Testing Simulation Theory in LLMs' Theory of Mind

Koshiro Aoki, Daisuke Kawahara
Waseda University, Tokyo, Japan
aokikoshiro@akane.waseda.jp, dkw@waseda.jp

Abstract

Theory of Mind (ToM) is the ability to understand others' mental states, which is essential for human social interaction. Although recent studies suggest that large language models (LLMs) exhibit human-level ToM capabilities, the underlying mechanisms remain unclear. "Simulation Theory" posits that we infer others' mental states by simulating their cognitive processes, which has been widely discussed in cognitive science. In this work, we propose a framework for investigating whether the ToM mechanism in LLMs is based on Simulation Theory by analyzing their internal representations. Following this framework, we successfully steered LLMs' ToM reasoning through modeled perspective-taking and counterfactual interventions. Our results suggest that Simulation Theory may partially explain the ToM mechanism in state-of-the-art LLMs, indicating parallels between human and artificial social reasoning.

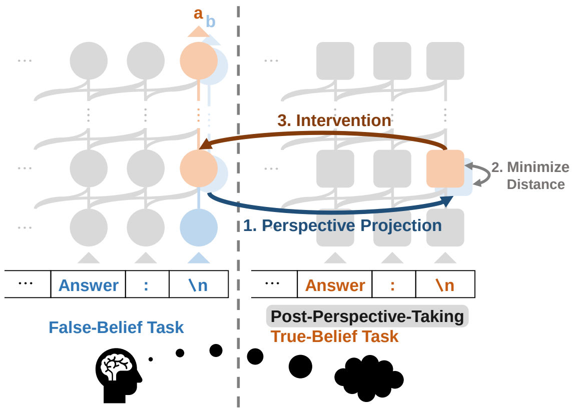


Figure 1: A schematic diagram of our experiment. Gray circles and squares denote the LLM's internal representations across layers. We intervene in the internal representation while the LLM is solving the false-belief task so that its perspective-projected representation approaches the representation of the post-perspective-taking true-belief task. We then observe changes in the answer.

1 Introduction

For large language models (LLMs) to communicate smoothly with users, they need to understand the users' knowledge, intentions, beliefs, and desires. This capability to infer the mental states of others is called Theory of Mind (ToM). ToM is pivotal for social interactions such as communication (Milligan et al., 2007), moral judgment (Moran et al., 2011), and cooperation (Markiewicz et al., 2024; Li et al., 2023a). One prominent account of ToM in cognitive science and psychology is **Simulation Theory** (Gordon, 1986), which posits that we understand others' minds by simulating their cognitive processes. This process of adopting the viewpoint of others is called **perspective-taking**, a foundational ability under Simulation Theory (Barlassina and Gordon, 2017). Such simulation need not be explicit; for instance, mirror neurons (Gallese and Goldman, 1998) activate both when performing an action and when observing someone else perform it, suggesting an implicit simulation process.

Meanwhile, recent work has found that some LLMs acquire ToM abilities comparable to those of humans (Strachan et al., 2024; Kosinski, 2024; Street et al., 2024). At the same time, the robustness of many ToM tests has been questioned, and there is ongoing debate about whether current models genuinely possess ToM or merely exploit artifacts of these benchmarks (Ullman, 2023; Shapira et al., 2024). This debate emphasizes the importance of not only evaluating their behavioral performance but also investigating the underlying mechanisms (Hu et al., 2025). Nevertheless, the mechanism of ToM in LLMs, particularly its relationship to Simulation Theory, remains poorly understood. In this work, we investigate whether the internal representations of LLMs align with Simulation Theory by proposing a framework for modeling perspective-taking. We use counterfactual inter-

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 96–104
December 20-24, 2025 ©2025 Association for Computational Linguistics

