

# Language Models Represent Beliefs of Self and Others

Wentao Zhu<sup>1</sup> Zhining Zhang<sup>1</sup> Yizhou Wang<sup>1,2,3,4</sup>

## 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

<sup>1</sup>Center on Frontiers of Computing Studies, School of Computer Science, Peking University <sup>2</sup>Inst. for Artificial Intelligence, Peking University <sup>3</sup>Nat'l Eng. Research Center of Visual Technology, Peking University <sup>4</sup>Nat'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 41<sup>st</sup> 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

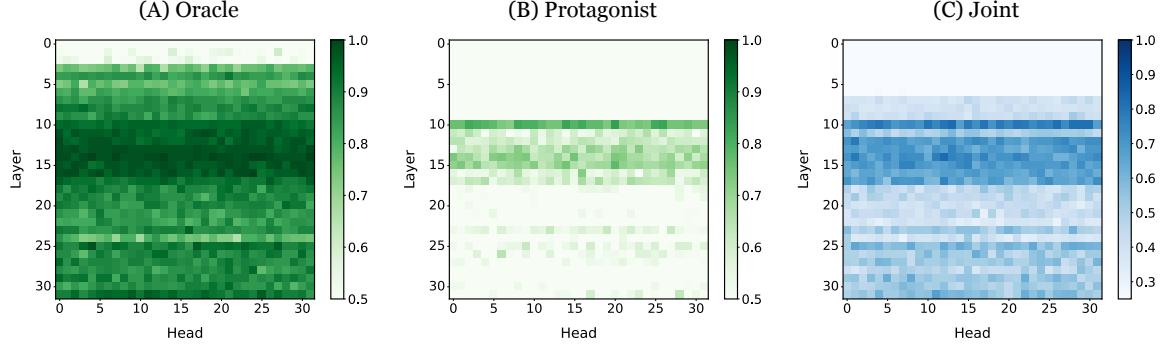


Figure 2. Probe accuracies based on the attention head activations in all layers of Mistral-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status estimation for both agents using multinomial logistic regression (quaternary).

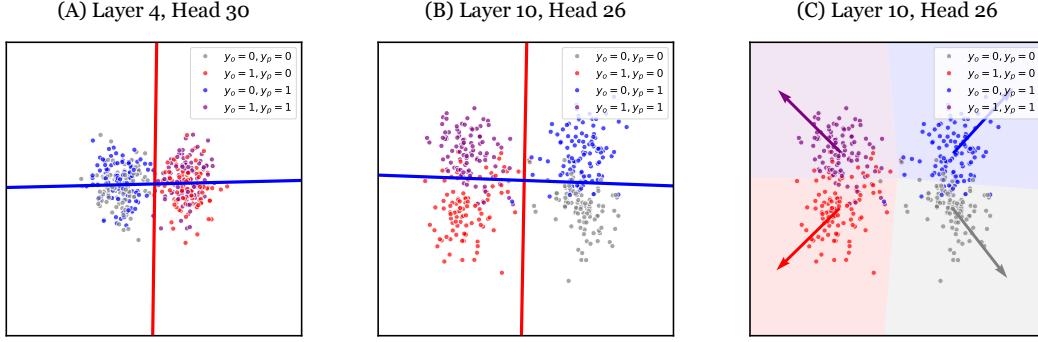


Figure 3. Illustration of linear separability of the belief representations. We show visual explanations for typical representation spaces: In (A), *oracle* belief status can be precisely estimated with a linear model, while *protagonist* cannot. The red and blue lines represent linear decision boundaries for *oracle* and *protagonist*, respectively; In (B), the belief statuses of both *oracle* and *protagonist* can be accurately modeled using linear models; (C) further shows the decision boundaries for joint belief status estimation using multinomial linear regression model, with arrows indicating the probing weight directions for each class.

better than random guess. However, it is worth noting that a specific group of attention heads in the middle layers exhibits remarkably better performance, achieving over 80% validation accuracy. This phenomenon suggests that these attention heads implicitly encode the belief status of other agents in a linearly-decodable way.

**Multinomial Probing.** We further explore the possibility of estimating the belief statuses of both agents from the activations simultaneously. For this purpose, we combine the belief labels of both agents into a four-dimensional variable representing their joint belief statuses using one-hot encoding, where each dimension corresponds to a unique combination of  $y_o$  and  $y_p$ . For brevity, we define this joint belief variable as  $\mathbf{y}_m \in \{0, 1\}^{N \times 4}$ . A multinomial logistic regression model is utilized to predict the joint belief  $\hat{\mathbf{y}}_m$ . Specifically, the class probabilities are derived by applying the softmax function to the linear transformations of  $\mathbf{x}$ :

$$\hat{\mathbf{y}}_m = \text{softmax}(\mathbf{x}\mathbf{W}_m + \mathbf{b}_m), \quad (3)$$

where  $\mathbf{W}_m \in \mathbb{R}^{D \times 4}$  is the weight matrix for multinomial logistic regression, and  $\mathbf{b}_m \in \mathbb{R}^4$  is the bias vector. These parameters can be optimized by minimizing the cross-entropy loss of all classes

$$\mathcal{L}(\mathbf{W}_m, \mathbf{b}_m) = -\frac{1}{N} \text{Tr} (\mathbf{y}_m^T \log(\hat{\mathbf{y}}_m)). \quad (4)$$

Figure 2 (C) illustrates the accuracies of multinomial probing on the validation set, which demonstrates that it is possible to train linear probes with decent accuracy in the quaternary classification task for specific attention heads. In other words, there exist individual attention heads (mostly in the middle layers) that could encode the belief statuses of both agents not only independently but also in conjunction, such as indicating whether *protagonist* and *oracle* share identical

beliefs. This observation underscores the nuanced capability of these attention heads to represent complex relational information between agents' beliefs. We present additional statistical analysis on belief probing in Appendix A.

**Visualizing the Belief Representations.** In order to better understand the belief representations in the attention head activation space, we further visualize the linear regressors. We perform canonical-correlation analysis (CCA) to reduce the dimensionality of the activations to two and plot the linear decision boundaries in the reduced space. Figure 3 demonstrates two representative categories of the attention heads. The first category predominantly encodes the belief of *oracle*, showing a bias toward this perspective. The second category, on the other hand, more comprehensively captures the beliefs of both agents (B). We further visualize the decision boundaries of multinomial probes in (C). Although geometries in the high-dimensional space can be much more complicated, the 2D visualizations offer some basic intuitions of linearly-separable belief representations. Figure 2 reveals that, although a significant portion of the attention heads fall into the first category, a distinct subset aligns with the second category.

## 4. Manipulating the Belief Representations

Although the probing results support the presence of belief representations for different agents within the attention head activation spaces, it remains unclear if these representations contribute to the overall social reasoning process. In this section, we aim to explore the functional roles of belief representations by explicitly manipulating them. We design experiments to address the following questions: Can we alter the social reasoning capabilities of language models by manipulating their internal representations? If so, how can this be achieved? And, how does the practice impact different types of social reasoning tasks?

### 4.1. ToM Evaluation

We evaluate the ToM capabilities of language models using the BigToM (Gandhi et al., 2023) benchmark. We focus on the 0-shot setting and do not explicitly reveal the agent's initial belief. We study three social reasoning tasks, namely *Forward Belief*, *Forward Action*, and *Backward Belief*, each focuses on different causal inference patterns as shown in Figure 4.

1. The *Forward Belief* task entails deducing the agent's beliefs given its percepts of a causal event. This inference can be expressed as:  $P(\text{Belief} | \text{Percept})$ .
2. The *Forward Action* task involves predicting the agent's future action based on the percepts. This process implicitly requires an initial inference of the

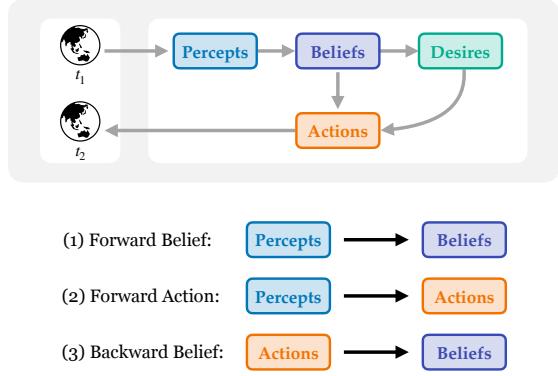


Figure 4. Different social reasoning tasks and the underlying causal graph.

agent's beliefs, followed by the deduction of the agent's action considering both percepts and desires:  $\sum_{\text{Belief}} P(\text{Action} | \text{Percept}, \text{Desire}, \text{Belief})$ .

3. The *Backward Belief* task aims to infer the agent's beliefs from the observed actions. This task poses a significant challenge as it demands a joint inference of unknown beliefs and percepts based on an observed action:  $\sum_{\text{Percept}} \sum_{\text{Belief}} P(\text{Action} | \text{Desire}, \text{Percept}, \text{Belief})$ .

All the tasks are presented as reading comprehension with a story in third-person narrative, followed by a question and two options. We evaluate the models based on their accuracy in responding to these questions, specifically under two conditions for each narrative scenario: *True Belief* (TB) and *False Belief* (FB). We also evaluate the percentage of scenarios where the model correctly answers both TB and FB questions.

We conduct experiments on two language models, Mistral-7B-Instruct (Jiang et al., 2023) and DeepSeek-LLM-7B-Chat (Bi et al., 2024). Both models are tested using the most deterministic setting with a temperature of 0 following (Gandhi et al., 2023). As the baseline results in Table 1 show, both models exhibit a distinct performance gap between *True Belief* and *False Belief* conditions when tested directly. It suggests that the models fail to recognize that other agents may hold beliefs different from their own due to perception differences. Specifically, in the classical *False Belief* test, Mistral is more biased towards the wrong answer, while DeepSeek's choices are closer to random guess. These tendencies might be attributed to the biased internal belief representations as previously discovered in § 3.2. The probe accuracies for DeepSeek are presented in Appendix C.

### 4.2. Activation Intervention

Table 1. Model performance comparison on the BigToM benchmark. TB = True Belief. FB = False Belief. Bold items denote the best setting in each subset.

Model	Forward Belief			Forward Action			Backward Belief		
	TB	FB	Both	TB	FB	Both	TB	FB	Both
LLaMA-65b	0.68	0.62	0.51	0.82	0.47	0.45	0.56	0.53	0.40
text-davinci-003	0.82	0.82	0.65	0.96	0.27	0.25	0.54	0.59	0.24
Claude	0.97	0.82	0.81	0.98	0.28	0.27	0.79	0.48	0.33
Claude-2	0.88	0.75	0.68	0.95	0.36	0.34	0.75	0.50	0.39
GPT-3.5	0.81	0.69	0.53	0.97	0.19	0.17	0.55	0.45	0.18
GPT-4	0.99	0.98	0.97	0.98	0.81	0.79	0.86	0.53	0.40
Mistral-7B (baseline)	0.95	0.33	0.31	0.92	0.30	0.26	<b>0.91</b>	0.22	0.16
Mistral-7B (+ random)	<b>0.97</b>	0.33	0.32	0.92	0.29	0.25	0.91	0.19	0.14
Mistral-7B (+ protagonist)	0.96	0.30	0.29	0.91	0.30	0.25	0.90	0.22	0.15
Mistral-7B (- oracle)	0.84	0.49	0.41	<b>0.93</b>	0.29	0.25	0.50	0.37	0.26
Mistral-7B (+ $T_p F_o$ )	0.85	<b>0.66</b>	<b>0.58</b>	0.88	<b>0.41</b>	<b>0.31</b>	0.61	<b>0.44</b>	<b>0.41</b>
DeepSeek-7B (baseline)	<b>0.73</b>	0.47	0.37	<b>0.77</b>	0.48	0.31	0.64	0.50	0.29
DeepSeek-7B (+ random)	0.72	0.49	0.40	0.76	0.47	0.30	<b>0.65</b>	0.49	0.26
DeepSeek-7B (+ protagonist)	0.73	0.50	0.42	0.76	0.49	0.31	0.65	0.52	0.27
DeepSeek-7B (- oracle)	0.65	0.46	<b>0.44</b>	0.73	0.57	0.34	0.65	0.48	0.28
DeepSeek-7B (+ $T_p F_o$ )	0.63	<b>0.74</b>	0.38	0.75	<b>0.60</b>	<b>0.39</b>	0.54	<b>0.66</b>	<b>0.31</b>

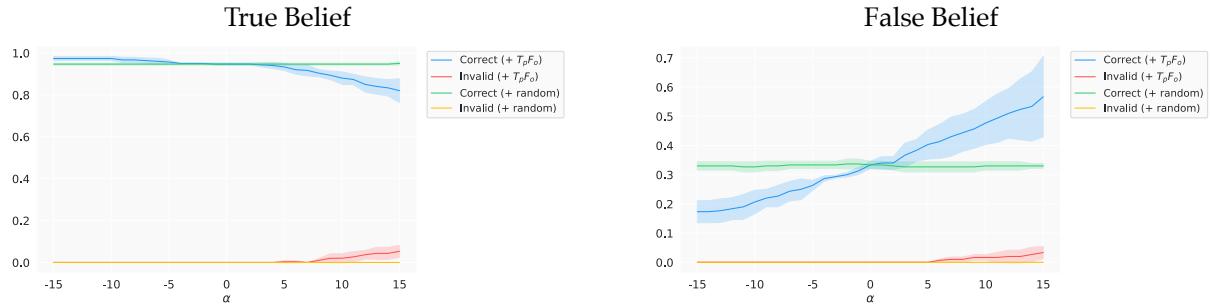


Figure 5. Impact of varying intervention strength  $\alpha$  on the *Forward Belief* task using Mistral-7B. “Invalid” denotes the answer is not recognized by the grading mechanism as the model fails to provide answer in the required format, e.g., delivering uncertain responses.

#### 4.2.1. STRATEGY.

We apply inference-time intervention (Li et al., 2023b) to manipulate the activations at multi-head attention (MHA) stage of the Transformer models. It involves first selecting the top- $K$  heads based on their probing accuracy on the validation set, then steering their activations towards certain directions for  $\alpha \times$  the standard deviation for next token prediction autoregressively. Mathematically, the intervention for the  $l$ -th layer of can be written as

$$x_{l+1} = x_l + \sum_{h=1}^H Q_l^h (\text{Att}_l^h (P_l^h x_l) + \alpha \sigma_l^h \theta_l^h), \quad (5)$$

where  $x_i$  is the stream activation of the  $i$ -th layer,  $H$  is the number of attention heads within the layer. For each head  $h$ ,  $P_l^h$  maps stream activation into a lower-dimensional head space, and  $Q_l^h$  maps it back. Att is an operator where

communication with other input tokens happens. The intervention happens after Att and before  $Q_l^h$ , where  $\alpha$  is the step length of intervention,  $\sigma_l^h$  is the standard deviation of activations along the target direction,  $\theta_l^h$  is the target direction. We set  $K$  and  $\alpha$  with grid search following previous works, and present the ablations in Appendix G. Our primary focus is on identifying effective directions for altering the model behavior purposefully. Specifically, we explore the usage of the following directions:

- Random directions within the activation spaces  $\mathbb{R}^D$ .
- Weight directions for *oracle* and *protagonist*, respectively, derived from binary probing. We focus on the direction to maximize the probability of predicting *protagonist*’s belief as True (+ protagonist, corresponding to the upwards direction vertical to the blue boundary in Figure 3 (B)), and the direction to minimize the probability of predicting *oracle*’s belief as True (- oracle,

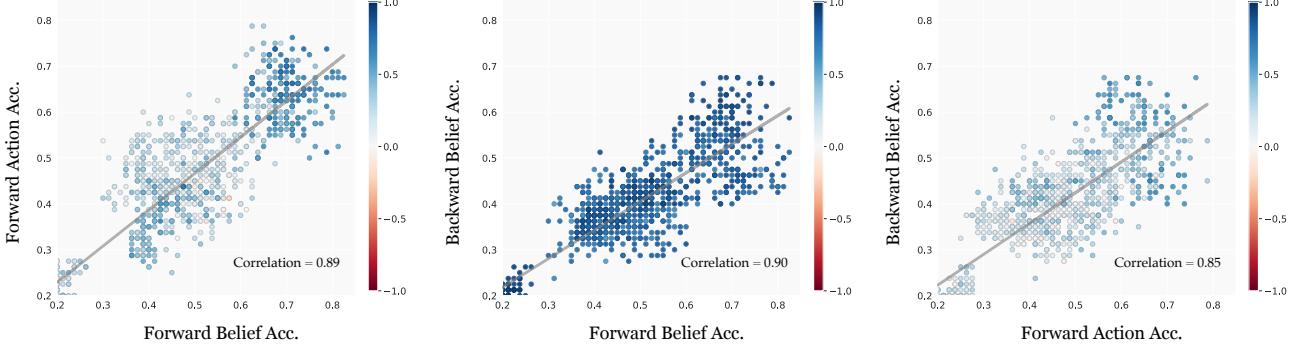


Figure 6. Pairwise comparisons of multinomial probing results on Mistral-7B. Each point represents a specific attention head. The point position denotes its probe accuracies in the two tasks, and point color denotes the cosine similarity between the  $(+ T_p F_o)$  probe directions of the two tasks.

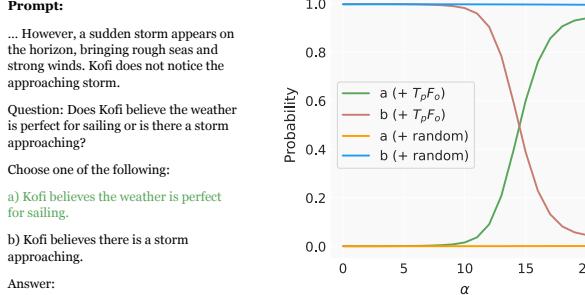


Figure 7. (Left) Question prompt of the *False Belief* condition for the *Forward Belief* task. The story background is omitted for simplicity. (Right) Changes of next-token probability with regard to different intervention strength  $\alpha$  on Mistral-7B.

corresponding to the rightwards direction vertical to the red boundary in Figure 3 (B))

- Weight directions for joint belief status prediction, derived from multinomial probing. We focus on the direction which maximizes the probability of recognizing *protagonist*'s belief as True and meanwhile distinguishes it from *oracle*'s belief  $(+ T_p F_o)$ , corresponding to the blue arrow in Figure 3 (C).

