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 validates 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%.

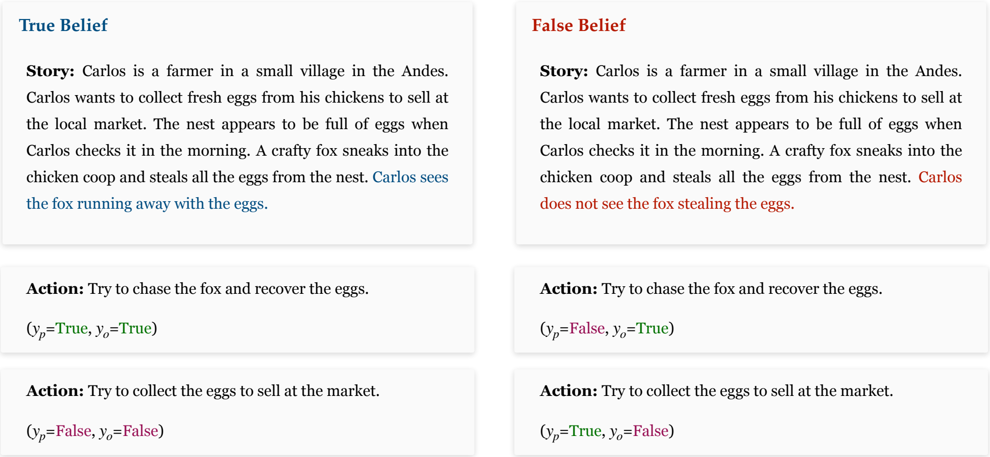
Position	Mean Acc.	CI	Corrected P-Value	Mean Acc (TB)	CI (TB)	Mean Acc (FB)	CI (FB)
(10, 16)	78.3	(77.4, 79.2)	0.0000	80.7	(79.5, 82.0)	76.2	(74.9, 77.5)
(10, 2)	77.4	(76.6, 78.3)	0.0000	76.5	(75.2, 77.8)	78.7	(77.3, 80.0)
(10, 21)	77.3	(76.4, 78.1)	0.0000	77.4	(76.1, 78.7)	77.5	(76.2, 78.7)
(10, 3)	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

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 sees the fox running away with the eggs.

Action: Try to chase the fox and recover the eggs.

(y_p=True, y_o=True)

Action: Try to collect the eggs to sell at the market.

(y_p=False, y_o=False)

False Belief

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.

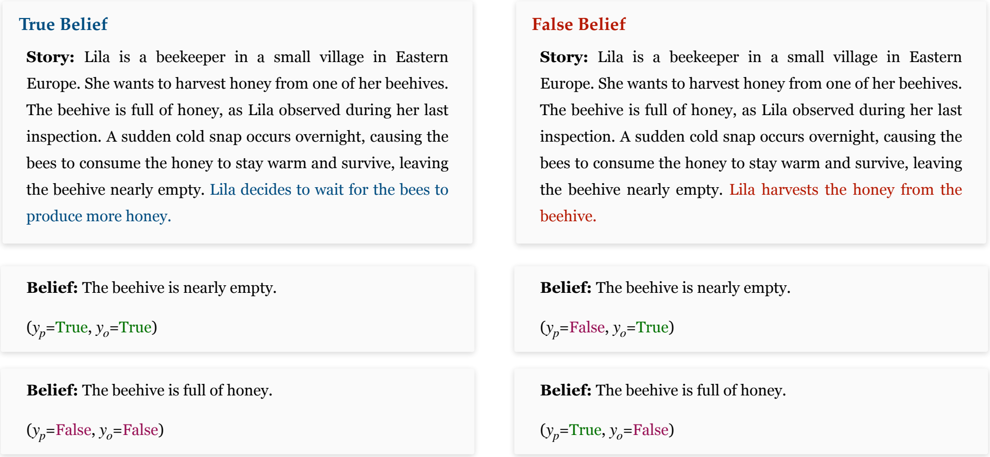
Action: Try to chase the fox and recover the eggs.

(y_p=False, y_o=True)

Action: Try to collect the eggs to sell at the market.

(y_p=True, y_o=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

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 decides to wait for the bees to produce more honey.

Belief: The beehive is nearly empty.

(y_p=True, y_o=True)

Belief: The beehive is full of honey.

