Do Large Language Models Simulate Minds? A Representation Analysis of Theory of Mind

Anonymous ACL submission

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 controlled LLMs' ToM reasoning through modeled perspective-taking and counterfactual interventions. Our results provide initial evidence that state-of-the-art LLMs implement an emergent ToM partially based on Simulation Theory, suggesting parallels between human and artificial social reasoning.

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.

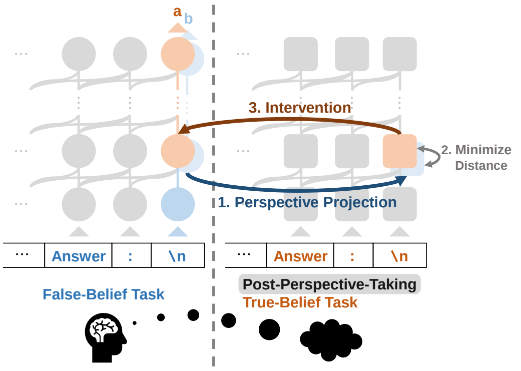


Figure 1: A schematic diagram of our intervention. Gray circles and squares denote the LLM's internal representations across layers. We intervene in the internal representation while the LLM is solving the false-belief task so that its perspective-projected representation becomes closer to that of the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

Meanwhile, recent work finds that some LLMs have acquired ToM abilities comparable to that of humans (Strachan et al., 2024; Kosinski, 2024; Street et al., 2024). 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 interventions in these internal representations to assess their causal effect on the model's outputs. An overview of our intervention process is illustrated in Figure 1.

2 Related Work

Several studies have examined whether LLMs can solve false-belief tasks and other ToM-related tasks, revealing high levels of performance on certain benchmarks (Strachan et al., 2024; Kosinski, 2024;

Street et al., 2024). These findings imply that LLMs encode latent structures analogous to human ToM. Moreover, Wilf et al. (2024) have shown that explicitly prompting an LLM to take others' perspective, based on Simulation Theory, can improve its ToM performance. However, these studies focus on the model's behavior and do not investigate its internal mechanisms.

Recently, some studies have shown that internal representations in LLMs encode information about beliefs, especially for tracking reality versus false beliefs (Zhu et al., 2024; Bortoletto et al., 2024; Jamali et al., 2023). While these analyses hint at the presence of ToM-relevant structures, they do not draw strong connections to Simulation Theory.

3 Setup for Verifying Simulation Theory in LLMs

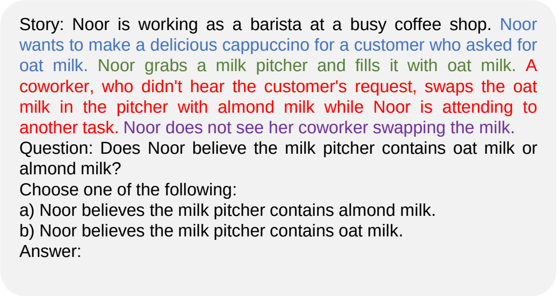
Model. The LLM used in this study is Llama-3.1-70B-Instruct (Grattafiori et al., 2024). This is a Transformer-based autoregressive language model with 80 Transformer blocks. We set the temperature to 0 to ensure deterministic outputs.

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. From the BigToM benchmark, we select 198 of the 200 false-belief tasks which Llama-3.1-70B-Instruct answered correctly. We split this into training and test subsets at a ratio of 8:2. The training tasks are used to train the perspective projection (§ 4.3), and the test tasks are reserved for the intervention experiments (§ 4.4).

4 Framework for Verifying Simulation Theory in LLMs

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



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.
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) Noor believes the milk pitcher contains almond milk.
b) Noor believes the milk pitcher contains oat milk.
Answer:

Figure 2: An example of a false-belief task from the BigToM benchmark. Each false-belief story consists of five sentences in the following order: *Context*, *Desire*, *Action*, *Causal Event*, and *Percept*. A question and two choices follow these sentences. The model's response appears after "Answer:\n." In this example, "b" is based on a false belief and thus the correct answer, though the order of the choices is randomized in each task.

