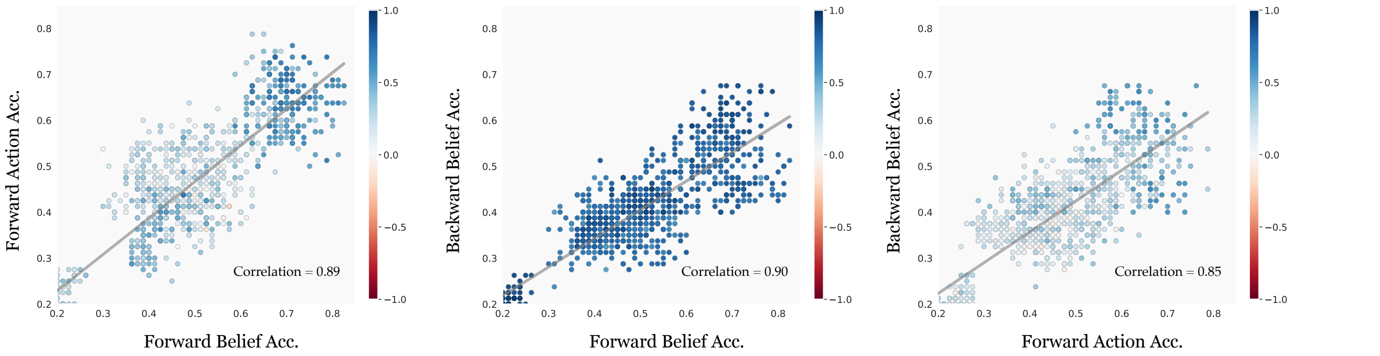

Figure 6. Pairwise comparisons of multinomial probing results on Mistral-7B. Each point represents a specific attention head. The point position denotes its probe accuracies in the two tasks, and point color denotes the cosine similarity between the (+ T_pF_o) probe directions of the two tasks.

Prompt:

... 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.

Answer:

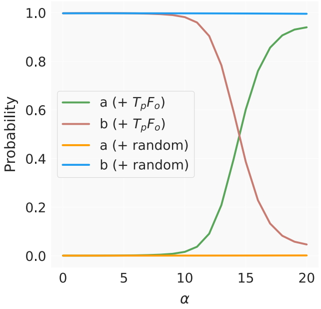

Figure 7. (Left) Question prompt of the *False Belief* condition for the *Forward Belief* task. The story background is omitted for simplicity. (Right) Changes of next-token probability with regard to different intervention strength α on Mistral-7B.

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_pF_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 perform intervention based on their respective probing results. The (+ 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 (- 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_pF_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_pF_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

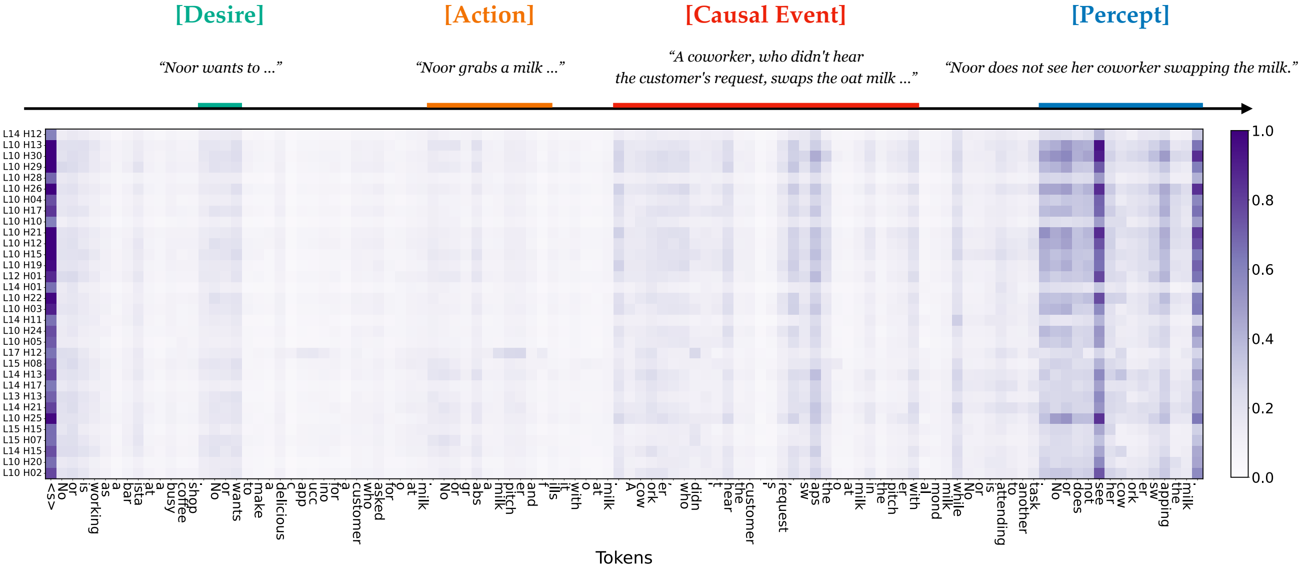

Figure 8. Magnitudes of gradients on the token embeddings with respect to the projection of attention head activations over the corresponding joint belief directions. Each line represents a specific attention head in Mistral-7B. We highlight the prominent segments and the corresponding causal variables.

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	**0.91**	0.22	0.16
+ random	**0.93**	0.31	0.26	0.90	0.22	0.16
+ transferred	0.90	**0.40**	**0.33**	0.73	**0.38**	**0.22**
DeepSeek-7B	**0.77**	0.48	0.31	0.64	0.50	0.29
+ random	0.76	0.42	0.29	**0.65**	0.47	**0.32**
+ transferred	0.75	**0.57**	**0.37**	0.57	**0.57**	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 attention 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 (+ $\mathrm{T}_p\mathrm{F}_o$) 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 (+ $\mathrm{T}_p\mathrm{F}_o$) 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 variables 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.