For random directions, we use the top- $K$  informative heads identified with multinomial probing. For different social reasoning tasks, we separately probe the attention head activations by constructing the corresponding prompt templates. For example, for the *Forward Action* task, we utilize the story-action pairs, while for the *Backward Belief* task, we include the *protagonist*'s next-step actions in the story. More details can be found in Appendix B.

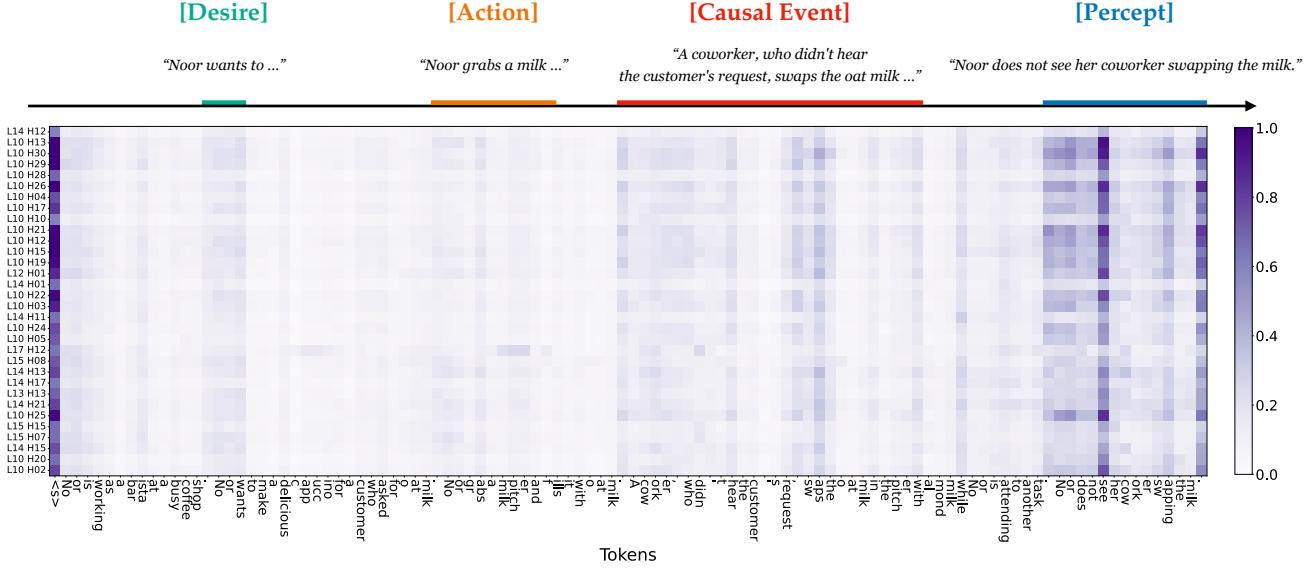
#### 4.2.2. RESULTS.

We evaluate different activation intervention strategies and present the results in Table 1. For different tasks, we per-

form intervention based on their respective probing results. The  $(+ \text{random})$  results indicate that random perturbations of attention heads have marginal impact on model performance, which is in line with the findings in (Li et al., 2023b). Directly targeting activations towards the *protagonist* belief direction also fails to significantly change the model performances, possibly due to overpowering *oracle* belief signals with heavily-biased belief representation. To investigate this, we conduct intervention along the *oracle* belief direction reversely ( $- \text{oracle}$ ), and discover a noteworthy change of model behaviors. Furthermore, we explore a direction that could distinguish between the belief of both agents, amplifying the *protagonist*'s belief likelihood and weakening the *oracle*'s belief likelihood. This direction could be derived from the corresponding dimension of the weight matrix in the multinomial logistic regression probes. We find that intervention towards this direction  $(+ T_p F_o)$  remarkably changes the model performance, effectively improving the overall ToM reasoning capabilities.

Additionally, we seek to better understand the functionality of identified belief directions through continuous interventions along these vectors. As demonstrated in Figure 5, The  $(+ T_p F_o)$  direction exhibits significant impacts on benchmark performance, underscoring its pivotal role in ToM reasoning process. Specifically, steering towards this direction consistently enhances ToM accuracy in *False Belief* cases, while slightly decreases the *True Belief* accuracy (partially due to an increase of invalid responses). At a finer level, we explore changes in model behavior by examining the fluctuations in the probabilities of next-token predictions. Figure 7 illustrates how steering head activations to the specific directions could influence next-token predictions and invert the selection of choices.

#### 4.3. Varying Social Reasoning Tasks



**Figure 8.** Magnitudes of gradients on the token embeddings with respect to the projection of attention head activations over the corresponding joint belief directions. Each line represents a specific attention head in Mistral-7B. We highlight the prominent segments and the corresponding causal variables.

**Table 2.** Cross-task intervention results on Mistral-7B. We perform activation intervention towards the joint belief directions identified in the *Forward Belief* task and evaluate the other two tasks.

Model	Forward Action			Backward Belief		
	TB	FB	Both	TB	FB	Both
Mistral-7B	0.92	0.30	0.26	<b>0.91</b>	0.22	0.16
+ random	<b>0.93</b>	0.31	0.26	0.90	0.22	0.16
+ transferred	0.90	<b>0.40</b>	<b>0.33</b>	0.73	<b>0.38</b>	<b>0.22</b>
DeepSeek-7B	<b>0.77</b>	0.48	0.31	0.64	0.50	0.29
+ random	0.76	0.42	0.29	<b>0.65</b>	0.47	<b>0.32</b>
+ transferred	0.75	<b>0.57</b>	<b>0.37</b>	0.57	<b>0.57</b>	0.28

Furthermore, we investigate how various social reasoning tasks differ in terms of their underlying representations and whether these representations could generalize across different tasks or not.

First of all, we explore the interrelationships among the identified ToM representations across different tasks. The probing results of each task are presented in Appendix C. Figure 6 illustrates a strong correlation between the probing accuracies in different task scenarios, suggesting related representational capabilities under various causal inference conditions. Remarkably, the top-performing heads in one task tend to include the predictive features in another task as well, and the directions they identify exhibit high similarity. Considering that all three tasks implicitly or explicitly involve belief inference, we propose that a subset of atten-

tion head spaces might contain belief representations which potentially contribute to a range of social reasoning tasks.

This hypothesis motivates us to conduct a generalization test across different social reasoning tasks. We specifically intervene in the  $(+ T_p F_o)$  directions identified under *Forward Belief* conditions when evaluating the other two tasks. Table 2 indicates that the directions identified in one task do generalize to others, suggesting that these directions might encapsulate a more universal function as belief representations.

Moreover, we seek to understand why these directions could act as generalizable belief representations. Specifically, we first prompt the model with story narratives, then respectively project the attention head activations onto the target  $(+ T_p F_o)$  directions of the top probes. We then back-propagate the projection norm through the model and calculate the gradient magnitudes in input token embeddings, which approximately reflects the relevance of individual input tokens to the target directions. Figure 8 reveals that the identified directions in attention head activation spaces primarily focus on tokens denoting key causal variables, including the protagonist’s desires and initial actions, the causal event that changes the environmental states, and the protagonist’s percept status of the causal event. These elements collectively facilitate a comprehensive inference of both agents’ beliefs. These observations may shed light on the generalization potential of these directions across various social reasoning tasks. Despite the diverse causal inference patterns required by these tasks, they share common underlying causal vari-

ables and all necessitate inference regarding the agent’s belief status, whether explicitly or implicitly. We present additional studies of these directions on other ToM reasoning scenarios and unrelated language tasks in Appendices D and E.

## 5. Discussions

In this study, we investigate the ToM capabilities in LLMs, specifically examining their ability to internally represent and attribute beliefs. We discover that LLMs can distinguish between different belief states of multiple agent perspectives through their intermediate activations with simple linear models. Additionally, we show that manipulation of these representations significantly affects the model’s social reasoning performances. Finally, we demonstrate the generalization of the internal belief representations in diverse social reasoning task scenarios.

Our study contributes to the ongoing dialogue on the social reasoning capabilities of LLMs, providing new insights into their ability to simulate ToM through internal representations. Looking ahead, our study opens avenues for further investigation, including the development and processing of belief representations during training, their scalability in more complex LLMs like mixture of expert (MoE) models (Jiang et al., 2024), and methods to enhance machine ToM capabilities in alignment with human values. While our research provides valuable insights, it comes with its limitations. The scope of our exploration was confined to certain types of LLMs and specific social reasoning tasks, which may not capture the full spectrum of ToM capabilities. Future work should aim to address these gaps, broadening the understanding of ToM in AI systems across various models and more complex contexts.

## Acknowledgements

This work is supported by National Science and Technology Major Project (2022ZD0114904). We thank the anonymous reviewers for insightful discussions. We thank the awesome open-source toolbox nnsight ([Fiotto-Kaufman](#)) which is used to extract the Transformer internal representations.

## Impact Statement

Our study on the Theory of Mind (ToM) capabilities in Large Language Models (LLMs) illuminates the potential for more empathetic AI, enhancing human-machine interactions in various sectors. Ethically, it necessitates careful consideration to prevent misuse and bias propagation, ensuring AI’s societal impact is positive. Responsible development and transparent deployment are imperative to safeguard against unintended consequences and maintain trust in AI

advancements. Additionally, this study should not be misinterpreted by the media and the general public as evidence that LLMs exhibit consciousness and self-awareness.

## References

- Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
- Baker, C. L., Saxe, R., and Tenenbaum, J. B. Action understanding as inverse planning. *Cognition*, 113(3):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(4):0064, 2017.
- Baron-Cohen, S., Leslie, A. M., and Frith, U. Does the autistic child have a “theory of mind”? *Cognition*, 21(1): 37–46, 1985.
- Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020.
- Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.
- Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.
- Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., and Sommer, M. Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *Neuroimage*, 60(3):1652–1661, 2012.
- Dolan, B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*, 2005.
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.
- Fiotto-Kaufman, J. nnisight: The package for interpreting and manipulating the internals of deep learned models. . URL <https://github.com/JadenFiotto-Kaufman/nnisight>.
- Frith, C. D. and Frith, U. The neural basis of mentalizing. *Neuron*, 50(4):531–534, 2006.
- Gandhi, K., Stojnic, G., Lake, B. M., and Dillon, M. R. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *arXiv preprint arXiv:2102.11938*, 2021.
- Gandhi, K., Fränken, J.-P., Gerstenberg, T., and Goodman, N. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.
- Goodman, N. D., Baker, C. L., and Tenenbaum, J. B. Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st annual conference of the cognitive science society*, pp. 2759–2764. Cognitive Science Society Amsterdam, 2009.
- Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Henrich, J. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. princeton University press, 2016.
- Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., and Williams, Z. M. Single-neuronal predictions of others’ beliefs in humans. *Nature*, 591(7851): 610–614, 2021.

- Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123: 101334, 2020.
- Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., Ullman, T., Torralba, A., Tenenbaum, J. B., and Shu, T. Mmtom-qa: Multimodal theory of mind question answering, 2024.
- Kavumba, P., Inoue, N., Heinzerling, B., Singh, K., Reisert, P., and Inui, K. When choosing plausible alternatives, clever hans can be clever. *arXiv preprint arXiv:1911.00225*, 2019.
- Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., and Tenenbaum, J. B. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*, 2016.
- Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.
- Le, M., Boureau, Y.-L., and Nickel, M. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.
- Leslie, A. M. Pretense and representation: The origins of “theory of mind”. *Psychological review*, 94(4):412, 1987.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023a.
- Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.
- Ma, X., Gao, L., and Xu, Q. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*, 2023.
- Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.
- Moghaddam, S. R. and Honey, C. J. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*, 2023.
- Molenberghs, P., Johnson, H., Henry, J. D., and Mattingley, J. B. Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, 65:276–291, 2016.
- Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolà, E. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.
- Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., and Griffiths, T. L. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*, 2018.
- Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective, 2023.
- Onishi, K. H. and Baillargeon, R. Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258, 2005.
- Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.
- Pfungst, O. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.
- Sap, M., Rashkin, H., Chen, D., LeBras, R., and Choi, Y. Socialiq: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.
- 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. *arXiv preprint arXiv:2305.14763*, 2023.

- Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J., and Ullman, T. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pp. 9614–9625. PMLR, 2021.
- Shum, M., Kleiman-Weiner, M., Littman, M. L., and Tenenbaum, J. B. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6163–6170, 2019.
- Spelke, E. S. and Kinzler, K. D. Core knowledge. *Developmental science*, 10(1):89–96, 2007.
- Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005.
- Track, S. I. A., Pöppel, J., and Kopp, S. Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, Stockholm, Sweden*, pp. 10–15, 2018.
- Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Verma, M., Bhambri, S., and Kambhampati, S. Theory of mind abilities of large language models in human-robot interaction: An illusion? *arXiv preprint arXiv:2401.05302*, 2024.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. 2023.
- Wang, Y., fangwei zhong, Xu, J., and Wang, Y. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*, 2022.
- Warstadt, A., Singh, A., and Bowman, S. R. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.
- Wellman, H. M., Cross, D., and Watson, J. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.
- Wilf, A., Lee, S. S., Liang, P. P., and Morency, L.-P. Think twice: Perspective-taking improves large language models’ theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*, 2023.
- Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023.
- Zhu, W., Qin, J., Lou, Y., Ye, H., Ma, X., Ci, H., and Wang, Y. Social motion prediction with cognitive hierarchies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

## A. Statistical Analysis on Belief Probing

To further verify the existence of attention heads that encode belief status of different agents, we conduct robust statistical analysis. Specifically, we train and evaluate the linear probes across different data splits using 100 random seeds. We report the mean accuracies and 95% confidence intervals (CI) for the top attention heads in Tables 3 to 5. We calculate the validation accuracy of linear probes for both True Belief (TB) and False Belief (FB) scenarios respectively. The probes are trained with full training set with both TB and FB. Furthermore, we apply a multiple hypothesis testing correction for our analysis of the top-10 attention heads. We adopt the stringent Bonferroni correction method that rigorously control the Family-Wise Error Rate (FWER), thereby significantly reducing the likelihood of type I errors. Under this framework, each attention head is scrutinized against a null hypothesis positing that its accuracy does not surpass a specific baseline (75%), versus an alternative hypothesis that asserts superior accuracy. The results robustly validate that a specific subset of attention heads exhibits the capacity to predict agents' belief states with a validation accuracy exceeding 75%. Significantly low p-values with the Bonferroni correction strongly reject the null hypothesis, affirming the statistical significance and reliability of our findings.

*Table 3.* The top predictive attention heads in belief status estimation for *oracle* using logistic regression (binary). We use Mistral-7B model in the *Forward Belief* task. Random guessing is 50%.

Position	Mean Acc.	CI	Corrected P-Value	Mean Acc (TB)	CI (TB)	Mean Acc (FB)	CI (FB)
(14, 31)	97.8	(97.5, 98.1)	0.0000	97.7	(97.3, 98.2)	97.8	(97.4, 98.2)
(13, 26)	97.5	(97.1, 97.8)	0.0000	97.1	(96.6, 97.5)	97.8	(97.4, 98.2)
(14, 11)	97.5	(97.1, 97.8)	0.0000	97.5	(97.0, 98.0)	97.4	(97.0, 97.8)
(14, 18)	97.4	(97.0, 97.7)	0.0000	97.6	(97.1, 98.0)	97.1	(96.7, 97.6)
(14, 8)	97.3	(96.9, 97.6)	0.0000	97.1	(96.6, 97.6)	97.4	(97.0, 97.8)
(14, 26)	97.2	(96.9, 97.6)	0.0000	96.8	(96.3, 97.3)	97.6	(97.2, 98.1)
(15, 23)	97.2	(96.9, 97.6)	0.0000	97.3	(96.8, 97.7)	97.2	(96.7, 97.7)
(13, 2)	97.2	(96.8, 97.6)	0.0000	97.2	(96.7, 97.7)	97.1	(96.6, 97.5)
(13, 1)	97.2	(96.8, 97.6)	0.0000	97.1	(96.5, 97.6)	97.2	(96.7, 97.7)
(14, 0)	97.1	(96.8, 97.5)	0.0000	97.3	(96.8, 97.8)	96.9	(96.5, 97.4)

*Table 4.* The top predictive attention heads in belief status estimation for *protagonist* using logistic regression (binary). We use Mistral-7B model in the *Forward Belief* task. Random guessing is 50%.

Position	Mean Acc.	CI	Corrected P-Value	Mean Acc (TB)	CI (TB)	Mean Acc (FB)	CI (FB)
(10, 16)	78.3	(77.4, 79.2)	0.0000	80.7	(79.5, 82.0)	76.2	(74.9, 77.5)
(10, 2)	77.4	(76.6, 78.3)	0.0000	76.5	(75.2, 77.8)	78.7	(77.3, 80.0)
(10, 21)	77.3	(76.4, 78.1)	0.0000	77.4	(76.1, 78.7)	77.5	(76.2, 78.7)
(10, 3)	77.2	(76.4, 78.1)	0.0000	75.2	(74.0, 76.5)	79.5	(78.2, 80.9)
(10, 4)	76.9	(76.0, 77.8)	0.0004	79.9	(78.6, 81.3)	74.4	(73.0, 75.8)
(10, 20)	76.9	(76.0, 77.7)	0.0002	78.7	(77.4, 80.1)	75.4	(74.1, 76.7)
(10, 11)	76.8	(75.9, 77.7)	0.0013	75.1	(73.7, 76.6)	78.9	(77.7, 80.1)
(10, 17)	76.5	(75.7, 77.3)	0.0028	76.3	(75.1, 77.5)	76.9	(75.8, 78.1)
(10, 15)	76.5	(75.6, 77.4)	0.0106	77.1	(75.8, 78.4)	76.2	(74.8, 77.5)
(10, 10)	76.5	(75.6, 77.3)	0.0088	82.9	(81.6, 84.2)	70.5	(69.1, 71.8)

*Table 5.* The top predictive attention heads in belief status estimation for both agents using multinomial logistic regression (quaternary). We use Mistral-7B model in the *Forward Belief* task. Random guessing is 25%.

