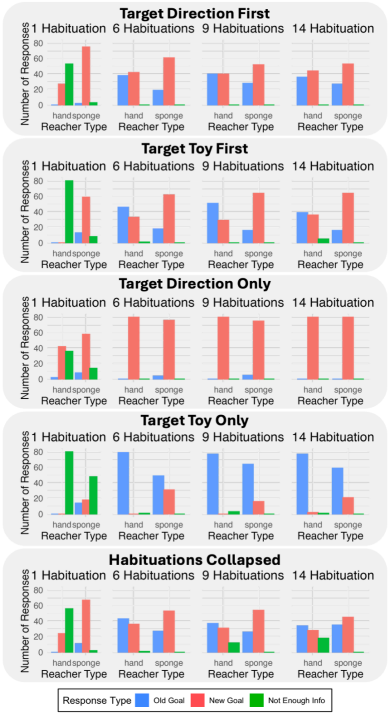

Figure 2: Results from Experiment 1

We conducted several generalized linear models for each prompt version separately, predicting GPT-4's responses as outlined above with the Reacher (hand vs. sponge), Number of Habituation Trials (1, 6, 9, 14), Target Toy (white teddy bear vs. multi-colored ball), Target Location (left vs. right), Response Order (A vs. B), and all interactions. All results are reported in Table 2.

Notably, when predicting Old Goal responses, the effect of Reacher and any interactions including Reacher were only significant in the Toy First variation, with the only exceptions being a significant Reacher x Target Direction interaction and a significant Reacher x Number of Habituation Trials x Target Direction interaction for the Habituations Collapsed variation. This suggests that GPT-4 did not reliably differentiate between reachers.

Additionally, for responses reporting not enough info, there was a significant effect of Number of Habituation trials for all variations except for the Direction First variation, with fewer "not enough info" responses with more habituation trials.

Table 2: Experiment 1 results from GLMs exploring Old/New Goal and "Not Enough Info" responses

	Direction First		**Toy First**		**Direction Only**		**Toy Only**		**Habituations Collapsed**	
Effect/Interaction	χ^2	*p*	χ^2	*p*	χ^2	*p*	χ^2	*p*	χ^2	*p*
Old vs. New Goal Responses										
Reacher	9.66	0.09	36.63	<.001***	2.14	0.14	1.91	0.17	6.42	0.27
Number Habituation Trials	25.27	<.001***	0.04	0.84	5.08	0.02*	19.62	0.01**	17.44	<.001***
Target Toy	11.36	0.04*	11.7	0.02*	0	1	22.9	<.001***	2.99	0.7
Response Order	2.08	0.56	6.23	0.1	3.1	0.08	21.05	0.01**	1.97	0.58
Number Habituation Trials x Target Toy	0.25	0.62	6.77	0.01**	0	1	9.81	0.04*	0.62	0.43
Reacher x Target Direction	.75	.69	17.86	<.001***	0	1	0	1	7.64	0.02*
Target Toy x Response Order	1.44	0.49	2.54	0.28	0	1	17.78	<.001***	1.19	0.55
Reacher x Number Habituation Trials x Target Toy	0.53	0.47	8.07	<.001***	0	1	0	1	1.56	0.21
Reacher x Number Habituation Trials x Target Direction	0.12	0.73	0.18	0.67	0	1	0	1	5.73	0.02*
Reacher x Target Toy x Target Direction	0.44	0.51	5.19	0.02*	0	1	0	1	1.85	0.17
"Not Enough Info" Responses										
Reacher	5.73	0.02*	17.64	<.001***	2.14	0.71	7.37	0.06	0.16	0.69
Number Habituation Trials	0	0.99	29.94	<.001***	0	0.01**	29.7	<.001***	10.4	0.01**
Target Toy	1.87	0.6	2.74	0.25	2.3	<.001***	7.07	0.22	3.63	0.06
Target Direction	1.16	0.56	2.48	0.29	0.12	0.58	6.06	0.42	4.12	0.04*
Response Order	1.98	0.37	1.04	0.59	4.02	0.01**	8.22	0.22	1.18	1
Reacher x Target Toy	0	1	0	1	0.22	1	0	1	0	0.01**
Number Habituation Trials x Target Toy	0	1	2.49	0.11	0	0.04*	0	1	7.56	1
Reacher x Target Direction	0.58	0.45	0	1	0	1	2.5	0.11	0	0.03*
Target Toy x Response Order	0.12	0.3	0	1	0.98	<.001***	5.03	0.08	0.57	0.17

Note: No effects or interactions other than those listed were significant (*p*s > .1).