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

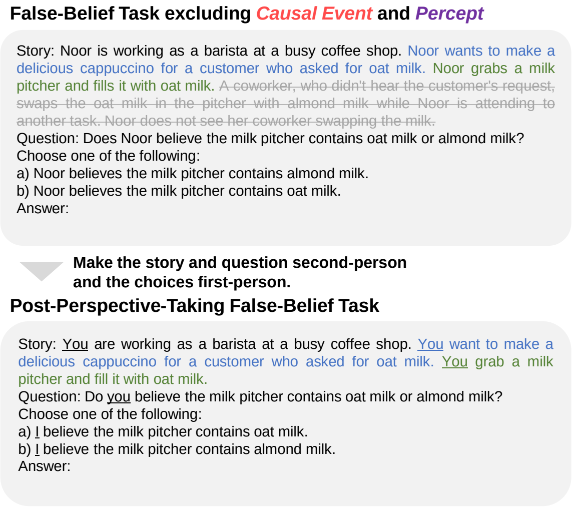
We adapt these steps for LLMs as follows:

1. **Modeling Perspective-Taking:** We generate **post-perspective-taking (PPT) tasks** to simulate the LLM "stepping into others' shoes" (§ 4.1). Using the internal representation 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 whether 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. We call the text used for deriving this representation the **post-perspective-taking (PPT) task**. Specifically, we generate two types of PPT tasks, a PPT false-belief task and a PPT true-belief task.



False-Belief Task excluding *Causal Event* and *Percept*

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.~~
Question: Does Noor believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) Noor believes the milk pitcher contains almond milk.
b) Noor believes the milk pitcher contains oat milk.
Answer:

Make the story and question second-person and the choices first-person.

Post-Perspective-Taking False-Belief Task

Story: You are working as a barista at a busy coffee shop. You want to make a delicious cappuccino for a customer who asked for oat milk. You grab a milk pitcher and fill it with oat milk.
Question: Do you believe the milk pitcher contains oat milk or almond milk?
Choose one of the following:
a) I believe the milk pitcher contains oat milk.
b) I believe the milk pitcher contains almond milk.
Answer:

Figure 3: Overview of generating post-perspective-taking tasks. We remove sentences containing information that the protagonist does not know and then rewrite the text from a protagonist's perspective to a second/first-person perspective so that the other person's situation is simulated as the reader's own.

As shown in Figure 3, 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 LLM's own[1].

From these steps, we obtain a dataset of size N

$$\{(f_1, p_1, \widetilde{p}_1), \ldots, (f_N, p_N, \widetilde{p}_N)\},$$

where each triple consists of a false-belief task f_i, the corresponding PPT false-belief task p_i, and a PPT true-belief task $\widetilde{p}_i$.

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 tasks and take the average of the resulting residual streams across the original and reversed versions. This averaging mitigates ordering biases in the choices.

Let $\boldsymbol{x}_i, \boldsymbol{y}_i, \widetilde{\boldsymbol{y}}_i \in \mathbb{R}^d$ denote the representations for f_i, p_i, and $\widetilde{p}_i$, respectively. Here, d is the dimension of the residual stream. The PPT false-belief representation $\boldsymbol{y}_i$ is used as the gold standard data 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[2] that takes $\boldsymbol{x}_i$ (the false-belief representation) as input and predicts $\boldsymbol{y}_i$ (the PPT false-belief representation), similar to the approaches of probing (Alain and Bengio, 2017; Belinkov, 2022). 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:

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

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

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

4.4 Counterfactual Representation Intervention

Perspective projection can show correlation but not causation between PPT representation and 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.

[1] We use `gpt-4o-mini-2024-07-18` for these transformations.

[2] 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.

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

$$\begin{aligned} \widetilde{\boldsymbol{x}}_i &= \arg\min_{\boldsymbol{x}} \left\{ \|\boldsymbol{x}\boldsymbol{W} - \widetilde{\boldsymbol{y}}_i\|_2^2 + \alpha\|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2 \right\} && (3) \\ &= \left(\widetilde{\boldsymbol{y}}_i \boldsymbol{W}^\top + \alpha \boldsymbol{x}_i\right)\left(\boldsymbol{W}\boldsymbol{W}^\top + \alpha \boldsymbol{I}\right)^{-1}, && (4) \end{aligned}$$

where α is a 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” → “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. Intervening with $\boldsymbol{y}_i$ should produce little change in the model's final answer if perspective projection generalizes well to the test data.

