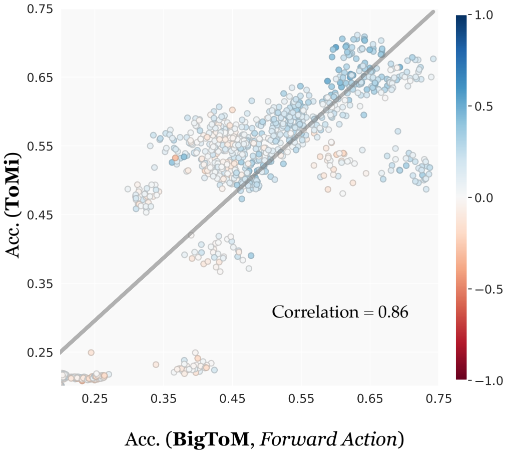


Figure 15. Pairwise comparisons of multinomial probing results on Mistral-7B between ToMi and BigToM. Each point represents a specific attention head, with its position denoting the probe accuracies in the two datasets (multinomial), and its color denoting the cosine similarity between the $(+\mathrm{T}_p\mathrm{F}_o)$ 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	**66.4**	48.3	65.5
Mistral-7B (+random)	57.6	52.7	66.3	48.4	**65.6**
Mistral-7B (+$\mathrm{T}_p\mathrm{F}_o$)	**57.7**	**53.0**	66.3	**48.7**	65.5

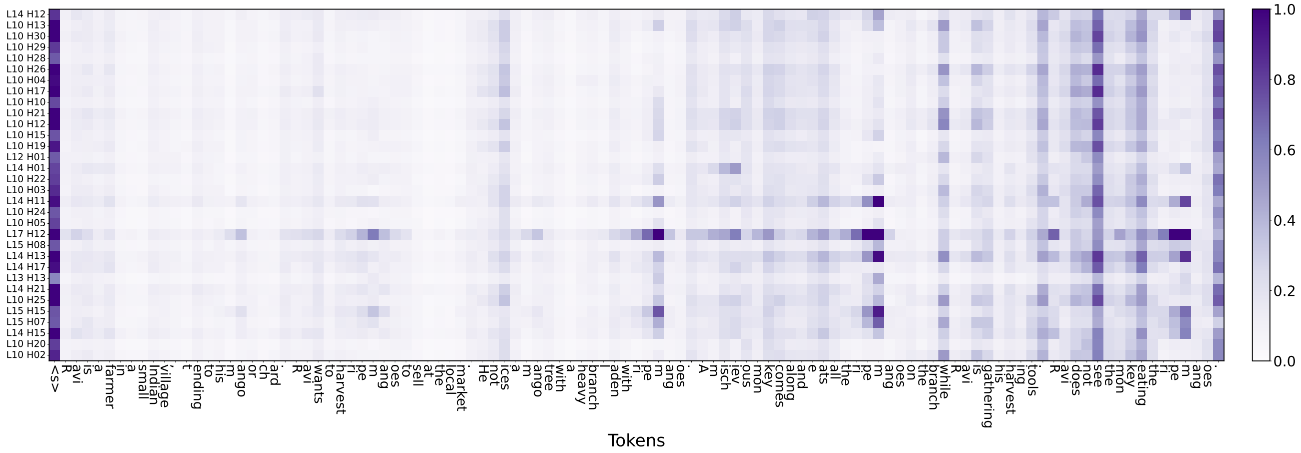
Table 8. Model performance comparison on CoLA, MRPC, and QNLI benchmarks.

Dataset	CoLA	MRPC	QNLI
Mistral-7B (baseline)	70.0	58.8	62.5
Mistral-7B (+random)	70.1	**59.2**	**63.1**
Mistral-7B (+$\mathrm{T}_p\mathrm{F}_o$)	**71.1**	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 acquicision, 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.



