Benchmarking Mental State Representations in Language Models

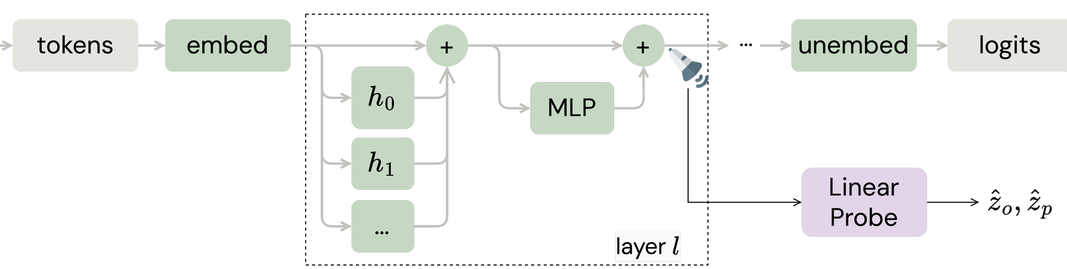
PUI Perceptual User Interfaces
imprs-is
SimTech
University of Stuttgart
Germany

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

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 used **probing** to show 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.

Our Research Questions

RQ1. What is the relation between model size and probing accuracy?

RQ2. Does fine-tuning have an effect on probing accuracy?

RQ3. Are models' internal representations of beliefs sensitive to prompt variations?

RQ4. Is there a risk of probes memorising training data due to the large dimensionality of LM representations?

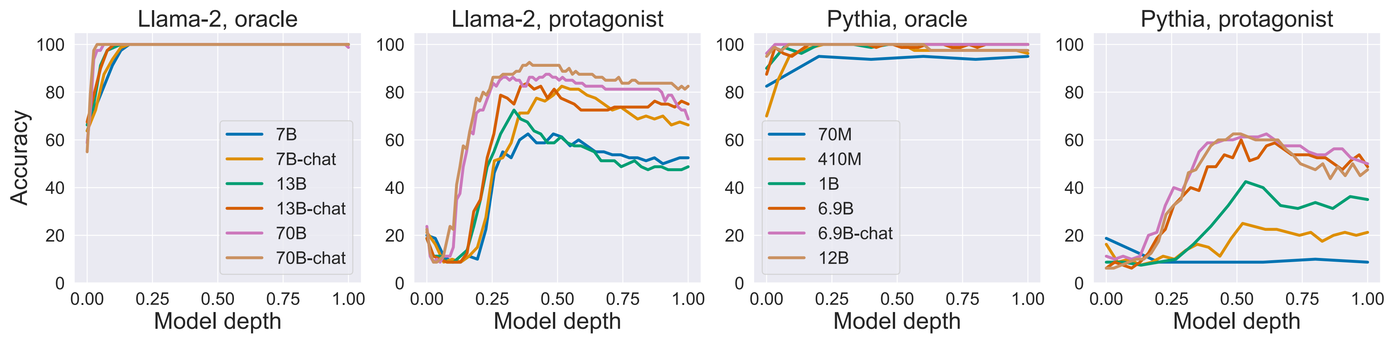
RQ5. Can we enhance LMs' performance by editing their activations without training dedicated probes?

Our Contributions

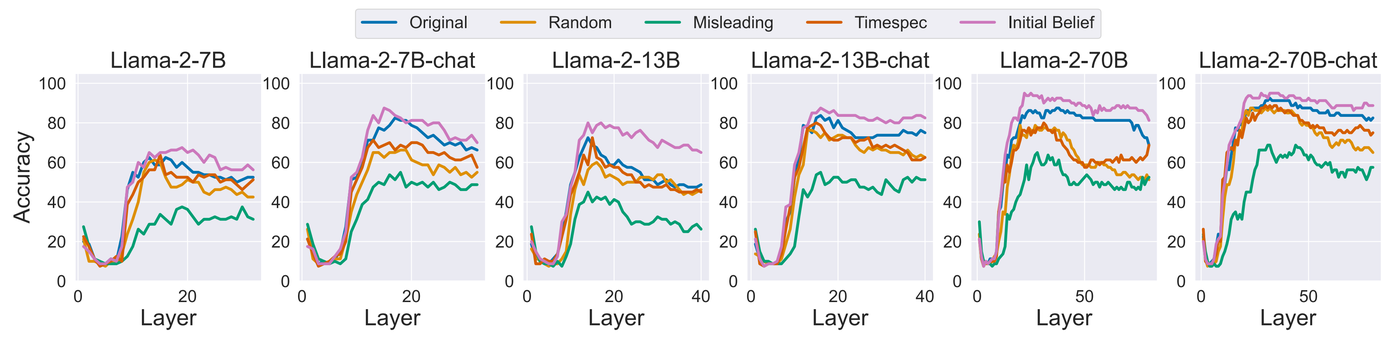
1. Extensive probing experiments with various types of LMs showing that the quality of models' internal representations of the beliefs of others increases with model size and, more crucially for smaller models, fine-tuning.
2. We are the first to study how prompt variations impact belief probing performance, showing that models' representations are sensitive to prompt variations, even when such variations should be beneficial.
3. 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