Net Intervention Effect. Finally, for each layer l and regularization strength α, we compute the “net intervention effect” as:

$$\mathrm{Flip}_{\mathrm{true}}(l, \alpha) - \mathrm{Flip}_{\mathrm{false}}(l, \alpha),$$

where $\mathrm{Flip}_{\mathrm{true}}$ and $\mathrm{Flip}_{\mathrm{false}}$ 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 Analysis. Figure 4 indicates that the net intervention effect increases in the later layers. This suggests these later layers encode perspective-taking information, i.e., representations of the simulated others' mental states.

Effect of Regularization Strength. Figure 5 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 ($\alpha \leq 10^{-4}$). 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}.

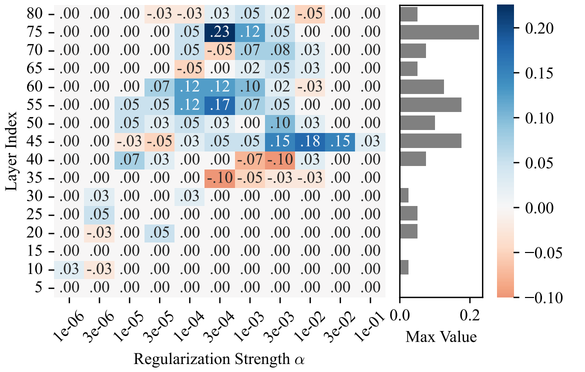


Figure 4: Net intervention effect across model layers and regularization strengths. The heatmap shows the difference in proportions of flipped answers between true-belief and false-belief intervention (true-belief − false-belief). The bar plot on the right shows the maximum difference in each layer.

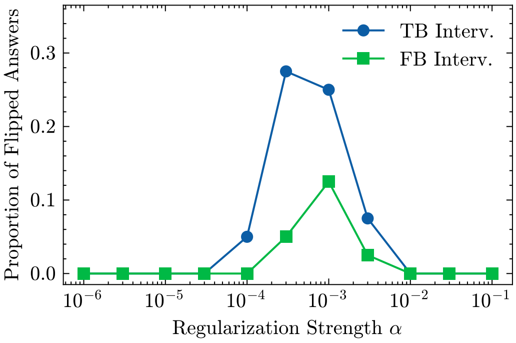


Figure 5: The proportion of tasks where the LLM's answer flips from the false-belief to the true-belief choice under intervention in the 75th layer. The “TB Interv.” line shows the result of the intervention with the PPT true-belief representation; the “FB Interv.” line shows the result with the PPT false-belief representation.

6 Conclusion

In this work, we developed a framework for investigating whether the LLMs' Theory of Mind aligns with Simulation Theory. As a result of applying this framework to Llama-3.1-70B-Instruct, we found that the later layers encode representations of the simulated mental states of others. This suggests that state-of-the-art LLMs have acquired a Theory of Mind partially based on Simulation Theory. The proposed framework can be applied to future, more powerful LLMs and will also provide insights into ToM mechanisms in these 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.

Scope of Evaluation. Our study primarily focuses on false-belief tasks within a single benchmark (BigToM) and experiments on a single model (Llama-3.1-70B-Instruct). Although false-belief tasks are standard in assessing ToM, they represent only a narrow slice of real-world social reasoning. Extending our approach to more diverse models and tasks (e.g., second-order beliefs, deception detection, or cooperative tasks) could provide a more comprehensive view of the ToM capabilities of LLMs.

Limited Net Intervention Effect. The maximum net intervention effect observed in Llama-3.1-70B-Instruct 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 Llama-3.1-70B-Instruct, we cannot claim that it fully accounts for the mechanism. The model may utilize additional mechanisms for ToM reasoning.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

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.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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.

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*.

