

PICTURE: Enhancing Theory-of-Mind in Large Language Models by Revealing, Not Hiding, Characters’ Lack of Knowledge

Eojin Jeon¹, SangKeun Lee^{1,2}

¹Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea

²Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
{skdlcm456, yalphy}@korea.ac.kr

Abstract

Simulating human-like Theory of Mind (ToM) has been a longstanding problem in natural language processing (NLP). To address this, existing works introduce a reasoning step of event hiding (a.k.a. perspective-taking), where events unknown to a character are removed before question answering. However, resorting to event hiding for ToM reasoning presents a performance degradation issue due to the strict output format constraints involved in event hiding. To mitigate this issue, we propose generating perspective-taking outputs as free-form explanations without event hiding, but this poses a notable yet underexplored challenge: LLMs need to inhibit responses to events unknown to characters, because the absence of event hiding exposes LLMs to these events throughout reasoning. To address this challenge, we hypothesize and empirically verify that LLMs can achieve such inhibition if a character’s *lack of knowledge* about events is made explicit during reasoning. Based on this finding, we introduce PICTURE, a new prompting method that enables LLMs to generate a character’s lack of knowledge within free-form CoT. Experimental results show that PICTURE outperforms existing prompting methods by an average of 7.3% on false-belief tasks.¹

1 Introduction

Humans utilize a remarkable ability, called Theory of Mind (ToM), to attribute mental states (e.g., beliefs) to others (Baron-Cohen et al., 1985; Premack and Woodruff, 1978). Despite the advancement of Large Language Models (LLMs), simulating human-like ToM using them has been a longstanding problem in natural language processing (NLP) (Sclar et al., 2023; Wilf et al., 2024; Jung et al., 2024). A key bottleneck of this challenge lies in

LLMs’ failure to inhibit² responses to events unknown to characters in false-belief tasks. As illustrated in Figure 1(a), LLMs tend to answer the false-belief question with respect to the true state of reality rather than the character’s own perspective (Wilf et al., 2024; Jung et al., 2024).

To address this bottleneck, recent studies (Wilf et al., 2024; Jung et al., 2024; Hou et al., 2024; Xu et al., 2025; Sarangi et al., 2025a) introduce a preliminary reasoning step called perspective-taking, which is typically implemented via event hiding. In this reasoning step, LLMs are first instructed to remove events unknown to a character from an input story. Then, LLMs answer the question based solely on this filtered story. This allows LLMs to bypass the need to inhibit responses to events unknown to a character during question answering.

While event hiding is a standard practice to implement perspective-taking, resorting to event hiding presents a critical issue that can limit LLMs’ ToM capabilities. Specifically, event hiding typically relies on strict output format constraints—such as requiring JSON structures (Jung et al., 2024), graph representations (Xu et al., 2025), or subsets of the story (Wilf et al., 2024; Hou et al., 2024; Sarangi et al., 2025a)—to ensure that events unknown to the character are cleanly separated. However, due to the well-documented trade-off between format adherence and reasoning performance in LLMs (Tam et al., 2024; Banerjee et al., 2025), these format constraints may inadvertently hinder perspective-taking, potentially degrading LLMs’ ToM capabilities. This issue is exemplified in Figure 1(b) where the second event in the story is incorrectly removed during perspective-taking.

²In this work, we use the term inhibition as a behavior-level notion, consistent with psychology literature (Carlson and Moses, 2001). Specifically, LLMs are said to inhibit if they correctly answer a ToM question despite the presence of events unknown to the character in the prompt. This definition does not make any claim about inhibition at the level of internal mechanisms (e.g., neurons within LLMs).

¹Our code is available at <https://github.com/jej127/PICTURE>.

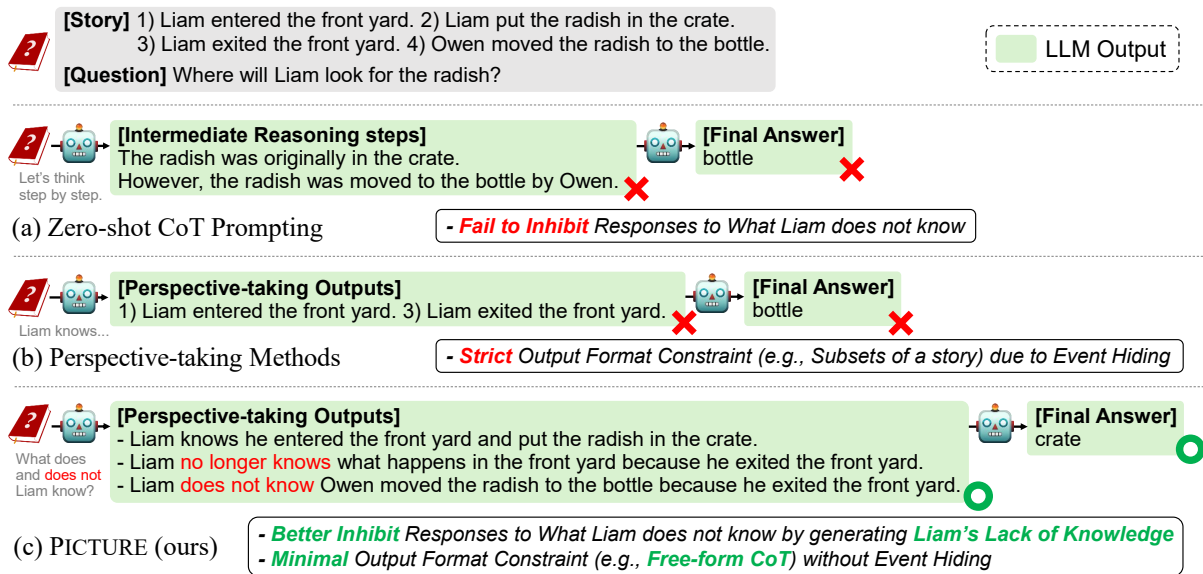


Figure 1: Comparison between (a) zero-shot CoT prompting, (b) the existing perspective-taking methods with event hiding, and (c) our proposed method. Green areas denote LLM outputs.

To alleviate this issue, instead of relying on event hiding, we propose generating perspective-taking outputs as free-form explanations. By omitting event hiding and by expressing the character’s perspective in a free-form manner, we impose minimal constraints on the output format for perspective-taking. In doing so, we can effectively mitigate the performance degradation stemming from rigid output format constraints. However, eliminating event hiding poses a notable yet underexplored challenge: LLMs need to inhibit responses to events unknown to characters, because the absence of event hiding exposes LLMs to these events throughout reasoning. From a psychology standpoint, addressing this challenge can be understood as eliciting the cognitive ability known as *inhibitory control*, which involves inhibiting responses to irrelevant stimuli while pursuing a specific goal (Carlson and Moses, 2001).