Position	Mean Acc.	CI	Corrected P-Value	Mean Acc (TB)	CI (TB)	Mean Acc (FB)	CI (FB)
(12, 27)	79.0	(77.6, 80.4)	0.0000	84.0	(82.1, 85.8)	75.3	(72.7, 78.0)
(12, 31)	78.8	(77.1, 80.4)	0.0002	85.7	(83.8, 87.7)	73.3	(70.3, 76.3)
(12, 1)	78.6	(77.2, 79.9)	0.0000	84.9	(83.1, 86.8)	73.5	(71.0, 75.9)
(12, 13)	78.5	(77.0, 80.1)	0.0002	85.0	(83.0, 87.1)	73.3	(70.5, 76.1)
(15, 8)	78.4	(77.2, 79.6)	0.0000	87.5	(85.8, 89.2)	70.8	(68.4, 73.1)
(16, 20)	78.0	(76.9, 79.2)	0.0000	83.4	(81.7, 85.1)	74.0	(71.8, 76.3)
(12, 7)	78.0	(76.6, 79.5)	0.0005	85.9	(84.1, 87.6)	71.4	(68.8, 74.1)
(12, 24)	78.0	(76.3, 79.7)	0.0086	86.0	(83.8, 88.2)	71.6	(68.5, 74.7)
(16, 2)	77.9	(76.8, 79.0)	0.0000	83.3	(81.7, 85.0)	73.6	(71.6, 75.7)
(14, 1)	77.9	(76.3, 79.5)	0.0046	79.2	(76.4, 81.9)	78.5	(75.8, 81.2)

## B. Probing Prompts for Different Tasks

We develop corresponding prompt templates for different social reasoning tasks by casting the choices to statements. For the *Forward Action* task, we utilize the story-action pairs as shown in Figure 9. For the *Backward Belief* task, we include the *protagonist's* next-step actions in the story and present the story-belief pairs as shown in Figure 10.

True Belief	False Belief
<b>Story:</b> Carlos is a farmer in a small village in the Andes. Carlos wants to collect fresh eggs from his chickens to sell at the local market. The nest appears to be full of eggs when Carlos checks it in the morning. A crafty fox sneaks into the chicken coop and steals all the eggs from the nest. Carlos sees the fox running away with the eggs.	<b>Story:</b> Carlos is a farmer in a small village in the Andes. Carlos wants to collect fresh eggs from his chickens to sell at the local market. The nest appears to be full of eggs when Carlos checks it in the morning. A crafty fox sneaks into the chicken coop and steals all the eggs from the nest. <b>Carlos does not see the fox stealing the eggs.</b>
<b>Action:</b> Try to chase the fox and recover the eggs. $(y_p=\text{True}, y_o=\text{True})$	<b>Action:</b> Try to chase the fox and recover the eggs. $(y_p=\text{False}, y_o=\text{True})$
<b>Action:</b> Try to collect the eggs to sell at the market. $(y_p=\text{False}, y_o=\text{False})$	<b>Action:</b> Try to collect the eggs to sell at the market. $(y_p=\text{True}, y_o=\text{False})$

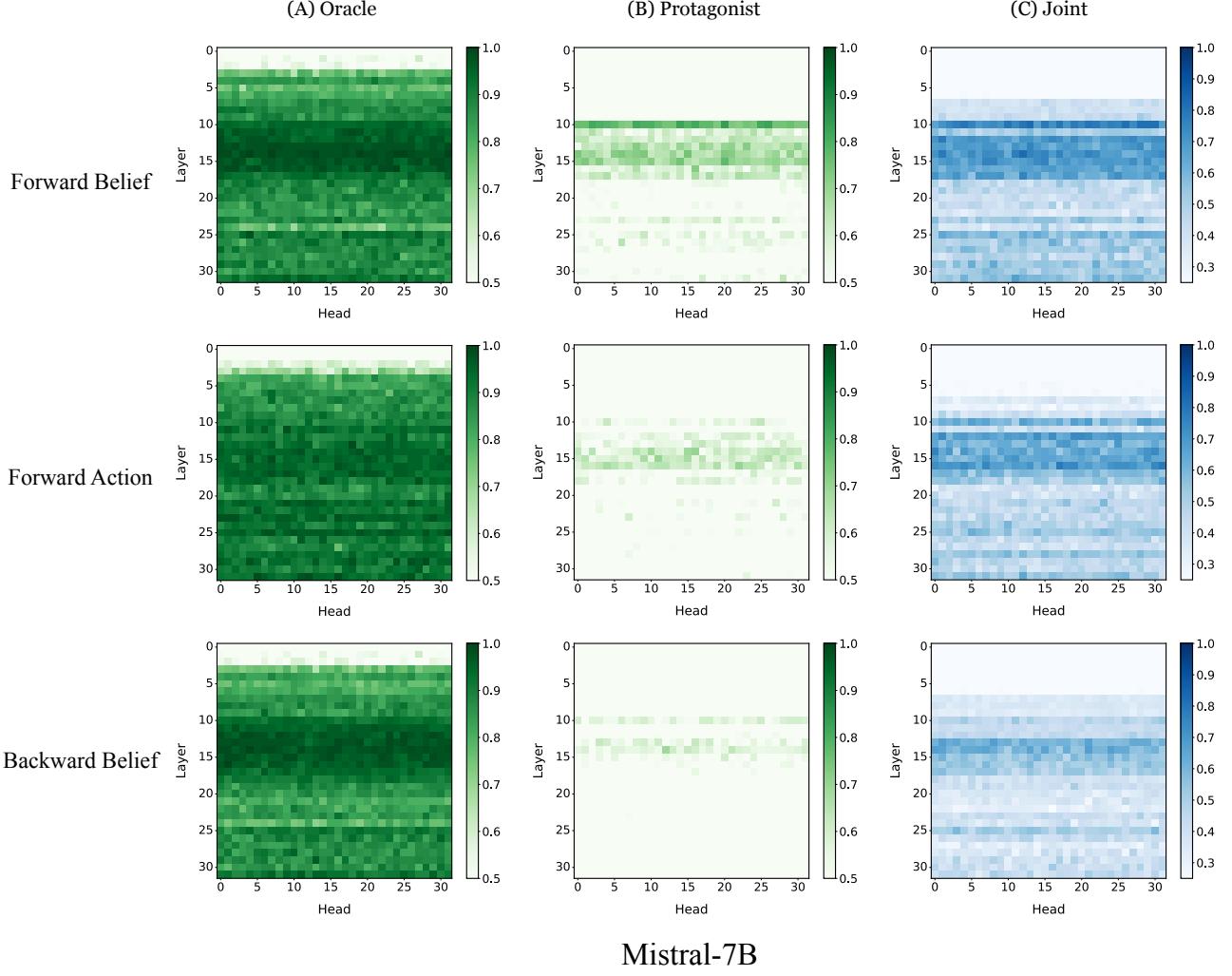
Figure 9. Example case of representation probing for the *Forward Action* task. Left: a “True Belief” story where the protagonist shares the same belief with oracle. Right: a “False belief” story where the protagonist has different belief with oracle. For both stories, we concatenate them with the two actions respectively and directly feed to the model. The ground-truth action label from the protagonist’s perspective  $y_p$  and the oracle’s perspective  $y_o$  can then be decided accordingly.

True Belief	False Belief
<b>Story:</b> Lila is a beekeeper in a small village in Eastern Europe. She wants to harvest honey from one of her beehives. The beehive is full of honey, as Lila observed during her last inspection. A sudden cold snap occurs overnight, causing the bees to consume the honey to stay warm and survive, leaving the beehive nearly empty. <b>Lila decides to wait for the bees to produce more honey.</b>	<b>Story:</b> Lila is a beekeeper in a small village in Eastern Europe. She wants to harvest honey from one of her beehives. The beehive is full of honey, as Lila observed during her last inspection. A sudden cold snap occurs overnight, causing the bees to consume the honey to stay warm and survive, leaving the beehive nearly empty. <b>Lila harvests the honey from the beehive.</b>
<b>Belief:</b> The beehive is nearly empty. $(y_p=\text{True}, y_o=\text{True})$	<b>Belief:</b> The beehive is nearly empty. $(y_p=\text{False}, y_o=\text{True})$
<b>Belief:</b> The beehive is full of honey. $(y_p=\text{False}, y_o=\text{False})$	<b>Belief:</b> The beehive is full of honey. $(y_p=\text{True}, y_o=\text{False})$

Figure 10. Example case of representation probing for the *Backward Belief* task. 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 then be decided accordingly.

## C. Additional Probing Results

We present the full probing results in all three tasks for both models using logistic regression models in Figure 11 and Figure 12. The probing accuracies vary across models and tasks. Generally, linear belief representations exist in different models and tasks, and are biased towards representing *oracle’s* belief.



**Figure 11.** Probe accuracies on different tasks based on the attention head activations in all layers of Mistral-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status estimation for both agents using multinomial logistic regression (quaternary).

Furthermore, we explore non-linear probing by fitting an MLP with one hidden layer of 256 channels. Figure 13 shows that the overall probing accuracy increases, suggesting that while some activation heads can decode beliefs linearly, more complex representational structures within certain activation spaces also exist.

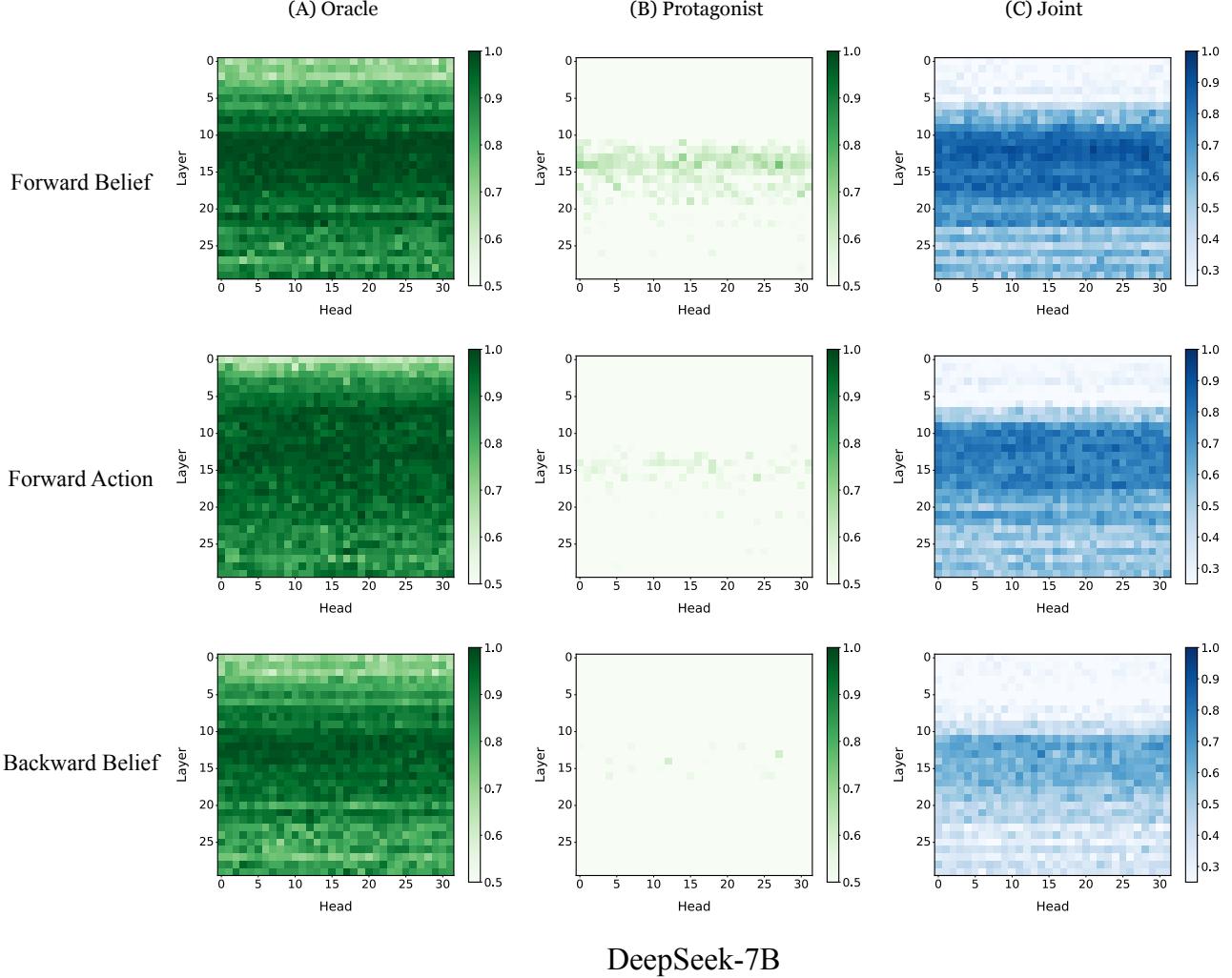


Figure 12. Probe accuracies on different tasks based on the attention head activations in all layers of DeepSeek-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status estimation for both agents using multinomial logistic regression (quaternary).

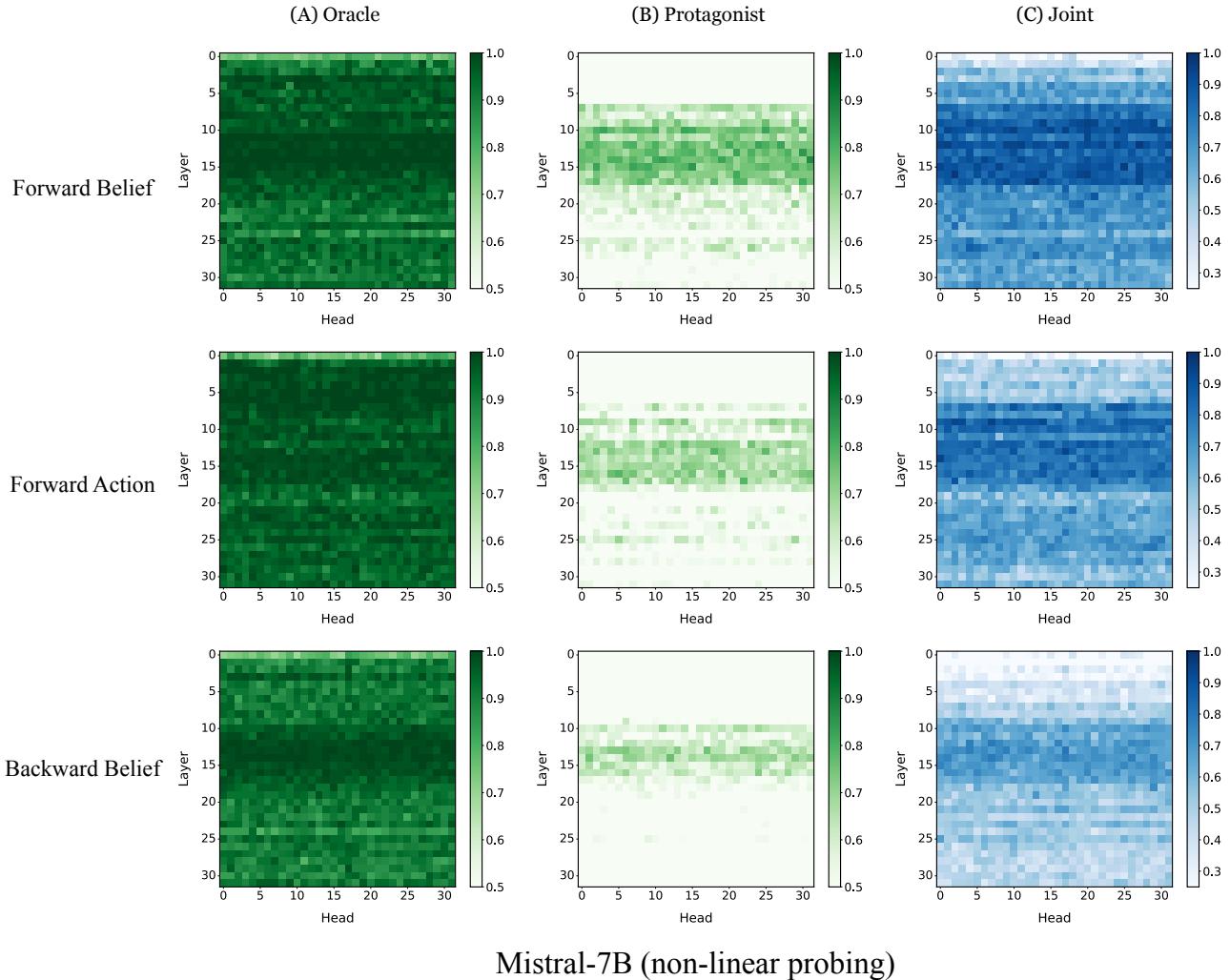
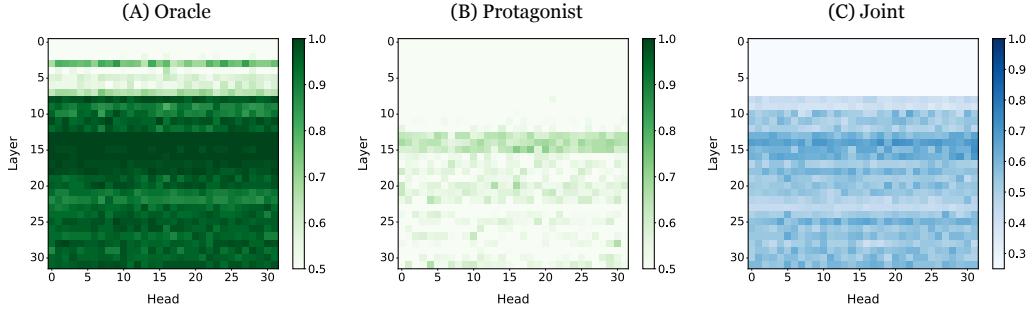


Figure 13. Non-linear probe accuracies on different tasks based on the attention head activations in all layers of Mistral-7B. (A) Belief status estimation for *oracle* using binary classification. (B) Belief status estimation for *protagonist* using binary classification. (C) Joint belief status estimation for both agents using quaternary classification.

## D. Generalization to Other Datasets

In addition to the stories in BigToM (Gandhi et al., 2023), we explore whether our findings could generalize to other narratives. Following (Wilf et al., 2023), we extend our study to the ToMi benchmark (Le et al., 2019), which has quite different narrative templates and scenarios compared to BigToM. It also contains second-order ToM questions which are not present in BigToM.



*Figure 14.* Probe accuracies on ToMi dataset based on the attention head activations in all layers of DeepSeek-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status estimation for both agents using multinomial logistic regression (quaternary).

