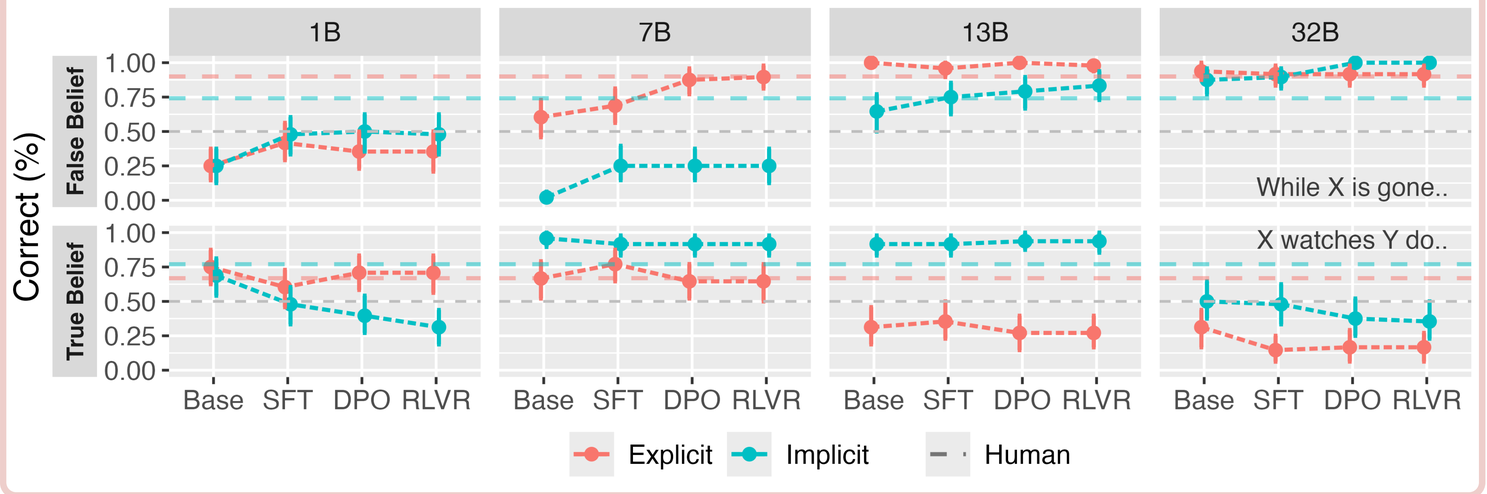
Contributions:

→ Model **scale benefits FBT performance**, but not **strictly**. False Belief scenarios benefit from scale but True Belief scenarios do not.

→ Models are sensitive to two interacting factors: (1) **learned associations** with stereotypical False Belief scenarios, and (2) the presence of **explicit propositional attitudes** themselves.

→ Post-training (SFT, DPO) improves FBT performance overall, but the gains are modulated by stronger interaction effects.

→ **Vector steering** suggests that answer **patterns** are influenced by information encoded in the **propositional attitude** verb 'to **think**'.



FBT performance of LLMs shows **nuanced associations** with **scale** and **post-training**, but isn't robust to **False vs. True** and **explicit vs. implicit variations**. Learned scenario patterns and information contained in the **propositional attitude verb 'to think'** interferes with genuine social reasoning, sometimes productively and sometimes in detrimental ways.

Traces of Social Competence in Large Language Models

Universiteit Leiden
The Netherlands

S I M L A B

Tom Kouwenhoven, **Michiel van der Meer,** Max van Duijn

Correspondence: t.kouwenhoven@liacs.leidenuniv.nl

Method:

Systematic comparison for a total of **17** open-source and open-weight **model variants** based on computed probabilities (Pimentel and Meister, 2024) of completions of 192 FBT scenario (Trott et al. (2023). The location word with the **highest probability** acts as the model's prediction.

In case study using OLMo 2, we verify that the stimuli are **not present in the training data** and apply **vector steering** (Subramani et al., 2022; Rimsky et al., 2024) to develop a new approach for distinguishing between **different drivers** of observed FBT performance or failure.

Analyses through Bayesian Regression Models:

```
correct ~ model_size * variant
          * knowledge_state * knowledge_cue
          + (1 | model_family_version)
```

The False Belief Task

Premise
Ed arrives home after a long day at work. He puts his keys in the hall and leaves his bag in the study. Seana arrives home a few minutes later.

TRUE BELIEF

Ed watches as Seana moves the keys from the hall to the study. Afterwards, Ed goes to the bathroom.