To address this challenge, we establish the following hypothesis: if a character’s *lack of knowledge* about events is made explicit during reasoning—for example, through statements such as “Liam does not know Owen moved the radish to the bottle” as in Figure 1(c)—, then LLMs can inhibit responses to those events unknown to the character, using the character’s lack of knowledge as hints for inhibition. Empirically, we validate this hypothesis through a preliminary study described in Section 3.

Based on this finding, we propose **Perspective-taking with Generated Lack of Knowledge in Chain-of-Thought Reasoning (PICTURE)**, a new

prompting method that instructs LLMs to generate characters’ lack of knowledge within free-form Chain-of-Thought (CoT), as illustrated in Figure 1(c). Specifically, PICTURE first prompts LLMs to infer what the character does and does not know. To complete the reasoning path that leads to the final answer, we include a trigger sentence “*Think step by step*”, motivated by Kojima et al. (2022).

To show the effectiveness of PICTURE, we conduct an extensive evaluation on ToM benchmarks. The results demonstrate that PICTURE achieves a 7.3% improvement on false-belief tasks compared to existing zero-shot prompting methods. Furthermore, through analysis, we show that LLMs using PICTURE generate characters’ lack of knowledge within free-form CoT as expected.

Our contributions are summarized as follows:

- We are the first to show that LLMs can inhibit responses to events unknown to a character for ToM reasoning by making the character’s lack of knowledge explicit during reasoning.
- We propose PICTURE, a prompting method to aid LLMs in inhibition for ToM reasoning, which instructs them to generate a character’s lack of knowledge within free-form CoT.
- Extensive evaluation on ToM benchmarks demonstrates that PICTURE improves ToM reasoning performance of LLMs over existing zero-shot prompting methods.

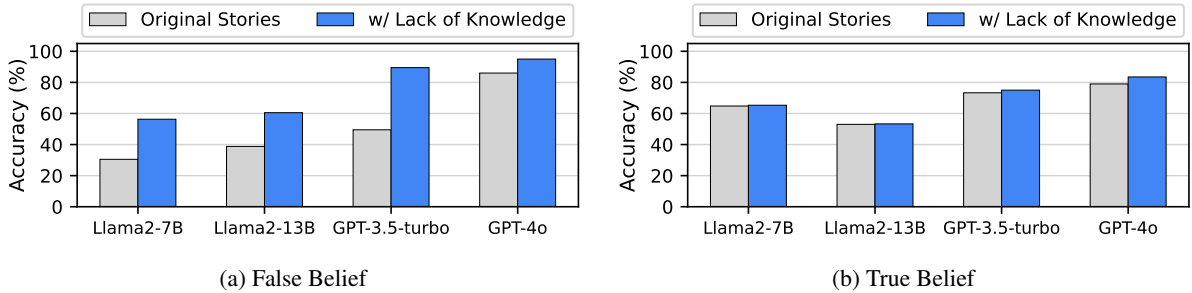


Figure 2: LLM performances on Percept-ToMi. “w/ Lack of Knowledge” denotes stories that are augmented with characters’ lack of knowledge—i.e., for each event unknown to a character, we prepend a phrase “{character} does not know that” to that event, explicitly indicating that the event is unknown to the character.

2 Related Work

2.1 Improving LLMs’ ToM with Event Hiding

Simulating human-like ToM has long been a central challenge in NLP (Sclar et al., 2023; Wilf et al., 2024; Jung et al., 2024). In particular, a crucial bottleneck of this challenge lies in LLMs’ failure to inhibit responses to events unknown to characters in false-belief tasks (Wilf et al., 2024; Jung et al., 2024). Consequently, LLMs often answer false-belief questions with respect to the true state of reality rather than a character’s own perspective. To address this, a line of work introduces a reasoning step of event hiding (a.k.a. perspective-taking), where events unknown to a character are removed from an input story before question answering.

Specifically, SimToM (Wilf et al., 2024) and PerceptToM (Jung et al., 2024) ask LLMs to filter out events using two-step prompting. Other works focus on improving the event hiding results by using external tools or iterative refinement algorithms. For example, SymbolicToM (Sclar et al., 2023) iteratively updates the characters’ belief states by representing them as a set of graphs. TimeToM (Hou et al., 2024) and Decompose-ToM (Sarangi et al., 2025a) refine characters’ beliefs by introducing an external tool called belief solver and by decomposing ToM tasks into recursive simulations, respectively. EnigmaToM (Xu et al., 2025) trains a neural knowledge base to infer the state information for entities. Agentic-ToM (Sarangi et al., 2025b) employs autonomously invocable cognitive tools to analyze agents’ perspectives.

One notable characteristic of these methods is that LLMs lack access to events unknown to characters during question answering, which is facilitated by event hiding. In doing so, LLMs no longer need to inhibit responses to those events for ToM

reasoning.³ In contrast, our method does not rely on event hiding and encourages LLMs to inhibit responses to events unknown to characters, maintaining LLMs’ access to those events throughout reasoning.

2.2 Improving LLMs’ ToM without Event Hiding

A few works attempt to enhance the ToM capability of LLMs without explicitly hiding events. Specifically, DWM (Huang et al., 2024) segments the context into multiple state events and updates characters’ beliefs for each state event. ThoughtTracing (Kim et al., 2025) tracks the mental states of characters, following the sequential Monte Carlo structure. Multi-agent ToM prompting (Li et al., 2023) incorporates explicit belief state representations for the cooperative text game. However, these methods primarily focus on generating characters’ knowledge, as reflected in their instructions, with limited explicit emphasis on generating characters’ lack of knowledge as intermediate reasoning steps. In contrast, our work aims to generate not only the characters’ knowledge but also their lack of knowledge for effective ToM reasoning.

3 Preliminary Study

As discussed in Section 1, our goal is to develop a prompting method for ToM reasoning that generates perspective-taking outputs as free-form explanations without event hiding. By discarding event hiding and by describing characters’ perspectives in a free-form manner, we impose minimal constraints on the output format for perspective-taking. This

³From a psychology standpoint, event hiding is conceptually distinct from inhibitory control and therefore can hardly be viewed as an implementation of inhibitory control in LLMs. See Appendix A for more discussion.