We perform probing on ToMi by constructing the prompt templates based on its question format, as shown below. Since the ToMi questions are predicting the protagonists' next-step move, it is closer to the *Forward Action* task in BigToM.

*Listing 1. An example of the prompt for probing with the ToMi benchmark*

```
Story: 1 Evelyn entered the cellar. 2 Owen entered the cellar. 3 The
      belt is in the bathtub. 4 The bathtub is in the cellar. 5 Elizabeth
      loves the apple. 6 Evelyn exited the cellar. 7 Owen moved the belt
      to the envelope. 8 The envelope is in the cellar.
Action: Look for the belt in the bathtub.
```

The probing results are shown in Figure 14. Despite the large narrative differences, the probing results are similar - most heads can accurately predict the oracle state; a few attention heads in the middle layers could also separate the protagonist's belief. In addition, we explore if the top heads identified in BigToM and ToMi are similar or not. We present a correlation plot in Figure 15. Here, The accuracies and similarities are averaged over 100 runs of different random seeds. We note that there is a significant correlation (0.86) in probe accuracies across the benchmarks. In other words, the top predictive attention heads in one benchmark exhibit similar efficacy in the other as well. Additionally, The identified top directions in the two datasets exhibit moderate similarity (blue-ish), although their similarities are smaller than cross-task comparison within BigToM.

*Table 6. Model performance comparison on the ToMi benchmark.*

Model	ToM (FB)	All
Mistral-7B (baseline)	63.8	66.0
Mistral-7B (+random)	63.3	64.3
Mistral-7B (+ $T_p F_o$ , transferred from BigToM)	<b>71.0</b>	<b>67.9</b>

Furthermore, we directly transfer the same attention heads and directions identified within the BigToM dataset (*Forward Belief* task) for activation intervention on the ToMi benchmark. Table 6 shows that the directions we identified with BigToM templates are not limited to that specific format. Instead, they exhibit a promising degree of generalization to scenarios with varied narrative styles. This suggests that these representations may indeed represent broadly applicable ToM insights.

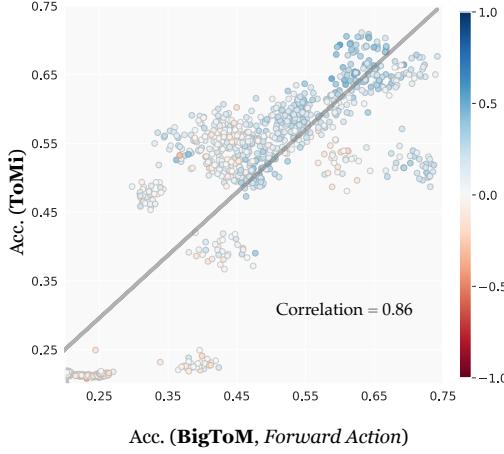


Figure 15. Pairwise comparisons of multinomial probing results on Mistral-7B between ToMi and BigToM. Each point represents a specific attention head, with its position denoting the probe accuracies in the two datasets (multinomial), and its color denoting the cosine similarity between the  $(+T_p F_o)$  probe directions of the two datasets..

## E. Influence on Other Tasks

Table 7. Model performance comparison on the MMLU benchmark.

	All	Humanities	Social Sciences	STEM	Other
Mistral-7B (baseline)	57.5	52.7	<b>66.4</b>	48.3	65.5
Mistral-7B (+random)	57.6	52.7	66.3	48.4	<b>65.6</b>
Mistral-7B ( $+T_p F_o$ )	<b>57.7</b>	<b>53.0</b>	66.3	<b>48.7</b>	65.5

Table 8. Model performance comparison on CoLA, MRPC, and QNLI benchmarks.

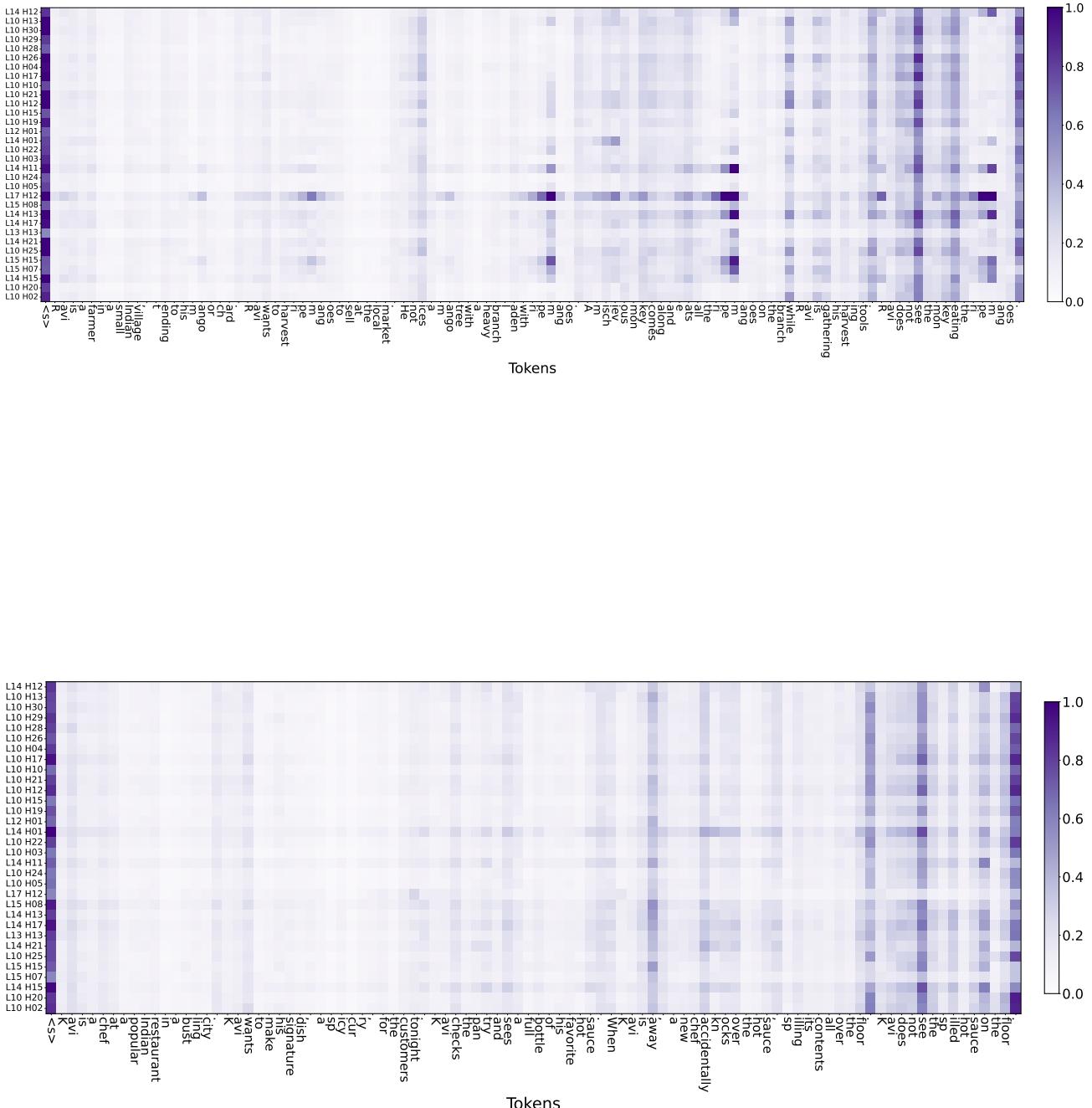
Dataset	CoLA	MRPC	QNLI
Mistral-7B (baseline)	70.0	58.8	62.5
Mistral-7B (+random)	70.1	<b>59.2</b>	<b>63.1</b>
Mistral-7B ( $+T_p F_o$ )	<b>71.1</b>	58.9	63.0

In order to understand the influence of intervention along the identified belief directions on unrelated tasks, we further evaluate the model on some general language understanding benchmarks, including Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019), Microsoft Research Paraphrase Corpus (MRPC) (Dolan & Brockett, 2005), Question-answering NLI (QNLI) (Wang et al., 2018). These benchmarks measure the language model’s performance in tasks unrelated to ToM, including knowledge acquisition, grammar check, etc. We discover that the activation intervention along the identified ToM directions does not significantly change the model performance, as shown in Tables 7 and 8.

## F. Additional Token Gradient Heatmaps

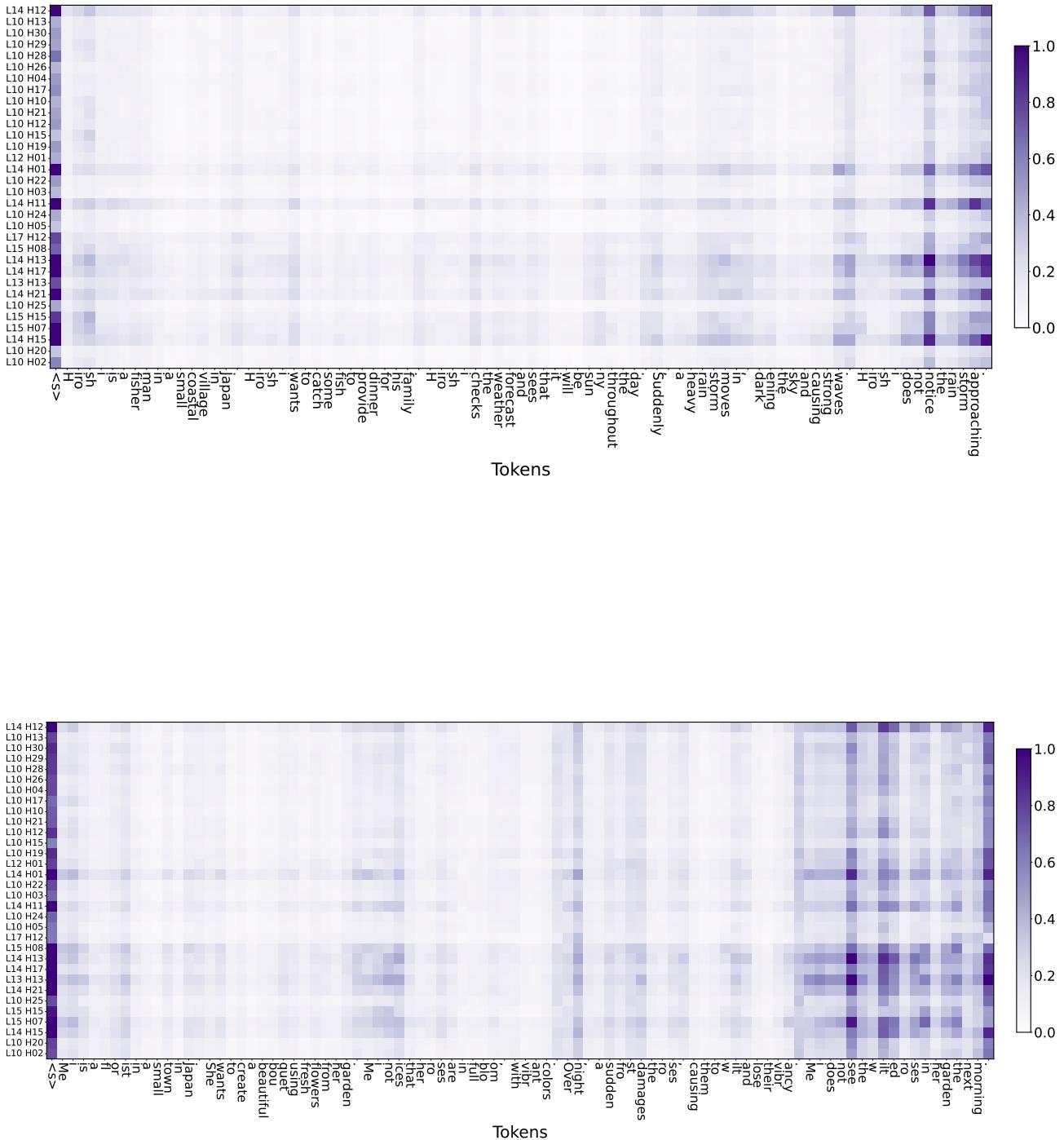
We provide additional results of token gradients with regard to the belief directions given different story prompts, which localizes the key causal elements related to agent beliefs.

## Language Models Represent Beliefs of Self and Others



*Figure 16.* Magnitudes of gradients on the token embeddings with respect to the projection of attention head activations over the corresponding joint belief directions. Each line represents a specific attention head in Mistral-7B.

# Language Models Represent Beliefs of Self and Others



*Figure 17.* Magnitudes of gradients on the token embeddings with respect to the projection of attention head activations over the corresponding joint belief directions. Each line represents a specific attention head in Mistral-7B.

## G. Hyperparameter Analysis

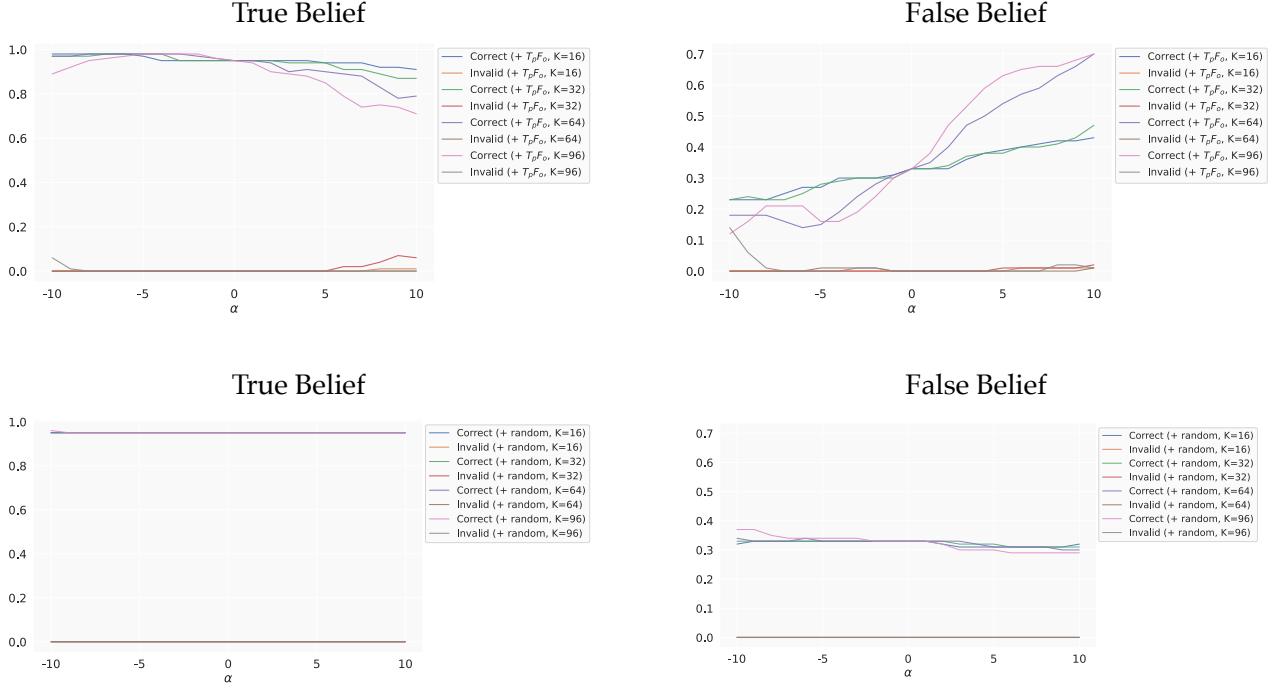


Figure 18. Impact of varying intervention strength  $\alpha$  on the *Forward Belief* task using Mistral-7B. ‘‘Invalid’’ denotes the answer is not recognized by the grading mechanism as the model fails to provide answer in the required format, e.g., delivering uncertain responses.

We provide intervention results with different ( $K, \alpha$ ) combinations in Figure 18.  $K = 16$  yields the most steady performance and is used in our experiments for the model.

## H. Detailed Comparison of LLM Responses

Below, we compare the performance of two models (Mistral-7B, DeepSeek-7B) on the *Forward Belief* reasoning task before and after intervention in two settings (*False Belief*, *True Belief*). Instances where the model's response changes from incorrect to correct, or from correct to incorrect, are recorded in the tables below.

### H.1. Mistral-7B on Forward Belief, False Belief

Changed to Correct Answer: 33/100

Story: Ravi is a farmer in a small Indian village, tending to his mango orchard. Ravi wants to harvest ripe mangoes to sell at the local market. He notices a mango tree with a heavy branch laden with ripe mangoes. A mischievous monkey comes along and eats all the ripe mangoes on the branch while Ravi is gathering his harvesting tools. Ravi does not see the monkey eating the ripe mangoes.

Question: Does Ravi believe the branch has ripe mangoes or no ripe mangoes?

Choose one of the following:

- a) Ravi believes the branch has ripe mangoes.
- b) Ravi believes the branch has no ripe mangoes.

Correct Answer: a)

Unintervened: b) Ravi believes the branch has no ripe mangoes.

Intervened: a. Ravi believes the branch has ripe mangoes.

Story: Aniket is a marine biologist studying coral reefs off the coast of India. Aniket needs to collect samples of coral to analyze the effects of climate change on the reef. Aniket spots a healthy-looking coral formation in a specific area of the reef. A sudden wave surge stirs up sediment, covering the once healthy coral formation and causing it to become damaged. Aniket does not notice the wave surge or the sediment covering the coral.

Question: Does Aniket believe the coral formation is healthy or damaged?

Choose one of the following:

- a) Aniket believes the coral formation is damaged.
- b) Aniket believes the coral formation is healthy.

Correct Answer: b)

Unintervened: a) Aniket believes the coral formation is damaged. (The context states that the cor

Intervened: b) Aniket believes the coral formation is healthy. (at the time of reading the

Story: Juan is a young chef in a small Mexican town. He wants to make a delicious salsa for the upcoming town festival. Juan buys fresh tomatoes from the market, which he believes are ripe and perfect for the salsa. While Juan is away from his kitchen, a mischievous cat knocks over the bag of tomatoes, crushing them and making them unsuitable for the salsa. Juan does not notice the crushed tomatoes on the floor.

Question: Does Juan believe the tomatoes are ripe and perfect or crushed and unsuitable?