Discussion

These results suggest that GPT-4 does not reason about goal-directed actions in this scenario in the same way that infants from a very young age are capable of. Unlike infants, who hold the strongest expectations for goal-directed behavior of the hand (representing a human agent who should have preferences and goals), *GPT-4 fails to consistently distinguish between the hand and the inanimate reacher, a hand-like sponge*.

Additionally, our results suggest that GPT-4 is most sensitive to lower-level features when reasoning about goal-directed actions. From our variations manipulating the frequency of either the intended target direction or the intended target toy, a clear pattern emerges such that *GPT-4 will answer with the response that was presented most frequently, regardless of the type of reacher*.

Lastly, our results may indicate some effects of in-context learning based on the number of habituation trials, as exhibited by the increase in reporting that the prompt does not contain enough information to make a response for prompts with fewer habituation trials.

There are, however, some differences between how infants were tested in Woodward (1998) and how the ChatGPT was prompted. Given that infants have direct experience with hands and goal-directed actions before the study even begins (e.g. reaching for toys when playing), we had reasoned that infants may have inadvertently been given extensive prior context which may have facilitated performance that the model did not receive. In Experiment 2, we explore whether providing additional context and more natural scenarios improved performance.

Experiment 2

Methods

We adapted our prompt-generation process from Experiment 1 to create new prompts using the following scenarios. These scenarios were designed to explore whether additional and more naturalistic context may be necessary for GPT-4 to recognize goal-directed action and overcome its bias towards lower-level cues. We additionally manipulated several of the same variables as in Experiment 1 (Reacher, number of habituation trials, randomizations). We again tested GPT-4 on these three prompts with the model set to a moderately high sampling temperature of 0.7.

Low-Context Version. Adapting the prompts from Experiment 1, we first set the scene in the context of watching a play and being presented with toys on a stage (much like in the original infant paradigm; "*Pretend you are watching a play…*"). To avoid failure due to perceiving the habituation trial repetition as an unintentional error, we provided additional context that we expected would make the repetition seem more natural, by describing sequences of observations as multiple scenes of a play (e.g. *"In the first scene…*").

High-Context Low-Naturalistic Version. To provide even more context that reinforces the potential for goal-directed, preference-based action, we described the context in terms of viewing desserts being selected from inside a refrigerator over a sequence of days (e.g. *"pretend you are inside a refrigerator watching desserts be selected to eat…*"). Similarly to the Low-Context version, we additionally provided more context for habituation trial repetitions (e.g. *"On the first day…*"). However, because this scenario is still somewhat unnatural, we created an additional scenario to provide a more naturalistic context. Additionally, we examined this version with an additional reacher from Woodward (1998), a claw, as this may be more natural for the given context (e.g. a claw vending machine).

High-Context High-Naturalistic Version. In an effort to provide the most naturalistic setting that the model may be familiar with, we ran an additional version taking place watching a child select toys to play with. We reasoned that this language and this type of context may be the most familiar to the model. This version was only run with the hand as the reacher, as failure to attribute goal-directed action in this version would strongly suggest a difference in social reasoning compared to infants (who show the strongest goal-directed expectations for the hand).

Past-Oriented Versions. The previous prompts all required GPT-4 to reason about the future, as they ask GPT-4 to predict what event will happen next. However, this introduces other factors that may influence GPT4's responses (e.g. the model may think after 14 days of selecting the same dessert, the agent now wants something different). To control for this, we examined additional variations of the three previous scenarios but asked the model to reason about what happened in the past (e.g. *"On the last day, the lights on the stage went out…what toy did the hand grasp?"*).

