Brittle Minds, Fixable Activations: Understanding Belief Representations in Language Models

Matteo Bortoletto **Constantin Ruhdorfer** **Lei Shi** **Andreas Bulling**
University of Stuttgart, Germany
matteo.bortoletto@vis.uni-stuttgart.de

Abstract

Despite growing interest in Theory of Mind (ToM) tasks for evaluating language models (LMs), little is known about how LMs *internally represent mental states* of self and others. Understanding these internal mechanisms is critical – not only to move beyond surface-level performance, but also for model alignment and safety, where subtle misattributions of mental states may go undetected in generated outputs. In this work, we present the first systematic investigation of belief representations in LMs by probing models across different scales, training regimens, and prompts – using control tasks to rule out confounds. Our experiments provide evidence that both model size and fine-tuning substantially improve LMs' internal representations of others' beliefs, which are structured – not mere by-products of spurious correlations – yet brittle to prompt variations. Crucially, we show that these representations can be strengthened: targeted edits to model activations can correct wrong ToM inferences.

1 Introduction

Language models (LMs) trained on next token prediction have demonstrated impressive capabilities across various tasks, spanning coding, math, and embodied interaction (Wei et al., 2022; Bubeck et al., 2023). As these models are designed with the ultimate goal of collaborating with humans, it becomes imperative that they complement these skills with an understanding of humans. Core to this understanding is *Theory of Mind* (ToM) – the ability to attribute mental states to oneself and others (Premack and Woodruff, 1978). ToM is essential for effective communication and cooperation with other agents, facilitating interaction and learning from feedback and demonstrations (Saha et al., 2023). Given its significance, computational ToM has emerged as a key capability when evaluating cutting-edge LMs (Ma et al., 2023; Shapira et al., 2024; Chen et al., 2025).

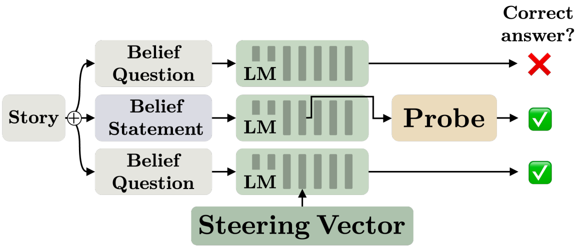


Figure 1: ToM tasks are challenging for LMs, but correct predictions can sometimes be recovered by *probing* their internal representations. We study how internal representations of beliefs of self and others emerge in 12 LMs, and show that these representations are structured yet brittle to prompts, and can be strengthened with a steering vector to fix incorrect ToM inferences.

Despite the improved performance on ToM benchmarks compared to earlier models, modern LMs are still far from perfect (Sap et al., 2022). Text generated by LMs often contains errors that limit their performance on ToM tasks. Zhu et al. (2024) showed that *probing* LMs' internal representations can sometimes recover correct belief inferences, with models like Mistral-7B-Instruct (Jiang et al., 2023) and DeepSeek-7B-Chat (Bi et al., 2024) capturing beliefs from both their own and others' perspectives. While promising, this remains a preliminary step: it examines only single-sized, fine-tuned models, leaves possible confounds uncontrolled, and ignores how subtle changes in prompting affect belief representations. As a result, we still lack a clear understanding of how internal belief representations differ across models, whether they reflect true ToM or spurious patterns, and how robust they are to prompts.

To address these gaps, we pose four key research questions and present evidence for each. We begin by studying how emergence scales across models:

RQ1 *Do internal belief representations* ***emerge*** *similarly in different LMs, and are they* ***affected*** *by model size and training regime?*

Finding training regimes or scales that are more conducive to belief reasoning can guide future model development toward more reliable ToM behaviour. However, it is also crucial to verify if representations are structured, indicating genuine modelling of mental states, or spurious:

RQ2 *Are LMs' internal belief representations* ***structured*** *or the result of spurious correlations?*

This distinction is essential for determining if representations reflect a genuine understanding of beliefs or only exploit statistical patterns that happen to correlate with correct answers in the training data. This is also crucial for alignment and safety, as misaligned mental state attributions may not appear overtly in text – leading to false signals of understanding. Equally important is that models can maintain robust belief attributions:

RQ3 *Are LMs' internal belief representations* ***robust****?*

Fragile representations may break under slight variations, leading to inconsistent or unsafe behaviour in real-world applications involving social reasoning or user interaction. Strengthening these representations, then, offers a promising path toward improving their reliability:

RQ4 *Can we* ***strengthen*** *LMs' internal belief representations to improve their performance?*

To answer these research questions, we perform probing and activation editing experiments using **12 LMs** (Figure 1). We first compare base models with those fine-tuned via SFT and/or RLHF (Ouyang et al., 2022)(**RQ1**), finding that belief representations emerge in consistent patterns across models, improve with model size, and – especially in smaller models – benefit significantly from fine-tuning. To provide evidence that LMs' belief representations are structured (**RQ2**), we show that (1) probes trained on randomly permuted labels perform at chance – confirming selectivity, and (2) probes trained on top-k principal components still recover most accuracy for $k \ll d_{model}$. Next, we test robustness (**RQ3**) using varied prompts. Surprisingly, semantically neutral changes can reduce accuracy, revealing that representations of others' beliefs are brittle to prompts. However, we show that it is possible to strengthen models' representation by using contrastive activation addition (Rimsky et al., 2023, CAA), obtaining significant performance improvements across different ToM tasks (**RQ4**).

In summary, our work makes the following contributions:

1. We provide extensive probing experiments across 12 LMs, suggesting that representations of others' beliefs improve with size and fine-tuning, and that these representations are structured yet brittle to prompt variations.
2. We show that we can strengthen models' representations by using contrastive activation addition and improve their ToM performance.

2 Related Work

Machine Theory of Mind Theory of mind has been studied in AI for more than a decade (Baker et al., 2009; Rabinowitz et al., 2018; Bara et al., 2021; Bortoletto et al., 2024a,b,c). Various benchmarks have been proposed, aiming to measure LMs' ability to understand and reason about the beliefs, goals, and intentions of others (Le et al., 2019; He et al., 2023; Kim et al., 2023; Gandhi et al., 2023; Xu et al., 2024; Tan et al., 2024; Sclar et al., 2023; Ma et al., 2023; Wu et al., 2023). Additionally, efforts have been made to enhance LMs' ToM through prompting techniques (Zhou et al., 2023b; Moghaddam and Honey, 2023; Wilf et al., 2023). Our work dives deeper into LMs' internal belief representations, offering a broader insight into these mechanisms that go beyond surface-level performance.

Probing Neural Representations Initially proposed by Alain and Bengio (2017), probing is a widely used method for determining if models represent particular features or concepts. In the realm of LMs, numerous works used probing to demonstrate that these models acquire rich linguistic representations – spanning semantic concepts such as syntactic categories, dependency relations, coreference, and word meaning (Conneau et al., 2018; Tenney et al., 2018, 2019; Rogers et al., 2021; Li et al., 2021; Hernandez and Andreas, 2021; Marks and Tegmark, 2023; Liu et al., 2023). A separate line of work explored if LMs possess a *world model* (Li et al., 2021; Abdou et al., 2021; Patel and Pavlick, 2022; Li et al., 2023a; Nanda et al., 2023). An emergent line of work that is relevant to our work used probing to explore if LMs have *agent models*, for example, if they can represent beliefs of self and others (Zhu et al., 2024; Bortoletto et al., 2024a). In this work, we contribute with extensive experiments that characterise models' representations of beliefs along different axes: emergence, structure, robustness, and steerability.

