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

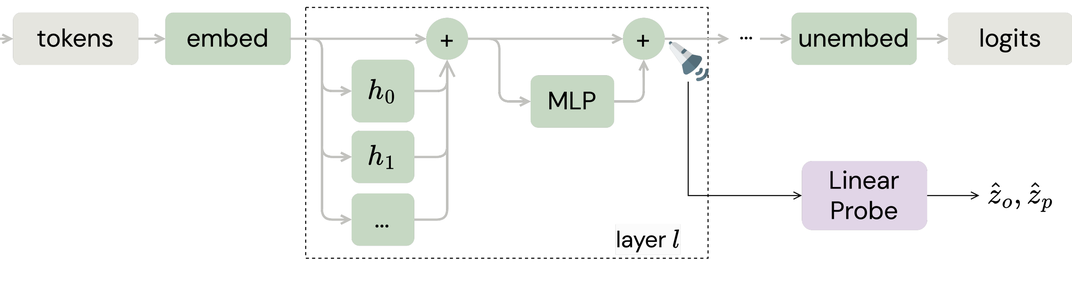
Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, Andreas Bulling
matteo.bortoletto@vis.uni-stuttgart.de

University of Stuttgart
Germany

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



Motivation

- **Theory of Mind** (ToM) is the ability to attribute mental states to oneself and others [1].
- Recent interest in evaluating **Language Models'** (LMs) generative performance on ToM tasks [2].
- Previous work suggests that LMs can represent beliefs of self and others [3].
- Experiments are limited in the number of models and settings studied, leaving several questions unanswered.

Research Questions

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

RQ2. Are internal belief representations structured or the results of spurious correlations?

RQ3. Are internal belief representations robust?

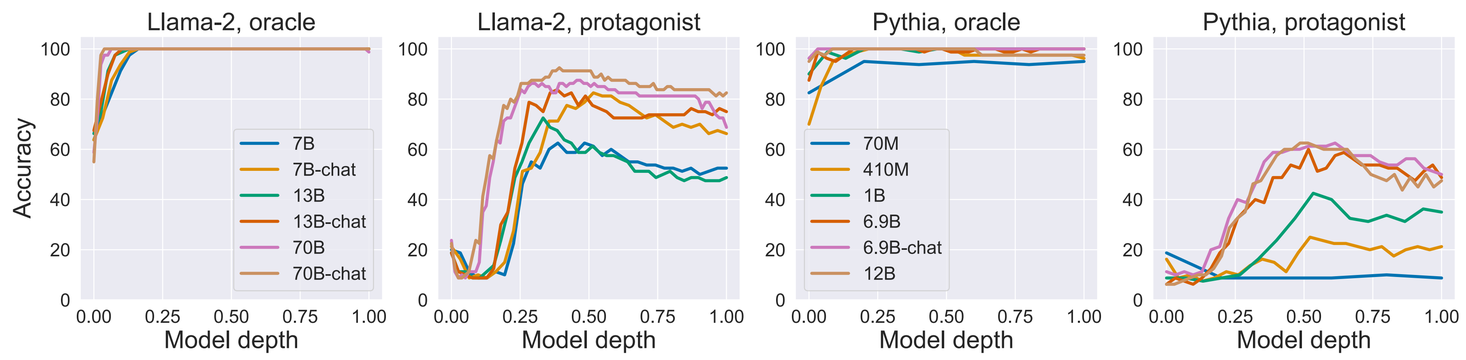
RQ4. Can we enhance LMs' performance by editing their activations without training dedicated probes?