Choose one of the following:

- a) Juan believes the tomatoes are ripe and perfect.
- b) Juan believes the tomatoes are crushed and unsuitable.

Correct Answer: a)

Unintervened: b) Juan believes the tomatoes are ripe and perfect, but they are actually crushed and un

Intervened: a) Juan believes the tomatoes are ripe and perfect.

Story: Cheng is a teacher in a rural Chinese village. Cheng wants to use a projector to show a documentary about Chinese history to his students. The projector is plugged in and turned on, as Cheng has prepared it beforehand. As the wind blows outside, the power goes out, causing the projector to turn off. Cheng does not notice the power outage.

Question: Does Cheng believe the projector is ready to use or turned off?

Choose one of the following:

- a) Cheng believes the projector is turned off.
- b) Cheng believes the projector is ready to use.

Correct Answer: b)

Unintervened: a) Cheng believes the projector is ready to use. (However, it is actually turned

Intervened: Cheng believes the projector is ready to use. (before the power outage)

## Language Models Represent Beliefs of Self and Others

Story: Isabela is a baker in a small Argentinian town known for its delicious pastries. Isabela wants to make a batch of alfajores, a traditional Argentinian dessert, for her bakery's customers. Isabela gets a bag of powdered sugar from her pantry to dust the alfajores. Her assistant, who was organizing the pantry earlier, accidentally placed a bag of flour in the powdered sugar's spot, and Isabela unknowingly grabs the bag of flour instead. Isabela does not notice the flour.

Question: Does Isabela believe the bag she took from the pantry contains powdered sugar or flour?

Choose one of the following:

- a) Isabela believes the bag she took from the pantry contains powdered sugar.
- b) Isabela believes the bag she took from the pantry contains flour.

Correct Answer: a)

Unintervened: b) Isabela believes the bag she took from the pantry contains flour.

Intervened: a) Isabela believes the bag she took from the pantry contains powdered sugar.

Story: Mei is a pearl diver in a small coastal village in Japan. Mei wants to find a perfect pearl to give to her grandmother for her birthday. Mei spots an oyster at the bottom of the sea that looks to be the right size and age to contain a pearl. A curious octopus opens the oyster, revealing that there is no pearl inside, and then swims away. Mei does not notice the octopus opening the oyster.

Question: Does Mei believe the oyster she spotted contains a pearl or that it is empty?

Choose one of the following:

- a) Mei believes the oyster she spotted is empty.
- b) Mei believes the oyster she spotted contains a pearl.

Correct Answer: b)

Unintervened: a) Mei believes the oyster she spotted[PAD]is empty.

Intervened: b) Mei believes the oyster she spotted contains a pearl.

Story: Heiwa is a skilled fisherman in a coastal village in Japan. Heiwa wants to catch a large number of fish to sell at the local market. The sea appears calm and full of fish, making it an ideal time for fishing. Suddenly, a massive school of fish moves away from the area, leaving very few fish near Heiwa's fishing spot. Heiwa does not notice the fish moving away.

Question: Does Heiwa believe the sea is calm and full of fish or that the fish have moved away?

Choose one of the following:

- a) Heiwa believes the sea is calm and full of fish.
- b) Heiwa believes the fish have moved away.

Correct Answer: a)

Unintervened: b) Heiwa believes the sea is calm and full of fish, but the fish have moved

Intervened: a) Heiwa believes the sea is calm and full of fish.

Story: Fatima is a potter in a small Moroccan village known for its vibrant ceramics. Fatima wants to create a beautiful vase as a wedding gift for her cousin. Fatima has left her kiln turned off overnight to cool down. A power surge during the night causes the kiln to turn on unexpectedly, making it dangerously hot. Fatima does not feel the heat emanating from the kiln.

Question: Does Fatima believe her kiln is off and cool or unexpectedly hot?

Choose one of the following:

- a) Fatima believes her kiln is unexpectedly hot.
- b) Fatima believes her kiln is off and cool.

Correct Answer: b)

Unintervened: a) Fatima believes her kiln is unexpectedly hot.

Intervened: b) Fatima believes her kiln is off and cool.

Story: Aiko is a florist in a small Japanese town. Aiko wants to create an exquisite flower arrangement for a customer's wedding. Aiko sees a vase filled with fresh, colorful flowers in her shop. Overnight, a sudden drop in temperature causes the flowers in the vase to wilt. Aiko does not notice the wilted flowers in the vase.

Question: Does Aiko believe the vase contains fresh, colorful flowers or wilted flowers?

Choose one of the following:

- a) Aiko believes the vase contains fresh, colorful flowers.
- b) Aiko believes the vase contains wilted flowers.

Correct Answer: a)

Unintervened: b) Aiko believes the vase contains wilted flowers. (She did not notice the

Intervened: Aiko believes the vase contains fresh, colorful flowers. (At the time of reading the

## Language Models Represent Beliefs of Self and Others

Story: Carlos is a farmer in a small village in the Andes. Carlos wants to collect fresh eggs from his chickens to sell at the local market. The nest appears to be full of eggs when Carlos checks it in the morning. A crafty fox sneaks into the chicken coop and steals all the eggs from the nest. Carlos does not see the fox stealing the eggs.

Question: Does Carlos believe the nest contains eggs or that the fox has stolen them?

Choose one of the following:

- a) Carlos believes the fox has stolen the eggs.
- b) Carlos believes the nest contains eggs.

Correct Answer: b)

Unintervened: a) Carlos believes the nest contains eggs.

Intervened: b) Carlos believes the nest contains eggs. (at the time of asking)

Story: Amara is an artist living in a bustling city in Nigeria. She wants to create a beautiful painting for her friend's birthday present. She finds a smooth, flawless canvas in her art studio. Later, her mischievous little brother accidentally spills black paint all over the canvas, leaving it ruined. Amara does not discover the ruined canvas.

Question: Does Amara believe the canvas is smooth and flawless or ruined with black paint?

Choose one of the following:

- a) Amara believes the canvas is ruined with black paint.
- b) Amara believes the canvas is smooth and flawless.

Correct Answer: b)

Unintervened: a) Amara believes the canvas is smooth and flawless (but it is actually ruined).

Intervened: Amara believes the canvas is smooth and flawless. (At the time of the story, she

Story: Fatima is a talented artist living in a small Moroccan village. Fatima wants to create a beautiful painting to donate for the local charity auction. She has a brand-new set of high-quality paintbrushes in her studio. While she is away, her younger brother accidentally ruins the paintbrushes by using them for a school project with glue. Fatima does not notice the ruined paintbrushes.

Question: Does Fatima believe she has high-quality paintbrushes or ruined paintbrushes in her studio?

Choose one of the following:

- a) Fatima believes the paintbrushes are ruined.
- b) Fatima believes she has high-quality paintbrushes in her studio.

Correct Answer: b)

Unintervened: a) Fatima believes the paintbrushes are ruined. (The context states that the paintbr

Intervened: b) Fatima believes she has high-quality paintbrushes in her studio. (At the

Story: Fatima is an archaeologist working at an excavation site in Egypt. She wants to carefully remove an ancient artifact from the ground to study and preserve it. The artifact appears to be intact and in good condition. Suddenly, a sandstorm blows through the area, burying the artifact in sand and causing some damage to it. Fatima does not witness the sandstorm and its effect on the artifact.

Question: Does Fatima believe the artifact is intact and in good condition or damaged by the sandstorm?

Choose one of the following:

- a) Fatima believes the artifact is intact and in good condition.
- b) Fatima believes the artifact is damaged by the sandstorm.

Correct Answer: a)

Unintervened: b) Fatima believes the artifact is damaged by the sandstorm.

Intervened: a. (Fatima believes the artifact is intact and in good condition.)

Story: Juan is a gardener in a small village in Mexico, tending to a flower garden near the town plaza. Juan wants to water a specific flower bed to ensure its health and beauty. Juan sees that the garden hose is neatly coiled and ready for use. Unexpectedly, a group of children playing nearby accidentally puncture the hose with a sharp stick while running around. Juan does not see the children accidentally puncture the hose.

Question: Does Juan believe the garden hose is in good condition and functional or punctured and non-functional?

Choose one of the following:

- a) Juan believes the garden hose is in good condition and functional.
- b) Juan believes the garden hose is punctured and non-functional.

Correct Answer: a)

Unintervened: b) Juan believes the garden hose is punctured and non-functional. (Juan

Intervened: a) Juan believes the garden hose is in good condition and functional. (At the moment,

## Language Models Represent Beliefs of Self and Others

Story: Hiro is a bonsai enthusiast in a small village in Japan. He wants to trim his favorite bonsai tree to maintain its shape and health. Hiro sees that the tree's branches are overgrown and need pruning. A gust of wind blows through Hiro's garden, snapping off several of the overgrown branches from the bonsai tree. Hiro does not see the gust of wind snapping off the branches.

Question: Does Hiro believe the bonsai tree has overgrown branches or that they have been snapped off by the wind?

Choose one of the following:

- a) Hiro believes the bonsai tree has overgrown branches.
- b) Hiro believes the branches have been snapped off by the wind.

Correct Answer: a)

Unintervened: b) Hiro believes the branches have been snapped off by the wind.

Intervened: a. Hiro believes the bonsai tree has overgrown branches.

Story: Tumelo is a skilled fisherman in a coastal village in South Africa. He wants to catch fish using his large, sturdy fishing net. Tumelo inspected the net before leaving home and saw that it was in perfect condition. As he walks to the shore, a mischievous seagull pecks at the net, causing a large tear in it. Tumelo does not see the seagull damaging the net.

Question: Does Tumelo believe the fishing net is in perfect condition or torn?

Choose one of the following:

- a) Tumelo believes the fishing net is torn.
- b) Tumelo believes the fishing net is in perfect condition.

Correct Answer: b)

Unintervened: a) Tumelo believes the fishing net is in perfect condition.

Intervened: Before leaving[PAD]=>Tumelo believes the fishing net is in perfect condition.

Answer:

Story: Kofi is a fisherman in a small coastal village in Ghana. He wants to set sail on his boat to catch fish for his family and the village market. Kofi checks the weather and sees clear skies and calm waters. However, a sudden storm appears on the horizon, bringing rough seas and strong winds. Kofi does not notice the approaching storm.

Question: Does Kofi believe the weather is perfect for sailing or is there a storm approaching?

Choose one of the following:

- a) Kofi believes the weather is perfect for sailing.
- b) Kofi believes there is a storm approaching.

Correct Answer: a)

Unintervened: b) Kofi believes the weather is perfect for sailing (but he is wrong).

Intervened: a) Kofi believes the weather is perfect for sailing.

Story: Mei is a fisherman in a small coastal village in China. She wants to go fishing early in the morning when the tide is high. Mei checks the tide chart and sees that the tide will be high at 5 AM. However, a strong storm far out at sea has caused the tide to come in much earlier, at 3 AM. Mei does not look out the window and does not notice the high tide.

Question: Does Mei believe the tide is high at 5 AM or 3 AM?

Choose one of the following:

- a) Mei believes the tide is high at 5 AM.
- b) Mei believes the tide is high at 3 AM.

Correct Answer: a)

Unintervened: b) Mei believes the tide is high at 3 AM (based on the actual situation).

Intervened: Mei believes the tide is high at 5 AM. (This is what Mei currently believes)

Story: Fatima is an archaeologist working on a dig site in Egypt, searching for ancient artifacts. She wants to carefully excavate a small, fragile artifact she has discovered. Fatima observes that the soil around the artifact is soft and loose. Just as Fatima is about to begin excavating, a strong gust of wind blows through the site, hardening the soil around the artifact. Fatima does not notice the gust of wind or the change in the soil's condition.

Question: Does Fatima believe the soil around the artifact is soft and loose or hardened?

Choose one of the following:

- a) Fatima believes the soil around the artifact has hardened.
- b) Fatima believes the soil around the artifact is soft and loose.

Correct Answer: b)

Unintervened: a) Fatima believes the soil around the artifact has hardened.

Intervened: b) Fatima believes the soil around the artifact is soft and loose.

## Language Models Represent Beliefs of Self and Others

Story: Rohan is a potter in a small Indian village. Rohan wants to make a beautiful ceramic vase as a gift for his friend's housewarming party. Rohan selects a bag of high-quality clay in his workshop that appears to be soft and malleable. While Rohan is away for lunch, the hot sun dries the clay, making it hard and unworkable. Rohan does not notice the dried, hard clay.

Question: Does Rohan believe the bag of clay is soft and malleable or hard and unworkable?

Choose one of the following:

- a) Rohan believes the bag of clay is hard and unworkable.
- b) Rohan believes the bag of clay is soft and malleable.

Correct Answer: b)

Unintervened: a) Rohan believes the bag of clay is soft and malleable (but it's

Intervened: Rohan believes the bag of clay is soft and malleable (before the incident), but later

Story: Rohit is a renowned painter from India who has been commissioned to create a mural on a city building. Rohit wants to paint an intricate scene depicting the vibrant culture and history of India. Rohit selects a set of high-quality, long-lasting paint colors for the mural. While Rohit is out for lunch, a leak in the ceiling drips water onto the paint colors, diluting them and making them less durable. Rohit does not notice the water-damaged paint colors.

Question: Does Rohit believe the paint colors are of high quality and long-lasting or water-damaged and less durable?

Choose one of the following:

- a) Rohit believes the paint colors are water-damaged and less durable.
- b) Rohit believes the paint colors are of high quality and long-lasting.

Correct Answer: b)

Unintervened: a) Rohit does not know the true condition of the paint colors. However, due to the

Intervened: b) Rohit believes the paint colors are of high quality and long-lasting. (at

Story: Amara is a talented henna artist in a bustling Indian market. Amara wants to create intricate henna designs for a group of bridesmaids attending a wedding. Amara checks her supply of henna paste and finds it to be smooth and consistent. While Amara is busy servicing another client, a sudden rainstorm causes the henna paste to become diluted and runny. Amara does not realize the rain has affected the henna paste.

Question: Does Amara believe the henna paste is smooth and consistent or diluted and runny?

Choose one of the following:

- a) Amara believes the henna paste is smooth and consistent.
- b) Amara believes the henna paste is diluted and runny.

Correct Answer: a)

Unintervened: b) Amara believes the henna paste is smooth and consistent (but it is actually dilut

Intervened: a) Amara believes the henna paste is smooth and consistent.

Story: Kalyani is a traditional Indian dancer in a temple in Kerala, India. Kalyani wants to perform a special dance for the upcoming festival to pay homage to the temple's deity. The temple floor is clean and free of any obstacles, as Kalyani has prepared the space for her dance practice. An accidental spill of oil from a temple lamp causes a slippery patch on the floor. Kalyani does not notice the oil spill.

Question: Does Kalyani believe the temple floor is clean and perfect for dancing or slippery due to the oil spill?

Choose one of the following:

- a) Kalyani believes the temple floor is clean and perfect for dancing.
- b) Kalyani believes the floor is slippery due to the oil spill.

Correct Answer: a)

Unintervened: b) Kalyani believes the floor is slippery due to the oil spill.

Intervened: a) Kalyani believes the temple floor is clean and perfect for dancing.

Story: Naledi is a young astronomer in a remote observatory in South Africa. Naledi wants to observe a distant supernova to gather data for her research. The telescope is pointed at the coordinates of the supernova, ready for observation. A minor earthquake occurs, causing the telescope to shift its position slightly. Naledi does not notice the earthquake and its effect on the telescope.

Question: Does Naledi believe the telescope is aimed at the correct location or has shifted due to the earthquake?

Choose one of the following:

- a) Naledi believes the telescope is aimed at the correct location.
- b) Naledi believes the telescope has shifted due to the earthquake.

Correct Answer: a)

Unintervened: b) Naledi is unaware that the telescope has shifted due to the earthquake.

Intervened: a) Naledi believes the telescope is aimed at the correct location.

## Language Models Represent Beliefs of Self and Others

Story: Emilia is a talented baker in a small Italian town. She plans to bake a large batch of biscotti to sell at the local market. Emilia inspects the oven's temperature, which appears to be set at the correct temperature for baking. A power surge causes the oven's temperature to spike, making it too hot for baking biscotti. Emilia does not notice the oven's temperature increase.

Question: Does Emilia believe the oven is set at the correct temperature for baking biscotti or too hot?

Choose one of the following:

- a) Emilia believes the oven is set at the correct temperature for baking biscotti.
- b) Emilia believes the oven is too hot for baking biscotti.

Correct Answer: a)

Unintervened: b) Emilia believes the oven is too hot for baking biscotti (even though she doesn

Intervened: a) Emilia believes the oven is set at the correct temperature for baking biscotti. (

Story: Emeka is a skilled fisherman in a small Nigerian village. He wants to catch fish to feed his family and sell at the market. Emeka prepares his fishing net, which appears to be strong and without any holes. However, while Emeka is not looking, a mischievous child cuts a large hole in the net with a knife. Emeka does not see the child cutting the hole in the fishing net.

Question: Does Emeka believe the fishing net is strong and without holes, or that it has a large hole in it?

Choose one of the following:

- a) Emeka believes the fishing net is strong and without holes.
- b) Emeka believes that the fishing net has a large hole in it.

Correct Answer: a)

Unintervened: b) Emeka believes that the fishing net has a large hole in it (without knowing it

Intervened: a) Emeka believes the fishing net is strong and without holes.

Story: Aria is a talented baker from Italy who has been asked to bake a cake for a friend's birthday party. She wants to create a delicious, multi-layered cake that will impress everyone. Aria has a fresh batch of ingredients ready to use for the cake. However, a mischievous cat sneaks into her kitchen and nibbles on some of the ingredients, making them no longer fresh. Aria does not notice the cat or the nibbled ingredients.

Question: Does Aria believe her ingredients are fresh and suitable or nibbled on and no longer fresh?

Choose one of the following:

- a) Aria believes the ingredients are nibbled on and no longer fresh.
- b) Aria believes her ingredients are fresh and suitable for baking the cake.

Correct Answer: b)

Unintervened: a) Aria believes her ingredients are fresh and suitable for baking the cake (but they are actually

Intervened: b) Aria believes her ingredients are fresh and suitable for baking the cake.

Story: Ailani is a skilled Hawaiian chef preparing a traditional luau feast for her community. She wants to make a delicious laulau dish using fresh taro leaves. Ailani examines the taro leaves, which look fresh and green. However, a sudden downpour soaks the taro leaves, causing them to become wilted and less ideal for the dish. Ailani does not notice the downpour and the effect it has on the taro leaves.