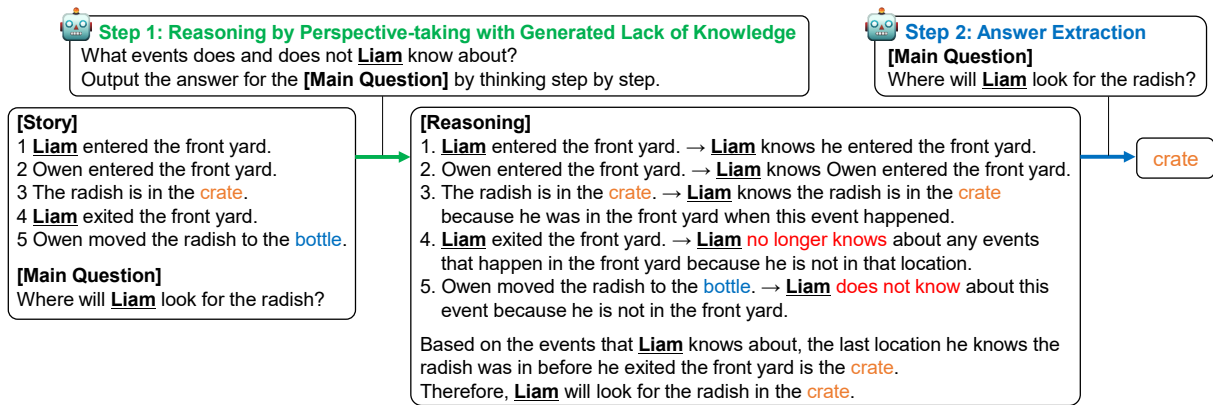


Figure 3: An overview of PICTURE. In the reasoning by perspective-taking with generated lack of knowledge (Step 1), we first ask LLMs to infer what the character does and does not know within a story. We then include a trigger sentence “Think step by step”, to generate further reasoning steps that lead to the final answer. In the answer extraction step (Step 2), LLMs extract the answer from the reasoning generated in the previous step.

design choice helps mitigate the performance degradation arising from rigid output format constraints (Tam et al., 2024). However, removing event hiding introduces a remarkable challenge: LLMs need to inhibit responses to events unknown to characters, since the absence of event hiding exposes LLMs to these events throughout reasoning. In this section, we present and empirically validate the hypothesis for addressing this challenge.

To begin with, we formulate the following hypothesis: if a character’s *lack of knowledge* about events is made explicit during reasoning—for instance, through statements such as “Liam does not know Owen moved the radish to the bottle”—, then LLMs can inhibit responses to those events during reasoning. To verify this hypothesis, we conduct a preliminary study using the Percept-ToMi (Jung et al., 2024) dataset. This dataset contains not only the ToM questions and answers, but also the perceiver information for each event (i.e., the information of who perceived the event). To indicate the character’s lack of knowledge, we use this perceiver information and append “{character} does not know that” to each event that the character fails to perceive. We evaluate LLMs separately on the original and modified stories.

The results are presented in Figure 2. We observe a substantial improvement in the performance of LLMs on the false-belief task when stories are augmented with the characters’ lack of knowledge. In addition, the performances on the true-belief task do not diminish across all LLMs. This suggests that LLMs can effectively inhibit responses to events unknown to characters in false-belief tasks by making the characters’ lack of knowledge explicit, thereby

confirming our hypothesis. Motivated by this observation, we detail the prompting method that generates characters’ lack of knowledge to achieve inhibition in LLMs in Section 4.

4 PICTURE: Prompting to Generate Characters’ Lack of Knowledge

Based on the hypothesis verified in Section 3, we aim to devise a prompting method for ToM reasoning that instructs LLMs to generate characters’ lack of knowledge, especially via free-form explanations. To this end, we propose PICTURE that consists of two steps: 1) reasoning by perspective-taking with generated lack of knowledge (Section 4.1), and 2) answer extraction (Section 4.2). Figure 3 illustrates an overview of PICTURE. The prompts for PICTURE are available in Appendix C.

4.1 Reasoning by Perspective-taking with Generated Lack of Knowledge

To solve a given question with a story, LLMs in this step are expected to 1) provide a free-form CoT that leads to the final answer, and to 2) generate a character’s lack of knowledge within that CoT. We adopt CoT as the generation strategy because it is a well-established and effective method for producing free-form reasoning explanations (Chu et al., 2024). To this end, LLMs are first prompted to infer what the character does and does not know within a story. Different from prior methods which instruct LLMs to hide a character’s lack of knowledge (e.g., SimToM), our method instructs LLMs to reveal a character’s lack of knowledge during reasoning. To complete the CoT that leads to the final answer, we include a trigger sentence “Think

step by step”, inspired by Kojima et al. (2022). This facilitates CoT reasoning of LLMs that explicitly involves a character’s lack of knowledge.

4.2 Answer Extraction

In this step, similar to Kojima et al. (2022), we prompt LLMs to extract the final answer from the reasoning generated in the previous step. We implement this prompt by asking LLMs to answer the same question based on the generated reasoning. In doing so, LLMs are expected to isolate the final answer for evaluation.

5 Experimental Settings

5.1 Datasets

In the experiment, we primarily aim to evaluate how effectively LLMs can reason about characters’ false beliefs using each prompting method. To this end, we use three popular ToM benchmarks that include false-belief questions. Specifically, we use 800 questions from BigToM (Gandhi et al., 2023) and 1,000 questions from ToMi (Le et al., 2019) used in Wilf et al. (2024). In addition, we use 1,540 belief questions in FANToM (Kim et al., 2023). ToMi consists of structured stories generated based on templates, whereas BigToM features more natural and less structured narratives. FANToM involves multi-party conversations where the specific speaker leaves and rejoins. We use accuracy as the evaluation metric for these benchmarks.

5.2 Models

Similar to Wilf et al. (2024), we evaluate each method using four widely used LLMs. Among the open-source LLMs, we employ Llama2-7B-chat and Llama2-13B-chat (Touvron et al., 2023). For the closed-source models, we use GPT-3.5-Turbo (OpenAI, 2023) and GPT-4o (OpenAI, 2024). We perform inference with the open-source models using Hugging Face and access the closed-source models through the OpenAI API.⁴ We set the temperature of $\tau = 0.0$ and top- p of 1.0 for the closed-source models.

5.3 Baselines

We compare PICTURE with two categories of zero-shot prompting methods: (1) methods to improve general reasoning capabilities, and (2) methods ded-

icated to improving ToM capabilities. Prompts for each method are provided in Appendix D.

