

Language Models Represent Beliefs of Self and Others

Wentao Zhu¹ Zhining Zhang¹ Yizhou Wang^{1,2,3,4}

Abstract

Understanding and attributing mental states, known as Theory of Mind (ToM), emerges as a fundamental capability for human social reasoning. While Large Language Models (LLMs) appear to possess certain ToM abilities, the mechanisms underlying these capabilities remain elusive. In this study, we discover that it is possible to linearly decode the belief status from the perspectives of various agents through neural activations of language models, indicating the existence of internal representations of self and others' beliefs. By manipulating these representations, we observe dramatic changes in the models' ToM performance, underscoring their pivotal role in the social reasoning process. Additionally, our findings extend to diverse social reasoning tasks that involve different causal inference patterns, suggesting the potential generalizability of these representations. *

1. Introduction

Developing machine systems that can engage in sophisticated social reasoning in a human-like manner represents one of the paramount goals in artificial intelligence. At the core of such an endeavor is the necessity for these systems to possess a “*Theory of Mind*” (ToM) capability, which involves recognizing and attributing mental states — such as beliefs, desires, intentions, and emotions — to oneself and others, while acknowledging that others may possess mental states distinct from one’s own (Leslie, 1987; Wellman et al., 2001). This foundational capability is crucial not only

¹Center on Frontiers of Computing Studies, School of Computer Science, Peking University ²Inst. for Artificial Intelligence, Peking University ³Nat'l Eng. Research Center of Visual Technology, Peking University ⁴Nat'l Key Lab of General Artificial Intelligence, Peking University. Correspondence to: Wentao Zhu <wtzhu@pku.edu.cn>, Yizhou Wang <yizhou.wang@pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

*Project page: <https://walter0807.github.io/RepBelief/>

for the nuanced navigation of human social interactions but also for enabling machines to engage in cooperative, adaptive, and sympathetic behaviors in diverse social environments (Kleiman-Weiner et al., 2016; Rabinowitz et al., 2018; Zhu et al., 2023).

The recent advancements in Large Language Models (LLMs) appear to be a promising approach towards this objective, as emerging research indicates that LLMs exhibit reasonable ToM capabilities (Kosinski, 2023; Bubeck et al., 2023). These studies suggest that LLMs could, to some extent, predict and understand human intentions and beliefs, thereby demonstrating a foundational level of social reasoning. Meanwhile, some other research underscores that these capabilities tend to be superficial and fragile (Shapira et al., 2023; Ullman, 2023; Ma et al., 2023; Verma et al., 2024). Critics argue that while LLMs may mimic the outward appearance of understanding social contexts and mental states, akin to the “Clever Hans” (Pfungst, 1911; Kavumba et al., 2019) and “Stochastic Parrot” (Bender et al., 2021) analogies, this performance may not stem from a deep, genuine comprehension similar to human ToM. Instead, it may simply reflect the models’ ability to replicate patterns observed in their training data.

These observations highlight a critical gap in our understanding of LLM social reasoning capabilities extending beyond mere black-box tests. Key questions remain unanswered, such as whether LLMs develop an internal representation of others’ mental states, and whether it is feasible to distinguish between the mental states of others and those of the LLMs when the two have a conflict due to reasons such as information mismatch. Addressing these questions not only helps us gain a deeper insight on how LLMs understand others’ mental states and perform social reasoning, but is also meaningful for the trustworthiness and alignment of AI systems (Wang et al., 2023; Ngo et al., 2023; Ji et al., 2023).

In this work, we undertake a preliminary exploration to understand the ToM capabilities of LLMs by studying their internal representations, going beyond merely analyzing the text responses they generate. Firstly, we seek to identify if LLMs have internal representations of others’ beliefs and their own (§ 3). If the answer is true, models potentially possess the ability to recognize others’ mental states and differentiate them from their own prior to generating a final

response. Specifically, we prompt the model with a short story in third-person narrative paired with a belief, which may or may not hold true, and attempt to classify the belief’s validity from both the main character’s perspective and the model’s (“God’s-eye view”), based on the model’s intermediate activations. Furthermore, we explore the possibility of modifying the internal representations to steer the model’s behavior towards or away from reflecting others’ mental states (§ 4.2). Lastly, we assess how our findings generalize across various social reasoning tasks with different causal inference patterns (§ 4.3).

