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).

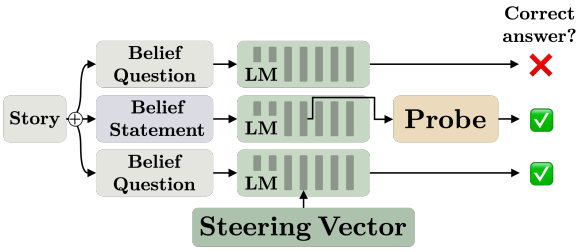

Figure 1: ToM tasks are challenging for LMs, but correct predictions can sometimes be recovered by *probing* their internal representations. We study how internal representations of beliefs of self and others emerge in 12 LMs, and show that these representations are structured yet brittle to prompts, 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: $\mathcal{D}_p^P = \{x_p^{(i)}, z_p^{(i)}\}$, where the labels $z_p^{(i)}$ correspond to ground-truth beliefs from the *protagonist* perspective, and $\mathcal{D}_o^P = \{x_o^{(i)}, z_o^{(i)}\}$, where the labels $z_o^{(i)}$ reflect the perspective of an omniscient *oracle*. $\mathcal{D}_p^P$ and $\mathcal{D}_o^P$ 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