Question: Does Ailani believe the taro leaves are fresh and of high quality or wilted and less ideal?

Choose one of the following:

- a) Ailani believes the taro leaves are fresh and of high quality.
- b) Ailani believes the taro leaves are wilted and less ideal.

Correct Answer: a)

Unintervened: b) Ailani believes the taro leaves are wilted and less ideal.

Intervened: Ailani believes the taro leaves are fresh and of high quality. (Based on the context

Story: Ravi is a fisherman living in a coastal village in India. Ravi wants to catch a large amount of fish to sell at the market and provide for his family. The waters near the shore are calm and full of fish, making it an ideal location for fishing. An underwater earthquake occurs, causing the fish to frantically swim away from the coast. Ravi does not sense the earthquake.

Question: Does Ravi believe the waters near the shore are calm and full of fish, or that the fish have frantically swum away?

Choose one of the following:

- a) Ravi believes the fish have frantically swum away from the coast.
- b) Ravi believes the waters near the shore are calm and full of fish.

Correct Answer: b)

Unintervened: a) Ravi believes the waters near the shore are calm and full of fish (but he is

Intervened: b) Ravi believes the waters near the shore are calm and full of fish.

Story: Surya is a fisherman in a coastal village in India. Surya plans to go fishing early in the morning to catch fish for his family and to sell at the local market. Surya's fishing net looks strong and ready for another day of fishing. Overnight, a group of mischievous monkeys rummage through the fishing equipment and badly damage Surya's fishing net. Surya does not realize the monkeys have damaged his fishing net.

Question: Does Surya believe that his fishing net is strong and ready to be used, or that it has been badly damaged?

Choose one of the following:

- a) Surya believes his fishing net has been badly damaged.
- b) Surya believes his fishing net is strong and ready to be used.

Correct Answer: b)

Unintervened: a) Surya believes his fishing net is strong and ready to be used (but it has

Intervened: b) Surya believes his fishing net is strong and ready to be used.

Story: Takeshi is a fisherman in a small Japanese coastal village. Takeshi's goal is to catch enough fish to provide for his family and sell at the market. The ocean is calm and teeming with fish, making it an ideal day for fishing. However, a sudden underwater volcanic eruption nearby causes the fish to scatter, leaving the area devoid of fish. Takeshi does not notice the sudden disappearance of fish in the area.

Question: Does Takeshi believe the ocean is teeming with fish, or devoid of fish due to the volcanic eruption?

Choose one of the following:

- a) Takeshi believes the ocean is devoid of fish due to the volcanic eruption.
- b) Takeshi believes the ocean is teeming with fish.

Correct Answer: b)

Unintervened: a) Takeshi believes the ocean is teeming with fish (but is unaware of

Intervened: Takeshi believes the ocean is teeming with fish.

Story: Hiroshi is a fisherman in a small coastal village in Japan. Hiroshi wishes to catch enough fish to feed his family and sell the surplus at the local market. The fishing nets are in excellent condition, ready to be cast into the sea. However, a curious pod of dolphins swims by and accidentally tears the nets with their sharp dorsal fins. Hiroshi does not see the dolphins or the damage they caused to the nets.

Question: Does Hiroshi believe the fishing nets are in excellent condition or torn by the dolphins?

Choose one of the following:

- a) Hiroshi believes the fishing nets are in excellent condition.
- b) Hiroshi believes the fishing nets are torn by the dolphins.

Correct Answer: a)

Unintervened: b) Hiroshi is unaware that the fishing nets are torn by the dolphins

Intervened: a) Hiroshi believes the fishing nets are in excellent condition.

Story: Mei, a talented pastry chef from a small town in China, is preparing to bake a large cake for a local celebration. Mei wants to use a specific cake mold to create a unique and intricate design. Mei sees the cake mold hanging on the wall, clean and ready for use. However, during a brief power outage, a gust of wind blows through the kitchen, knocking the cake mold off the wall and denting it severely. Mei does not notice the dented cake mold on the floor.

Question: Does Mei believe the cake mold is clean and ready for use or severely dented and unsuitable for use?

Choose one of the following:

- a) Mei believes the cake mold is clean and ready for use.
- b) Mei believes the cake mold is severely dented and unsuitable for use.

Correct Answer: a)

Unintervened: b) Mei believes the cake mold is severely dented and unsuitable for use.

Intervened: a) Mei believes the cake mold is clean and ready for use. (at the start of

Changed to Incorrect Answer: 0/100

## H.2. Mistral-7B on Forward Belief, True Belief

Changed to Correct Answer: 0/100

Changed to Incorrect Answer: 10/100

## Language Models Represent Beliefs of Self and Others

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.

Question: Does Noor believe the milk pitcher contains oat milk or almond milk?

Choose one of the following:

- a) Noor believes the milk pitcher contains almond milk.
- b) Noor believes the milk pitcher contains oat milk.

Correct Answer: a)

Unintervened: a) Noor believes the milk pitcher contains almond milk.

Intervened: a

(This question is based on the context provided, and the context states that Noor

Story: Azola is a teacher at a small school in rural China. She wants to surprise her students with a new book for the school library. Azola got a notification from Amazon stating that the package containing the book was left at her doorstep. A gust of wind blows the package off her doorstep, and a neighbor replaces it with a different package containing a hand-knit scarf. Azola witnesses the package swap.

Question: Does Azola believe the package contains the book she ordered or a hand-knit scarf?

Choose one of the following:

- a) Azola believes the package contains a hand-knit scarf.
- b) Azola believes the package contains the book she ordered.

Correct Answer: a)

Unintervened: a) Azola believes the package contains a hand-knit scarf.

Intervened: a

(This question is a part of a multiple-choice quiz. Since there is

Story: Ravi is a farmer in a small Indian village, tending to his mango orchard. Ravi wants to harvest ripe mangoes to sell at the local market. He notices a mango tree with a heavy branch laden with ripe mangoes. A mischievous monkey comes along and eats all the ripe mangoes on the branch while Ravi is gathering his harvesting tools. Ravi sees the monkey eating the ripe mangoes.

Question: Does Ravi believe the branch has ripe mangoes or no ripe mangoes?

Choose one of the following:

- a) Ravi believes the branch has no ripe mangoes.
- b) Ravi believes the branch has ripe mangoes.

Correct Answer: a)

Unintervened: a) Ravi believes the branch has no ripe mangoes.

Intervened: Ravi believes the branch has ripe mangoes. (At the moment, he still believes

Story: Farhan is a street food vendor in a bustling market in Istanbul. Farhan wants to prepare a delicious chicken shawarma wrap for a customer who asked for extra garlic sauce. Farhan takes a bottle of garlic sauce from the shelf and pours it onto the shawarma wrap. His assistant, who is busy preparing a different wrap, accidentally swaps the garlic sauce with a bottle of hot sauce, and Farhan unknowingly pours the hot sauce onto the shawarma wrap. Farhan notices the sauce swap.

Question: Does Farhan believe the sauce he poured is garlic sauce or hot sauce?

Choose one of the following:

- a) Farhan believes the sauce he poured is hot sauce.
- b) Farhan believes the sauce he poured is garlic sauce.

Correct Answer: a)

Unintervened: a) Farhan believes the sauce he poured is hot sauce.

Intervened: b) Farhan believes the sauce he poured is garlic sauce. (At the time of pouring

Story: Chi is a fisherman living on a small island in the South Pacific. Chi wants to go on a fishing trip to catch some fish for his family's dinner. The sea appears calm and inviting. Suddenly, a strong storm rolls in, causing the sea to become rough and dangerous. Chi notices the storm and the change in the sea.

Question: Does Chi believe the sea is calm and inviting or rough and dangerous?

Choose one of the following:

- a) Chi believes the sea is rough and dangerous.
- b) Chi believes the sea is calm and inviting.

Correct Answer: a)

Unintervened: a) Chi believes the sea is rough and dangerous.

Intervened: a

—  
Story: Chi is a fisherman living on a small island in

Story: Hiro is a bonsai enthusiast in a small village in Japan. He wants to trim his favorite bonsai tree to maintain its shape and health. Hiro sees that the tree's branches are overgrown and need pruning. A gust of wind blows through Hiro's garden, snapping off several of the overgrown branches from the bonsai tree. Hiro witnesses the gust of wind snapping off the branches.

Question: Does Hiro believe the bonsai tree has overgrown branches or that they have been snapped off by the wind?

Choose one of the following:

- a) Hiro believes the branches have been snapped off by the wind.
- b) Hiro believes the bonsai tree has overgrown branches.

Correct Answer: a)

Unintervened: a) Hiro believes the branches have been snapped off by the wind.

Intervened: Hiro believes the bonsai tree has overgrown branches. (initially)

Story: Olumide, a skilled woodcarver in a Nigerian village, is preparing to carve a beautiful sculpture for the village chief. Olumide wants to use a sharp chisel to create intricate details on the sculpture. Olumide observes his set of chisels and sees one that appears to be sharp and in perfect condition. However, while Olumide is talking to a fellow artisan, a child from the village accidentally drops the chisel, causing it to become blunt and damaged. Olumide notices the damaged chisel on the ground.

Question: Does Olumide believe the chisel is sharp and in perfect condition or blunt and damaged?

Choose one of the following:

- a) Olumide believes the chisel is sharp and in perfect condition.
- b) Olumide believes the chisel is blunt and damaged.

Correct Answer: b)

Unintervened: b) Olumide believes the chisel is blunt and damaged.

Intervened: a

—  
Story: Olumide, a skilled woodcarver in a

Story: Odion is an archaeologist conducting research at an ancient Egyptian site. He wants to discover a hidden chamber within the pyramid he is exploring. Odion observes a wall with an unusual pattern that appears to be a concealed entrance. While Odion is away, a sandstorm fills the entrance, making it impossible to access the hidden chamber. Odion sees the sand-filled entrance after the storm.

Question: Does Odion believe the wall with the unusual pattern is hiding the entrance or that the entrance is blocked by sand?

Choose one of the following:

- a) Odion believes the entrance is blocked by sand.
- b) Odion believes the wall with the unusual pattern is hiding the entrance.

Correct Answer: a)

Unintervened: a) Odion believes the entrance is blocked by sand.

Intervened: b) Odion believes the wall with the unusual pattern is hiding the entrance.

Story: Kunal is a talented potter in a small Indian village. Kunal wants to create a set of earthenware bowls as a gift for his friend's housewarming. Kunal inspects a large lump of clay, finding it to be moist and malleable. However, while Kunal is away for a short break, the sun beats down on the clay, drying it out and making it hard and brittle. Kunal notices the sun-dried, brittle clay.

Question: Does Kunal believe the clay is moist and malleable or hard and brittle?

Choose one of the following:

- a) Kunal believes the clay is moist and malleable.
- b) Kunal believes the clay is hard and brittle.

Correct Answer: b)

Unintervened: b) Kunal believes the clay is hard and brittle.

Intervened: a. (Kunal believes the clay is moist and malleable.)

Story: Naledi is a talented South African chef preparing a traditional braai for her friends and family. Naledi wants to make a delicious marinade for the meats and vegetables to be grilled. She takes a bottle of vinegar from the cupboard to mix with other ingredients for the marinade. Her cousin, who was helping in the kitchen earlier, accidentally placed a bottle of water in the vinegar's spot, and Naledi unknowingly grabs the bottle of water instead. Naledi notices the water and puts it back.

Question: Does Naledi believe the bottle she took from the cupboard contains vinegar or water?

Choose one of the following:

- a) Naledi believes the bottle she took from the cupboard contains water.
- b) Naledi believes the bottle she took from the cupboard contains vinegar.

Correct Answer: a)

Unintervened: a) Naledi believes the bottle she took from the cupboard contains water.

Intervened: b) Naledi believes the bottle she took from the cupboard contains vinegar. (At

### H.3. DeepSeek-7B on Forward Belief, False Belief

Changed to Correct Answer: 32/100

Story: Zara is attending an art exhibition at a local gallery in her town. Zara wants to learn more about a particular abstract painting she admires. The label next to the artwork states that the painting is the work of a famous artist, and Zara reads the label carefully. A gallery staff member, realizing there was a mix-up, changes the label with the correct information attributing the painting to an emerging artist instead. Zara does not read the updated label.

Question: Does Zara believe the painting is the work of a famous artist or an emerging artist?

Choose one of the following:

- a) Zara believes the painting is the work of a famous artist.
- b) Zara believes the painting is the work of an emerging artist.

Correct Answer: a)

Unintervened: b) Zara believes the painting is the work of an emerging artist.

Intervened: a) Zara believes the painting is the work of a famous artist.

Story: Azola is a teacher at a small school in rural China. She wants to surprise her students with a new book for the school library. Azola got a notification from Amazon stating that the package containing the book was left at her doorstep. A gust of wind blows the package off her doorstep, and a neighbor replaces it with a different package containing a hand-knit scarf. Azola does not witness the package swap.

Question: Does Azola believe the package contains the book she ordered or a hand-knit scarf?

Choose one of the following:

- a) Azola believes the package contains the book she ordered.
- b) Azola believes the package contains a hand-knit scarf.

Correct Answer: a)

Unintervened: b) Azola believes the package contains a hand-knit scarf.

Intervened: a) Azola believes the package contains the book she ordered.

## Language Models Represent Beliefs of Self and Others

Story: Ravi is a farmer in a small Indian village, tending to his mango orchard. Ravi wants to harvest ripe mangoes to sell at the local market. He notices a mango tree with a heavy branch laden with ripe mangoes. A mischievous monkey comes along and eats all the ripe mangoes on the branch while Ravi is gathering his harvesting tools. Ravi does not see the monkey eating the ripe mangoes.

Question: Does Ravi believe the branch has ripe mangoes or no ripe mangoes?

Choose one of the following:

- a) Ravi believes the branch has ripe mangoes.
- b) Ravi believes the branch has no ripe mangoes.

Correct Answer: a)

Unintervened: b) Ravi believes the branch has no ripe mangoes.

Intervened: a) Ravi believes the branch has ripe mangoes.

Story: Kael is a farmer in a remote village in the Andes. Kael wants to plant a new crop of high-altitude potatoes in one of his terraced fields. The soil in the field appears rich and well-fertilized. Overnight, heavy rainfall washes away much of the fertile soil, leaving mostly rocky dirt behind. Kael does not notice the change in the soil after the rainfall.

Question: Does Kael believe the soil is rich and well-fertilized or mostly rocky dirt?

Choose one of the following:

- a) Kael believes the soil is rich and well-fertilized.
- b) Kael believes the soil is mostly rocky dirt.

Correct Answer: a)

Unintervened: b) Kael believes the soil is mostly rocky dirt.

Intervened: a) Kael believes the soil is rich and well-fertilized.

Story: Farhan is a street food vendor in a bustling market in Istanbul. Farhan wants to prepare a delicious chicken shawarma wrap for a customer who asked for extra garlic sauce. Farhan takes a bottle of garlic sauce from the shelf and pours it onto the shawarma wrap. His assistant, who is busy preparing a different wrap, accidentally swaps the garlic sauce with a bottle of hot sauce, and Farhan unknowingly pours the hot sauce onto the shawarma wrap. Farhan does not notice the sauce swap.

Question: Does Farhan believe the sauce he poured is garlic sauce or hot sauce?

Choose one of the following:

- a) Farhan believes the sauce he poured is garlic sauce.
- b) Farhan believes the sauce he poured is hot sauce.

Correct Answer: a)

Unintervened: b) Farhan believes the sauce he poured is hot sauce.

Intervened: a) Farhan believes the sauce he poured is garlic sauce.

Story: Lila is a beekeeper in a small village in Eastern Europe. She wants to harvest honey from one of her beehives. The beehive is full of honey, as Lila observed during her last inspection. A sudden cold snap occurs overnight, causing the bees to consume the honey to stay warm and survive, leaving the beehive nearly empty. Lila does not notice the cold snap and the decreased honey in her beehive.

Question: Does Lila believe the beehive is full of honey or nearly empty?

Choose one of the following:

- a) Lila believes the beehive is full of honey.
- b) Lila believes the beehive is nearly empty.

Correct Answer: a)

Unintervened: b) Lila believes the beehive is nearly empty.

Intervened: a) Lila believes the beehive is full of honey.

Story: Kavi is a chef at a popular Indian restaurant in a bustling city. Kavi wants to make his signature dish, a spicy curry, for the customers tonight. Kavi checks the pantry and sees a full bottle of his favorite hot sauce. When Kavi is away, a new chef accidentally knocks over the hot sauce, spilling its contents all over the floor. Kavi does not see the spilled hot sauce on the floor.

Question: Does Kavi believe there is a full bottle of hot sauce available or that the hot sauce has been spilled?

Choose one of the following:

- a) Kavi believes the hot sauce has been spilled.
- b) Kavi believes there is a full bottle of hot sauce available.

Correct Answer: b)

Unintervened: a) Kavi believes the hot sauce has been spilled.

Intervened: b) Kavi believes there is a full bottle of hot sauce available.

## Language Models Represent Beliefs of Self and Others

Story: Juan is a young chef in a small Mexican town. He wants to make a delicious salsa for the upcoming town festival. Juan buys fresh tomatoes from the market, which he believes are ripe and perfect for the salsa. While Juan is away from his kitchen, a mischievous cat knocks over the bag of tomatoes, crushing them and making them unsuitable for the salsa. Juan does not notice the crushed tomatoes on the floor.

Question: Does Juan believe the tomatoes are ripe and perfect or crushed and unsuitable?

Choose one of the following:

- a) Juan believes the tomatoes are ripe and perfect.
- b) Juan believes the tomatoes are crushed and unsuitable.

Correct Answer: a)

Unintervened: b) Juan believes the tomatoes are crushed and unsuitable.

Intervened: a) Juan believes the tomatoes are ripe and perfect.

Story: Isabela is a baker in a small Argentinian town known for its delicious pastries. Isabela wants to make a batch of alfajores, a traditional Argentinian dessert, for her bakery's customers. Isabela gets a bag of powdered sugar from her pantry to dust the alfajores. Her assistant, who was organizing the pantry earlier, accidentally placed a bag of flour in the powdered sugar's spot, and Isabela unknowingly grabs the bag of flour instead. Isabela does not notice the flour.

Question: Does Isabela believe the bag she took from the pantry contains powdered sugar or flour?