Prompt Analysis Previous work has shown that LMs are vulnerable to prompt alterations like token deletion or reordering (Ishibashi et al., 2023), biased or toxic prompts (Shaikh et al., 2023) and similarity to training data (Razeghi et al., 2022). Other works have shown the importance of input-output format (Min et al., 2022) and of demonstration example ordering for few-shot performance (Zhao et al., 2021; Lu et al., 2022; Zhou et al., 2023a). In this work, *we shift our focus from analysing how sensitive model outputs are to how model representations change* (Gurnee and Tegmark, 2024). In particular, we explore for the first time the effect of prompt variations on how models internally represent mental states.

Activation Editing Activation editing has emerged as a way to influence model behaviour without any additional fine-tuning (Li et al., 2023a; Hernandez et al., 2023). One notable method in this domain is inference-time intervention (Li et al., 2023b, ITI), which involves training linear probes on contrastive question-answering datasets to identify "truthful" attention heads and then shifting their activations during inference along the identified truthful directions. In contrast, activation addition (Turner et al., 2023, AA) and contrastive activation addition (Rimsky et al., 2023, CAA) generate *steering vectors* by only using LMs' activations. Zhu et al. has used ITI to show that it is possible to manipulate LMs' internal representations of mental states. In this work, we show that using CAA can further improve LMs' ToM capabilities while eliminating the need for a fine-grained search over attention heads.

3 Experimental Setup

3.1 Probing

We linearly decode belief status from the perspective of different agents by using probing (Alain and Bengio, 2017). Probing involves localising specific concepts in a neural model by training a simple classifier (called a *probe*) on model activations to predict a target label associated with the input data. To provide a formal definition, we adopt a similar notation to the one introduced in (Belinkov, 2022). Consider an *original model* $f : x \mapsto \hat{y}$ that is trained on a dataset $\mathcal{D}^O = \{x^{(i)}, y^{(i)}\}$ to map input x to output $\hat{y}$. Model performance is evaluated by some measure, denoted $\textsc{Perf}(f, \mathcal{D}^O)$. A *probe* $g_l : f_l(x) \mapsto \hat{z}$ maps intermediate representations of x in f at layer l to some property $\hat{z}$, which is the label of interest. The probe g_l is trained on a *probing dataset* $\mathcal{D}^P = \{x^{(i)}, z^{(i)}\}$ and evaluated using some performance measure $\textsc{Perf}(g_l, f, \mathcal{D}^O, \mathcal{D}^P)$. In our case, f is an autoregressive language model that, given a sequence of tokens x, outputs a probability distribution over the token vocabulary to predict the next token in the sequence. Our probe is a logistic regression model $g_l : \hat{z} = Wa_l + b$ trained on neural activations $f_l(x) = a_l$ to predict binary belief labels $y = \{0, 1\}$.

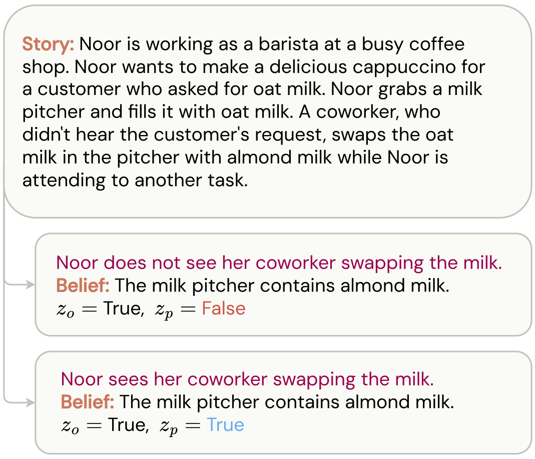
3.2 Dataset

We use BigToM (Gandhi et al., 2023), a question-answering dataset constructed by populating causal templates and combining elements from these templates. Each causal template is set up with a *context* and a description of the *protagonist* (e.g. *"Noor is working as a barista [...]"*, see Story in Figure 2), a *desire* (*"Noor wants to make a cappuccino"*), a *percept* (*"Noor grabs a milk pitcher and fills it with oat milk"*), and a *belief* (*"Noor believes that the pitcher contains oat milk"*). The state of the world is changed by a *causal event* (*"A coworker swaps the oat milk in the pitcher with almond milk"*). The dataset constructs different conditions by changing the percepts of the protagonist after the causal event, which will result in different beliefs. Similar to (Zhu et al., 2024), we focus on the *Forward Belief* setting in which models have to infer the belief of the protagonist given the percepts of the causal event, $P(\text{belief}|\text{percepts})$. We report additional details in Appendix A.1.1

Probing Datasets We consider two probing datasets: $\mathcal{D}^P_p = \{x^{(i)}_p, z^{(i)}_p\}$, where the labels $z^{(i)}_p$ correspond to ground-truth beliefs from the *protagonist* perspective, and $\mathcal{D}^P_o = \{x^{(i)}_o, z^{(i)}_o\}$, where the labels $z^{(i)}_o$ reflect the perspective of an omniscient *oracle*. $\mathcal{D}^P_p$ and $\mathcal{D}^P_o$ are built by pairing each story in BigToM with a belief statement, as shown in Figure 2. After prompting the model with a story-belief pair x we cache the residual stream activations $f_l(x)$ at the final token position for all residual streams (see Figure 6).

3.3 Models

We study two families of LMs that offer us options in model sizes and fine-tuning: Pythia (Biderman et al., 2023) and Llama-2 (Touvron et al., 2023) – for a total of **12 models**. While Llama-2 offers "chat" versions first trained with SFT and then RLHF, Pythia's open-source training set (Gao



Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task.

Noor does not see her coworker swapping the milk.
Belief: The milk pitcher contains almond milk.
$z_o =$ True, $z_p =$ False

Noor sees her coworker swapping the milk.
Belief: The milk pitcher contains almond milk.
$z_o =$ True, $z_p =$ True

Figure 2: Example of false belief from our probing datasets. The labels z_p and z_o correspond to $\mathcal{D}_p^P$ and $\mathcal{D}_o^P$, respectively. By manipulating the protagonist's percepts after the causal event, we obtain two scenarios: true belief and false belief.

et al., 2020) ensures that there is no data leakage[1]. Additionally, we consider a SFT version of Pythia-6.9B trained on open-source instruction datasets (Wang et al., 2024), which we refer to as Pythia-6.9B-chat.[2] We provide model details in Table 2.

3.4 Probing Experiments

To study how LMs represent beliefs of self and others, we propose a set of extensive probing experiments across LMs that differ in architecture, size, and fine-tuning regime. We train probes on the residual stream, as it integrates information from both the attention and feed-forward components, potentially encoding richer representations. Additionally, since the residual activations directly contribute to the final output predictions, probing them may better align with understanding the model's behaviour for downstream tasks.

Control Tasks Depending on the model dimension, the probes we train have a significant number of learnable parameters – up to $16,385$ for Llama-2-70B. This raises the concern that probes might learn to rely on irrelevant patterns in the data instead of capturing meaningful relationships. To account for the potential confounding effect of hidden state size, we include two controls. First, following Hewitt and Liang (2019), we train and evaluate probes on a version of $\mathcal{D}_p^P$ with randomly permuted labels – thus removing real input-label relationships. If a probe still performs well on the permuted data, this suggests it may be exploiting superficial correlations rather than capturing genuine structure. Second, we effectively reduce the number of learnable parameters in the probes by projecting $\mathcal{D}_p^P$ and $\mathcal{D}_o^P$ onto their k largest principal components using PCA before training. This minimises the risk of the probes relying on spurious patterns in the data.