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.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

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

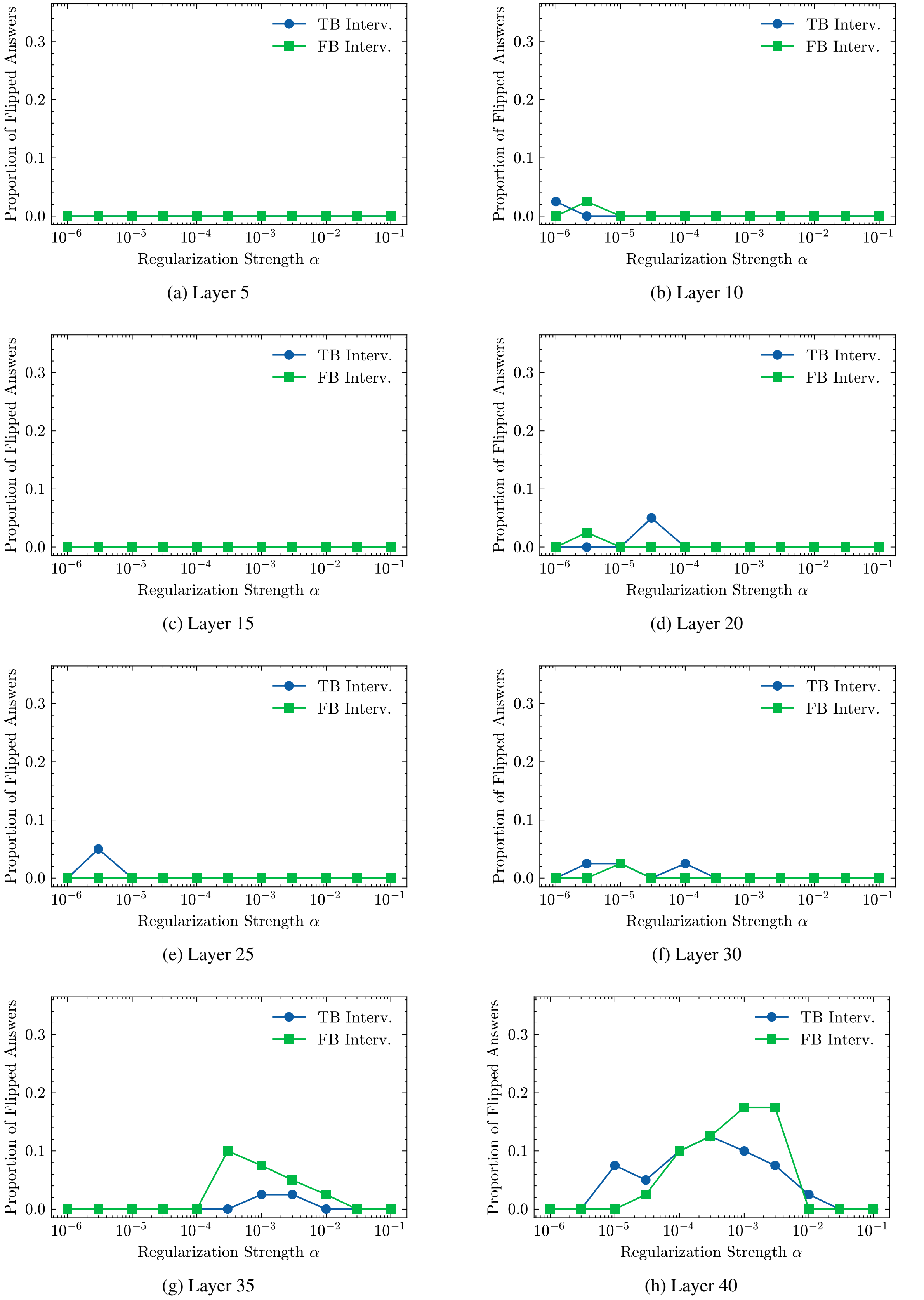
```
Text: {{text}}
Change “{{protagonist_name}}” to “I/me/my” in this text to make it first-person. Pay attention to verb conjugation and grammar to ensure the text is grammatically correct. Output only the converted text.
```

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 6 and 7 show the flip proportions for layers 5 through 80 besides layer 75, which was presented in Figure 5.