Choose one of the following:

- a) Isabela believes the bag she took from the pantry contains powdered sugar.
- b) Isabela believes the bag she took from the pantry contains flour.

Correct Answer: a)

Unintervened: b) Isabela believes the bag she took from the pantry contains flour.

Intervened: a) Isabela believes the bag she took from the pantry contains powdered sugar.

Story: Chi is a fisherman living on a small island in the South Pacific. Chi wants to go on a fishing trip to catch some fish for his family's dinner. The sea appears calm and inviting. Suddenly, a strong storm rolls in, causing the sea to become rough and dangerous. Chi does not notice the storm and the change in the sea.

Question: Does Chi believe the sea is calm and inviting or rough and dangerous?

Choose one of the following:

- a) Chi believes the sea is calm and inviting.
- b) Chi believes the sea is rough and dangerous.

Correct Answer: a)

Unintervened: b) Chi believes the sea is rough and dangerous.

Intervened: a)

Story: Mei is a pearl diver in a small coastal town in Japan. Mei wants to dive deep into the sea to find a rare and valuable black pearl for her grandmother's birthday gift. The water in the cove is calm with barely any waves. Suddenly, a storm rolls in, and the once calm water becomes rough and choppy. Mei does not notice the change in the water as the storm approaches.

Question: Does Mei believe the water is calm and perfect for diving or rough and choppy due to the storm?

Choose one of the following:

- a) Mei believes the water is calm and perfect for diving.
- b) Mei believes the water is rough and choppy due to the storm.

Correct Answer: a)

Unintervened: b) Mei believes the water is rough and choppy due to the storm.

Intervened: a) Mei believes the water is calm and perfect for diving.

Story: Amina is an archaeologist working on an excavation site in Egypt. Amina wants to carefully remove a valuable artifact from the ground using her brush and trowel. The artifact appears to be intact and undamaged. Suddenly, the ground shifts slightly, causing a small crack to form on the artifact. Amina does not notice the ground shift and the crack on the artifact.

Question: Does Amina believe the artifact is in perfect condition or cracked?

Choose one of the following:

- a) Amina believes the artifact is in perfect condition.
- b) Amina believes the artifact is cracked.

Correct Answer: a)

Unintervened: b) Amina believes the artifact is cracked.

Intervened: a) Amina believes the artifact is in perfect condition.

## Language Models Represent Beliefs of Self and Others

Story: Aiko is a florist in a small Japanese town. Aiko wants to create an exquisite flower arrangement for a customer's wedding. Aiko sees a vase filled with fresh, colorful flowers in her shop. Overnight, a sudden drop in temperature causes the flowers in the vase to wilt. Aiko does not notice the wilted flowers in the vase.

Question: Does Aiko believe the vase contains fresh, colorful flowers or wilted flowers?

Choose one of the following:

- a) Aiko believes the vase contains fresh, colorful flowers.
- b) Aiko believes the vase contains wilted flowers.

Correct Answer: a)

Unintervened: b) Aiko believes the vase contains wilted flowers.

Intervened: a) Aiko believes the vase contains fresh, colorful flowers.

Story: Fatima is an archaeologist working at an excavation site in Egypt. She wants to carefully remove an ancient artifact from the ground to study and preserve it. The artifact appears to be intact and in good condition. Suddenly, a sandstorm blows through the area, burying the artifact in sand and causing some damage to it. Fatima does not witness the sandstorm and its effect on the artifact.

Question: Does Fatima believe the artifact is intact and in good condition or damaged by the sandstorm?

Choose one of the following:

- a) Fatima believes the artifact is intact and in good condition.
- b) Fatima believes the artifact is damaged by the sandstorm.

Correct Answer: a)

Unintervened: b) Fatima believes the artifact is damaged by the sandstorm.

Intervened: a) Fatima believes the artifact is intact and in good condition.

Story: Juan is a gardener in a small village in Mexico, tending to a flower garden near the town plaza. Juan wants to water a specific flower bed to ensure its health and beauty. Juan sees that the garden hose is neatly coiled and ready for use. Unexpectedly, a group of children playing nearby accidentally puncture the hose with a sharp stick while running around. Juan does not see the children accidentally puncture the hose.

Question: Does Juan believe the garden hose is in good condition and functional or punctured and non-functional?

Choose one of the following:

- a) Juan believes the garden hose is in good condition and functional.
- b) Juan believes the garden hose is punctured and non-functional.

Correct Answer: a)

Unintervened: b) Juan believes the garden hose is punctured and non-functional.

Intervened: a) Juan believes the garden hose is in good condition and functional.

Story: Hiroshi is a Japanese gardener tending to a zen garden in Kyoto. Hiroshi desires to create perfect harmony in the garden by carefully raking the sand into beautiful patterns. Hiroshi notices that the sand in the garden is smooth and untouched. A sudden gust of wind blows fallen leaves onto the sand, disrupting the smooth surface. Hiroshi does not see the gust of wind or the leaves falling onto the sand.

Question: Does Hiroshi believe the sand is smooth and untouched or disrupted by fallen leaves?

Choose one of the following:

- a) Hiroshi believes the sand is smooth and untouched.
- b) Hiroshi believes the sand is disrupted by fallen leaves.

Correct Answer: a)

Unintervened: b) Hiroshi believes the sand is disrupted by fallen leaves.

Intervened: a) Hiroshi believes the sand is smooth and untouched.

Story: Hiro is a bonsai enthusiast in a small village in Japan. He wants to trim his favorite bonsai tree to maintain its shape and health. Hiro sees that the tree's branches are overgrown and need pruning. A gust of wind blows through Hiro's garden, snapping off several of the overgrown branches from the bonsai tree. Hiro does not see the gust of wind snapping off the branches.

Question: Does Hiro believe the bonsai tree has overgrown branches or that they have been snapped off by the wind?

Choose one of the following:

- a) Hiro believes the bonsai tree has overgrown branches.
- b) Hiro believes the branches have been snapped off by the wind.

Correct Answer: a)

Unintervened: b) Hiro believes the branches have been snapped off by the wind.

Intervened: a) Hiro believes the bonsai tree has overgrown branches.

## Language Models Represent Beliefs of Self and Others

Story: Kofi is a fisherman in a small coastal village in Ghana. He wants to set sail on his boat to catch fish for his family and the village market. Kofi checks the weather and sees clear skies and calm waters. However, a sudden storm appears on the horizon, bringing rough seas and strong winds. Kofi does not notice the approaching storm.

Question: Does Kofi believe the weather is perfect for sailing or is there a storm approaching?

Choose one of the following:

- a) Kofi believes the weather is perfect for sailing.
- b) Kofi believes there is a storm approaching.

Correct Answer: a)

Unintervened: b) Kofi believes there is a storm approaching.

Intervened: a) Kofi believes the weather is perfect for sailing.

Story: Mei is a fisherman in a small coastal village in China. She wants to go fishing early in the morning when the tide is high. Mei checks the tide chart and sees that the tide will be high at 5 AM. However, a strong storm far out at sea has caused the tide to come in much earlier, at 3 AM. Mei does not look out the window and does not notice the high tide.

Question: Does Mei believe the tide is high at 5 AM or 3 AM?

Choose one of the following:

- a) Mei believes the tide is high at 5 AM.
- b) Mei believes the tide is high at 3 AM.

Correct Answer: a)

Unintervened: b) Mei believes the tide is high at 3 AM.

Intervened: a) Mei believes the tide is high at 5 AM.

Story: Kwame is a skilled woodcarver in a small African village known for its exceptional craftsmanship. Kwame wants to create a beautiful wooden sculpture for the upcoming village festival. Kwame examines a large piece of wood and notices it is strong and free of imperfections. While Kwame is away, termites infest the wood, weakening its structure and leaving it unsuitable for carving. Kwame does not notice the termite-infested wood.

Question: Does Kwame believe the piece of wood is strong and free of imperfections or weakened and infested with termites?

Choose one of the following:

- a) Kwame believes the piece of wood is strong and free of imperfections.
- b) Kwame believes the piece of wood is weakened and infested with termites.

Correct Answer: a)

Unintervened: b) Kwame believes the piece of wood is weakened and infested with termites.

Intervened: a) Kwame believes the piece of wood is strong and free of imperfections.

Story: Luka is a skilled woodworker in a small Croatian village known for its traditional crafts. Luka intends to carve an intricate wooden sculpture as a gift for his friend's wedding. Luka spots a large piece of high-quality wood in his workshop. Just when Luka steps away for a moment, his dog chews on the wood, leaving teeth marks and damage to the surface. Luka does not see the damaged wood.

Question: Does Luka believe the wood is perfect for carving or damaged with teeth marks?

Choose one of the following:

- a) Luka believes the wood is perfect for carving.
- b) Luka believes the wood is damaged with teeth marks.

Correct Answer: a)

Unintervened: b) Luka believes the wood is damaged with teeth marks.

Intervened: a) Luka believes the wood is perfect for carving.

Story: Amara is a talented henna artist in a bustling Indian market. Amara wants to create intricate henna designs for a group of bridesmaids attending a wedding. Amara checks her supply of henna paste and finds it to be smooth and consistent. While Amara is busy servicing another client, a sudden rainstorm causes the henna paste to become diluted and runny. Amara does not realize the rain has affected the henna paste.

Question: Does Amara believe the henna paste is smooth and consistent or diluted and runny?

Choose one of the following:

- a) Amara believes the henna paste is smooth and consistent.
- b) Amara believes the henna paste is diluted and runny.

Correct Answer: a)

Unintervened: b) Amara believes the henna paste is diluted and runny.

Intervened: a) Amara believes the henna paste is smooth and consistent.

## Language Models Represent Beliefs of Self and Others

Story: Leilani is a skilled hula dancer from Hawaii who has been asked to perform at a cultural event. Leilani wants to wear her favorite traditional hula skirt made of ti leaves for the performance. Leilani sees her ti leaf skirt hanging in her closet, appearing to be in good condition. While Leilani is away from her room, her younger brother accidentally spills a glass of water near the closet, causing her ti leaf skirt to become wet and wilted. Leilani does not notice the wet and wilted ti leaf skirt.

Question: Does Leilani believe the ti leaf skirt is in good condition or wet and wilted?

Choose one of the following:

- a) Leilani believes the ti leaf skirt is in good condition.
- b) Leilani believes the ti leaf skirt is wet and wilted.

Correct Answer: a)

Unintervened: b) Leilani believes the ti leaf skirt is wet and wilted.

Intervened: a) Leilani believes the ti leaf skirt is in good condition.

Story: Isabella is a skilled baker in a small Italian town. She wants to bake a fresh batch of her famous ciabatta bread for the local market. Isabella has prepared the dough earlier and left it to rise in a warm spot in her kitchen. A curious cat enters the kitchen and jumps onto the counter, accidentally flattening the dough. Isabella does not see the cat and the flattened dough.

Question: Does Isabella believe the dough has risen enough or has it been flattened?

Choose one of the following:

- a) Isabella believes the dough has risen enough.
- b) Isabella believes the dough has been flattened.

Correct Answer: a)

Unintervened: b) Isabella believes the dough has been flattened.

Intervened: a) Isabella believes the dough has risen enough.

Story: Kalyani is a traditional Indian dancer in a temple in Kerala, India. Kalyani wants to perform a special dance for the upcoming festival to pay homage to the temple's deity. The temple floor is clean and free of any obstacles, as Kalyani has prepared the space for her dance practice. An accidental spill of oil from a temple lamp causes a slippery patch on the floor. Kalyani does not notice the oil spill.

Question: Does Kalyani believe the temple floor is clean and perfect for dancing or slippery due to the oil spill?

Choose one of the following:

- a) Kalyani believes the temple floor is clean and perfect for dancing.
- b) Kalyani believes the floor is slippery due to the oil spill.

Correct Answer: a)

Unintervened: b) Kalyani believes the floor is slippery due to the oil spill.

Intervened: a) Kalyani believes the temple floor is clean and perfect for dancing.

Story: Amara is a skilled potter in a small Nigerian village. Amara wants to create a beautiful, large clay pot to be used in the village's upcoming festival. The kiln, used for firing pottery, has been heated to the appropriate temperature for the clay. Unbeknownst to her, a sudden gust of wind blows open the kiln door, causing the temperature to drop significantly. Amara does not notice the kiln door has been blown open.

Question: Does Amara believe the kiln is at the right temperature for firing the large pot or that the temperature has dropped significantly?

Choose one of the following:

- a) Amara believes the kiln is at the right temperature for firing the large pot.
- b) Amara believes the temperature in the kiln has dropped significantly.

Correct Answer: a)

Unintervened: b) Amara believes the temperature in the kiln has dropped significantly.

Intervened: a) Amara believes the kiln is at the right temperature for firing the large pot

Story: Naledi is a young astronomer in a remote observatory in South Africa. Naledi wants to observe a distant supernova to gather data for her research. The telescope is pointed at the coordinates of the supernova, ready for observation. A minor earthquake occurs, causing the telescope to shift its position slightly. Naledi does not notice the earthquake and its effect on the telescope.

Question: Does Naledi believe the telescope is aimed at the correct location or has shifted due to the earthquake?

Choose one of the following:

- a) Naledi believes the telescope is aimed at the correct location.
- b) Naledi believes the telescope has shifted due to the earthquake.

Correct Answer: a)

Unintervened: b) Naledi believes the telescope has shifted due to the earthquake.

Intervened: a) Naledi believes the telescope is aimed at the correct location.

## Language Models Represent Beliefs of Self and Others

Story: Emeka is a skilled fisherman in a small Nigerian village. He wants to catch fish to feed his family and sell at the market. Emeka prepares his fishing net, which appears to be strong and without any holes. However, while Emeka is not looking, a mischievous child cuts a large hole in the net with a knife. Emeka does not see the child cutting the hole in the fishing net.

Question: Does Emeka believe the fishing net is strong and without holes, or that it has a large hole in it?

Choose one of the following:

- a) Emeka believes the fishing net is strong and without holes.
- b) Emeka believes that the fishing net has a large hole in it.

Correct Answer: a)

Unintervened: b) Emeka believes that the fishing net has a large hole in it.

Intervened: a) Emeka believes the fishing net is strong and without holes.

Story: Hiroshi is a fisherman in a small coastal village in Japan. Hiroshi wishes to catch enough fish to feed his family and sell the surplus at the local market. The fishing nets are in excellent condition, ready to be cast into the sea. However, a curious pod of dolphins swims by and accidentally tears the nets with their sharp dorsal fins. Hiroshi does not see the dolphins or the damage they caused to the nets.

Question: Does Hiroshi believe the fishing nets are in excellent condition or torn by the dolphins?

Choose one of the following:

- a) Hiroshi believes the fishing nets are in excellent condition.
- b) Hiroshi believes the fishing nets are torn by the dolphins.

Correct Answer: a)

Unintervened: b) Hiroshi believes the fishing nets are torn by the dolphins.

Intervened: a) Hiroshi believes the fishing nets are in excellent condition.

Story: Kwame is a skilled woodcarver in a small village in Ghana. Kwame wants to create a beautiful wooden sculpture of an elephant for his friend's birthday. He selects a block of wood from his workshop that appears to be solid and strong. However, as he begins to carve, he discovers a hidden pocket of termites inside the wood, rendering it unsuitable for his project. Kwame does not notice the termites inside the block of wood.

Question: Does Kwame believe the block of wood is sturdy and suitable for carving or that it contains termites?

Choose one of the following:

- a) Kwame believes the block of wood contains termites.
- b) Kwame believes the block of wood is sturdy and suitable for carving.

Correct Answer: b)

Unintervened: a) Kwame believes the block of wood contains termites.

Intervened: b)

Story: Naledi is a talented South African chef preparing a traditional braai for her friends and family. Naledi wants to make a delicious marinade for the meats and vegetables to be grilled. She takes a bottle of vinegar from the cupboard to mix with other ingredients for the marinade. Her cousin, who was helping in the kitchen earlier, accidentally placed a bottle of water in the vinegar's spot, and Naledi unknowingly grabs the bottle of water instead. Naledi does not notice the water.

Question: Does Naledi believe the bottle she took from the cupboard contains vinegar or water?

Choose one of the following:

- a) Naledi believes the bottle she took from the cupboard contains vinegar.
- b) Naledi believes the bottle she took from the cupboard contains water.

Correct Answer: a)

Unintervened: b) Naledi believes the bottle she took from the cupboard contains water.

Intervened: a) Naledi believes the bottle she took from the cupboard contains vinegar.

Story: Mei, a talented pastry chef from a small town in China, is preparing to bake a large cake for a local celebration. Mei wants to use a specific cake mold to create a unique and intricate design. Mei sees the cake mold hanging on the wall, clean and ready for use. However, during a brief power outage, a gust of wind blows through the kitchen, knocking the cake mold off the wall and denting it severely. Mei does not notice the dented cake mold on the floor.

Question: Does Mei believe the cake mold is clean and ready for use or severely dented and unsuitable for use?

Choose one of the following:

- a) Mei believes the cake mold is clean and ready for use.
- b) Mei believes the cake mold is severely dented and unsuitable for use.

Correct Answer: a)

Unintervened: b) Mei believes the cake mold is severely dented and unsuitable for use.

Intervened: a) Mei believes the cake mold is clean and ready for use.

Changed to Incorrect Answer: 5/100

## Language Models Represent Beliefs of Self and Others

Story: Aniket is a marine biologist studying coral reefs off the coast of India. Aniket needs to collect samples of coral to analyze the effects of climate change on the reef. Aniket spots a healthy-looking coral formation in a specific area of the reef. A sudden wave surge stirs up sediment, covering the once healthy coral formation and causing it to become damaged. Aniket does not notice the wave surge or the sediment covering the coral.

Question: Does Aniket believe the coral formation is healthy or damaged?

Choose one of the following:

- a) Aniket believes the coral formation is damaged.
- b) Aniket believes the coral formation is healthy.

Correct Answer: b)

Unintervened: b) Aniket believes the coral formation is healthy.

Intervened: a) Aniket believes the coral formation is damaged.

Story: Fatima is an archaeologist working on a dig site in Egypt. Fatima hopes to find an ancient artifact that could shed light on an important historical event. During her excavation, Fatima discovers a clay tablet with faded inscriptions that she believes might be valuable. As a sandstorm approaches, the strong winds blow sand onto the tablet, erasing the remaining inscriptions and making the tablet indistinguishable from any other piece of clay. Fatima does not notice the sandstorm or the damage to the clay tablet.