Robustness Tests Previous work left the impact of prompting on belief probing accuracy unexplored. Our second set of experiments aims to study whether belief representations are robust to different prompts. Research on prompt robustness in language models focused mainly on revealing vulnerability to prompt alterations on *downstream performance* (Min et al., 2022; Ishibashi et al., 2023; Shaikh et al., 2023; Leidinger et al., 2023; Sclar et al., 2024). In contrast, we study how different prompt alterations influence *probing performance*, i.e. models' internal representations. Unlike model outputs that are shaped by decoding strategies, which act as confounders, models' activations are more abstract and offer a better lens into how robust or brittle internal representations are. We define four prompt variations:

- *Random*: Following Gurnee and Tegmark (2024), we add 10 random tokens to the belief statement.
- *Misleading*: Each story is followed by two belief statements, one pertinent to the story and one randomly chosen from another.
- *Time Specification*: The prompt specifies that the belief statement refers to the end of the story. We include this variation because some belief statements can be true (false) at the story's beginning but false (true) at the end. For example, consider the story in Figure 2: if Noor does not witness the swap, in the end, she will believe the pitcher contains almond milk ($z_p =$ True). However, if the same belief is referred to the beginning of the story, then it is false ($z_p =$ False).
- *Initial Belief*: We explicitly reveal the protagonist's initial belief (e.g. *"Noor believes that the pitcher contains oat milk"*) in the story to test whether it biases the representations of LMs.

While all maintain conceptual and semantic parity with the *Original* prompt used in (Zhu et al., 2024), *Random* and *Misleading* are expected to negatively impact LMs' representations, while *Time Specifi-*

[1] Llama-2 was released later than BigToM.

[2] https://huggingface.co/allenai/open-instruct-pythia-6.9b-tulu

cation and *Initial Belief* are supposed to have a positive influence. Robust representations of beliefs should exhibit minimal sensitivity to these alterations. Our experiments compare probe accuracy across different model sizes, fine-tuning, and prompt variations. Examples of prompts are reported in Appendix A.1.4.

3.5 Activation Editing

Prior work found that it is possible to manipulate models' representations of beliefs by using (Li et al., 2023b, ITI), and that such interventions can improve LMs' performance on ToM tasks. We take this further by asking whether a general "belief vector" can be distilled and *injected* into the models' activations to *strengthen* their ToM abilities. To this end, we use contrastive activation addition (Rimsky et al., 2023, CAA), an extension of activation addition (Turner et al., 2023, AA) that computes *steering vectors* to control LMs' behaviour. Steering vectors are computed as the average difference in residual stream activations between pairs of positive and negative instances of a specific behaviour. Formally, given a dataset $\mathcal{D}$ of triplets (p, c_p, c_n), where p is a prompt, c_p is a positive completion, and c_n is a negative completion, CAA computes a *mean difference* vector v_l^{md} for layer l as:

$$v_l^{md} = \frac{1}{|\mathcal{D}|} \sum_{p,c_p,c_n \in \mathcal{D}} a_l(p, c_p) - a_l(p, c_n) \quad (1)$$

For example, in Figure 2, p is the *Story*, c_p could be the true belief, and c_n the false belief. During inference, these steering vectors are multiplied by an appropriate coefficient α and added at every token position of the generated text after the prompt. CAA has two main advantages over ITI: First, it eliminates the need to train probes, making it *computationally cheap*. For example, for Llama2 70B, ITI needs to train 5,120 probes while CAA only needs to compute 80 vectors. Second, it operates at the residual stream level, making it easier to use than methods that intervene on specific attention heads like ITI. While CAA has been used to control alignment-relevant behaviour, such as hallucinations, refusal, and sycophancy (Rimsky et al., 2023), we are the first to apply it to enhance LMs' ToM reasoning. The "belief vectors" (i.e. steering vectors) we obtain can be understood as isolating the direction in the LMs' latent space corresponding to taking the perspective of another agent. To evaluate both base and fine-tuned LMs, we rank their answers to the ToM questions according to $p_{LM}(a|q)$ (Petroni et al., 2019). For a fair comparison, we adopt the train/test *Forward Belief* split used in (Zhu et al., 2024) to compute and evaluate the steering vectors. Additionally, we evaluate the transferability of the CAA steering vectors by applying them to two other BigToM tasks: *Forward Action* and *Backward Belief*. We provide details about these tasks in Appendix A.1.1, and a more detailed explanation ITI in Appendix A.5.

4 Results

Effect of Model Size and Fine-tuning Results from our study on model size and fine-tuning are shown in Figure 3. For *oracle* beliefs, probing accuracy rapidly converges to 100, with larger models showing faster convergence. Even the smallest Pythia-70m achieves 95% accuracy. For *protagonist* beliefs, we notice a similar pattern across most models, where accuracy at early layers is particularly low and then increases at the intermediate layers. What happens at early layers is overfitting, which may be caused by spurious features introduced by the initial coding strategy of language models, where individual token representations are mixed together (Gurnee et al., 2023). We further discuss this in Appendix A.2.1.

In general, probing accuracy increases with model size, although there is a performance gap between Llama-2 and Pythia. For example, Llama2-13B reaches around 80% accuracy, while Pythia-12B achieves approximately 60%. This gap is likely due to Llama-2 being trained on nearly seven times more tokens than Pythia (cf. Table 2). Probes from fine-tuned LMs show significantly better accuracy, with improvements of up to +29% for Llama2-7B-chat (SFT + RLHF) and +26% for Pythia-6.9B-chat (SFT) compared to probes from their base version. The same probes outperform (Llama-2) or are on par (Pythia) with probes trained on twice as large base models (12/13B). This highlights a key role of fine-tuning in shaping belief representations in smaller LMs. The performance gap closes for the largest Llama2-70B, for which the improvements from fine-tuning are marginal.

We characterise the relationship between probe accuracy and model size in Figure 7, using the *best* accuracy for each LM – i.e., the highest accuracy among probes g_l trained on activations a_l for model f. For Llama-2 base and Pythia base, probing accuracy scales logarithmically with model size (Fig-

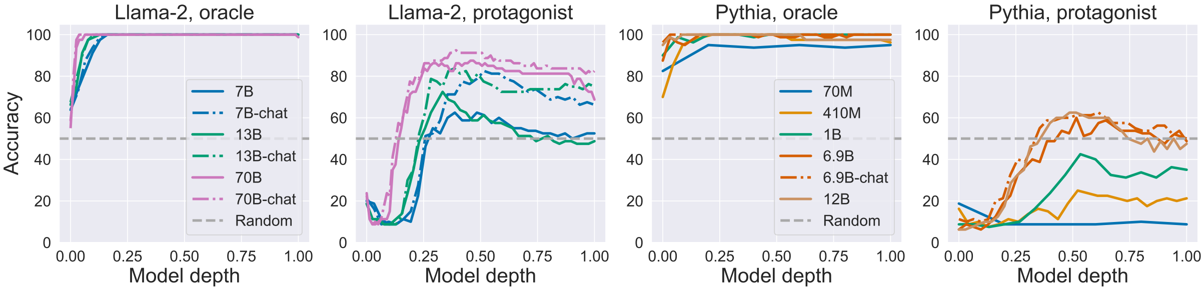


Figure 3: Belief probing accuracy show similar patterns across all models: *oracle* belief representations generally form already in the first layers, while *protagonist* belief representations emerge at the intermediate layers. Moreover, probing accuracy increases with model size and, more crucially for smaller models, with fine-tuning.