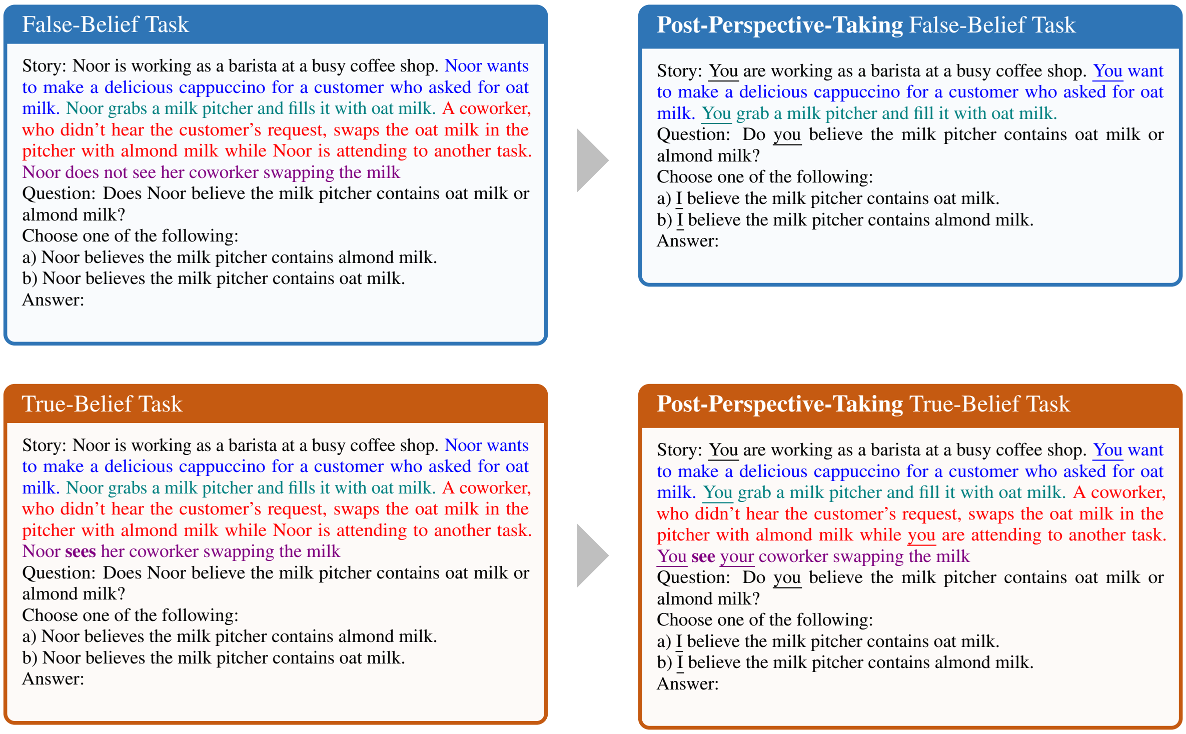


Figure 2: Examples of false-belief and true-belief tasks from the BigToM benchmark and their corresponding post-perspective-taking versions. (**Top Left**) A false-belief task consists of five sentences: *Context*, *Desire*, *Action*, *Causal Event*, and *Percept*. (**Top Right**) The post-perspective-taking false-belief task removes information unknown to the protagonist and rewrites the text in second/first person. (**Bottom Left**) A true-belief task differs from false-belief only in the *Percept*, where the protagonist is aware of the *Causal Event*. (**Bottom Right**) The post-perspective-taking true-belief task retains all sentences and rewrites them in second/first person.

ventions in these internal representations to assess their causal effect on the model's outputs. Figure 1 shows an overview of our experiment.

2 Related Work

Some studies have shown that internal representations in LLMs encode information about beliefs, especially for dissociating reality from false belief (Zhu et al., 2024; Bortoletto et al., 2024; Jamali et al., 2023). While these analyses suggest the presence of ToM-relevant structures, they do not establish explicit links to Simulation Theory.

3 Setup for Verifying Simulation Theory in LLMs

Model. We evaluate two instruction-tuned LLMs: Llama-3.1-70B-Instruct (Grattafiori et al., 2024) and Qwen2.5-72B-Instruct (Qwen et al., 2024). Both are Transformer-based autoregressive language models with 80 Transformer blocks. We set the temperature to 0 to ensure deterministic outputs.

Dataset. In this work, we use the false-belief tasks from the social reasoning benchmark BigToM (Gandhi et al., 2023). A false-belief task assesses whether an individual recognizes that others may hold beliefs different from their own, serving as a test for ToM. As shown in Figure 2, each BigToM benchmark item comprises five elements: *Context*, *Desire*, *Action*, *Causal Event*, and *Percept*. We also use the true-belief tasks from BigToM. The false-belief and true-belief tasks are identical except for the *Percept*. In a false-belief task, the *Percept* contains information indicating that the protagonist is unaware of the *Causal Event*. In contrast, the *Percept* in a true-belief task indicates that the protagonist is aware of the *Causal Event*.