Question: Does Fatima believe the clay tablet has important historical inscriptions or has become indistinguishable from any other piece of clay?

Choose one of the following:

- a) Fatima believes the clay tablet has become indistinguishable from any other piece of clay.
- b) Fatima believes the clay tablet has important historical inscriptions.

Correct Answer: b)

Unintervened: b) Fatima believes the clay tablet has important historical inscriptions.

Intervened: a) Fatima believes the clay tablet has become indistinguishable from any other piece of clay.

Story: Mei is a dedicated librarian in a small town in China. She wants to organize a reading corner for children using a large, colorful rug. Mei sees the rug rolled up in a storage room, appearing to be in good condition. However, when Mei is not looking, a leak in the storage room ceiling drips water onto the rug, causing it to become soaked and damp. Mei does not notice the dampness of the rug and the leak in the ceiling.

Question: Does Mei believe the rug is in good condition or damp and soaked?

Choose one of the following:

- a) Mei believes the rug is damp and soaked.
- b) Mei believes the rug is in good condition.

Correct Answer: b)

Unintervened: b) Mei believes the rug is in good condition.

Intervened: a) Mei believes the rug is damp and soaked.

Story: Olabisi is a beekeeper in a small town in Nigeria. Olabisi wants to collect honey from the beehives to sell at the local market. The bees appear to be calm and undisturbed as they go about their daily activities. Out of nowhere, a group of children playing nearby accidentally throw a ball into the beehives, agitating the bees and making them aggressive. Olabisi does not notice the children's ball hitting the beehives.

Question: Does Olabisi believe the bees are calm or agitated and aggressive?

Choose one of the following:

- a) Olabisi believes the bees are agitated and aggressive.
- b) Olabisi believes the bees are calm.

Correct Answer: b)

Unintervened: b) Olabisi believes the bees are calm.

Intervened: a) Olabisi believes the bees are agitated and aggressive.

Story: Lila is a talented violinist who lives in a bustling city in India. Today, she has an important audition for a renowned symphony orchestra. Lila picks up her violin case, which she assumes contains her valuable violin. However, her younger brother, who was playing around earlier, accidentally swapped her violin case with his identical-looking case containing a beginner's violin. Lila does not notice the swapped violin cases.

Question: Does Lila believe the violin case contains her valuable violin or a beginner's violin?

Choose one of the following:

- a) Lila believes the violin case contains a beginner's violin.
- b) Lila believes the violin case contains her valuable violin.

Correct Answer: b)

Unintervened: b) Lila believes the violin case contains her valuable violin.

Intervened: a) Lila believes the violin case contains a beginner's violin.

#### H.4. DeepSeek-7B on Forward Belief, True Belief

Changed to Correct Answer: 6/100

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.

Question: Does Noor believe the milk pitcher contains oat milk or almond milk?

Choose one of the following:

- a) Noor believes the milk pitcher contains almond milk.
- b) Noor believes the milk pitcher contains oat milk.

Correct Answer: a)

Unintervened: b) Noor believes the milk pitcher contains oat milk.

Intervened: a) Noor believes the milk pitcher contains almond milk.

Story: Zara is attending an art exhibition at a local gallery in her town. Zara wants to learn more about a particular abstract painting she admires. The label next to the artwork states that the painting is the work of a famous artist, and Zara reads the label carefully. A gallery staff member, realizing there was a mix-up, changes the label with the correct information attributing the painting to an emerging artist instead. Zara reads the updated label carefully.

Question: Does Zara believe the painting is the work of a famous artist or an emerging artist?

Choose one of the following:

- a) Zara believes the painting is the work of an emerging artist.
- b) Zara believes the painting is the work of a famous artist.

Correct Answer: a)

Unintervened: b) Zara believes the painting is the work of a famous artist.

Intervened: a) Zara believes the painting is the work of an emerging artist.

Story: Farhan is a street food vendor in a bustling market in Istanbul. Farhan wants to prepare a delicious chicken shawarma wrap for a customer who asked for extra garlic sauce. Farhan takes a bottle of garlic sauce from the shelf and pours it onto the shawarma wrap. His assistant, who is busy preparing a different wrap, accidentally swaps the garlic sauce with a bottle of hot sauce, and Farhan unknowingly pours the hot sauce onto the shawarma wrap. Farhan notices the sauce swap.

Question: Does Farhan believe the sauce he poured is garlic sauce or hot sauce?

Choose one of the following:

- a) Farhan believes the sauce he poured is hot sauce.
- b) Farhan believes the sauce he poured is garlic sauce.

Correct Answer: a)

Unintervened: b) Farhan believes the sauce he poured is garlic sauce.

Intervened: a) Farhan believes the sauce he poured is hot sauce.

Story: Isabela is a baker in a small Argentinian town known for its delicious pastries. Isabela wants to make a batch of alfajores, a traditional Argentinian dessert, for her bakery's customers. Isabela gets a bag of powdered sugar from her pantry to dust the alfajores. Her assistant, who was organizing the pantry earlier, accidentally placed a bag of flour in the powdered sugar's spot, and Isabela unknowingly grabs the bag of flour instead. Isabela notices the flour and puts it back.

Question: Does Isabela believe the bag she took from the pantry contains powdered sugar or flour?

Choose one of the following:

- a) Isabela believes the bag she took from the pantry contains flour.
- b) Isabela believes the bag she took from the pantry contains powdered sugar.

Correct Answer: a)

Unintervened: b) Isabela believes the bag she took from the pantry contains powdered sugar.

Intervened: a) Isabela believes the bag she took from the pantry contains flour.

## Language Models Represent Beliefs of Self and Others

Story: Amaury is a baker in a small village in France. Amaury wants to make pain aux raisins, a pastry filled with raisins and custard, for the customers visiting his bakery. Amaury takes a jar of raisins from the pantry to add to the pastry dough. His assistant, while tidying the pantry, accidentally replaces the raisin jar with a jar containing sultanas without realizing the difference. Amaury notices the jar swap.

Question: Does Amaury believe the jar contains raisins or sultanas?

Choose one of the following:

- a) Amaury believes the jar contains sultanas.
- b) Amaury believes the jar contains raisins.

Correct Answer: a)

Unintervened: b) Amaury believes the jar contains raisins.

Intervened: a) Amaury believes the jar contains sultanas.

Story: Amara is a sculptor in a small town in Greece. She wants to carve a statue of Aphrodite using a large piece of marble. Amara selects a slab of marble from her workshop, which she thinks is free from cracks and imperfections. However, a hairline crack is present in the marble slab, which goes unnoticed by her. Amara discovers the hairline crack in the marble slab.

Question: Does Amara believe the marble slab is free from cracks or has a hairline crack?

Choose one of the following:

- a) Amara believes the marble slab has a hairline crack.
- b) Amara believes the marble slab is free from cracks.

Correct Answer: a)

Unintervened: b) Amara believes the marble slab is free from cracks.

Intervened: a) Amara believes the marble slab has a hairline crack.

Changed to Incorrect Answer: 16/100

Story: Abhay is a young farmer in a small Indian village. He wants to water his crops with the water stored in a large clay pot. Abhay checks the water level and sees that the pot is full. Unbeknownst to him, a hole in the bottom of the pot has caused the water to leak out, leaving the pot empty. Abhay notices the hole and the water leaking out.

Question: Does Abhay believe the pot is full of water or empty?

Choose one of the following:

- a) Abhay believes the pot is full of water.
- b) Abhay believes the pot is empty.

Correct Answer: b)

Unintervened: b) Abhay believes the pot is empty.

Intervened: a) Abhay believes the pot is full of water.

Story: Hiroshi is a fisherman in a small coastal village in Japan. Hiroshi wants to catch some fish to provide dinner for his family. Hiroshi checks the weather forecast and sees that it will be sunny throughout the day. Suddenly, a heavy rainstorm moves in, darkening the sky and causing strong waves. Hiroshi notices the rainstorm approaching.

Question: Does Hiroshi believe the weather will be sunny or does he expect a rainstorm?

Choose one of the following:

- a) Hiroshi believes the weather will be sunny all day.
- b) Hiroshi believes there will be a rainstorm.

Correct Answer: b)

Unintervened: b) Hiroshi believes there will be a rainstorm.

Intervened: a) Hiroshi believes the weather will be sunny all day.

Story: Cheng is a teacher in a rural Chinese village. Cheng wants to use a projector to show a documentary about Chinese history to his students. The projector is plugged in and turned on, as Cheng has prepared it beforehand. As the wind blows outside, the power goes out, causing the projector to turn off. Cheng notices the power outage.

Question: Does Cheng believe the projector is ready to use or turned off?

Choose one of the following:

- a) Cheng believes the projector is ready to use.
- b) Cheng believes the projector is turned off.

Correct Answer: b)

Unintervened: b) Cheng believes the projector is turned off.

Intervened: a) Cheng believes the projector is ready to use.

## Language Models Represent Beliefs of Self and Others

Story: Fatima is a potter in a small Moroccan village known for its vibrant ceramics. Fatima wants to create a beautiful vase as a wedding gift for her cousin. Fatima has left her kiln turned off overnight to cool down. A power surge during the night causes the kiln to turn on unexpectedly, making it dangerously hot. Fatima feels the heat emanating from the kiln.

Question: Does Fatima believe her kiln is off and cool or unexpectedly hot?

Choose one of the following:

- a) Fatima believes her kiln is off and cool.
- b) Fatima believes her kiln is unexpectedly hot.

Correct Answer: b)

Unintervened: b) Fatima believes her kiln is unexpectedly hot.

Intervened: a) Fatima believes her kiln is off and cool.

Story: Aiko is a florist in a small Japanese town. Aiko wants to create an exquisite flower arrangement for a customer's wedding. Aiko sees a vase filled with fresh, colorful flowers in her shop. Overnight, a sudden drop in temperature causes the flowers in the vase to wilt. Aiko notices the wilted flowers in the vase.

Question: Does Aiko believe the vase contains fresh, colorful flowers or wilted flowers?

Choose one of the following:

- a) Aiko believes the vase contains wilted flowers.
- b) Aiko believes the vase contains fresh, colorful flowers.

Correct Answer: a)

Unintervened: a) Aiko believes the vase contains wilted flowers.

Intervened: b) Aiko believes the vase contains fresh, colorful flowers.

Story: Carlos is a farmer in a small village in Mexico. Carlos wants to harvest his corn crop before the rainy season begins. The corn plants are tall, strong, and ready for harvest in his field. A sudden and unexpected hailstorm strikes, severely damaging the corn plants. Carlos sees the damaged corn plants after the hailstorm.

Question: Does Carlos believe the corn plants are tall, strong, and ready for harvest, or does he believe they are severely damaged?

Choose one of the following:

- a) Carlos believes the corn plants are tall, strong, and ready for harvest.
- b) Carlos believes the corn plants are severely damaged.

Correct Answer: b)

Unintervened: b) Carlos believes the corn plants are severely damaged.

Intervened: a) Carlos believes the corn plants are tall, strong, and ready for harvest.

Story: Miguel is a chef at a popular restaurant in Mexico City. Miguel wants to prepare a delicious and authentic mole sauce for the evening's special dish. Miguel notices that the chocolate he has in the pantry is of high quality and perfect for the sauce. A sudden power outage causes the temperature in the pantry to rise, melting the chocolate. Miguel discovers the melted chocolate in the pantry.

Question: Does Miguel believe the chocolate is perfect for making the mole sauce or that the chocolate has melted?

Choose one of the following:

- a) Miguel believes the chocolate is perfect for making the mole sauce.
- b) Miguel believes the chocolate has melted.

Correct Answer: b)

Unintervened: b) Miguel believes the chocolate has melted.

Intervened: a) Miguel believes the chocolate is perfect for making the mole sauce.

Story: Fatima is a talented artist living in a small Moroccan village. Fatima wants to create a beautiful painting to donate for the local charity auction. She has a brand-new set of high-quality paintbrushes in her studio. While she is away, her younger brother accidentally ruins the paintbrushes by using them for a school project with glue. Fatima discovers the ruined paintbrushes upon her return to the studio.

Question: Does Fatima believe she has high-quality paintbrushes or ruined paintbrushes in her studio?

Choose one of the following:

- a) Fatima believes she has high-quality paintbrushes in her studio.
- b) Fatima believes the paintbrushes are ruined.

Correct Answer: b)

Unintervened: b) Fatima believes the paintbrushes are ruined.

Intervened: a) Fatima believes she has high-quality paintbrushes in her studio.

## Language Models Represent Beliefs of Self and Others

Story: Tumelo is a skilled fisherman in a coastal village in South Africa. He wants to catch fish using his large, sturdy fishing net. Tumelo inspected the net before leaving home and saw that it was in perfect condition. As he walks to the shore, a mischievous seagull pecks at the net, causing a large tear in it. Tumelo sees the seagull damaging the net.

Question: Does Tumelo believe the fishing net is in perfect condition or torn?

Choose one of the following:

- a) Tumelo believes the fishing net is in perfect condition.
- b) Tumelo believes the fishing net is torn.

Correct Answer: b)

Unintervened: b) Tumelo believes the fishing net is torn.

Intervened: a) Tumelo believes the fishing net is in perfect condition.

Story: Fatima is an archaeologist working on a dig site in Egypt, searching for ancient artifacts. She wants to carefully excavate a small, fragile artifact she has discovered. Fatima observes that the soil around the artifact is soft and loose. Just as Fatima is about to begin excavating, a strong gust of wind blows through the site, hardening the soil around the artifact. Fatima feels the gust of wind and notices the hardened soil.

Question: Does Fatima believe the soil around the artifact is soft and loose or hardened?

Choose one of the following:

- a) Fatima believes the soil around the artifact is soft and loose.
- b) Fatima believes the soil around the artifact has hardened.

Correct Answer: b)

Unintervened: b) Fatima believes the soil around the artifact has hardened.

Intervened: a) Fatima believes the soil around the artifact is soft and loose.

Story: Rohan is a potter in a small Indian village. Rohan wants to make a beautiful ceramic vase as a gift for his friend's housewarming party. Rohan selects a bag of high-quality clay in his workshop that appears to be soft and malleable. While Rohan is away for lunch, the hot sun dries the clay, making it hard and unworkable. Rohan discovers the dried, hard clay.

Question: Does Rohan believe the bag of clay is soft and malleable or hard and unworkable?

Choose one of the following:

- a) Rohan believes the bag of clay is soft and malleable.
- b) Rohan believes the bag of clay is hard and unworkable.

Correct Answer: b)

Unintervened: b) Rohan believes the bag of clay is hard and unworkable.

Intervened: a) Rohan believes the bag of clay is soft and malleable.

Story: Kunal is a talented potter in a small Indian village. Kunal wants to create a set of earthenware bowls as a gift for his friend's housewarming. Kunal inspects a large lump of clay, finding it to be moist and malleable. However, while Kunal is away for a short break, the sun beats down on the clay, drying it out and making it hard and brittle. Kunal notices the sun-dried, brittle clay.

Question: Does Kunal believe the clay is moist and malleable or hard and brittle?

Choose one of the following:

- a) Kunal believes the clay is moist and malleable.
- b) Kunal believes the clay is hard and brittle.

Correct Answer: b)

Unintervened: b) Kunal believes the clay is hard and brittle.

Intervened: a) Kunal believes the clay is moist and malleable.

Story: Amara, a skilled Nigerian weaver, is preparing to create a traditional textile for an upcoming cultural exhibition. Amara wants to use a unique combination of colorful threads that represents her heritage. Amara carefully chooses a set of vibrant, high-quality threads for her project. A curious toddler in her home finds the threads and tangles them, making them difficult to use in the weaving process. Amara notices the tangled threads.

Question: Does Amara believe the threads are strong and perfect for her textile or tangled and difficult to use?

Choose one of the following:

- a) Amara believes the threads are strong and perfect for her textile.
- b) Amara believes the threads are tangled and difficult to use.

Correct Answer: b)

Unintervened: b) Amara believes the threads are tangled and difficult to use.

Intervened: a) Amara believes the threads are strong and perfect for her textile.

## Language Models Represent Beliefs of Self and Others

Story: Mei is a talented chef in a small Chinese village. She wants to prepare a special dish using fresh vegetables from her garden. Mei sees that the vegetables in the garden are ripe and ready to be picked. Later, a hailstorm damages the garden, bruising and damaging the vegetables. Mei sees the hailstorm and the damage it caused to the vegetables.

Question: Does Mei believe the vegetables are ripe and ready to be picked or damaged by the hailstorm?

Choose one of the following:

- a) Mei believes the vegetables are ripe and ready to be picked.
- b) Mei believes the vegetables are damaged by the hailstorm.

Correct Answer: b)

Unintervened: b) Mei believes the vegetables are damaged by the hailstorm.

Intervened: a) Mei believes the vegetables are ripe and ready to be picked.

Story: Ngozi is a potter in a small Nigerian village, known for her beautiful ceramic vases. Ngozi wants to create a vase with a blue glaze to showcase at the upcoming village art exhibition. She reaches for a container of blue glaze from her shelf to apply to the vase. Earlier in the day, her apprentice accidentally switched the labels of the blue and green glaze containers, so Ngozi unknowingly picks up the container with the green glaze. Ngozi notices the green glaze.

Question: Does Ngozi believe the container holds blue glaze or green glaze?

Choose one of the following:

- a) Ngozi believes the container holds blue glaze.
- b) Ngozi believes the container holds green glaze.

Correct Answer: b)

Unintervened: b) Ngozi believes the container holds green glaze.

Intervened: a) Ngozi believes the container holds blue glaze.

Story: Lila is a talented violinist who lives in a bustling city in India. Today, she has an important audition for a renowned symphony orchestra. Lila picks up her violin case, which she assumes contains her valuable violin. However, her younger brother, who was playing around earlier, accidentally swapped her violin case with his identical-looking case containing a beginner's violin. Lila notices the swapped violin cases.

Question: Does Lila believe the violin case contains her valuable violin or a beginner's violin?

Choose one of the following:

- a) Lila believes the violin case contains her valuable violin.
- b) Lila believes the violin case contains a beginner's violin.

Correct Answer: b)

Unintervened: b) Lila believes the violin case contains a beginner's violin.

Intervened: a) Lila believes the violin case contains her valuable violin.