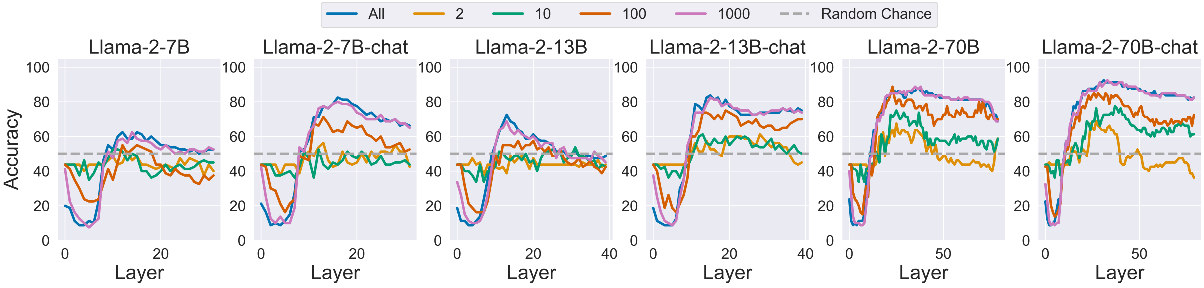


Figure 4: We compare the probing accuracy obtained by using the original set of activations (All) with the accuracy obtained by considering only the first $k = \{2, 10, 100, 1000\}$ principal components. Results are for *protagonist* beliefs (for *oracle* see Figure 12). In general, it is possible to recover most of the original accuracy by training probes on a smaller number k of principal components of the activations.

ure 7b, 7d), while for fine-tuned Llama-2 models, it scales linearly (Figure 7c).

Control Tasks Figure 8a shows that probes trained on the control task consistently perform at random chance, confirming that higher probing accuracy in larger models meaningfully reflects a greater ability to extract ToM representations, rather than simply being a by-product of spurious correlations. For Llama models, the probes generally exhibit selectivity: they achieve high accuracy when probing for beliefs but remain at chance level on control tasks. Pythia's overall accuracy is too low to allow for selectivity.

Figure 4 shows probing accuracy on *protagonist* when training the probes on the top k principal components of Llama-2's internal activations. We provide results for Pythia in Figure 11, and for all models on *oracle* settings in Figure 12. We consider $k = \{2, 10, 100, 1000\}$, spanning several orders of magnitude.[3] Results show that it is generally possible to recover most of the original accuracy by training probes on a smaller number k of principal components of the activations. We also performed the first control experiment, this time only using the first $k = \{100, 1000\}$ principal components. Figure 8b and 8c again show that probes trained on the control task consistently perform at random chance, confirming that probes are not fitting spurious patterns. Additionally, this suggests that belief representations are embedded in a low-dimensional subspace $\mathcal{B}$ spanned by the top k eigenvectors $\{v_1, \ldots, v_k\}$ of the covariance matrix $\mathsf{C} = \mathbb{E}[(a - \mathbb{E}[a])(a - \mathbb{E}[a])^\top]$.

Sensitivity to Prompting Figure 5 compares *protagonist* probe accuracy across various prompt variations for Llama-2 models. As can be seen from the figure, providing the protagonist's *Initial Belief* in the story yields higher probe accuracy compared to the *Original* prompt. Accuracy for all the other prompt variations is generally lower than *Original*. *Misleading* prompts hurt performance across all models. This finding resonates with Webson and Pavlick (2022), who found that instruction-tuned models, despite being more robust, are still sensitive to misleading prompts. On the other hand, *Time Specification* unexpectedly does not help in disambiguating belief states in different time frames, as we hypothesised in §3.4. Additionally, models

[3] For models with hidden dimensions smaller than 1000, we skip this value.

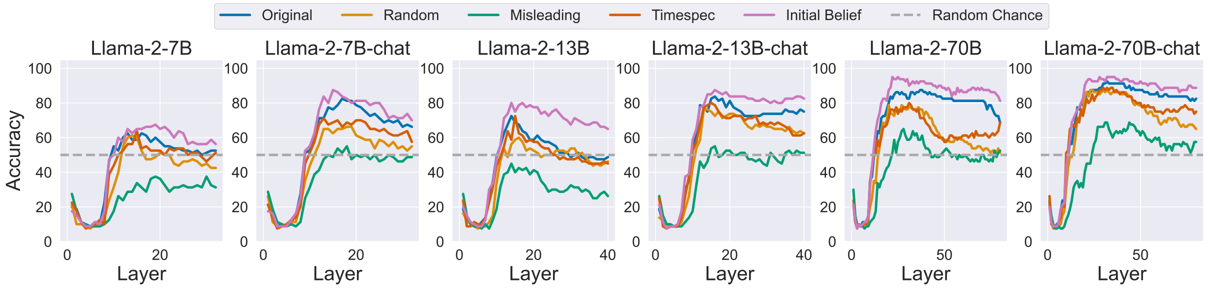


Figure 5: Sensitivity of *protagonist* belief probing accuracy to different prompt variations. Results for Pythia are shown in Figure 9. Representations are brittle to prompt variations.

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 × 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.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *International Conference on Learning Representations*.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, Online. Association for Computational Linguistics.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Evaluating the robustness of discrete prompts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2373–2384, Dubrovnik, Croatia. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4797.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations:

What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *Advances in Neural Information Processing Systems*, 37.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2024. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, and Minlie Huang. 2023. Coke: A cognitive knowledge graph for machine theory of mind. *arXiv preprint arXiv:2305.05390*.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023b. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496*.

A Appendix

A.1 Experimental setup

A.1.1 BigToM

BigToM (Gandhi et al., 2023) is constructed using GPT-4 (Achiam et al., 2023) to populate causal templates and combine elements from these templates. Each causal template is set up with a *context* and a description of the *protagonist* (e.g. *"Noor is working as a barista [...]"*), a *desire* (*"Noor wants to make a cappuccino"*), a *percept* (*"Noor grabs a milk pitcher and fills it with oat milk"*), and a *belief* (*"Noor believes that the pitcher contains oat milk"*). The state of the world is changed by a *causal event* (*"A coworker swaps the oat milk in the pitcher with almond milk"*). The dataset constructs different conditions by changing the percepts of the protagonist after the causal event, which will result in different beliefs – true or false. Gandhi et al. (2023) generated 200 templates and extracted 25 conditions from each template, resulting in 5,000 test samples. In this work, following Zhu et al. (2024) and Gandhi et al. (2023) we focused on the 6 most important conditions, corresponding to true and false beliefs on the following three tasks:

- *Forward Belief*: given the protagonist's percepts of the causal event, infer their belief: $P(\text{belief}|\text{percept})$.

- *Forward Action*: infer the protagonist's action given their desire and percepts of the causal event. Before inferring the action, one would need to first implicitly infer the protagonist's belief: $\sum_{\text{belief}} P(\text{action}|\text{percept}, \text{belief}, \text{desire})$.

- *Backward Belief*: infer the protagonist's belief from observed actions. This requires to first implicitly infer the protagonist's percepts: $\sum_{\text{percepts}} P(\text{belief}|\text{action}, \text{percept}, \text{desire})$.

The dataset was released under the MIT license and can be accessed at `https://github.com/cicl-stanford/procedural-evals-tom`. We report one example for each task in Example 1, 2, and 3, where the text defining true belief or false belief task is shown in blue and red, respectively.

A.1.2 Linear probes

Our probing approach is illustrated in Figure 6. For our experiments, we cache activations at the residual stream level. To perform ITI and compare it to CAA, we also cache attention heads activations. We trained the probes using the L-BFGS solver (Liu and Nocedal, 1989) with L2 penalty with inverse of regularisation strength 10 for a maximum of 1000 iterations. We use zero as random seed.