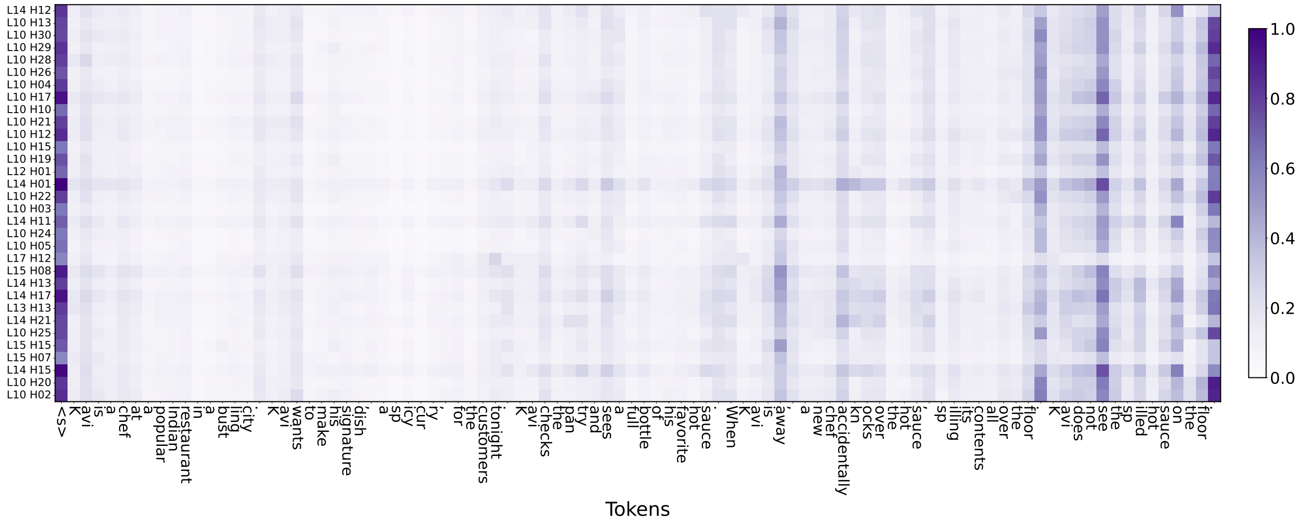


Figure 16. Magnitudes of gradients on the token embeddings with respect to the projection of attention head activations over the corresponding joint belief directions. Each line represents a specific attention head in Mistral-7B.

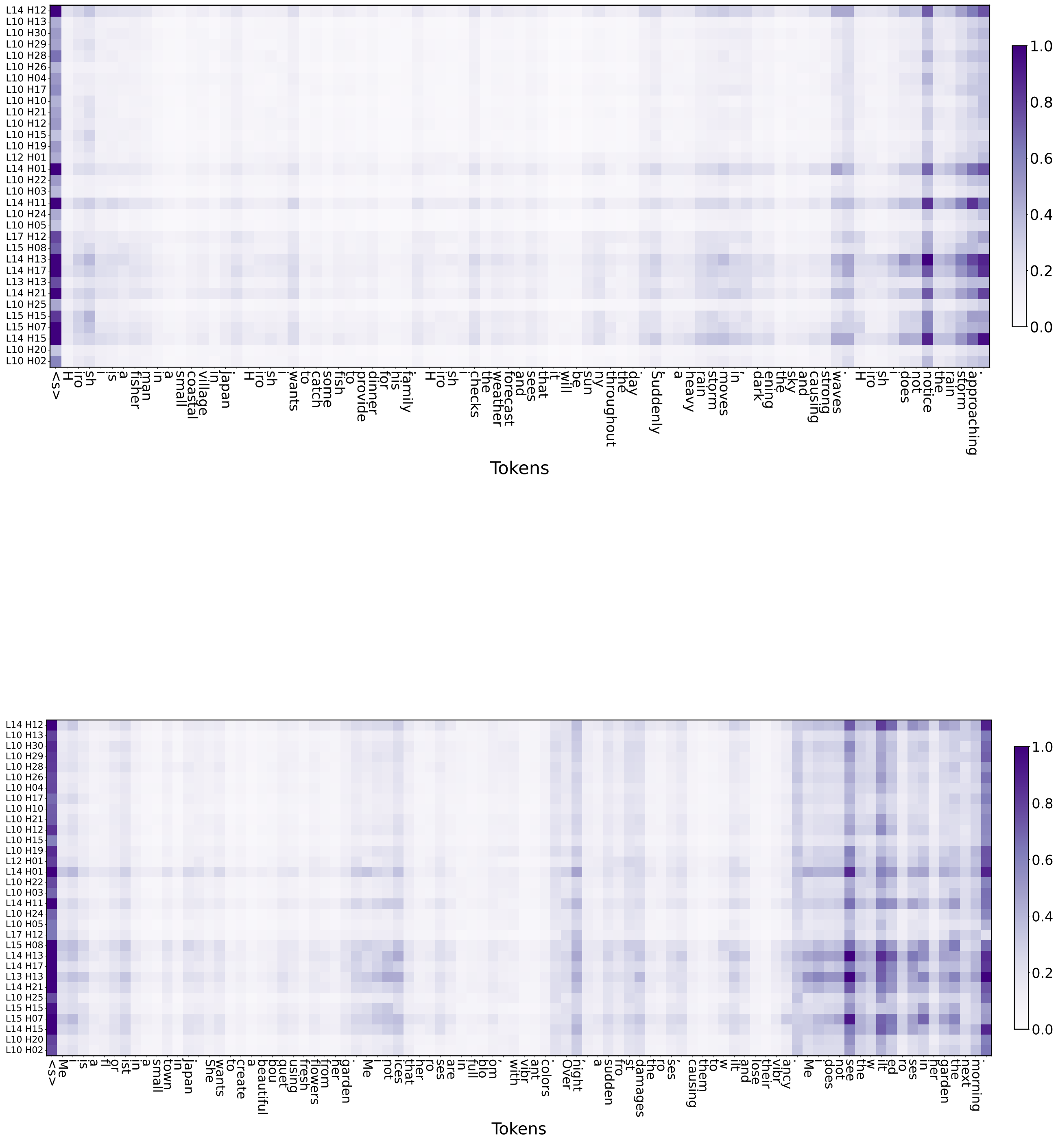


Figure 17. Magnitudes of gradients on the token embeddings with respect to the projection of attention head activations over the corresponding joint belief directions. Each line represents a specific attention head in Mistral-7B.