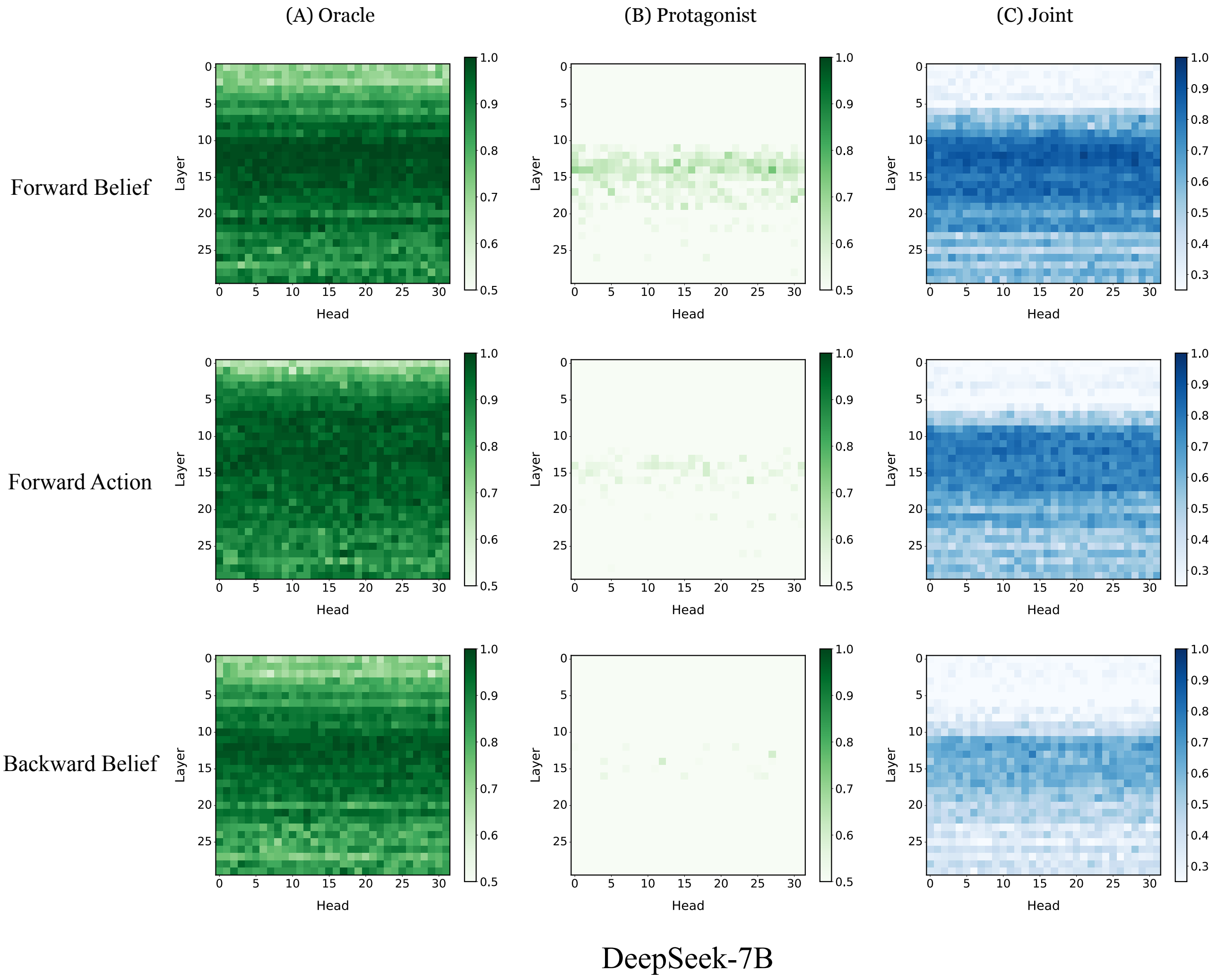

Figure 12. Probe accuracies on different tasks based on the attention head activations in all layers of DeepSeek-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status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