A.1.3 Language models

A detailed summary of the models we use in this work is shown in Table 2. Pythia was released under the Apache 2.0 license. Llama-2 is licensed by Meta for both researchers and commercial entities (Touvron et al., 2023). For all the models, we set the temperature to zero.

A.1.4 Examples of prompt variations

Example 4 shows an example of *Original* prompt. Examples of prompt variations are provided in Example 5 (*Random*), Example 6 (*Misleading*), Example 7 (*Time Specification*), and Example 8 (*Initial Belief*).

A.2 Model size and fine-tuning

To characterise the relationship between probe accuracy and model size we consider the *best* probe accuracy for every LM, i.e. the highest accuracy among probes $\{g_l\}$ trained on $\{a_l\}$ for a LM f. For Llama-2 base, the best probe accuracy scales logarithmically with model size ($R^2 = 0.98$, Figure 7b), whereas for fine-tuned models it scales linearly ($R = 1.0$, cf. Figure 7c). For Pythia base, the best probe accuracy also scales logarithmically with model size ($R^2 = 0.96$, Figure 7d).

A.2.1 Overfitting Issues

Figure 3 also that probing accuracy at early layers is particularly low across all models, performing even worse than random. This happens due to overfitting, which may be caused by spurious features introduced by the initial coding strategy of language models, where individual token representations are mixed together (Gurnee et al., 2023). We also identified the same issue when reproducing the results in Zhu et al. (2024), who address it by manually clipping all accuracies below random chance to 50%.[4] Since probing experiments require training a large number of probes for each model, both we and Zhu et al. (2024) trained each probe for the same fixed number of epochs (1,000). However, for activations from the earlier layers, overfitting occurs very quickly - often within the first 10 iterations.

We ran an experiment with Llama2-7B-chat, reducing training to fewer than 10 iterations, and found that the probes performed at random chance. Therefore, to fully resolve this issue, we would need to choose the number of training epochs for each probe individually. This would likely flatten the observed "U" shape in the results. However, this process would be computationally expensive and does not contribute to our main research questions. Rather than artificially adjusting accuracies to 50%, we prefer to present the results as they are.

A.3 Sensitivity to prompting

Accuracy on *protagonist* belief probing for Pythia models is shown in Figure 9.

Accuracy on *oracle* belief probing for different prompt variations are reported in Figure 10.

A.4 Dimensionality reduction

Probing accuracy obtained by Pythia models for the *protagonist* setting is reported in Figure 11.

Oracle probe accuracy obtained by considering only the first $n = \{2, 10, 100, 1000\}$ principal components are shown in Figure 12.

A.5 Inference-time intervention

Inference-time intervention (Li et al., 2023b, ITI) employs a two-step process. First, it trains a probe for each attention head across all layers of a LM. These probes are evaluated on a validation set, and the top-k heads with the highest accuracy are selected. Subsequently, during inference, ITI steers the activations of these top heads along the directions defined by their corresponding probes. Formally, ITI can be defined as an additional term to the multi-head attention:

$$x_{l+1} = x_l + \sum_{h=1}^{H} Q_l^h \left(\mathrm{Att}_l^h(P_l^h x_l) + \alpha \sigma_l^h \theta_l^h \right)$$

[4] `https://github.com/Walter0807/RepBelief/blob/0fc86396f2f0a998643ea01786eb3db4dd20ff9c/probe.py#L60`

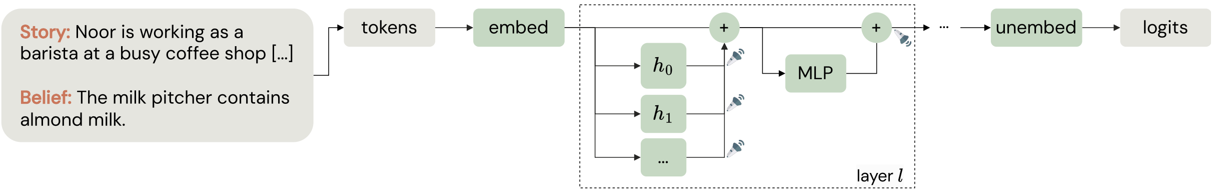

Story: Noor is working as a barista at a busy coffee shop [...]
Belief: The milk pitcher contains almond milk.
tokens
embed
+
h_0
h_1
...
MLP
+
...
unembed
logits
layer l

Figure 6: Given a tokenised input, we cache the internal activations for all attention heads h_i, $i = 0, \ldots, H-1$, and residual streams. In our experiments, we use residual stream activations.

LM	Size	+ SFT	+ RLHF	Tokens	d_{model}	Layers
Llama-2	7B			2T	4096	32
	13B			2T	5120	40
	70B			2T	8192	80
Llama-2-chat	7B	✓	✓	2T	4096	32
	13B	✓	✓	2T	5120	40
	70B	✓	✓	2T	8192	80
Pythia	70M			300B	512	6
	410M			300B	1024	24
	1B			300B	2048	16
	6.9B			300B	4096	32
	12B			300B	5120	36
	6.9B	✓		300B	4096	32

Table 2: The 12 models used in this work.

where x_l is the residual stream at layer l, H is the number of attention heads, $\alpha \in \mathbb{R}^+$ is a coefficient, σ_l^h is the standard deviation of activations along the direction identified by the probe trained on attention head h at layer l, and θ_l^h is zero ofr not-selected attention heads.

A.6 Activation editing

Table 3 reports results obtained on the three BigToM tasks with the corresponding hyperparameters used for ITI (Li et al., 2023b) and CAA (Rimsky et al., 2023). We report an example of prompt used for evaluation in Example 9. Table 4 shows the accuracy obtained by using CAA on the Forward Belief True Control task in BigToM. On this control task, CAA produced improved results for all model, proving that CAA not only improves performance on ToM tasks, but also does not degrades the models' ability to perform other tasks.

A.7 Compute resources

We ran our experiments on a server running Ubuntu 22.04, equipped with eight NVIDIA Tesla V100-SXM2 GPUs with 32GB of memory and Intel Xeon Platinum 8260 CPUs.

A.8 Code

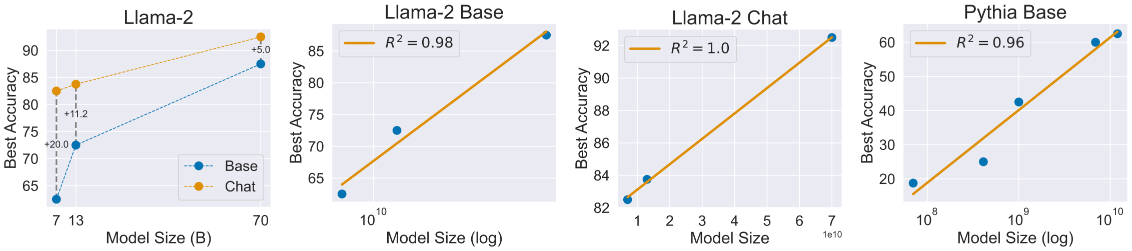
Our code is provided as supplementary material and it will be made public under the MIT licence at `www.this-is-a-placeholder.com`.

A.9 Societal impact

While our work is foundational and remains distant from specific applications with direct societal impact, it's important to recognise the ethical implications of predicting and editing mental state representations.

Handling sensitive aspects of individuals' inner experiences and emotions requires careful consideration to avoid reinforcing biases or misunderstanding psychological nuances. As LMs begin to encode aspects of ToM, there's a risk that overinterpreting these capabilities could lead to misplaced trust – especially in real-world applications requiring nuanced social reasoning, such as education, healthcare, or mental health support.