The methods in the first category have demonstrated potential across a wide range of reasoning tasks, such as arithmetic and commonsense reasoning. These methods include: (1) Vanilla prompting, where the LLM is asked to select an answer without rationales; (2) CoT prompting (Kojima et al., 2022), which decomposes a problem into multiple intermediate steps to be solved; and (3) Plan-and-Solve (PS) prompting (Wang et al., 2023a), which generates a plan that divides the task into subtasks. For Vanilla and CoT, we use the prompts used in Wilf et al. (2024). For PS, we use the prompts according to their published settings.

The methods in the second category are dedicated to improving ToM capabilities. They either hide events unknown to characters, or iteratively track characters’ knowledge without event hiding. We consider methods that, like ours, do not rely on external tools or iterative refinement algorithms for reasoning. These methods include: SimToM (Wilf et al., 2024) and PercepToM (Jung et al., 2024), both of which rely on event hiding. We use the prompts as described in their respective papers.

6 Results

In this section, we pose several key research questions (RQs) and answer each through experiments.

- **RQ1:** Do LLMs better inhibit responses to events unknown to characters using PICTURE? (§6.1)
- **RQ2:** Does PICTURE enable LLMs to perform perspective-taking better than baselines? (§6.2)
- **RQ3:** Does PICTURE encourage LLMs to generate characters’ lack of knowledge within free-form CoT as expected? (§6.2)
- **RQ4:** Can PICTURE generalize to different types of ToM questions? (§6.3)
- **RQ5:** Do existing methods based on event hiding perform comparably to PICTURE when additionally using CoT prompts? (§6.4)

6.1 Main Results

The performances of PICTURE and the baselines are presented in Table 1. Following Wilf et al. (2024), we report the results on the false-belief questions (False Belief) and those on the full set of questions (All). The results are averaged over four runs with random seeds {0, 111, 222, 333}. We do not report the performance of the Llama2-7B-chat model using PercepToM, as this model repetitively

⁴<https://platform.openai.com/docs/api-reference/>

Model	Method	False Belief			All		
		BigToM	ToMi	FANToM	BigToM	ToMi	FANToM
Llama2-7B-chat	Vanilla	47.3 \pm 1.5	32.8 \pm 0.9	15.3 \pm 0.4	53.6 \pm 1.2	45.3 \pm 0.8	42.2 \pm 0.5
	CoT (Kojima et al., 2022)	42.9 \pm 1.7	39.2 \pm 0.9	26.2 \pm 1.0	55.5 \pm 1.1	44.8 \pm 1.0	47.5 \pm 0.7
	PS (Wang et al., 2023a)	44.0 \pm 1.6	39.1 \pm 1.3	18.2 \pm 0.3	50.8 \pm 1.3	51.4 \pm 0.8	42.9 \pm 0.5
	SimToM (Wilf et al., 2024)	70.3 \pm 0.9	48.2 \pm 0.5	13.0 \pm 0.8	57.2 \pm 1.0	51.8 \pm 0.5	41.2 \pm 0.7
	PICTURE (Ours)	72.9 \pm 1.1	53.0 \pm 1.0	37.8 \pm 0.8	69.0 \pm 0.9	59.2 \pm 0.8	50.3 \pm 1.0
Llama2-13B-chat	Vanilla	40.9 \pm 1.3	39.3 \pm 1.1	18.4 \pm 0.5	52.0 \pm 0.5	50.9 \pm 1.3	43.9 \pm 0.3
	CoT (Kojima et al., 2022)	52.2 \pm 1.1	31.9 \pm 0.8	43.4 \pm 1.1	56.1 \pm 0.8	46.9 \pm 0.7	51.5 \pm 1.1
	PS (Wang et al., 2023a)	28.3 \pm 0.6	38.6 \pm 1.7	34.2 \pm 0.7	53.1 \pm 0.2	43.8 \pm 1.1	49.1 \pm 0.6
	PercepToM (Jung et al., 2024)	45.2 \pm 1.6	51.2 \pm 1.2	38.0 \pm 0.9	49.0 \pm 0.6	48.4 \pm 1.0	49.4 \pm 0.7
	SimToM (Wilf et al., 2024)	61.8 \pm 0.9	50.6 \pm 0.9	36.5 \pm 1.0	58.0 \pm 1.6	60.8 \pm 1.1	49.6 \pm 0.8
	PICTURE (Ours)	67.8 \pm 2.0	64.3 \pm 1.5	47.7 \pm 0.8	73.6 \pm 0.8	64.3 \pm 0.7	55.5 \pm 0.2
GPT-3.5-Turbo	Vanilla	41.1 \pm 0.9	64.6 \pm 0.8	10.8 \pm 0.4	66.6 \pm 0.7	68.5 \pm 0.6	40.9 \pm 0.4
	CoT (Kojima et al., 2022)	56.1 \pm 0.8	55.3 \pm 0.9	49.1 \pm 0.5	75.9 \pm 0.8	65.2 \pm 0.7	59.0 \pm 0.5
	PS (Wang et al., 2023a)	55.6 \pm 0.5	59.2 \pm 0.6	44.9 \pm 0.7	71.6 \pm 0.8	66.3 \pm 0.6	56.1 \pm 1.0
	PercepToM (Jung et al., 2024)	45.7 \pm 1.1	79.7 \pm 0.8	25.3 \pm 0.4	68.4 \pm 0.3	70.5 \pm 0.5	50.8 \pm 0.6
	SimToM (Wilf et al., 2024)	70.4 \pm 0.6	81.0 \pm 0.6	27.0 \pm 0.6	81.5 \pm 0.5	72.6 \pm 0.4	49.9 \pm 0.9
	PICTURE (Ours)	86.0 \pm 0.5	83.5 \pm 0.9	58.1 \pm 0.8	88.7 \pm 0.6	77.5 \pm 0.6	65.0 \pm 1.0
GPT-4o	Vanilla	88.6 \pm 0.3	80.2 \pm 0.6	52.5 \pm 0.5	92.2 \pm 0.4	74.6 \pm 0.7	65.4 \pm 0.4
	CoT (Kojima et al., 2022)	90.2 \pm 0.4	83.3 \pm 0.8	76.8 \pm 0.6	93.9 \pm 0.3	78.9 \pm 0.5	79.9 \pm 0.5
	PS (Wang et al., 2023a)	85.9 \pm 0.2	89.1 \pm 0.4	76.4 \pm 0.4	91.3 \pm 0.3	77.3 \pm 0.4	78.9 \pm 1.0
	PercepToM (Jung et al., 2024)	87.5 \pm 0.3	88.1 \pm 0.3	79.4 \pm 0.6	87.4 \pm 0.3	76.3 \pm 0.4	84.1 \pm 0.6
	SimToM (Wilf et al., 2024)	91.5 \pm 0.3	88.1 \pm 0.4	79.6 \pm 0.7	91.9 \pm 0.3	79.4 \pm 0.4	81.3 \pm 0.9
	PICTURE (Ours)	96.5 \pm 0.6	90.5 \pm 0.2	91.5 \pm 0.7	97.3 \pm 0.4	91.8 \pm 0.5	87.3 \pm 1.0

Table 1: Results on the BigToM, ToMi, and FANToM datasets. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**. All improvements achieved by PICTURE are statistically significant according to a t-test with $p < 0.05$.

fails to generate a valid JSON array for over 95% of the questions during the perception inference stage of PercepToM.