Contributions & Findings

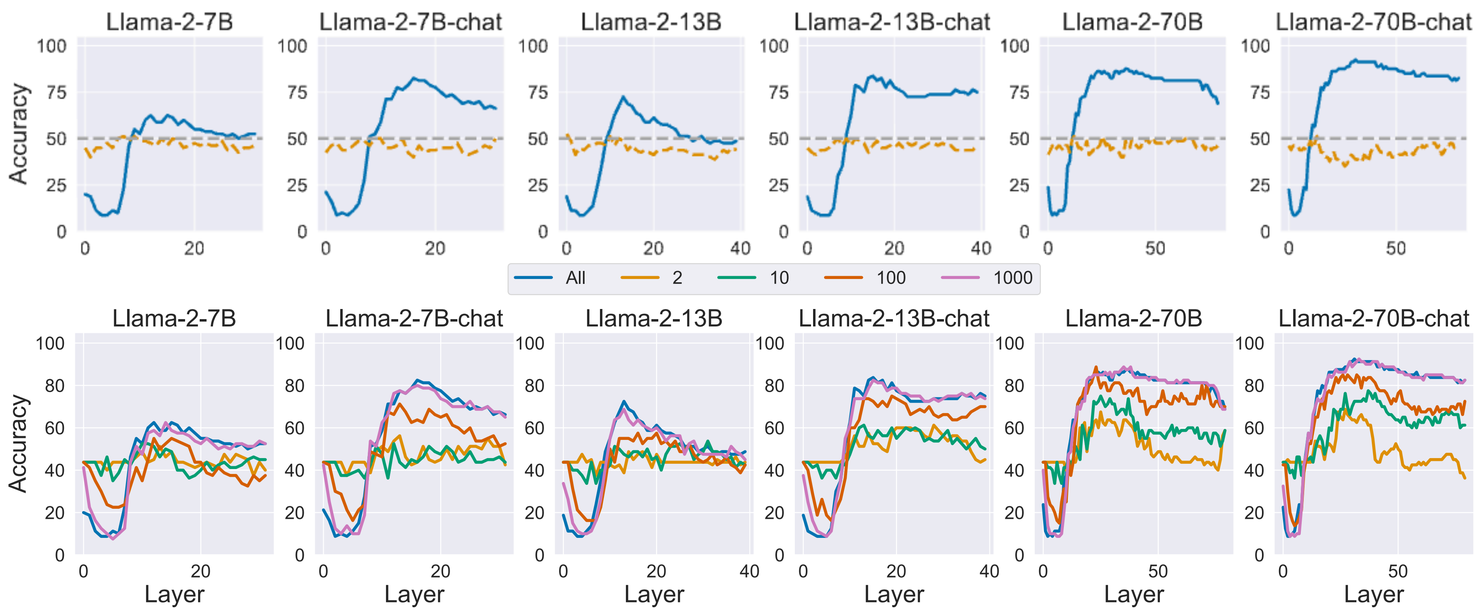
1. Extensive probing experiments across **12 models**, suggesting that:
   - The representations of others' beliefs of others improve with model size and, more crucially for smaller models, fine-tuning.
   - Probes capture structured belief representations rather than spurious correlations.
   - LMs' representations of others' beliefs are brittle to prompt variations.
2. We show that by using contrastive activation addition [4] it is possible to improve models' ToM performance by steering their activations without the need to train any probe.