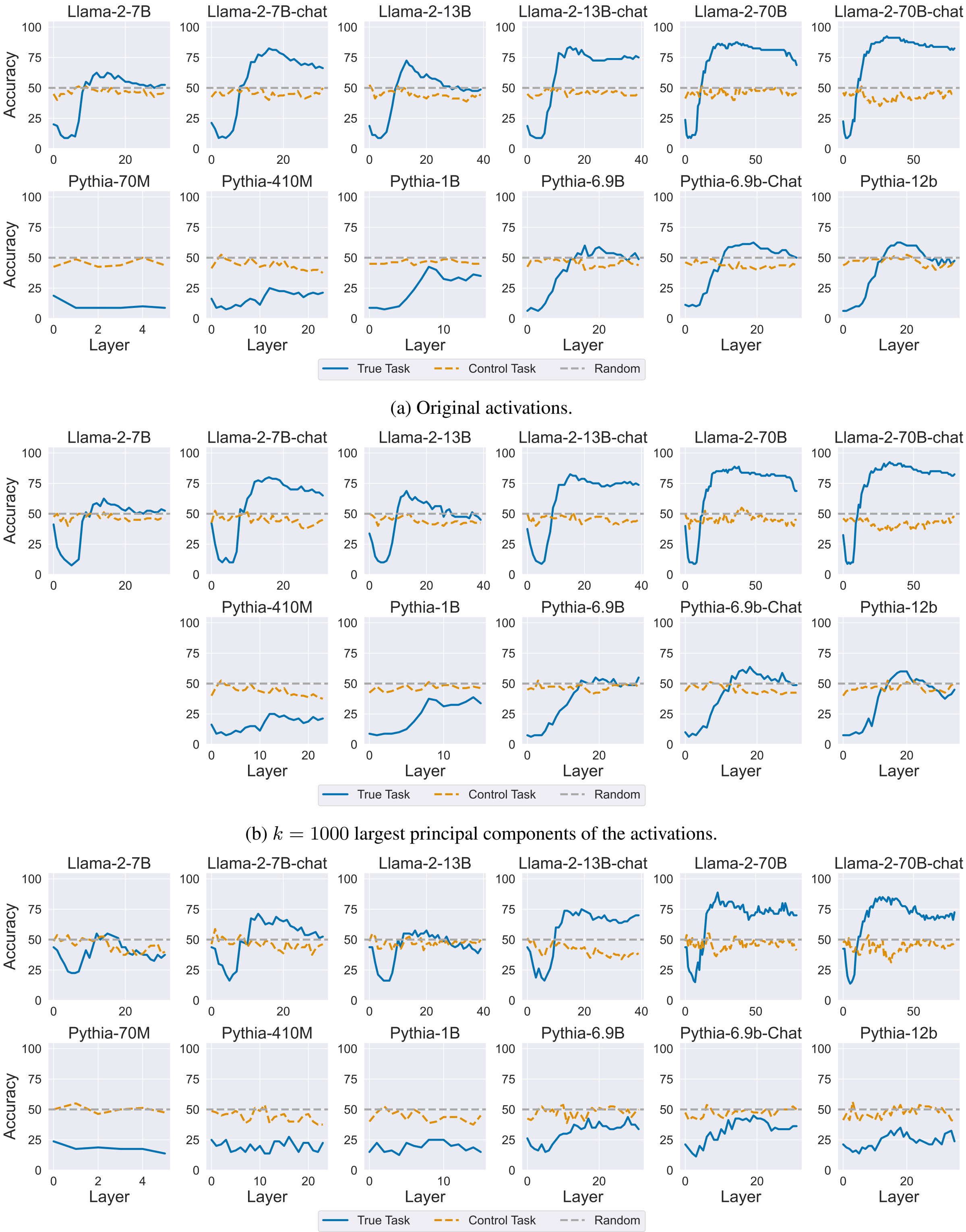
Furthermore, while techniques like CAA show promise for steering internal representations, they also potentially introduce new ethical challenges. Manipulating a model's internal states, especially in ways that affect social reasoning, requires trans-



(a) (b) (c) (d)

Figure 7: To characterise the relationship between probe accuracy and model size we consider the *best* probe accuracy for every LM, i.e. the highest accuracy among probes $\{g_l\}$ trained on $\{a_l\}$ for a LM f. **(a)** Best accuracy for Llama-2 models of different size. Numbers on the vertical dotted lines indicate the gain in accuracy between base and fine-tuned model of the same size. **(b)** Logarithmic fit for Llama-2 base. **(c)** Linear fit for Llama-2 fine-tuned (chat). **(d)** Logarithmic fit for Pythia base.

parency and caution to avoid unintended consequences such as bias amplification or fairness issues. Future work should consider not only improving technical performance but also developing safeguards and evaluation frameworks to ensure responsible use of ToM-like abilities in LMs.



(a) Original activations.

(b) $k = 1000$ largest principal components of the activations.

(c) $k = 100$ largest principal components of the activations.

Figure 8: Comparison between accuracy on belief probing and accuracy obtained on a control task.

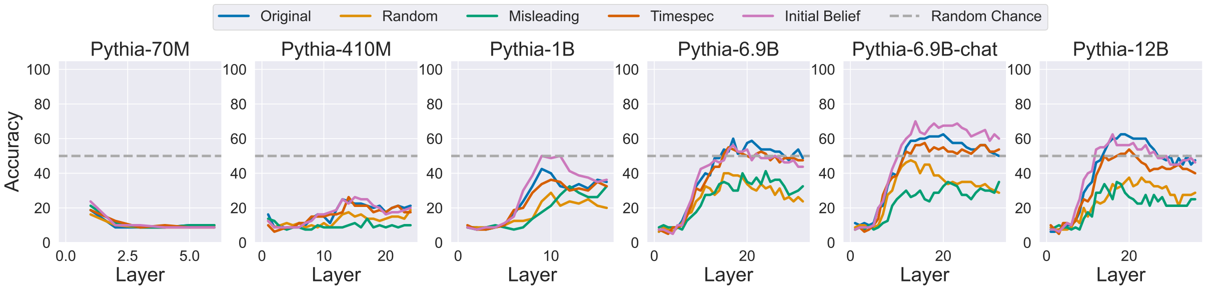


Figure 9: Sensitivity of protagonist belief probing accuracy to different prompt variations.

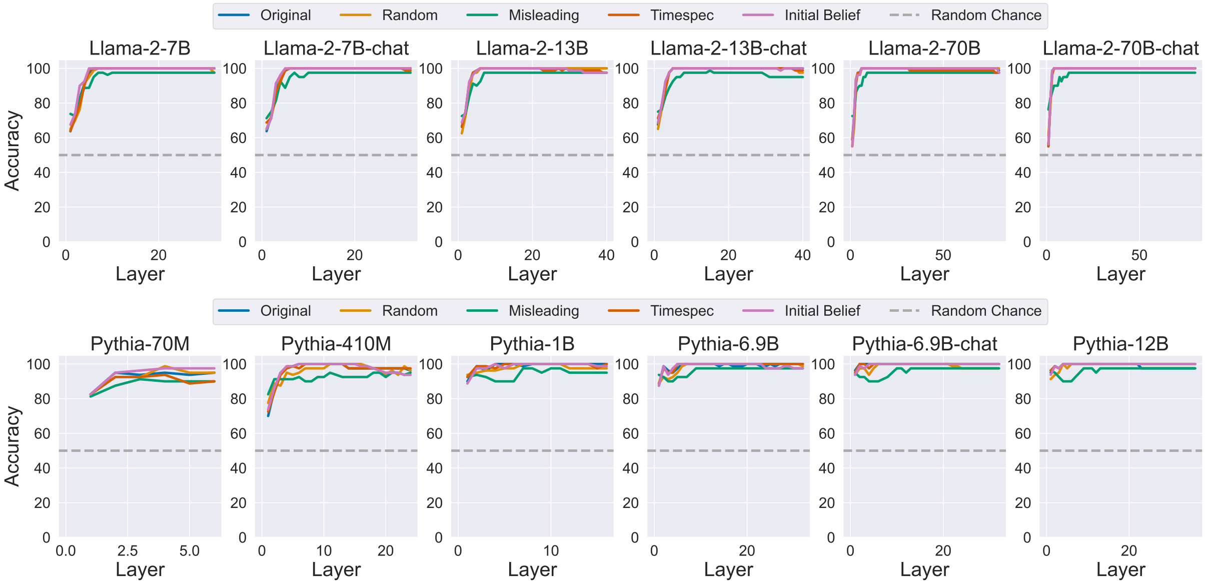


Figure 10: Sensitivity of protagonist belief probing accuracy to different prompt variations.