PICTURE outperforms all baselines across datasets and backbone LLMs. In particular, it achieves an average improvement of 7.3% on false-belief questions. While prior work has shown that LLMs without explicit perspective-taking prompts (e.g., CoT) often struggle to inhibit responses to events unknown to characters (Wilf et al., 2024), we demonstrate that PICTURE mitigates this issue and enhances inhibition in LLMs, outperforming all baselines that do not involve explicit perspective-taking prompts (i.e., Vanilla, CoT, and PS).

Furthermore, PICTURE surpasses perspective-taking methods based on event hiding, such as SimToM and PercepToM. Although event hiding allows LLMs to circumvent the need for inhibition during question answering, we show that moving beyond this paradigm is viable. In particular, LLMs can inhibit responses to events unknown to characters via PICTURE. Overall, LLMs using PICTURE better inhibit responses to events unknown to characters than baselines, showing performance improvements over them especially on false-belief

tasks. A detailed breakdown of the results across different question types can be found in Appendix B.8. In addition, we present additional results using more recent LLMs in Appendix B.1, and provide a comparison between PICTURE and more recent event hiding methods in Appendix B.6.

6.2 Case Study

To better understand why PICTURE enhances the ToM capabilities of LLMs, we manually inspect the perspective-taking outputs generated by GPT-3.5-Turbo using SimToM and PICTURE. Table 2 presents examples of the results on false-belief questions from the BigToM and ToMi datasets. For both examples, we observe that the LLM using SimToM makes errors in perspective-taking, either by incorrectly retaining events unknown to the character (upper), or by falsely removing events known to the character (lower). In contrast, the LLM using PICTURE provides the correct perspective-taking results, generating the characters’ lack of knowledge within free-form CoTs.

To further investigate these reasoning behaviors, we sample and analyze the results from 50 false-belief questions in ToMi. We observe that the LLM using SimToM misclassifies events unknown to

Example 1 (BigToM)
<p>Context: Amaury is a baker and wants to make a pastry for the customers visiting his bakery. Amaury takes a jar of raisins from the pantry to add to the pastry dough. His assistant, while tidying the pantry, accidentally replaces the raisin jar with a jar containing sultanas without realizing the difference. Amaury does not notice the jar swap.</p> <p>Question: Does Amaury believe the jar contains a) raisins or b) sultanas?</p>
<p>SimToM (baseline)</p> <p>Perspective-taking Outputs: Amaury is a baker and wants to make a pastry for the customers visiting his bakery. Amaury takes a jar of raisins from the pantry to add to the pastry dough. His assistant, while tidying the pantry, accidentally replaces the raisin jar with a jar containing sultanas without realizing the difference.</p> <p>Answer: Amaury believes the jar contains b) sultanas. (✗)</p>
<p>PICTURE (ours)</p> <p>Perspective-taking Outputs: Amaury knows that he took a jar of what he believed to be raisins from the pantry to add to the pastry dough. He is also aware that his assistant accidentally replaced the raisin jar with a jar containing sultanas without realizing the difference. However, the story mentions that Amaury did not notice the jar swap. Given this information, Amaury is not aware of the assistant's mistake in replacing the raisins with sultanas.</p> <p>Answer: Amaury believes the jar contains a) raisins. (✓)</p>
Example 2 (ToMi)
<p>Context: 1 Liam entered the front yard. 2 Jackson entered the front yard. 3 Owen entered the front yard. 4 The radish is in the crate. 5 Liam exited the front yard. 6 Owen moved the radish to the bottle.</p> <p>Question: Where will Liam look for the radish? Choose from the following: crate, bottle</p>
<p>SimToM (baseline)</p> <p>Perspective-taking Outputs: 1 Liam entered the front yard. 5 Liam exited the front yard.</p> <p>Answer: Liam will look for the radish in the bottle. (✗)</p>
<p>PICTURE (ours)</p> <p>Perspective-taking Outputs: Let's break down the events step by step to determine what Liam knows:</p> <ol style="list-style-type: none"> Liam entered the front yard. - Liam knows that he entered the front yard. Jackson entered the front yard. - Liam knows that Jackson entered the front yard. Owen entered the front yard. - Liam knows that Owen entered the front yard. The radish is in the crate. - Liam knows that the radish is in the crate. Liam exited the front yard. - Liam no longer knows about any events that happen in the front yard. Owen moved the radish to the bottle. - Liam does not know about this event. <p>Based on what Liam knows about, the last location he knows the radish was in before he exited the front yard is the crate.</p> <p>Answer: Liam will look for the radish in the crate. (✓)</p>

Table 2: Qualitative examples comparing SimToM and PICTURE on false-belief questions from the BigToM and ToMi datasets. **Red** text indicates the character’s lack of knowledge generated by PICTURE. Underlined text highlights events unknown to the character but are incorrectly retained by SimToM after perspective-taking.

the character as known for 18% of the questions, whereas the LLM using PICTURE makes the same type of error for only 2% of the questions. Moreover, the LLM using SimToM misclassifies events known to the character as unknown for 30% of the questions, whereas the LLM using PICTURE makes the same type of error for only 6% of the questions. Lastly, for all 50 questions, we observe that the LLM using PICTURE generates the characters’ lack of knowledge within free-form CoTs.

These results suggest that LLMs can perform perspective-taking better and more robustly through PICTURE compared to baselines, which in turn enhances subsequent ToM reasoning. Notably, PICTURE achieves these gains by expressing perspective-taking outputs as free-form explana-

tions. This observation aligns with prior work by Tam et al. (2024), which shows that relaxing output format constraints can improve LLM reasoning performance. Lastly, PICTURE helps LLMs to generate characters’ lack of knowledge within free-form CoT as expected. More qualitative results using different datasets and LLMs can be found in Appendix B.4. In addition, we present the error analysis of PICTURE in Appendix B.5.