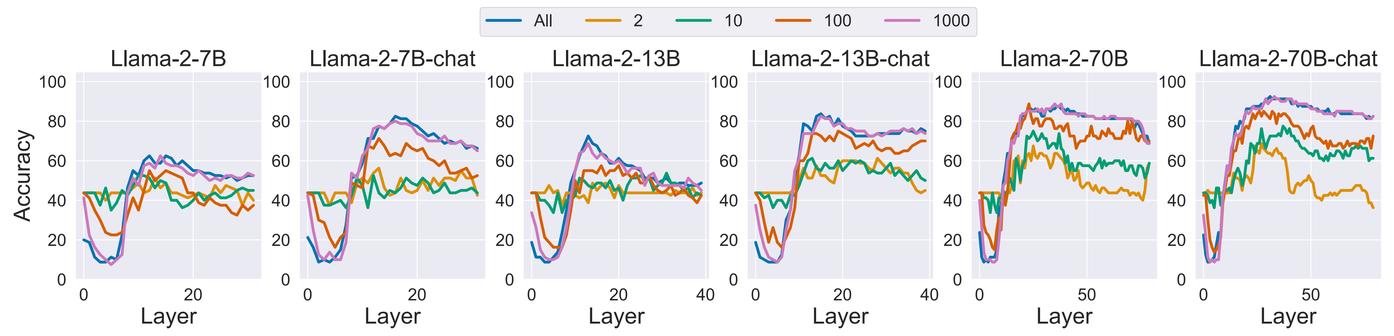
RQ1 & RQ2. Effect of Size and Fine-Tuning



RQ3. Sensitivity to Prompting

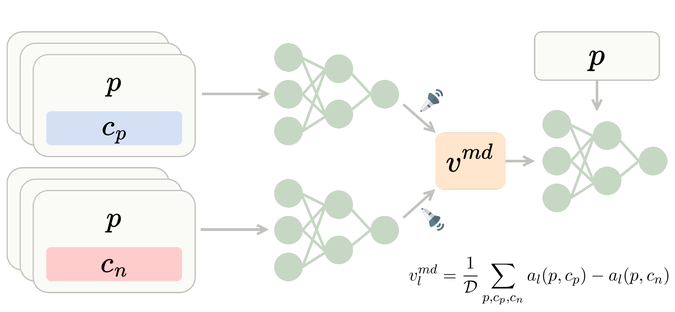


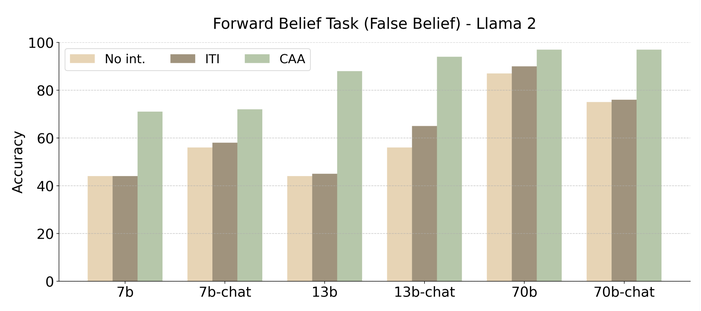
RQ4. Probe Memorisation

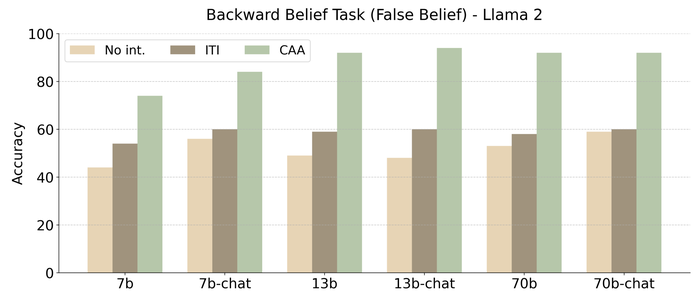


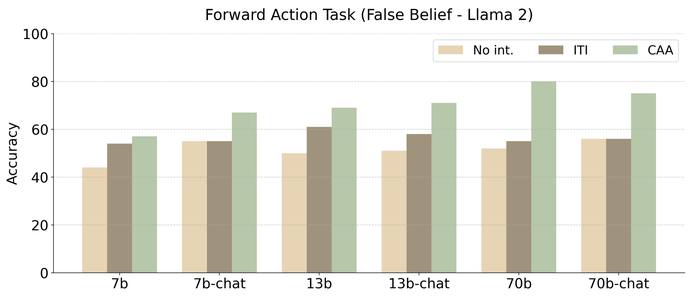
Activation Editing

RQ5. 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." *Advances in Neural Information Processing Systems* 36 (2024).
[3] Zhu, Wentao, Zhining Zhang, and Yizhou Wang. "Language Models Represent Beliefs of Self and Others." *Forty-first International Conference on Machine Learning*.
[4] Rimsky, Nina, et al. "Steering llama 2 via contrastive activation addition." *arXiv preprint arXiv:2312.06681* (2023).