Data Preprocessing. We split the false-belief tasks which the LLMs answered correctly[1] into training and test subsets at a ratio of 8:2. The training tasks are used to train the perspective projection

[1] Out of 200 questions, Llama-3.1-70B-Instruct answered 198 correctly, and Qwen2.5-72B-Instruct answered 196 correctly.

(§ 4.3), and the test tasks are reserved for the intervention experiments (§ 4.4).

4 Framework for Testing Simulation Theory in LLMs

Simulation Theory posits a two-step process for inferring others' mental states:

1. **Perspective-Taking:** Simulate being in another person's situation.
2. **Attribution:** Infer their mental state from that simulation.

We adapt these steps for LLMs as follows:

1. **Modeling Perspective-Taking:** We generate **post-perspective-taking (PPT) tasks** to simulate the LLM "stepping into others' shoes" (§ 4.1). Using the internal representations when the LLM solves the PPT tasks (§ 4.2), we train a linear transformation called **perspective projection** that projects the representations within the LLM into a hypothetical perspective-taking space, thereby modeling perspective-taking (§ 4.3).
2. **Testing Mental State Attribution:** We perform counterfactual interventions in the internal representations to test if the encoded PPT representations are used for ToM reasoning (§ 4.4).

Here, the internal representation refers to the residual stream, which denotes the output of each Transformer block in this paper.

4.1 Generating Post-Perspective-Taking Tasks

To model perspective-taking, we need the internal representation of the situation in which another person's perspective is replaced with the model's own. To derive this representation, we generate input texts, which we call **post-perspective-taking (PPT) tasks**. Specifically, we generate two types of PPT tasks, a PPT false-belief task and a PPT true-belief task.

As shown in Figure 2, each PPT task is generated by applying the following transformations to a false-belief or true-belief task:

1. Remove the information unknown to the protagonist from the original story. That is, for a false-belief task, remove the *Causal Event* and *Percept* (two sentences); for a true-belief task, keep all sentences unchanged.
2. Change the protagonist's name to the second person ("you/your") in the remaining story and question, and to the first person ("I/me/my") in the choices to make the protagonist's perspective the LLM's own[2].

From these steps, we obtain a dataset $\{(f_i, p_i, \widetilde{p}_i)\}_{i=1}^{N}$, where N is the dataset size, f_i denotes a false-belief task, p_i is the corresponding PPT false-belief task, and $\widetilde{p}_i$ is the PPT true-belief task.

4.2 Extracting Internal Representations

Next, we run the LLM on each task f_i, p_i, and $\widetilde{p}_i$ and extract the residual stream at the same specific layer for the final token position. We also prepare a variant with reversed choice ordering for the PPT false-belief and PPT true-belief tasks and take the average of the resulting residual streams across the original and reversed versions. This averaging ablates the information about choice symbols ("a", "b") from the representations.

Let $\boldsymbol{x}_i, \boldsymbol{y}_i, \widetilde{\boldsymbol{y}}_i \in \mathbb{R}^d$ (d is the residual stream dimension) denote the representations of f_i, p_i, and $\widetilde{p}_i$, respectively. The PPT false-belief representation $\boldsymbol{y}_i$ serves as the gold standard for the perspective projection (§ 4.3), while the PPT true-belief representation $\widetilde{\boldsymbol{y}}_i$ is used for intervention (§ 4.4).

4.3 Perspective Projection

According to Simulation Theory, if the model simulates others' minds through perspective-taking, then the internal representation when observing another's situation should contain the internal representation that would occur if one were in the same situation as that person. To verify this hypothesis, we train a linear transformation[3] to map $\boldsymbol{x}_i$ (the false-belief representation) to $\boldsymbol{y}_i$ (the PPT false-belief representation). We call this linear transformation **perspective projection**.