2. Related Work

2.1. Human Theory-of-Mind

Theory of Mind (ToM), recognized as a cornerstone of human social cognition, facilitates individuals to infer the mental states of others, including their beliefs, desires, and intentions. Research indicates that infants as young as 12 months exhibit the ability to ascribe mental states to others, showcasing early development of ToM (Onishi & Baillargeon, 2005; Spelke & Kinzler, 2007). The false-belief task (Baron-Cohen et al., 1985; Wellman et al., 2001) stands as a critical experimental approach for evaluating ToM. In this task, participants are required to predict a protagonist’s actions based on her incorrect beliefs, which are separate from the participant’s own knowledge. Cognitive scientists design meticulous experiments to dissect the nuances of reasoning related to agents’ desires and beliefs, employing rigorous control conditions to eliminate simplistic heuristic explanations (Goodman et al., 2009; Baker et al., 2009; 2017; Jara-Ettinger et al., 2020). Furthermore, neuroscientific studies seek to pinpoint the neural basis of social cognition, particularly highlighting the roles of the dorsal medial prefrontal cortex (dmPFC) and the temporoparietal junction (TPJ) (Frith & Frith, 2006; Döhnel et al., 2012; Molenberghs et al., 2016). Jamali et al. further reveal that single neurons in dmPFC could encode information about others’ beliefs. The significance of ToM extends beyond individual interactions, influencing the spread of culture and the unity of social groups. The cognitive mechanisms enabled by ToM play a crucial role in forming and sustaining social norms, fostering cooperative behavior, and perpetuating shared cultural practices (Tomasello et al., 2005; Henrich, 2016).

2.2. Machine Theory-of-Mind

Developing machine systems that exhibit human-like ToM ability has been a long-standing endeavor in artificial intelligence research. Notably, Rabinowitz et al. design a ToMnet that utilizes meta-learning to build models of the agents it encounters based solely on observations of their behaviors. Track et al. introduces the concept of Satisficing

Theory of Mind. Shum et al. explore the application of Bayesian inference to decipher group behaviors and anticipate group dynamics. Wang et al. propose to integrate ToM reasoning within multi-agent communication frameworks, enhancing the cooperative capabilities. These studies primarily focus on deducing the mental states of others explicitly and forming neural representations thereof. The remarkable achievements of LLMs have spurred further exploration into the ToM capabilities of these models. Research works in this field (Kosinski, 2023; Shapira et al., 2023; Ullman, 2023) predominantly evaluates model performance using various prompts related to false-belief tests, yielding diverse outcomes. Additional studies (Moghaddam & Honey, 2023; Wilf et al., 2023) advocate for enhancing the ToM performance of LLMs via strategic prompting. To facilitate a more uniform assessment of machine ToM abilities, numerous benchmarks have been established, including ToM-QA (Nematzadeh et al., 2018), ToMi (Le et al., 2019), SocialIQA (Sap et al., 2019), BIB (Gandhi et al., 2021), Agent (Shu et al., 2021), BigToM (Gandhi et al., 2023), ToMChallenges (Ma et al., 2023), MMToM-QA (Jin et al., 2024), and T4D (Zhou et al., 2023). Our research diverges from the prevailing focus by delving into the intrinsic mechanisms of LLM ToM reasoning, specifically through the examination of internal neural representations.

3. Belief Representations in Language Models

We first explore if and how LLMs characterize the beliefs of different agents. Previous works in neural network interpretability (Bau et al., 2020; Burns et al., 2022; Moschella et al., 2022; Li et al., 2023a) suggest that there often exist interpretable directions in the latent representation space of the model. Therefore, some research works propose to linearly project the learned representation to the target directions to uncover meaningful variables (Mikolov et al., 2013; Goh et al., 2021; Elhage et al., 2022; Gurnee & Tegmark, 2024; Park et al., 2023). Motivated by this insight, we start by training linear classifier probes (Alain & Bengio, 2016; Belinkov, 2022) on the latent representations of a language model to estimate the likelihood of a belief from a certain agent’s perspective.