6.3 Generalization to Different types of ToM Questions

While the benchmarks considered in Section 6.1 are representative ToM benchmarks, these benchmarks primarily focus on characters’ mental states associated with the physical world (e.g., the lo-

Method	Loc _c (F)	Loc _c (S)	Loc _f (F)	Loc _f (S)	MHop(F)	MHop(S)	Att
Vanilla	43.9 \pm 2.1	32.3 \pm 2.3	51.5 \pm 1.1	28.6 \pm 0.5	46.8 \pm 1.0	33.4 \pm 0.8	41.0 \pm 1.2
CoT (Kojima et al., 2022)	59.1 \pm 1.5	44.0 \pm 2.8	51.2 \pm 0.9	28.7 \pm 0.4	52.5 \pm 0.9	42.5 \pm 1.0	46.8 \pm 1.5
PS (Wang et al., 2023a)	63.3 \pm 1.8	54.0 \pm 2.0	65.7 \pm 1.5	33.5 \pm 0.6	64.3 \pm 1.5	46.6 \pm 0.8	49.3 \pm 1.3
PercepToM (Jung et al., 2024)	54.9 \pm 1.7	35.1 \pm 2.1	57.3 \pm 1.2	33.6 \pm 0.7	47.6 \pm 1.0	35.5 \pm 0.7	39.1 \pm 1.0
SimToM (Wilf et al., 2024)	62.5 \pm 1.4	45.8 \pm 2.4	53.3 \pm 1.3	28.7 \pm 0.5	53.9 \pm 1.3	34.4 \pm 0.8	41.6 \pm 1.4
PICTURE (Ours)	85.4 \pm 1.6	62.0 \pm 1.8	77.1 \pm 1.4	43.9 \pm 0.4	68.0 \pm 1.2	48.0 \pm 0.5	52.1 \pm 1.5

Table 3: Results on the OpenToM dataset, which includes questions about object locations (Loc), multi-hop reasoning (MHop), and characters’ attitudes (Att). We report F1 score as the evaluation metric following Xu et al. (2024), with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

Method	False Belief			All		
	BigToM	ToMi	FANToM	BigToM	ToMi	FANToM
CoT (Kojima et al., 2022)	56.1 \pm 0.8	55.3 \pm 0.9	49.1 \pm 0.5	75.9 \pm 0.8	65.2 \pm 0.7	59.0 \pm 0.5
SimToM (Wilf et al., 2024)	70.4 \pm 0.6	81.0 \pm 0.6	27.0 \pm 0.6	81.5 \pm 0.5	72.6 \pm 0.4	49.9 \pm 0.9
SimToM + CoT	74.2 \pm 0.3	81.5 \pm 0.5	52.8 \pm 1.0	82.4 \pm 0.7	73.0 \pm 0.9	58.8 \pm 0.7
PICTURE (Ours)	86.0 \pm 0.5	83.5 \pm 0.9	58.1 \pm 0.8	88.7 \pm 0.6	77.5 \pm 0.6	65.0 \pm 1.0

Table 4: Ablation study results for PICTURE. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

cation of an object). In contrast, they place less emphasis on mental states associated with the psychological world (e.g., characters’ attitudes towards situations). To verify the effectiveness of PICTURE across different types of questions, we evaluate our method on OpenToM (Xu et al., 2024), a comprehensive benchmark that covers questions involving both the physical and psychological worlds.

Specifically, OpenToM includes questions about object locations (Loc), multi-hop reasoning (MHop), and characters’ attitudes (Att). Questions about object locations and multi-hop reasoning are further divided into first-order (F) and second-order (S) questions. Additionally, questions about object locations are further categorized into coarse- and fine-grained levels, denoted by the subscripts *c* and *f*, respectively. We use GPT-3.5-Turbo as the backbone model and show the results in Table 3.

PICTURE outperforms the baselines across all question types. Notably, significant performance gains are observed on questions related to object locations (Loc). PICTURE is specifically designed to help LLMs explicitly represent a character’s knowledge state (i.e., what the character knows or does not know), to reduce errors arising from incorrectly using information unknown to the character. This design is most directly aligned with location-based questions, where performance critically depends on suppressing information unknown to the character. Consequently, the largest gains are observed in these tasks.

In contrast, for questions in multi-hop reasoning (MHop) and attitude (Att), the performance gain over the second-best method, PS, is relatively smaller. For MHop, this type of question requires additional compositional reasoning beyond knowledge state tracking. While the correct representation of knowledge states remains necessary, it is not sufficient for solving the task, which likely reduces the relative contribution of the inhibitory control trigger. Moreover, the questions in Att often assume that the character has access to all relevant events (e.g., "What would be Sam’s attitude towards Amy’s action, assuming he observed it?"). In such cases, epistemic conflict is weaker, and knowledge state reasoning plays a smaller role. Nonetheless, our method consistently outperforms these baselines by explicitly reasoning about what the character knows—or does not know—before answering the question. Additional results on ToM questions of more diverse types are available in Appendix B.2.

6.4 Ablation Study

While PICTURE involves CoT prompts (e.g., "Think step by step") to elicit free-form reasoning explanations, we are curious about whether performance comparable to PICTURE can be achieved by simply adding CoT prompts to existing methods based on event hiding. To investigate this, we introduce another baseline, SimToM + CoT, wherein 1) events unknown to a character are first removed as

in SimToM, and then 2) LLMs perform CoT reasoning based solely on events known to the character. Notably, unlike PICTURE, this baseline relies on event hiding and does not generate characters’ lack of knowledge, while other components such as CoT prompts are held constant across the two methods.

As shown in Table 4, SimToM + CoT does not yield consistent performance gains over SimToM and CoT prompting. This suggests that naively adding CoT prompts to existing perspective-taking methods provides limited and unstable benefit for ToM reasoning. In contrast, PICTURE significantly outperforms SimToM + CoT, highlighting that the advantage of PICTURE over SimToM is not attributed solely to the use of CoT prompts. Rather, its strength stems from omitting event hiding and thereby enabling LLMs to generate characters’ lack of knowledge. Additional results for further variants of PICTURE are in Appendix B.3.

7 Conclusion

In this paper, we explore how LLMs can inhibit responses to events unknown to characters for effective ToM reasoning. To this end, we first hypothesize and empirically verify that LLMs can achieve such inhibition if a character’s *lack of knowledge* about events is made explicit during reasoning. Based on this finding, we propose PICTURE, a new prompting method that enables LLMs to generate characters’ lack of knowledge within free-form CoT. Experimental results show that PICTURE outperforms existing prompting methods on ToM reasoning tasks, suggesting that LLMs using PICTURE can inhibit responses to events unknown to characters. Furthermore, analysis verifies that PICTURE helps LLMs to generate characters’ lack of knowledge within free-form CoT as expected.

