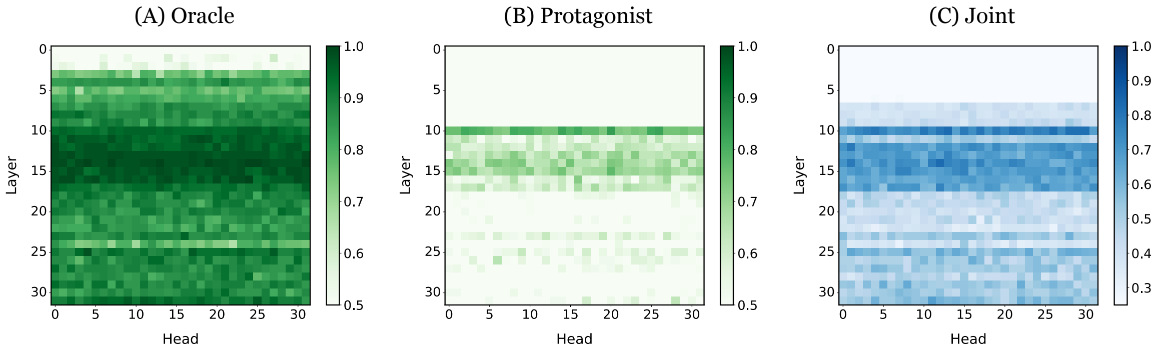

Figure 2. Probe accuracies based on the attention head activations in all layers of Mistral-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status estimation for both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

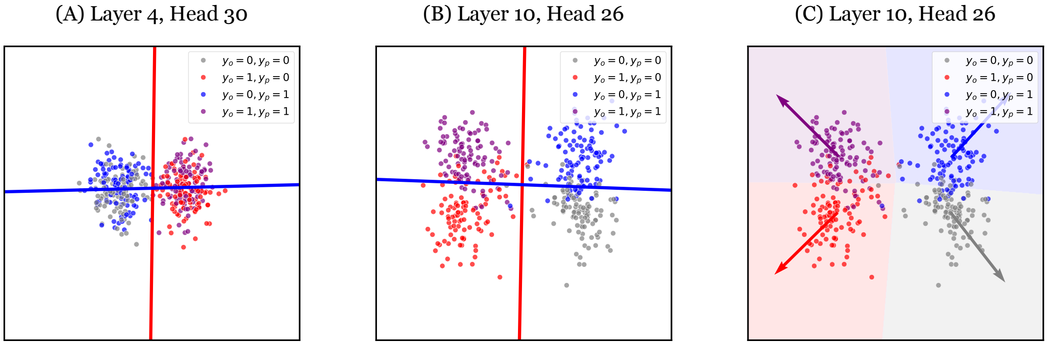

Figure 3. Illustration of linear separability of the belief representations. We show visual explanations for typical representation spaces: In (A), *oracle* belief status can be precisely estimated with a linear model, while *protagonist* cannot. The red and blue lines represent linear decision boundaries 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 $\boldsymbol{y}_m \in \{0,1\}^{N\times 4}$. A multinomial logistic regression model is utilized to predict the joint belief $\hat{\boldsymbol{y}_m}$. Specifically, the class probabilities are derived by applying the softmax function to the linear transformations of $\boldsymbol{x}$:

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

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

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

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

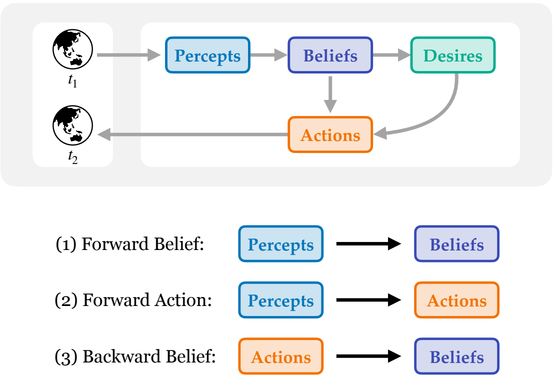

Figure 4. Different social reasoning tasks and the underlying causal graph.

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: $\sum_{\text{Percept}} \sum_{\text{Belief}} P(\text{Action} \mid \text{Desire}, \text{Percept}, \text{Belief})$.

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	**0.91**	0.22	0.16
Mistral-7B (+ random)	**0.97**	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	**0.93**	0.29	0.25	0.50	0.37	0.26
Mistral-7B (+ $\mathrm{T}_p\mathrm{F}_o$)	0.85	**0.66**	**0.58**	0.88	**0.41**	**0.31**	0.61	**0.44**	**0.41**
DeepSeek-7B (baseline)	**0.73**	0.47	0.37	**0.77**	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	**0.65**	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	**0.44**	0.73	0.57	0.34	0.65	0.48	0.28
DeepSeek-7B (+ $\mathrm{T}_p\mathrm{F}_o$)	0.63	**0.74**	0.38	0.75	**0.60**	**0.39**	0.54	**0.66**	**0.31**

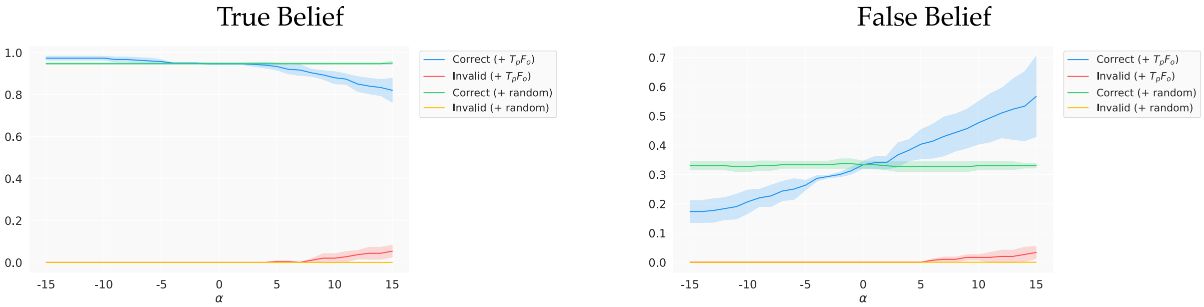

Figure 5. Impact of varying intervention strength α on the *Forward Belief* task using Mistral-7B. “Invalid” denotes the answer is not recognized by the grading mechanism as the model fails to provide answer in the required format, *e.g.*, delivering uncertain responses.

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

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

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 α is the step length of intervention, σ_l^h is the standard deviation of activations along the target direction, θ_l^h is the target direction. We set K and α 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,