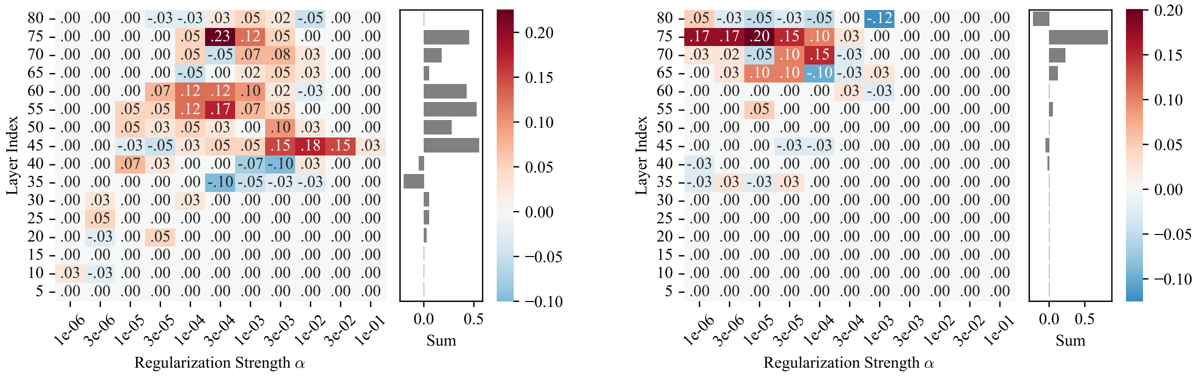
We derive the weight matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ of perspective projection by solving a ridge regression problem using input data $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^\top$ and target data $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)^\top$ as follows:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \left\{ \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_\mathrm{F}^2 + \lambda \|\boldsymbol{W}\|_\mathrm{F}^2 \right\} \quad (1)$$

$$= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \quad (2)$$

where λ is the regularization strength. We set $\lambda =$ 1e-4 in our experiments based on cross-validation.

[2] We use gpt-4o-mini-2024-07-18 for these transforma-



(a) Llama-3.1-70B-Instruct

(b) Qwen2.5-72B-Instruct

Figure 3: Net intervention effect across model layers and regularization strengths. The heatmap shows the difference in proportions of flipped answers between true-belief and false-belief interventions (true-belief $-$ false-belief). The bar plot on the right shows the sum of the difference in each layer.

4.4 Counterfactual Representation Intervention

Perspective projection can show correlation but not causation the between PPT representation and the LLM's answers. Simulation Theory requires, however, a causal link where the PPT representation is used to attribute mental states to others. We, therefore, perform counterfactual interventions (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Li et al., 2023b; Ghandeharioun et al., 2024) in the LLM's internal representations to test whether the PPT representations are indeed used in ToM reasoning.

True-Belief Intervention. As illustrated in Figure 1, we update the false-belief representation $\boldsymbol{x}_i$ such that its projection with $\boldsymbol{W}$ approaches the PPT true-belief representation $\widetilde{\boldsymbol{y}}_i$. We compute the updated representation $\widetilde{\boldsymbol{x}}_i$ by solving:

$$\widetilde{\boldsymbol{x}}_i = \arg\min_{\boldsymbol{x}} \left\{ \|\boldsymbol{W}\boldsymbol{x} - \widetilde{\boldsymbol{y}}_i\|_2^2 + \alpha \|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2 \right\} \tag{3}$$

$$= (\boldsymbol{W}^\top \boldsymbol{W} + \alpha \boldsymbol{I})^{-1} (\boldsymbol{W}^\top \widetilde{\boldsymbol{y}}_i + \alpha \boldsymbol{x}_i), \tag{4}$$

where α is the regularization strength to avoid ill-posed problems in which the updated representation diverges drastically from the original. If the LLM uses the PPT representation for ToM reasoning, then after this intervention, the LLM's response to the false-belief task should flip from the false-belief choice to the true-belief choice (e.g., "b" $\rightarrow$ "a").

False-Belief Intervention. We also perform a control experiment where we replace $\widetilde{\boldsymbol{y}}_i$ (the PPT true-belief representation) with $\boldsymbol{y}_i$ (the PPT false-belief representation) to study how the error in perspective projection affects the intervention. Ideally, intervening with $\boldsymbol{y}_i$ should produce little change in the model's answer if perspective projection generalizes well to the test data.

Net Intervention Effect. Finally, for each layer l and regularization strength α, we compute $\text{Flip}_{\text{true}}(l, \alpha) - \text{Flip}_{\text{false}}(l, \alpha)$ as the "net intervention effect," where $\text{Flip}_{\text{true}}$ and $\text{Flip}_{\text{false}}$ represent the proportion of tasks where the model's answer flips to the true-belief choice under the true-belief and false-belief intervention, respectively.

5 Results

Layer-wise Intervention Effect. Figure 3 presents the results of the net intervention effect. In both Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, the effect increases in the later layers. This suggests that these later layers encode perspective-taking information, i.e., representations of the simulated others' mental states.

