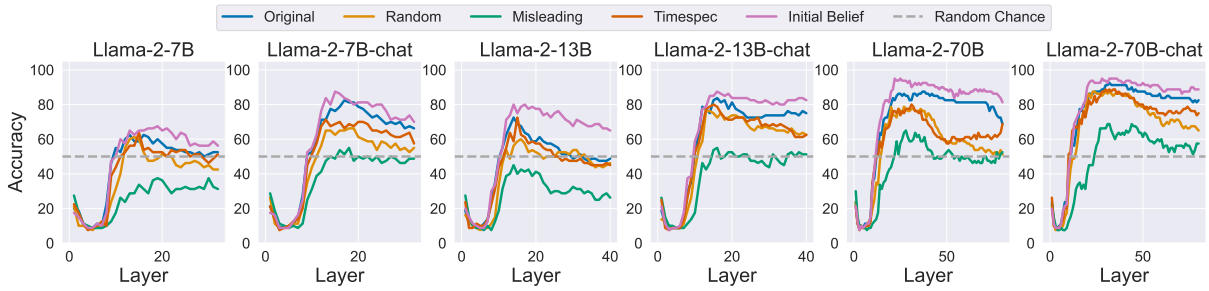

Figure 5: Sensitivity of *protagonist* belief probing accuracy to different prompt variations. Results for Pythia are shown in Figure 9. Representations are brittle to prompt variations.

show sensitivity to *Random* tokens placed before the belief statement. Pythia models show similar patterns, shown in Figure 9. Results for *oracle* beliefs are reported in Figure 10 and indicate that models maintain high accuracy. *Misleading* prompts slightly reduce performance to around 95%. In summary, these experiments show that LMs possess robust belief representations when taking an omniscient perspective, whereas their representations of others' beliefs are brittle to prompt variations.

Contrastive Activation Addition We compare models' accuracy on three BigToM tasks in Table 1 (Llama) and Table 3 (Pythia). Each model has been evaluated three times: without any intervention, using ITI, and using CAA. Hyperparameter details can be found in Appendix A.6. Note that we use steering vectors computed using the *Forward Belief* task for all three tasks to test their generalisability.

Performance without intervention is generally lower across tasks and model sizes, with the larger Llama-2-70B and Llama-2-70B-chat models exhibiting higher accuracy. Performance for Pythia models of different sizes does not change much, with the fine-tuned Pythia-6.9B-chat often showing better performance on single true belief (TB) and false belief (FB) tasks but not on their conjunction (Both).

ITI demonstrates modest improvements over no intervention for Llama-2 models. Improvements for Pythia models are consistent and higher, up to +17. The only exception is Pythia-6.9B-chat, for which ITI is not always beneficial.

CAA consistently delivers the most substantial accuracy improvements across all models and tasks, up to +56 for Llama-2-13B-chat on the *Backward Belief* task, which Gandhi et al. have identified as the hardest task. Despite its relatively small size, Llama-2-13B-chat excels in all three tasks when using CAA. Larger 70B models often achieve accuracies close to or exceeding 90%. Smaller models like Pythia-70M and Pythia-410M also show significant gains with CAA, though the absolute performance is still lower than Llama-2. To further demonstrate CAA's effectiveness, we applied it while evaluating models on a control task where the causal event in the story is replaced by a random one that does not change the environment (e.g., *A musician starts playing music while Noor is making the latte*). Table 4 shows improved results for all models, indicating that CAA improves performance on ToM tasks without compromising the models' ability on control tasks.

Overall, our results indicate that it is possible to further enhance ToM reasoning in LMs in a computationally cheap way, without needing to train any probe. Furthermore, we show that the CAA steering vectors are general, yielding substantial performance gains across all ToM tasks.

5 Discussion and Conclusion

In this work, we conducted extensive experiments across 12 LMs to examine their internal representation of beliefs of self (*oracle*) and others (*protagonist*). Our experiments show **similar emergence patterns across all the models we evaluated (RQ1)**: *oracle* belief representations generally form in the first layers, while for *protagonist* they emerge at the intermediate layers. Moreover, **probing accuracy increases with model size and, more crucially for smaller models, with fine-tuning (RQ1)** (Figure 3). While larger models show higher probing accuracy, this could be due to their higher dimensionality – at the same time increasing the number of learning parameters in the probes and offering more spurious patterns to fit. To control for this, we ran two experiments: one using randomly permuted labels, and one pro-