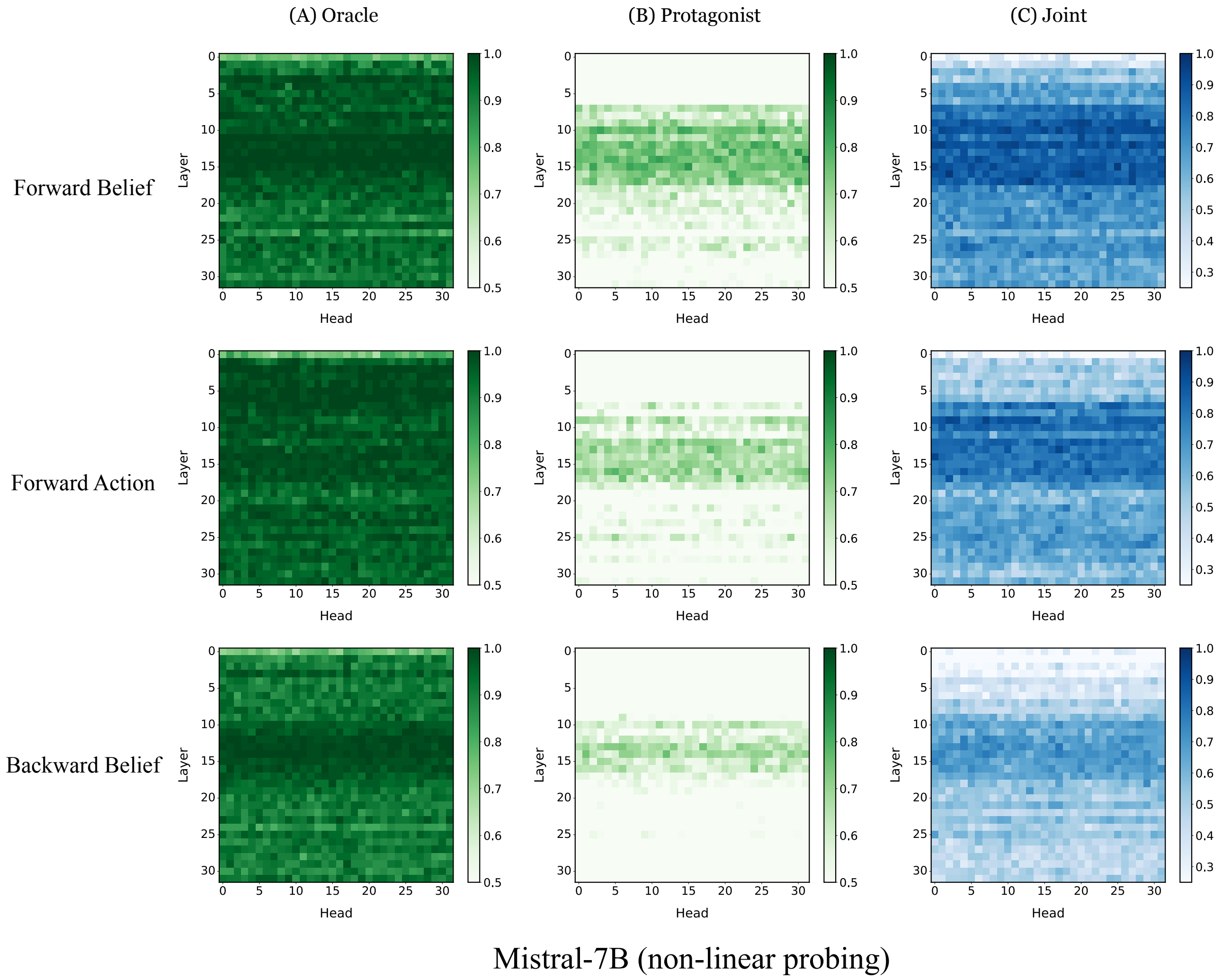

Figure 13. Non-linear probe accuracies on different tasks based on the attention head activations in all layers of Mistral-7B. (A) Belief status estimation for *oracle* using binary classification. (B) Belief status estimation for *protagonist* using binary classification. (C) Joint belief status estimation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

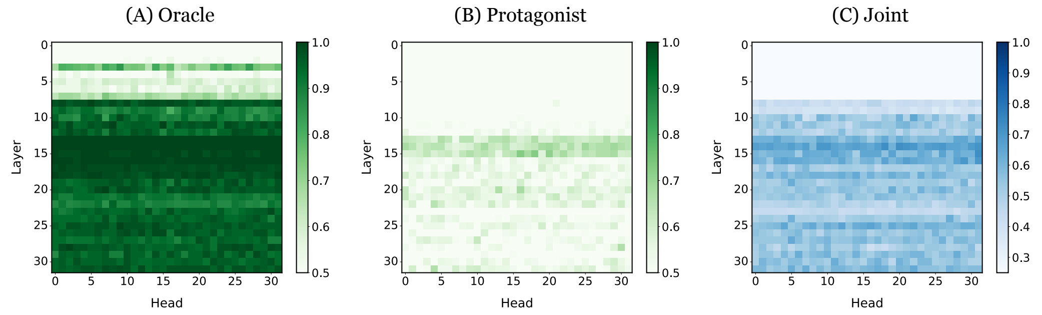

Figure 14. Probe accuracies on ToMi dataset based on the attention head activations in all layers of DeepSeek-7B. (A) Belief status estimation for *oracle* using logistic regression (binary). (B) Belief status estimation for *protagonist* using logistic regression (binary). (C) Joint belief status estimation 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_pF_o, transferred from BigToM)	**71.0**	**67.9**

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.