Data Coding and Analysis

Data coding and analysis procedures are identical to procedures from Experiment 1.

Results

We conducted several generalized linear models for each prompt version separately, predicting GPT-4's responses (Old/New Goal; Not Enough Info) with the Reacher (hand vs. sponge vs. claw where applicable), Number of Habituation Trials (1, 6, 9, 14), Target Object (bear/cake vs. ball/watermelon), Target Location (left vs. right), Response Order (A vs. B), and all interactions.

Low-Context Version. Results for the model exploring Old/New Goal responses revealed the following significant

effects and interactions: Reacher ($\chi^2(3)=17.78$, $p<.001$), Number of Habituation Trials ($\chi^2(6)=18.96$, $p=.004$), Target Object ($\chi^2(7)=19.37$, $p=.007$), Reacher x Number of Habituation Trials ($\chi^2(10)=13.53$, $p=.008$), Number of Habituation Trials x Target Object ($\chi^2(4)=14.53$, $p=.006$), and no other effects or interactions ($ps>.1$). A model exploring Not Enough Info responses revealed significant effects of Number of Habituation Trials ($\chi^2(3)=13.30$, $p=.004$), Target Object ($\chi^2(1)=5.45$, $p=.019$), and no other effects or interactions ($ps>.07$).

Low-Context Past Oriented Version. A model exploring Old/New Goal responses revealed no significant effects or interactions ($ps>.2$), with overall more "New Goal" responses. A model exploring Not Enough Info responses revealed the following significant effects and interactions: Number of Habituation Trials ($\chi^2(1)=76.38$, $p<.001$), Target Object ($\chi^2(1)=25.40$, $p<.001$), Target Direction ($\chi^2(1)=5.14$, $p=.023$), Response Order ($\chi^2(1)= 11.56$, $p<.001$), Reacher x Response Order ($\chi^2(1) =10.59$, $p=.001$), Target Object x Response Order ($\chi^2(1)=7.07$, $p=.008$), Reacher x Number of Habituation Trials x Response Order ($\chi^2(1)= 7.72$, $p=.005$), Reacher x Number of Habituation Trials x Target Object x Target Direction ($\chi^2(1)=4.93$, $p=.026$), Number of Habituation Trials x Target Object x Target Direction x Response Order ($\chi^2(1)=5.53$, $p=.019$), and no other effects or interactions ($ps>.1$).

High-Context Low-Naturalistic Version. Models exploring both Old/New Goal and Not Enough Info responses revealed no significant effects or interactions ($ps>.7$), with overall more "New Goal" responses.

High-Context Low-Naturalistic Past Oriented Version. Models exploring Old/New Goal responses and Not Enough Info responses revealed only a significant effect of Target Toy ($\chi^2(5)=12.79$, $p=.025$) for not enough info responses and no other significant effects or interactions ($ps>.1$), with overall more "New Goal" responses.

High-Context High-Naturalistic Version. Reacher was removed from these models as this version was run only with the hand. All other predictors and interactions were retained. A model exploring Old/New Goal responses revealed a significant effect of Number of Habituation Trials ($\chi^2(1)=4.15$, $p=.042$), and no other significant effects or interactions ($ps>.05$), with overall more "New Goal" responses. A model exploring Not Enough Info responses revealed no significant effects or interactions ($ps>.05$).

High-Context High-Naturalistic Past Oriented Version. Reacher was also removed from these models since this version was only run with the hand as the reacher. Models exploring both Old/New Goal and Not Enough Info responses revealed no significant effects or interactions ($ps>.1$), with overall more "New Goal" responses.

Discussion

Our results suggest that the context may not necessarily improve GPT-4's performance. Contrary to our expectations, the model performed significantly worse with additional context, such that it failed to distinguish between different reachers and also failed to attribute goal-directed behavior to the hand for all prompt versions, regardless of the degree of context, whether or not the context was naturalistic, or whether the model asked about past or future events.