Probing Language Models' Representations

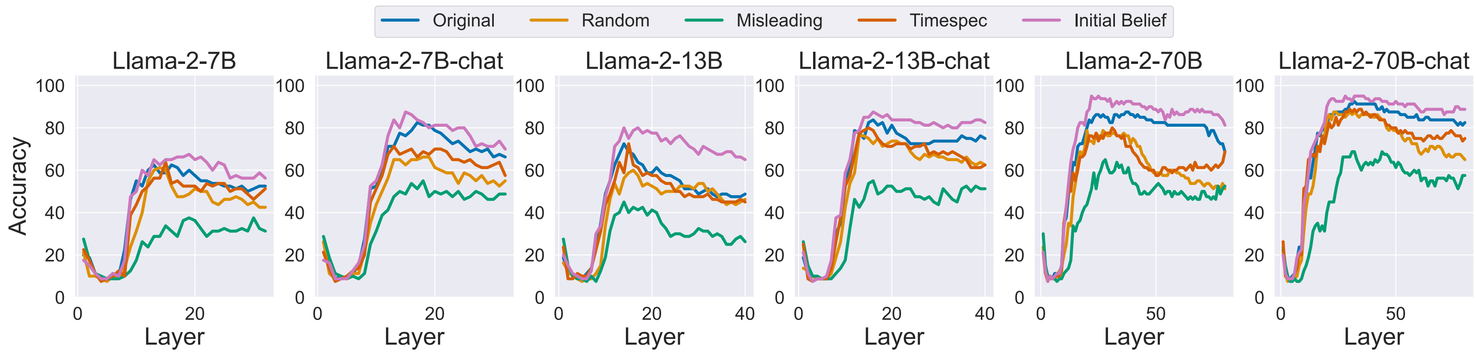
RQ1. Effect of size and fine-tuning



RQ2. Control tasks

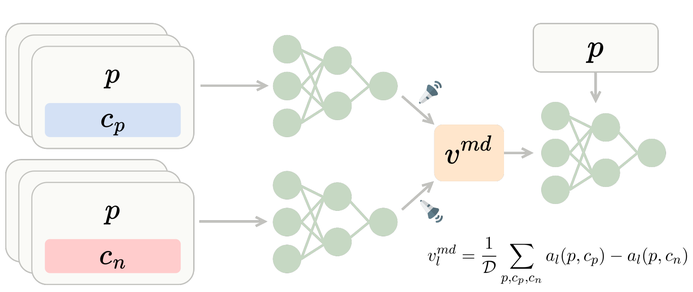


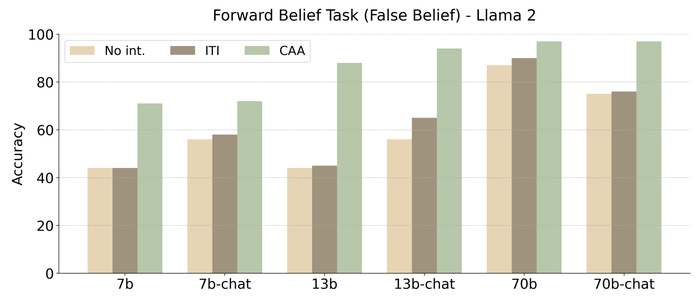
RQ3. Robustness to prompting

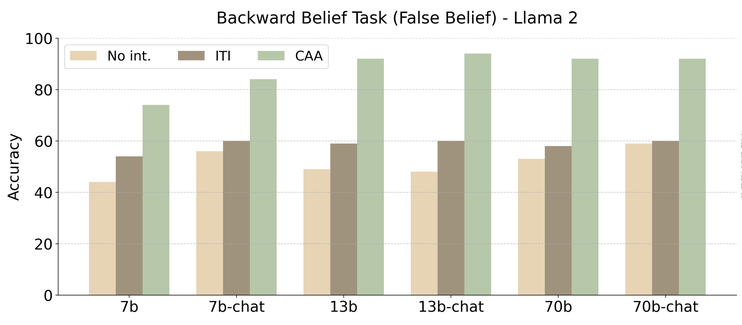


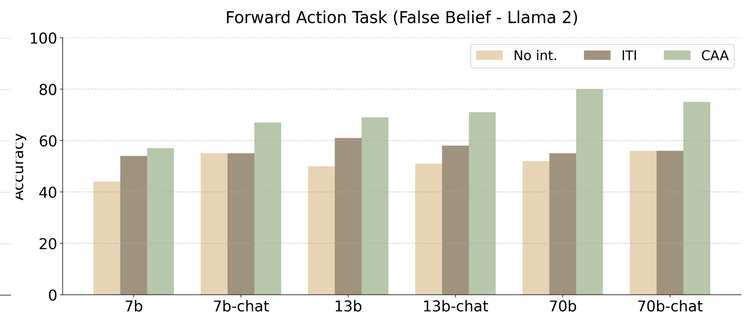
Activation Editing

RQ4. Contrastive Activation Addition [4]









References

[1] Premack, David, and Guy Woodruff. "Does the chimpanzee have a theory of mind?." *Behavioral and brain sciences* 1.4 (1978): 515-526.
[2] Gandhi, Kanishk, et al. "Understanding social reasoning in language models with language models." *NeurIPS* 2024.
[3] Zhu, Wentao, Zhining Zhang, and Yizhou Wang. "Language Models Represent Beliefs of Self and Others." *ICML* 2024.
[4] Rimsky, Nina, et al. "Steering llama 2 via contrastive activation addition." *ACL 2024*.

30th ANNIVERSARY EMNLP 2025
Suzhou, China 中国苏州
November 4-9 11月4日–9日