(a) Layer 5

(b) Layer 10

(c) Layer 15

(d) Layer 20

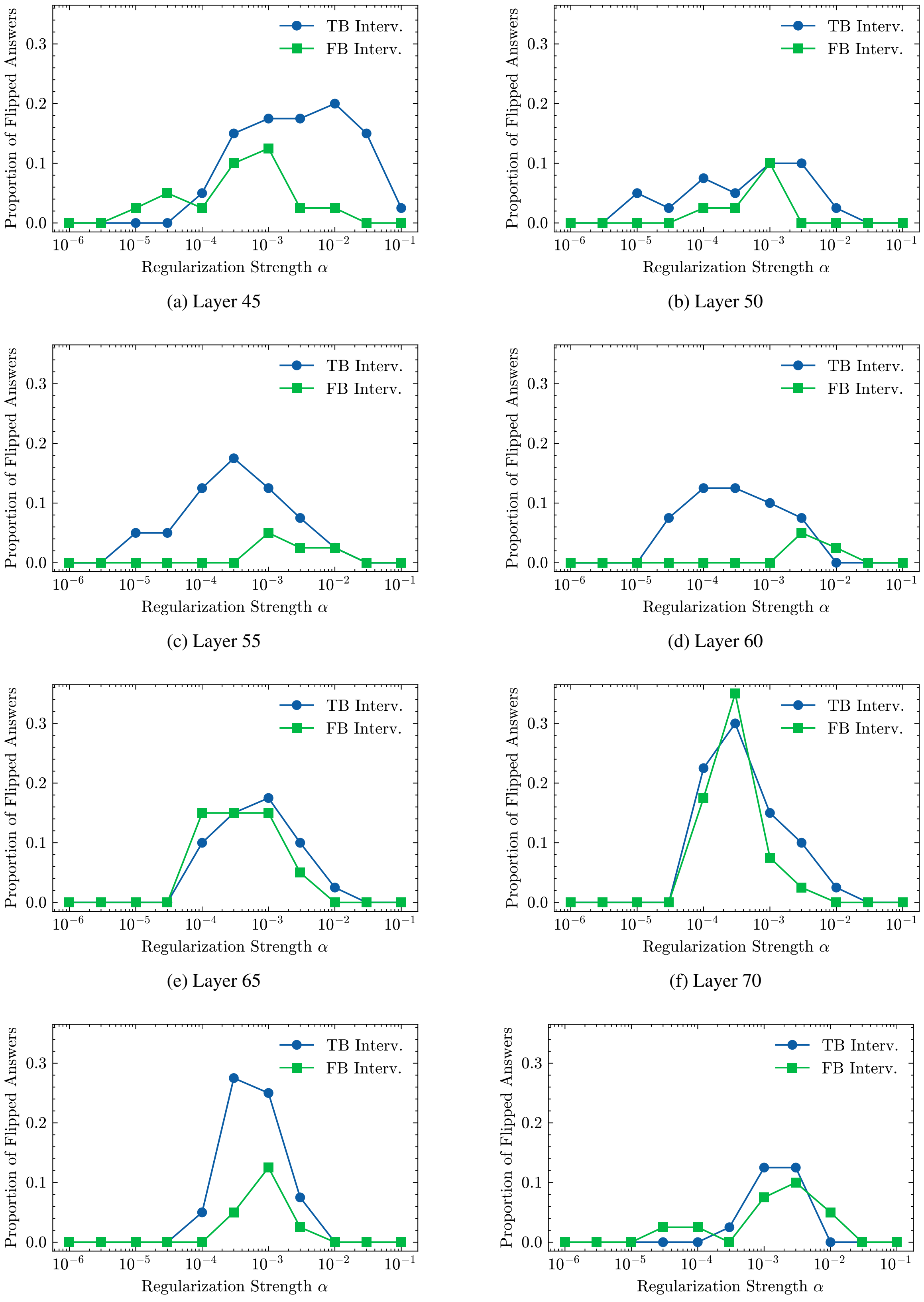
(e) Layer 25

(f) Layer 30

(g) Layer 35

(h) Layer 40

Figure 6: Proportion of flipped answers for layers 5 through 40 under intervention (see Figure 5 for a more detailed explanation).



(a) Layer 45

(b) Layer 50

(c) Layer 55

(d) Layer 60

(e) Layer 65

(f) Layer 70

(g) Layer 75

(h) Layer 80

Figure 7: Proportion of flipped answers for layers 45 through 80 under intervention (see Figure 5 for a more detailed explanation).