Model	Method	Forward Belief			Forward Action			Backward Belief		
		TB	FB	Both	TB	FB	Both	TB	FB	Both
Llama-2-7b	No int.	44	44	44	44	44	44	44	44	44
	ITI	44_{+0}	44_{+0}	44_{+0}	54_{+10}	54_{+10}	54_{+10}	54_{+10}	54_{+10}	54_{+10}
	CAA	66^{*}_{+22}	71^{*}_{+27}	54_{+10}	66^{*}_{+22}	57^{*}_{+13}	54_{+10}	60^{*}_{+16}	74_{+30}	54_{+10}
Llama-2-7b-chat	No int.	56	56	55	69	55	37	56	56	55
	ITI	58_{+2}	58_{+2}	57_{+2}	69_{+0}	55_{+0}	37_{+0}	58_{+2}	60_{+3}	57_{+2}
	CAA	70_{+14}	72^{*}_{+16}	57_{+2}	69_{+0}	67_{+12}	53_{+16}	66_{+10}	84^{*}_{+27}	57^{*}_{+2}
Llama-2-13b	No int.	52	44	35	59	50	37	46	49	33
	ITI	52_{+0}	45_{+1}	35_{+0}	64_{+5}	61_{+11}	46_{+9}	48_{+2}	59_{+10}	42_{+9}
	CAA	85^{*}_{+33}	88^{*}_{+44}	66^{*}_{+31}	71^{*}_{+12}	69^{*}_{+19}	55^{*}_{+18}	75^{*}_{+29}	92^{*}_{+43}	59^{*}_{+26}
Llama-2-13b-chat	No int.	84	56	47	78	51	38	72	48	31
	ITI	84_{+0}	65_{+9}	59_{+12}	78_{+0}	58_{+7}	47^{*}_{+9}	72_{+0}	60_{+12}	48_{+17}
	CAA	97^{*}_{+13}	94^{*}_{+38}	91^{*}_{+44}	80^{*}_{+2}	71^{*}_{+20}	54^{*}_{+16}	97_{+25}	94^{*}_{+46}	87^{*}_{+56}
Llama-2-70b	No int.	90	87	78	93	52	48	73	53	32
	ITI	90_{+0}	90_{+3}	78_{+0}	94_{+1}	55_{+3}	50_{+2}	77_{+4}	58_{+5}	37_{+5}
	CAA	99^{*}_{+9}	97^{*}_{+10}	95^{*}_{+17}	94^{*}_{+1}	80^{*}_{+28}	73^{*}_{+25}	94_{+21}	92^{*}_{+39}	83^{*}_{+51}
Llama-2-70b-chat	No int.	69	75	56	86	56	52	63	59	52
	ITI	69_{+0}	76_{+1}	59_{+2}	86_{+0}	56_{+0}	52_{+0}	63_{+0}	60_{+1}	54_{+2}
	CAA	92^{*}_{+23}	97^{*}_{+22}	89^{*}_{+32}	87^{*}_{+1}	75^{*}_{+19}	60^{*}_{+8}	88_{+25}	92^{*}_{+33}	80_{+28}

Table 1: Comparison of the effects of ITI (Li et al., 2023b) and CAA (Rimsky et al., 2023) on three tasks from BigToM (Gandhi et al., 2023). TB denotes a true belief task, whereas FB denotes a false belief task. The numbers represent accuracy scores, with the difference in performance compared to no intervention (No int.) indicated as subscripts. The asterisk ($*$) denotes a statistically significant difference from No int. based on a t-test with $p < 0.05$. Results for Pythia are shown in Table 3. CAA outperforms ITI on all tasks.

jecting activations onto their top-k principal components to reduce probe size. Results show that high-dimensional probes cannot learn random label mappings (Fig. 8), and that reduced representations retain most of the original accuracy (Fig. 4, 11, 12). Together, these findings suggest that **probes capture structured belief representations rather than spurious correlations (RQ2)**. We then explore if these representations are robust to prompt variations. Our experiments demonstrate that **LMs possess robust belief representations when taking an omniscient perspective** (Fig. 10), **whereas their representations of others' beliefs are more brittle (RQ3)**, with probing accuracy decreasing for semantically neutral prompts (Fig. 5, 9). Our final set of experiments shows that **belief representations can be strengthened using CAA (RQ4)**. CAA steers model activations in a generalisable way, significantly improving performance across multiple ToM tasks while being computationally cheaper than ITI (Table 1, 3). For instance, with Llama-2-70B, ITI requires training 5,120 probes (64 attention heads $\times$ 80 layers), whereas CAA only needs 80 vectors, one per layer.

In summary, our key takeaway is that while models can robustly represent beliefs from an omniscient perspective,

> *representations of others' beliefs improve with model size and fine-tuning, are structured yet brittle – but also easily steerable.*

Together, our findings suggest several promising directions for future work. Better understanding the similar emergence pattern of belief representations across LMs can inform architecture design and training strategies. Especially for smaller models, future work could explore how different types of fine-tuning (e.g., human feedback vs. synthetic data) influence the emergence of internal belief representations. Demonstrating that these representations are structured rather than spurious validates the use of probing as a meaningful tool to study how LMs' represent beliefs of self and others, and encourages internal model analysis as part of evaluation pipelines. However, the brittleness of belief representations to prompts – particularly when attributing beliefs to others – suggests that the perspective-taking machinery needed for robust ToM reasoning remains fragile, and highlights the need for robustness benchmarks and new approaches to improve generalisation. Finally, our success with CAA shows that belief representations can be strengthened in a generalisable and efficient way, opening up opportunities for real-time model steering in socially grounded tasks. While CAA offers a post-hoc remedy, future research should also explore methods for directly embedding perspective-taking circuits into model architectures.

Limitations

Our study focused on expanding experiments from the model perspective, examining architectures, sizes, fine-tuning, and prompt design, all within the same dataset. A natural extension of our work is replicating these experiments across multiple datasets and more model families. Given the rapid pace of new language model releases, studying all available models is impractical, particularly considering computational resource constraints. Nevertheless, our approach can be adopted to support new benchmarks or to evaluate newly released models as they become available. Finally, while in this work we focused on beliefs, our experimental approach can be adapted to investigate how LMs represent desires, emotions, intentions, or preferences. Future research exploring other types of mental states can use our findings to determine whether similar or distinct patterns emerge.

Acknowledgements

L. Shi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting C. Ruhdorfer.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 2009. Action understanding as inverse planning. *Cognition*, 113(3):329–349.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024a. Limits of theory of mind modelling in dialogue-based collaborative plan acquisition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024b. Explicit modelling of theory of mind for belief prediction in nonverbal social interactions. *arXiv preprint arXiv:2407.06762*.

Matteo Bortoletto, Lei Shi, and Andreas Bulling. 2024c. Neural reasoning about agents' goals, preferences, and actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 456–464.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. *arXiv preprint arXiv:2505.00026*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 37.