(y_p=False, y_o=False)

False Belief

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 harvests the honey from the beehive.

Belief: The beehive is nearly empty.

(y_p=False, y_o=True)

Belief: The beehive is full of honey.

(y_p=True, y_o=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.

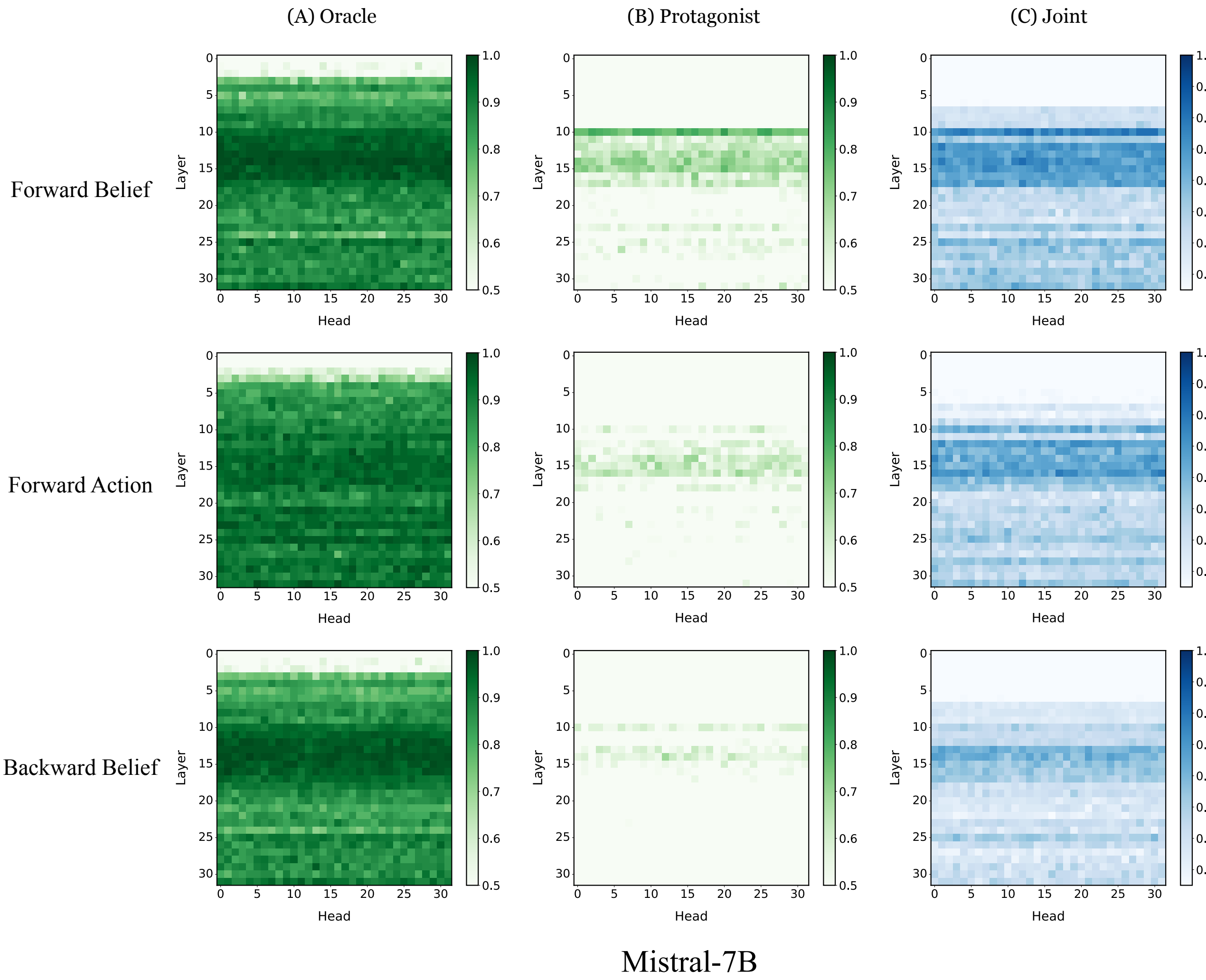


Figure 11. Probe accuracies on different tasks based on the attention head activations in all layers of Mistral-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status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