Effect of Regularization Strength. Figure 4 illustrates the effect of the regularization strength α on the intervention. The intervention, which is an inverse and ill-posed problem, causes catastrophic interference when α is excessively small

tions.

[3]This linear transformation approach is grounded in the linear representation hypothesis (Elhage et al., 2022; Park et al., 2024). Based on this hypothesis, we assume that two internal representations share a common linear subspace. Hence, these internal representations can be mapped to each other through an appropriate linear transformation.

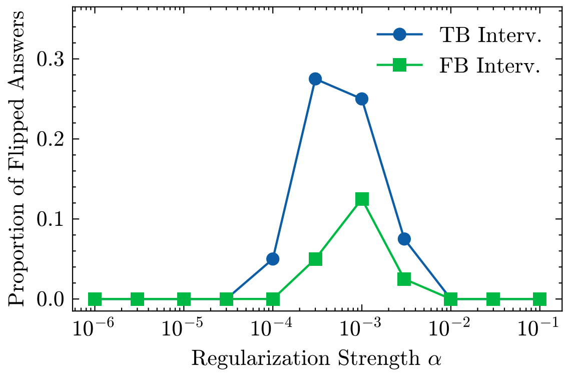


Figure 4: The proportion of tasks where the Llama's answer flips from the false-belief to the true-belief choice under intervention in the 75th layer. The "TB Interv." line shows the result of the intervention with the PPT true-belief representation; the "FB Interv." line shows the result with the PPT false-belief representation.

($\alpha \leq 10^{-4}$). This leads the model to output a token irrelevant to the choice symbols ("a", "b"), resulting in a low flip proportion. Conversely, when α is excessively large ($\alpha \geq 10^{-2}$), the intervention becomes too weak to change the model's response. As a result, the flip proportion reaches its maximum when α is between 10^{-4} and 10^{-2}.

6 Conclusion

In this work, we developed a framework for investigating whether LLMs' Theory of Mind aligns with Simulation Theory. Applying this framework to Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, we found evidence that later layers may encode representations consistent with perspective-taking. This suggests that Simulation Theory may partially explain the ToM mechanism in state-of-the-art LLMs.

Limitations

Potential Nonlinear Representations. We assumed a linear transformation to model perspective-taking. This is motivated by the linear representation hypothesis (Elhage et al., 2022; Park et al., 2024). However, mental-state representations could be distributed nonlinearly because some nonlinear representations have also been found (Engels et al., 2025). Our linear approach may therefore capture only a subset of the structures underlying ToM reasoning.

Limited Net Intervention Effect. The maximum net intervention effect observed in our experiments is still relatively small compared to the ideal value of 1, which would indicate perfect alignment with Simulation Theory. While our results suggest that Simulation Theory partially explains the ToM mechanism in LLMs, we cannot claim that it fully accounts for the mechanism. The model may use additional mechanisms for ToM reasoning, such as heuristics (Nikankin et al., 2025; Shapira et al., 2024).

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP24H00727.

References

Luca Barlassina and Robert M. Gordon. 2017. Folk Psychology as Mental Simulation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2017 edition. Metaphysics Research Lab, Stanford University.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024. Benchmarking mental state representations in language models. *arXiv preprint arXiv:2406.17513*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are linear. In *The Thirteenth International Conference on Learning Representations*.

Vittorio Gallese and Alvin Goldman. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501.

Kanishk Gandhi, J-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 13518–13529.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Robert M Gordon. 1986. Folk psychology as simulation. *Mind & language*, 1(2):158–171.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, 380(1932):20230499.

Mohsen Jamali, Ziv M. Williams, and Jing Cai. 2023. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. *arXiv preprint arXiv:2309.01660*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023a. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 41451–41530.

Roksana Markiewicz, Foyzul Rahman, Ian Apperly, Ali Mazaheri, and Katrien Segaert. 2024. It is not all about you: Communicative cooperation is determined by your partner's theory of mind abilities as well as your own. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(5):833–844.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. 2007. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2):622–646.

Joseph M. Moran, Liane L. Young, Rebecca Saxe, Su Mei Lee, Daniel O'Young, Penelope L. Mavros, and John D. Gabrieli. 2011. Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, 108(7):2688–2692.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *The Thirteenth International Conference on Learning Representations*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *Preprint*, arXiv:2302.08399.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62638–62681. PMLR.