General Discussion

The current study investigated whether LLMs perform comparably on implicit social reasoning tasks as infants do from a very young age, specifically using the case of goal-directed actions. Our results suggest that LLMs do not reason about goal-directed actions in the same way that infants do, but rather may use lower-level cues, such as the frequency of relevant terms in a given prompt, to reason about social scenarios. Additionally, contrary to our hypothesis, we did not find any evidence that increasing contextual factors improves the model's performance. However, we did find evidence of differences in responses based on the number of habituation trials, suggesting that GPT-4 was sensitive to in-context learning in our paradigm. Our findings provide two main contributions to the developmental cognitive science and artificial intelligence fields.

First, given GPT-4's failure to perform comparably to infants, this may suggest that implicit social reasoning abilities do not emerge from language alone. Although infants succeed on these tasks well before acquiring a natural language, there is evidence that infants even by 6 months already know many words, such as nouns referring to food and body-parts (Bergelson & Swingley, 2011). This makes disentangling language from other domains difficult, as language and precursors to language are present very early in development. LLMs, which are trained only on language and only have language capabilities, nonetheless provide a tool for exploring what abilities can emerge from language alone.

Second, GPT-4's failure may additionally provide insight into the importance of domain-specificity in infants' conceptual system. Given that LLMs may lack domain-specificity whereas infants' conceptual system is already domain-specific very early on, with a system specifically for representing agents, this may suggest that infants' domain-specificity enables more complex implicit social reasoning. In future work, we will explore other implicit reasoning abilities in LLMs to form a more comprehensive picture of LLMs' social reasoning capabilities, as well as investigate whether other LLMs perform comparably.

References

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). *Understanding Social Reasoning in Language Models with Language Models* (No. arXiv:2306.15448). arXiv. https://doi.org/10.48550/arXiv.2306.15448

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. https://doi.org/10.1016/0010-0277(95)00661-h

Gu, Y., Tafjord, O., Kim, H., Moore, J., Bras, R. L., Clark, P., & Choi, Y. (2024). *SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs* (No. arXiv:2410.13648). arXiv. https://doi.org/10.48550/arXiv.2410.13648

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559. https://doi.org/10.1038/nature06288

Kosoy, E., Reagan, E. R., Lai, L., Gopnik, A., & Cobb, D. K. (2023). *Comparing Machines and Children: Using Developmental Psychology Experiments to Assess the Strengths and Weaknesses of LaMDA Responses* (No. arXiv:2305.11243). arXiv. https://doi.org/10.48550/arXiv.2305.11243

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews. Cognitive Science*, *10*(6), e1503. https://doi.org/10.1002/wcs.1503

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041. https://doi.org/10.1126/science.aag2132

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science (New York, N.y.)*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Palmarini, A. B., & Mitchell, M. (2024, July 26). *Abstract Understanding of Core-Knowledge Concepts: Humans vs. LLMs*. ICML 2024 Workshop on LLMs and Cognition.

Qiu, L., Sha., F., Allen., K. R., Kim, Y., Linzen., T., & van Steenkiste, S. (2024). *Can Language Models Perform Implicit Bayesian Inference Over User Preference States?* First Workshop on System-2 Reasoning at Scale, NeurIPS'24.

Ruis, L., Findeis, A., Bradley, H., Rahmani, H. A., Choe, K. W., Grefenstette, E., & Rocktäschel, T. (2023, June 29). *Do LLMs selectively encode the goal of an agent's reach?* First Workshop on Theory of Mind in Communicating Agents. https://openreview.net/forum?id=KxvXjtyuYl#all

Spelke, E. S. (2022). *What babies know: Core knowledge and composition, Vol. 1*. Oxford University Press. https://doi.org/10.1093/oso/9780190618247.001.0001

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants Use Statistical Sampling to Understand the Psychological World. *Infancy*, *21*(5), 668–676. https://doi.org/10.1111/infa.12131

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Yiu, E., Qraitem, M., Wong, C., Majhi, A. N., Bai, Y., Ginosar, S., Gopnik, A., & Saenko, K. (2024). *KiVA: Kid-inspired Visual Analogies for Testing Large Multimodal Models* (No. arXiv:2407.17773). arXiv. https://doi.org/10.48550/arXiv.2407.17773