Limitations

While our method has demonstrated effectiveness in ToM reasoning of LLMs, it has three main limitations that should be addressed in future work.

(1) Following the direction of prior work aimed at improving LLMs’ ToM capabilities, we primarily focused on text-based ToM benchmarks. However, some recent benchmarks, such as MMTOM-QA (Jin et al., 2024), evaluate ToM in multimodal scenarios involving both text and video. Handling multimodal scenarios introduces additional challenges, such as grounding the meaning of natural language sentences in visual data. Therefore, ex-

tending PICTURE to multimodal ToM remains an important area for future exploration.

(2) Currently, we implement PICTURE such that LLMs perform perspective-taking within a single step of CoT reasoning. While our method shows the effectiveness across various ToM reasoning tasks, incorporating inference-time algorithms that involve advanced components such as complexity-aware decomposition (Huang et al., 2024) or hypothesis generation (Kim et al., 2025) may provide further benefit in modeling false beliefs of characters. Thus, we leave extensions of PICTURE in this direction to future work.

(3) PICTURE critically relies on correctly generated lack of knowledge, and errors in this intermediate step could affect downstream reasoning. While our analysis over 50 sampled questions in Section 6.2 provides empirical evidence regarding the robustness of this intermediate step, we acknowledge that this analysis relies on a manual inspection of a sampled subset. Extending robustness evaluation to a fully automated large-scale setting is therefore an important avenue for future work.

Ethical Considerations

We acknowledge that enhancing LLMs’ ability to reason about others’ beliefs and knowledge states may introduce societal risks if deployed irresponsibly. For instance, improved false-belief reasoning could be misused in deceptive dialog systems, in persuasive chatbots that exploit incorrect user beliefs, or in systems that attempt to infer mental states without consent. Such capabilities may also increase risks of manipulation in politically or financially motivated contexts.

In addition, benchmarking ToM reasoning on narrative datasets raises concerns about demographic bias, as stories may encode implicit assumptions about different social groups. If not carefully evaluated, improved ToM performance could inadvertently reinforce or amplify such biases in downstream applications. At the same time, perspective-aware reasoning may enable more empathetic and socially sensitive AI systems. Our experiments focus on offline benchmark evaluation rather than real-world deployment. We emphasize that future research involving interactive or user-facing systems should carefully assess potential misuse scenarios, bias amplification, and unintended social consequences.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-00517221 and No.RS-2024-00415812) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-III90079, Artificial Intelligence Graduate School Program (Korea University)).

References

- Debangshu Banerjee, Tarun Suresh, Shubham Ugare, Sasa Misailovic, and Gagandeep Singh. 2025. [CRANE: reasoning with constrained LLM generation](#). In *International Conference on Machine Learning, ICML 2025*. OpenReview.net.
- Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a “theory of mind”? *Cognition*, 21(1):37–46.
- Stephanie M Carlson and Louis J Moses. 2001. Individual differences in inhibitory control and children’s theory of mind. *Child development*, 72(4):1032–1053.
- Stephanie M Carlson, Louis J Moses, and Casey Breton. 2002. How specific is the relation between executive function and theory of mind? contributions of inhibitory control and working memory. *Infant and Child Development: An International Journal of Research and Practice*, 11(2):73–92.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. [ToMBench: Benchmarking theory of mind in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 15959–15983. Association for Computational Linguistics.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. [Navigate through enigmatic labyrinth A survey of chain of thought reasoning: Advances, frontiers and future](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 1173–1203. Association for Computational Linguistics.
- Adele Diamond. 2013. Executive functions. *Annual review of psychology*, 64(1):135–168.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. [Understanding social reasoning in language models with language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 13518–13529. Curran Associates, Inc.
- Cherie L Gerstadt, Yoon Joo Hong, and Adele Diamond. 1994. The relationship between cognition and action: performance of children 312–7 years old on a stroop-like day-night test. *Cognition*, 53(2):129–153.
- Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. [Timetom: Temporal space is the key to unlocking the door of large language models’ theory-of-mind](#). In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 11532–11547. Association for Computational Linguistics.
- X. Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony G. Cohn, and Michael J. Wooldridge. 2024. [A notion of complexity for theory of mind via discrete world models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2964–2983. Association for Computational Linguistics.
- Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer D. Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. 2024. [Mmtom-qa: Multimodal theory of mind question answering](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 16077–16102. Association for Computational Linguistics.
- Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. [Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 19794–19809. Association for Computational Linguistics.
- Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. 2025. [Hypothesis-driven theory-of-mind reasoning for large language models](#). In *Second Conference on Language Modeling*.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. [Fantom: A benchmark for stress-testing machine theory of mind in interactions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 14397–14413. Association for Computational Linguistics.

- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. [Revisiting the evaluation of theory of mind through question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 5871–5876. Association for Computational Linguistics.
- Huaoli Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia P. Sycara. 2023. [Theory of mind for multi-agent collaboration via large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 180–192. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-3.5 turbo fine-tuning and api updates](#).
- OpenAI. 2024. [Hello gpt-4o](#).
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.
- Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025a. [Decompose-tom: Enhancing theory of mind reasoning in large language models through simulation and task decomposition](#). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025*, pages 10228–10241. Association for Computational Linguistics.
- Sneheel Sarangi, Chetan Talele, and Hanan Salam. 2025b. [Agentic-ToM: Cognition-inspired agentic processing for enhancing theory of mind reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25645–25661. Association for Computational Linguistics.
- Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. [Minding language models’ \(lack of\) theory of mind: A plug-and-play multi-character belief tracker](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 13960–13980. Association for Computational Linguistics.
- J Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643.
- Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. [Let me speak freely? a study on the impact of format restrictions on large language model performance](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236. Association for Computational Linguistics.
- Gemma Team. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 2609–2634. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net.
- Alex Wilf, Sihyun Shawn 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, ACL 2024*, pages 8292–8308. Association for Computational Linguistics.
- Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. [Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706. Association for Computational Linguistics.
- Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, Caroline Catmur, and Yulan He. 2025. [Enigmatom: Improve llms’ theory-of-mind reasoning capabilities with neural knowledge base of entity states](#). In *Findings of the Association for Computational Linguistics, ACL 2025*, pages 13598–13622. Association for Computational Linguistics.
- Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. [Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 8593–8623. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with](#)

large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

A Distinction between Event Hiding and Inhibitory Control

In psychology, inhibitory control refers to the ability to inhibit responses to irrelevant stimuli while pursuing a cognitively represented goal (Carlson and Moses, 2001). As one of the core executive functions (Diamond, 2013), inhibitory control is closely associated with human ToM development (Carlson and Moses, 2001; Carlson et al., 2002). Existing methods for ToM reasoning typically implement perspective-taking through event hiding, wherein events unknown to a character are removed from the input prior to question answering. However, we argue that event hiding is conceptually distinct from inhibitory control and therefore can hardly be viewed as an implementation of inhibitory control in LLMs.