3.1. Setup

Model. We employ Mistral-7B-Instruct (Jiang et al., 2023) which is an instruction fine-tuned autoregressive language model with state-of-the-art performance. We focus on the activations of self-attention heads that enable Transformer-based language models to transfer information across various token positions (Vaswani et al., 2017; Elhage et al., 2021; Todd et al., 2023).

True Belief

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. [Noor sees her coworker swapping the milk.](#)

Belief: The milk pitcher contains almond milk.

(y_p =True, y_o =True)

Belief: The milk pitcher contains oat milk.

(y_p =False, y_o =False)

False Belief

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. [Noor does not see her coworker swapping the milk.](#)

Belief: The milk pitcher contains almond milk.

(y_p =False, y_o =True)

Belief: The milk pitcher contains oat milk.

(y_p =True, y_o =False)

Figure 1. Example case of belief representation probing. Left: a “True Belief” story where the protagonist shares the same belief with oracle. Right: a “False belief” story where the protagonist has different belief with oracle. For both stories, we concatenate them with the two beliefs respectively and directly feed to the model. The ground-truth belief label from the protagonist’s perspective y_p and the oracle’s perspective y_o can be decided accordingly.

Dataset. We utilize the BigToM dataset (Gandhi et al., 2023) which is constructed with a causal template and an example scenario including prior desires, actions, beliefs, and a causal event that changes the state of the environment. The protagonist could be aware or unaware of the causal event, which results in different beliefs. In this section, we focus on the “Forward Belief” setting, where the model infers the belief of the agent given the agent’s percepts of the causal event. We train and evaluate the probes on a held-out subset without access to the stories in the test set of the benchmark.

3.2. Probing

Feature Extraction. Our goal is to decode the belief status of different agents from the activations of attention heads, given a narrative and a corresponding belief statement. Specifically, we focus on two agents, namely *protagonist*, the central figure of the narrative, and *oracle*, which represents an omniscient spectator’s perspective. By prompting the model with pairs of story and belief without explicit directives, we capture the attention head activations at the final token position, denoted as $X \in \mathbb{R}^{L \times H \times D}$. Here, L , H , and D represent the number of layers, the number of attention heads per layer, and the dimensionality of the attention head features, respectively. Concurrently, we acquire the corresponding ground-truth belief labels y_p and y_o , as illustrated in Figure 1.

Binary Probing. We first train individual linear probes for each attention head at every layer to fit the belief labels y_p and y_o separately. For ease of explanation, we denote the activation of a particular attention head as $\mathbf{x} \in \mathbb{R}^{N \times D}$ where N is the size of the dataset, and the ground-truth belief labels as $\mathbf{y} \in \{0, 1\}^N$. We employ a logistic regression model to predict the probability of the belief being true:

$$\hat{\mathbf{y}} = \sigma(\mathbf{x}\mathbf{W} + b), \quad (1)$$

where $\sigma(\cdot)$ is the logistic sigmoid function, $\mathbf{W} \in \mathbb{R}^D$ is the weight vector, $b \in \mathbb{R}$ is the bias. The optimization of parameters \mathbf{W} and b is achieved through minimizing the cross-entropy loss

$$\mathcal{L}(\mathbf{W}, b) = -\frac{1}{N} \left(\mathbf{y}^T \log(\hat{\mathbf{y}}) + (1 - \mathbf{y})^T \log(1 - \hat{\mathbf{y}}) \right). \quad (2)$$

Figure 2 (A) and (B) display the validation accuracies of the linear probes. It reveals that a large number of attention heads can accurately capture the *oracle*’s belief status. These informative attention heads are distributed across various layers, particularly excluding the initial layers, with those in the middle layers demonstrating superior accuracy. It implies that the language model indeed develops intermediate representations that reflect its own belief status based on the full information provided. In contrast, the majority of attention heads only reach baseline accuracy in predicting the *protagonist*’s belief status, performing no