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

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. Llms achieve adult human performance on higher-order theory of mind tasks, 2024. URL https://arxiv.org/abs/2405.18870.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Yiwei Wu, Atticus Geiger, and Raphaël Millière. How do transformers learn variable binding in symbolic programs? *arXiv preprint arXiv:2505.20896*, 2025.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.466.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496*, 2024.

## A Full prompt

**No Visibility**

**Instruction:** 1. Track the belief of each character as described in the story. 2. A character’s belief is formed only when they perform an action themselves or can observe the action taking place. 3. A character does not have any beliefs about the container and its contents which they cannot observe. 4. To answer the question, predict only what is inside the queried container, strictly based on the belief of the character, mentioned in the question. 5. If the queried character has no belief about the container in question, then predict ‘unknown’. 6. Do not predict container or character as the final output.
**Story:** **Bob** and **Carla** are working in a busy restaurant. To complete an order, **Bob** grabs an opaque **bottle** and fills it with **beer**. Then **Carla** grabs another opaque **cup** and fills it with **coffee**.
**Question:** What does **Bob** believe the **bottle** contains?
**Answer:**

**Explicit Visibility**

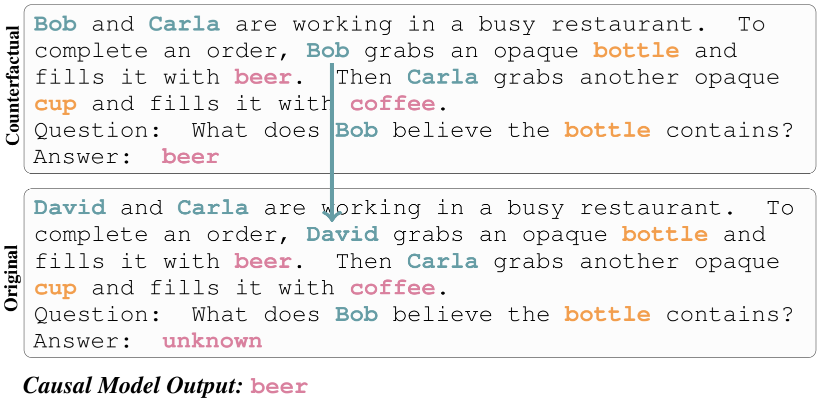
**Instruction:** 1. Track the belief of each character as described in the story. 2. A character’s belief is formed only when they perform an action themselves or can observe the action taking place. 3. A character does not have any beliefs about the container and its contents which they cannot observe. 4. To answer the question, predict only what is inside the queried container, strictly based on the belief of the character, mentioned in the question. 5. If the queried character has no belief about the container in question, then predict ‘unknown’. 6. Do not predict container or character as the final output.
**Story:** **Bob** and **Carla** are working in a busy restaurant. To complete an order, **Bob** grabs an opaque **bottle** and fills it with **beer**. Then **Carla** grabs another opaque **cup** and fills it with **coffee**. **Bob** can observe **Carla**’s actions. **Carla** cannot observe **Bob**’s actions.
**Question:** What does **Bob** believe the **cup** contains?
**Answer:**

## B The CausalToM Dataset

We needed to construct a new dataset because we required a task that models could reliably solve. In contrast, most existing ToM datasets remain challenging for LMs. Additionally, we needed a dataset in which each sample is paired with multiple counterfactuals, enabling causal computations and the extraction of the underlying mechanism. The only dataset that met both criteria was BigToM, which we used in our study. However, even BigToM was insufficient for investigating the full range of factors influencing the mechanism, such as the relationship between a character and their object. Hence, we needed to simplify the task to allow for additional counterfactuals. To test the effect of a specific element, we required the ability to modify only that element without altering the rest of the story or creating an incoherent scenario. For example, consider a BigToM story where a flood occurs, and opening a gate releases the water. In the counterfactual scenario where the gate remains closed, the story’s continuation becomes unintelligible, with the occurrence of a flood.

To address this, we developed CausalToM, which features simple stories accompanied by a range of counterfactuals. Key features include: (1) two characters, objects, and states, (2) the ability to modify each of them independently, and (3) control over whether characters witness each other’s actions. The dataset comprises four templates, one without visibility statements and three with explicit visibility statements. Each template supports four types of questions (e.g., “CharacterX asked about ObjectY”). We used lists of 103 characters, 21 objects, and 23 states. For our interchange intervention experiments, we randomly sampled 80 pairs of original and counterfactual stories.

## C Causal Mediation Analysis



Bob and Carla are working in a busy restaurant. To complete an order, Bob grabs an opaque bottle and fills it with beer. Then Carla grabs another opaque cup and fills it with coffee.
Question: What does Bob believe the bottle contains?
Answer: beer

David and Carla are working in a busy restaurant. To complete an order, David grabs an opaque bottle and fills it with beer. Then Carla grabs another opaque cup and fills it with coffee.
Question: What does Bob believe the bottle contains?
Answer: unknown

***Causal Model Output:*** beer

Figure 9: **Causal Mediation Analysis**: The original example produces the output *unknown* because *Bob* is not mentioned in the story, leaving the model without any information about his beliefs. However, when the residual stream vectors corresponding to *Bob* from the counterfactual run are patched into the original run, the model acquires the necessary information about that character and consequently updates its output to *beer*.

In addition to the experiment shown in Fig.9, we conduct similar experiments for the object and state tokens by replacing them in the story with random tokens, which alters the original example’s final output. However, patching the residual stream vectors of these tokens from the counterfactual run restores the relevant information, enabling the model to predict the causal model output. The results of these experiments are collectively presented in Fig.2, with separate heatmaps shown in Fig. 10, 11, 12.

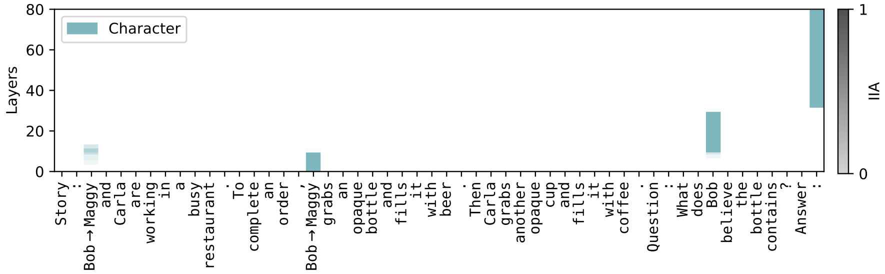


Figure 10: Information flow of character input tokens using causal mediation analysis.