FALSE BELIEF

Afterwards, Ed goes to the bathroom. **Ed doesn't see Seana move the keys from the hall to the study.**

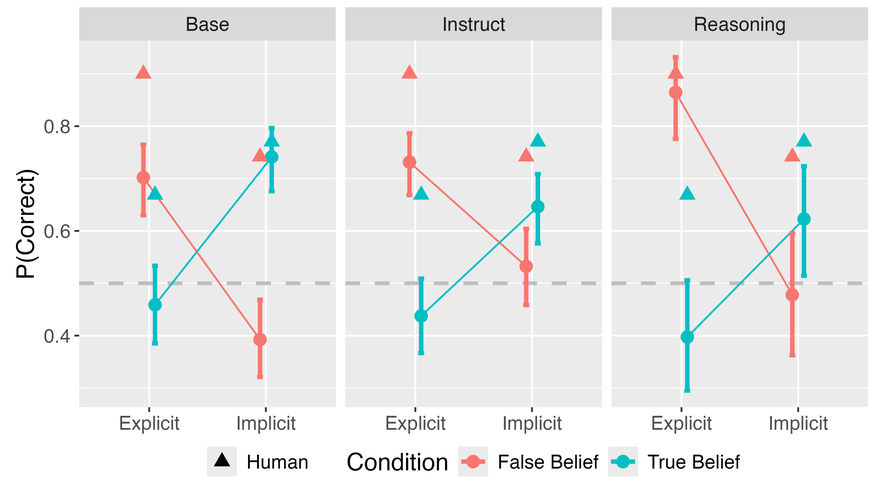
When Ed gets back from the bathroom, he realizes he needs his keys.
Implicit Cue: "Ed goes to get the keys from the... (*study / hall*)"
Explicit Cue: "Ed thinks the keys are in the... (*study / hall*)"

Steering vectors: difference in activations between paired scenarios. We inject scaled (λ) vectors during inference. λ = 1 pushes toward explicit-like behaviour, λ = −1 favours implicit-like behaviour.

Prositional attitudes

Implicit cues ('X goes to') impair FB performance but facilitate TB performance.

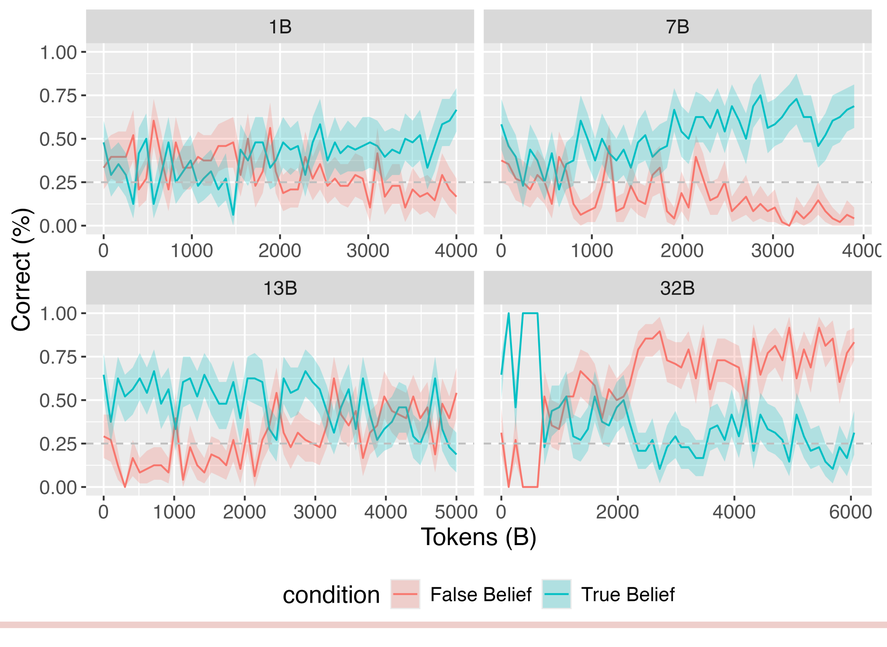
The other way around, explicit cues ('X thinks') facilitate FB but hurt TB performance.



Traces

Traces of OLMo 2 pre-training. TB and FB scores differ and show a cross-pattern.

OLMo 2 may acquire stereotypical responses tied to mental-state vocabulary that outweig scenario semantics.



Vector steering

Steering at layers 10 to 12 is effective. The propositional attitude verb 'to think', likely contains epistemic information and can influence predictions.