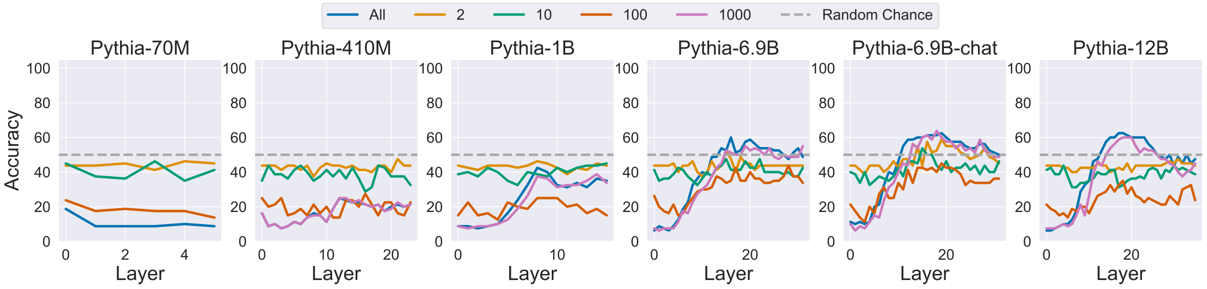


Figure 11: We compare the probing accuracy obtained by using the original set of activations (All) with the accuracy obtained by considering only the first $n = \{2, 10, 100, 1000\}$ principal components. For Pythia: All(70m) = 512, All(410m) = 1024, All(1b) = 2048, All(6.9b) = 4096, All(12b) = 5120. Results for *oracle* are shown in Figure 12.

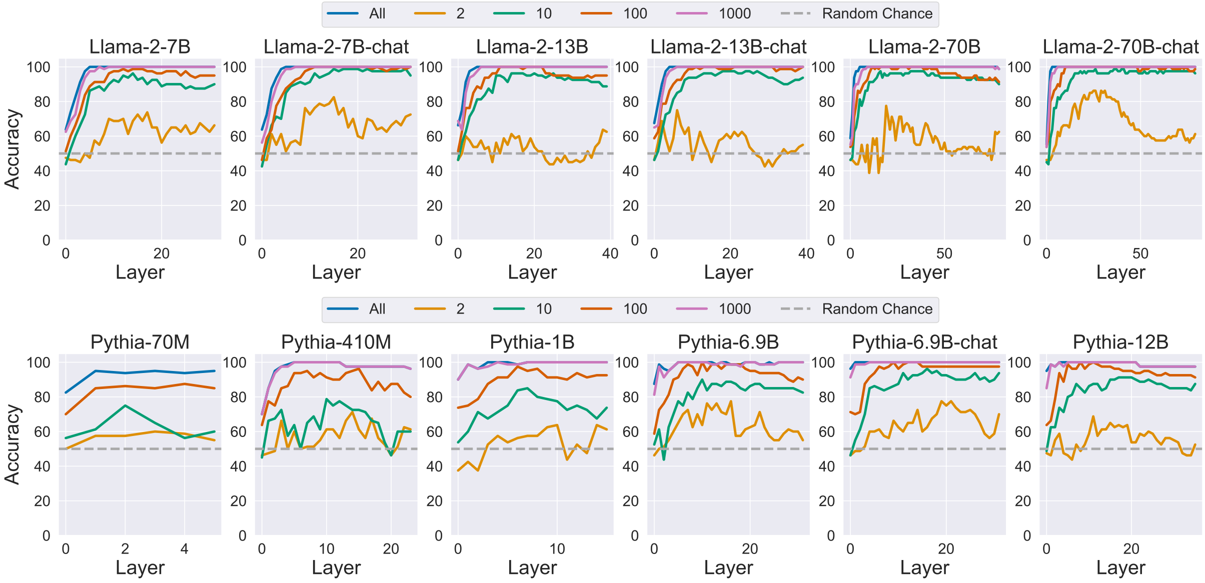


Figure 12: **(Oracle)** To investigate potential memorisation in the probes, we compare the probing accuracy obtained by using the original set of activations (All) with the accuracy obtained by considering only the first $n = \{2, 10, 100, 1000\}$ principal components. For Llama2: All(7b) = 4096, All(13b) = 5120, All(70b) = 8192. For Pythia: All(70m) = 512, All(410m) = 1024, All(1b) = 2048, All(6.9b) = 4096, All(12b) = 5120.

