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

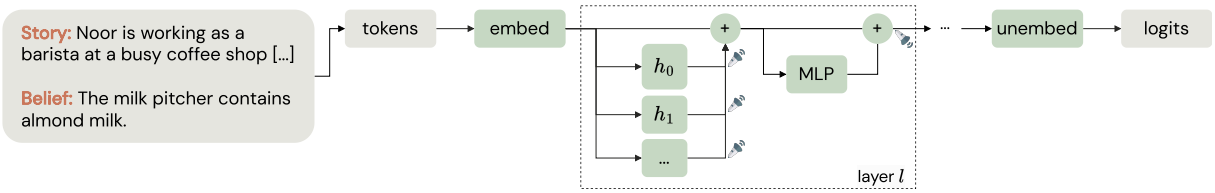



Figure 6: Given a tokenised input, we cache the internal activations for all attention heads $h_i$, $i = 0, \ldots, H-1$, and residual streams. In our experiments, we use residual stream activations.

| 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, $\sigma_l^h$ is the standard deviation of activations along the direction identified by the probe trained on attention head $h$ at layer $l$, and $\theta_l^h$ 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

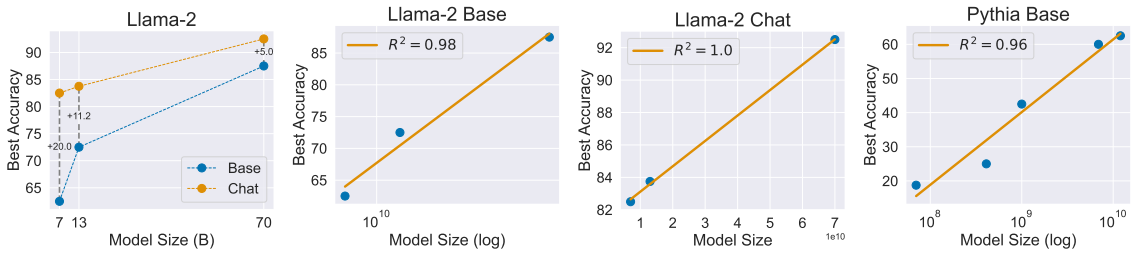
Our code is provided as supplementary material and it will be made public under the MIT licence at `www.this-is-a-placeholder.com`.

### 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 over-interpreting 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.