A Prompts for Generating Post-Perspective-Taking Tasks

Below is a template of the prompts used to convert the original text to second-person or first-person narratives. Here, `{{text}}` is replaced with the text to be converted, and `{{protagonist_name}}` is replaced with the protagonist's name.

Prompt for converting story and question to second person

```
Text: {{text}}
Change "{{protagonist_name}}" to "you/your" in this text to make it second-person. Pay attention to verb conjugation and grammar to ensure the text is grammatically correct. Output only the converted text.
```

Prompt for converting multiple-choice options to first person

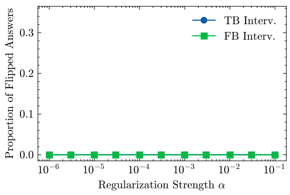
```
Text: {{text}}
Change "{{protagonist_name}}" to "I/me/my" in this text to make it first-person. Pay attention to verb conjugation and grammar to ensure the text is grammatically correct. Output only the converted text.
```

B Connection to Mirror Neurons

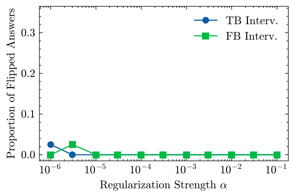
Perspective projection is inspired by mirror neurons, which respond similarly when performing an action and when observing another individual perform that action (Gallese and Goldman, 1998). Mirror neuron studies, however, focus on local neuronal activity correlations, whereas our approach considers linear correspondences across entire layers of neuron activations in an LLM.

C Flip Proportion for Each Layer

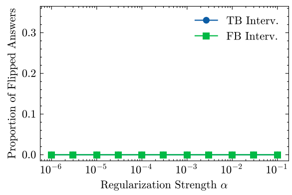
Figures 5 and 6 show the proportion of tasks where the LLM's answer flips from the false-belief to the true-belief choice under intervention