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	$44_{0.0}$	$44_{0.0}$	$44_{0.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$66_{2.0,11}$	$71_{1.0,31}$	$54_{2.0,0}$	$66_{2.0,11}$	$57_{2.0,12}$	$54_{2.0,2}$	$60_{2.0,11}$	$74_{1.0,31}$	$54_{2.0,2}$
Llama-2-7b-chat	No int.	56	56	55	69	55	37	56	56	55
	ITI	$58_{15.0}$	$58_{15.0}$	$57_{15.0}$	$69_{0.0}$	$55_{0.0}$	$37_{0.0}$	$58_{10.0}$	$60_{10.0}$	$57_{10.0}$
	CAA	$70_{1.0,11}$	$72_{1.5,10}$	$57_{1.0,1}$	$69_{0.0,0}$	$67_{1.5,11}$	$53_{1.5,12}$	$66_{1.0,11}$	$84_{1.5,10}$	$57_{1.0,0}$
Llama-2-13b	No int.	52	44	35	59	50	37	46	49	33
	ITI	$52_{0.0}$	$45_{15.0}$	$35_{0.0}$	$64_{15.0}$	$61_{20.0}$	$46_{20.0}$	$48_{20.0}$	$59_{20.0}$	$42_{20.0}$
	CAA	$85_{2.0,12}$	$88_{2.0,14}$	$66_{2.0,12}$	$71_{1.5,10}$	$69_{2.0,13}$	$55_{1.0,39}$	$75_{2.0,10}$	$92_{2.0,13}$	$59_{1.5,12}$
Llama-2-13b-chat	No int.	84	56	47	78	51	38	72	48	31
	ITI	$84_{0.0}$	$65_{15.0}$	$59_{15.0}$	$78_{0.0}$	$58_{15.0}$	$47_{15.0}$	$72_{0.0}$	$60_{15.0}$	$48_{15.0}$
	CAA	$97_{1.0,12}$	$94_{1.0,12}$	$91_{1.0,12}$	$80_{1.5,11}$	$71_{1.0,13}$	$54_{1.5,13}$	$97_{1.5,10}$	$94_{1.5,12}$	$87_{1.5,12}$
Llama-2-70b	No int.	90	87	78	93	52	48	73	53	32
	ITI	$90_{0.0}$	$90_{20.0}$	$78_{0.0}$	$94_{15.0}$	$55_{20.0}$	$50_{15.0}$	$77_{10.0}$	$58_{15.0}$	$37_{10.0}$
	CAA	$99_{2.0,16}$	$97_{1.5,19}$	$95_{1.5,18}$	$94_{1.5,2}$	$80_{2.0,19}$	$73_{1.5,18}$	$94_{2.0,18}$	$92_{2.0,19}$	$83_{1.5,19}$
Llama-2-70b-chat	No int.	69	75	56	86	56	52	63	59	52
	ITI	$69_{0.0}$	$76_{10.0}$	$59_{10.0}$	$86_{0.0}$	$56_{0.0}$	$52_{0.0}$	$63_{0.0}$	$60_{10.0}$	$54_{10.0}$
	CAA	$92_{1.5,18}$	$97_{1.5,25}$	$89_{1.5,18}$	$87_{1.5,17}$	$75_{1.0,19}$	$60_{1.0,19}$	$88_{1.5,18}$	$92_{1.0,19}$	$80_{1.5,18}$
Pythia-70m	No int.	41	41	37	46	45	41	44	41	37
	ITI	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$62_{1.0,2}$	$56_{1.0,1}$	$54_{1.5,1}$	$59_{1.0,2}$	$60_{1.0,3}$	$58_{1.0,2}$	$63_{1.0,2}$	$56_{1.0,2}$	$54_{1.5,1}$
Pythia-410m	No int.	48	45	45	44	44	44	44	47	44
	ITI	$55_{20.0}$	$62_{20.0}$	$52_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$60_{20.0}$	$63_{20.0}$	$56_{20.0}$
	CAA	$67_{2.0,4}$	$64_{2.0,4}$	$61_{2.0,0}$	$56_{2.0,6}$	$63_{1.5,12}$	$56_{2.0,6}$	$69_{2.0,4}$	$63_{2.0,0}$	$60_{2.0,0}$
Pythia-1b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$59_{2.0,8}$	$62_{2.0,5}$	$54_{2.0,0}$	$57_{2.0,4}$	$59_{2.0,10}$	$56_{2.0,4}$	$57_{2.0,3}$	$60_{2.0,5}$	$54_{2.0,0}$
Pythia-6.9b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$45_{20.0}$	$54_{20.0}$	$44_{0.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$56_{1.5,12}$	$71_{1.5,9}$	$55_{2.0,23}$	$55_{2.0,4}$	$63_{1.5,11}$	$55_{2.0,4}$	$55_{2.0,23}$	$71_{1.5,9}$	$55_{2.0,23}$
Pythia-6.9b-chat	No int.	55	54	28	36	64	20	44	67	30
	ITI	$57_{15.0}$	$54_{0.0}$	$28_{0.0}$	$44_{15.0}$	$71_{15.0}$	$32_{15.0}$	$44_{0.0}$	$67_{0.0}$	$30_{0.0}$
	CAA	$68_{1.5,15}$	$65_{1.5,12}$	$57_{1.5,11}$	$54_{1.5,10}$	$75_{1.5,5}$	$48_{1.5,10}$	$58_{1.5,15}$	$67_{0.0,0}$	$54_{1.5,10}$
Pythia-12b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$54_{2.0,0}$	$64_{2.0,9}$	$54_{2.0,0}$	$60_{2.0,11}$	$58_{2.0,11}$	$55_{2.0,12}$	$54_{2.0,0}$	$67_{2.0,10}$	$54_{2.0,0}$

Table 3: Activation intervention: comparison between ITI (Li et al., 2023b) and CAA (Rimsky et al., 2023). For ITI, the subscript indicates the value of the coefficient α_{ITI} used: $\text{Acc}_{\alpha_{\text{ITI}}}$. For CAA, the subscript indicates first the value of the coefficient α used and second the layer l at which intervention takes place: $\text{Acc}_{\alpha_{\text{CAA}},l}$.

Model	**Method**	**Control**	**CAA Parameters**
Llama-2-7b	No int.	44	
	CAA	66_{+22}	2.0, 11
Llama-2-7b-chat	No int.	56	
	CAA	70_{+14}	1.0, 11
Llama-2-13b	No int.	52	
	CAA	85_{+33}	2.0, 12
Llama-2-13b-chat	No int.	84	
	CAA	97_{+13}	1.0, 12
Llama-2-70b	No int.	90	
	CAA	99_{+9}	2.0, 16
Llama-2-70b-chat	No int.	69	
	CAA	92_{+23}	1.5, 18
Pythia-70m	No int.	41	
	CAA	62_{+21}	1.0, 2
Pythia-410m	No int.	48	
	CAA	67_{+19}	2.0, 4
Pythia-1b	No int.	44	
	CAA	59_{+15}	2.0, 8
Pythia-6.9b	No int.	44	
	CAA	56_{+12}	1.5, 12
Pythia-6.9b-chat	No int.	55	
	CAA	68_{+13}	1.5, 15
Pythia-12b	No int.	44	
	CAA	54_{+10}	2.0, 0

Table 4: Results for CAA (Rimsky et al., 2023) on the Forward Belief True Control task in BigToM (Gandhi et al., 2023). Numbers indicate accuracy scores, with differences (CAA − No int.) as subscripts.

Example 1: Forward Belief

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk. Noor does not see her coworker swapping the milk.
Does Noor believe the milk pitcher contains oat milk or almond milk?
a) Noor believes the milk pitcher contains oat milk.
b) Noor believes the milk pitcher contains almond milk.

Example 2: Forward Action

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk. Noor does not see her coworker swapping the milk.
What will Noor do?
a) Noor will make the cappuccino using the milk in the pitcher.
b) Noor will open the fridge once again to take out the oat milk and replace the almond milk with oat milk.

Example 3: Backward Belief

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor opens the fridge again and reaches for the oat milk. Noor makes the cappuccino using the milk in the pitcher.
Does Noor believe the milk pitcher contains oat milk or almond milk?
a) Noor believes the milk pitcher contains oat milk.
b) Noor believes the milk pitcher contains almond milk.

Example 4: Default prompt

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Belief: Noor believes the milk pitcher contains oat milk.

Example 5: Prompt variation – Random

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn’t hear the customer’s request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Belief: 장士 decided [_ countries sections behoSOUR gminy bef Noor believes the milk pitcher contains oat milk.

Example 6: Prompt variation – Misleading

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn’t hear the customer’s request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Belief: Noor believes the milk pitcher contains oat milk.
Belief: The water valve is closed.

Example 7: Prompt variation – Time specification

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn’t hear the customer’s request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Belief: In the end, Noor believes the milk pitcher contains oat milk.

Example 8: Prompt variation – Initial belief

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. Noor believes that the milk pitcher contains oatmilk. A coworker, who didn’t hear the customer’s request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Belief: Noor believes the milk pitcher contains oat milk.

Example 9: Prompt used for downstream evaluation

```
Answer the questions based on the context. Keep your answer concise, few words
are enough, maximum one sentence. Answer as 'Answer:<option>)<answer>'.

Story: Noor is working as a barista at a busy coffee shop. Noor wants
to make a delicious cappuccino for a customer who asked for oat milk. Noor
grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the
customer's request, swaps the oat milk in the pitcher with almond milk while
Noor is attending to another task. Noor does not see her coworker swapping the
milk.
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) Noor believes the milk pitcher contains oat milk.
b) Noor believes the milk pitcher contains almond milk.
Answer:
```