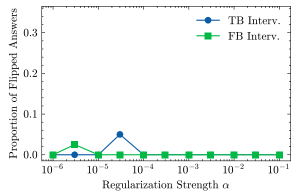
(a) Layer 5



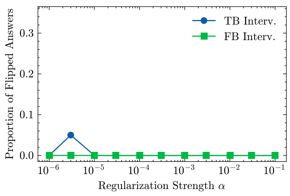
(b) Layer 10



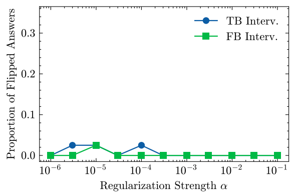
(c) Layer 15



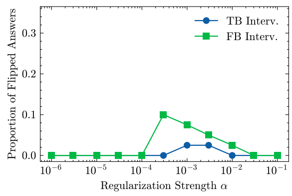
(d) Layer 20



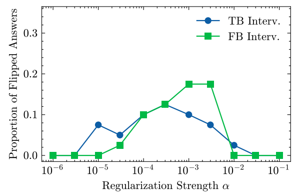
(e) Layer 25



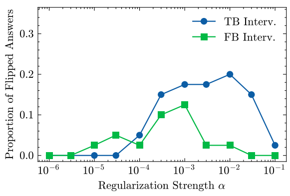
(f) Layer 30



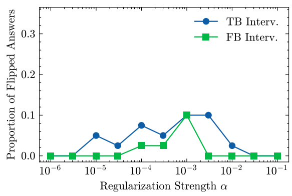
(g) Layer 35



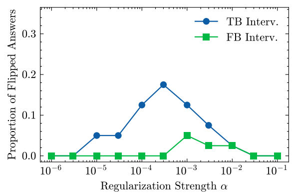
(h) Layer 40



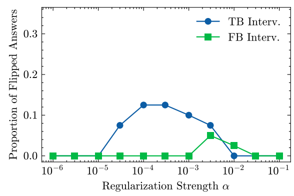
(i) Layer 45



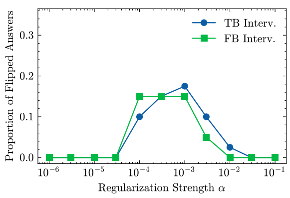
(j) Layer 50



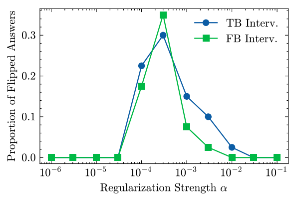
(k) Layer 55



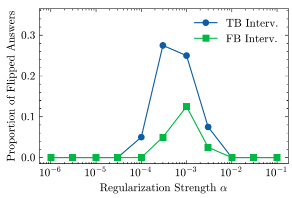
(l) Layer 60



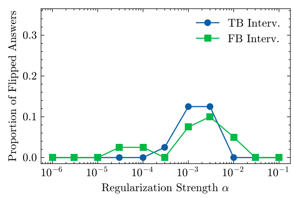
(m) Layer 65



(n) Layer 70

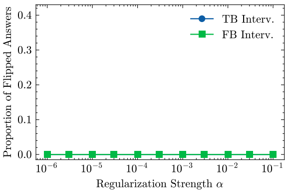


(o) Layer 75

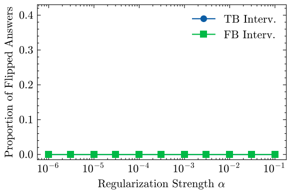


(p) Layer 80

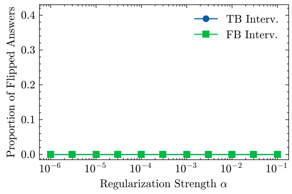
Figure 5: Proportion of flipped answers for layers 5 through 80 under intervention in Llama-3.1-70B-Instruct. The "TB Interv." line shows the result of the intervention with the PPT true-belief representation; the "FB Interv." line shows the result with the PPT false-belief representation.



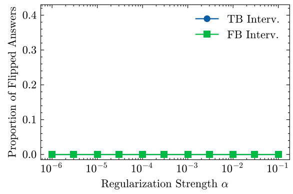
(a) Layer 5



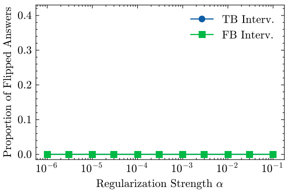
(b) Layer 10



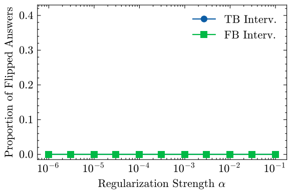
(c) Layer 15



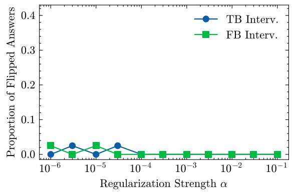
(d) Layer 20



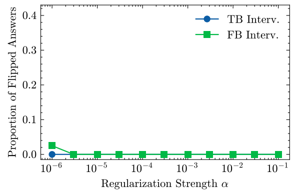
(e) Layer 25



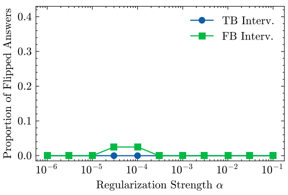
(f) Layer 30



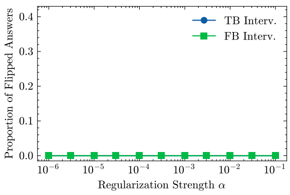
(g) Layer 35



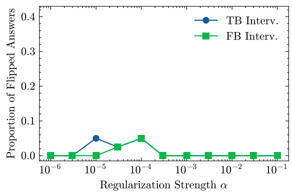
(h) Layer 40



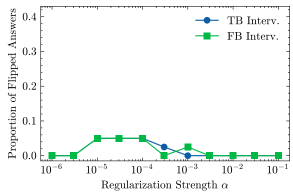
(i) Layer 45



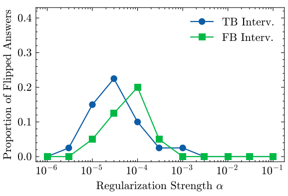
(j) Layer 50



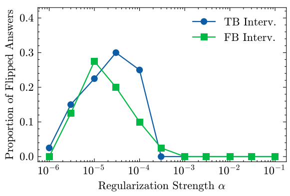
(k) Layer 55



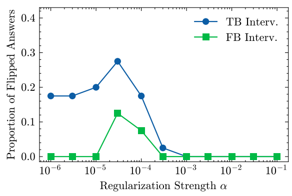
(l) Layer 60



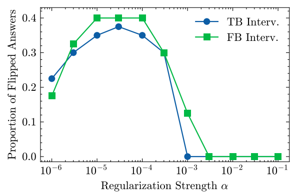
(m) Layer 65



(n) Layer 70



(o) Layer 75



(p) Layer 80

Figure 6: Proportion of flipped answers for layers 5 through 80 under intervention in Qwen2.5-72B-Instruct (see Figure 5 for a more detailed explanation).