First, according to the aforementioned definition of inhibitory control, irrelevant stimuli themselves are not eliminated for inhibition; rather, it is a response to those stimuli that is suppressed when inhibitory control operates. In contrast, event hiding prompts LLMs to perform perspective-taking by eliminating irrelevant stimuli altogether—namely, events unknown to the character. Therefore, this process of elimination is fundamentally misaligned with the core concept of inhibitory control.

Moreover, event hiding differs in principle from how inhibitory control is operationalized in standard psychological assessments such as the Stroop test (Stroop, 1935) or the day-night test (Gerstadt et al., 1994). In the Stroop test, for example, participants are presented with color words printed in incongruent ink colors (e.g., “blue”). Their task is to report the ink color (e.g., red) rather than reading the word itself (e.g., blue), as in Figure 4(a). Crucially, task-irrelevant stimuli—such as the meaning of words—remain present and unaltered, and participants are not allowed to remove or modify the stimuli to simplify the task (e.g., by rewriting the word as “red,” as in Figure 4(b)). In this sense, event hiding is analogous to such impermissible simplification of this classic test, since it attempts to remove task-irrelevant stimuli, i.e., events unknown to the character, before answering the question.

In summary, event hiding is fundamentally different from inhibitory control because it seeks to eliminate irrelevant stimuli altogether rather than requiring the inhibition of responses to them. Consequently, while event hiding may be an effective strategy for ToM reasoning, it circumvents the need

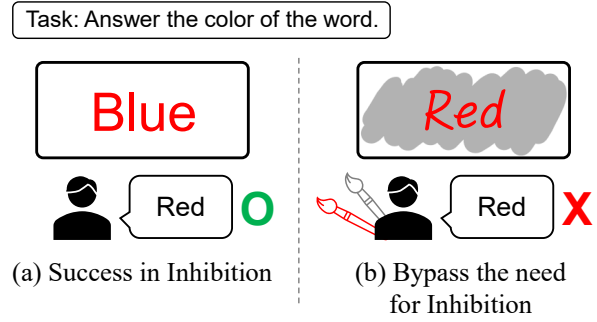


Figure 4: Comparison between (a) a correct response in the Stroop test and (b) an impermissible simplification of the test where the original color word “Blue” is removed by the test participant. Existing methods based on event hiding are analogous to case (b), since these methods attempt to remove task-irrelevant stimuli (i.e., events unknown to a character) before answering the question.

for inhibitory control rather than implementing it. Motivated by this distinction, we aim to implement inhibitory control in LLMs directly, rather than relying on methods that avoid the need for inhibition through event hiding.

B Additional Experiments

B.1 Results on More Recent LLMs

To further demonstrate the effectiveness of PICTURE using more recent LLMs, we employ Llama3.1-8B-Instruct (Dubey et al., 2024), Qwen3-8B (Yang et al., 2025), and Gemma3-12B-Instruct (Team, 2025) for additional evaluations. We use the same benchmarks as in Section 5.1 and show the results in Table 5. Overall, PICTURE achieves consistent improvements over the competitive methods across all three LLMs. This further indicates the efficacy of PICTURE on even more recent LLMs, similar to the results in Section 6.

B.2 Results on Additional Types of ToM Questions

In Section 6, we have shown the effectiveness of PICTURE on ToM questions regarding 1st-/2nd-order belief, knowledge, and emotion. However, the efficacy of PICTURE on ToM questions of other types (e.g., higher-order belief, intention, desire) is not explored yet. Motivated by this, we evaluate PICTURE on ToM questions whose types are not covered in Section 6. Specifically, we adopt ToM questions regarding characters’ intention, desire, and non-literal communication from ToMBench (Chen et al., 2024). In addition, we employ 3rd-

Model	Method	False Belief			All		
		BigToM	ToMi	FANToM	BigToM	ToMi	FANToM
Llama3.1-8B-Inst.	Vanilla	58.8 \pm 1.6	54.7 \pm 1.3	41.5 \pm 0.4	68.1 \pm 1.1	70.4 \pm 1.4	53.4 \pm 0.3
	CoT (Kojima et al., 2022)	61.8 \pm 1.5	53.0 \pm 1.2	52.7 \pm 0.8	73.9 \pm 1.3	63.9 \pm 1.1	55.3 \pm 0.5
	PS (Wang et al., 2023a)	65.3 \pm 1.7	62.1 \pm 1.0	49.9 \pm 0.5	73.3 \pm 1.1	65.1 \pm 0.9	57.6 \pm 0.4
	PercepToM (Jung et al., 2024)	67.3 \pm 1.2	72.4 \pm 1.4	53.0 \pm 0.7	70.8 \pm 0.5	74.8 \pm 1.3	57.7 \pm 0.6
	SimToM (Wilf et al., 2024)	73.9 \pm 0.7	69.5 \pm 1.1	53.4 \pm 0.7	62.5 \pm 1.0	76.4 \pm 0.6	59.3 \pm 0.5
	PICTURE (Ours)	76.5 \pm 1.0	82.3 \pm 1.2	55.5 \pm 0.6	79.8 \pm 0.8	79.6 \pm 0.7	61.4 \pm 0.8
Qwen3-8B	Vanilla	61.3 \pm 1.0	62.3 \pm 1.2	28.7 \pm 0.7	69.8 \pm 0.6	67.1 \pm 1.1	50.7 \pm 0.5
	CoT (Kojima et al., 2022)	73.3 \pm 1.4	72.0 \pm 1.2	59.9 \pm 1.0	77.6 \pm 0.6	70.7 \pm 0.9	66.5 \pm 1.3
	PS (Wang et al., 2023a)	82.0 \pm 0.8	69.5 \pm 1.3	65.0 \pm 0.5	82.8 \pm 0.8	68.2 \pm 1.4	66.9 \pm 0.9
	PercepToM (Jung et al., 2024)	63.0 \pm 1.4	76.8 \pm 1.0	57.9 \pm 0.8	66.4 \pm 0.8	77.4 \pm 1.3	65.9 \pm 0.9
	SimToM (Wilf et al., 2024)	74.3 \pm 1.1	75.0 \pm 0.5	63.7 \pm 1.4	77.9 \pm 1.2	76.3 \pm 1.3	65.4 \pm 0.9
	PICTURE (Ours)	95.0 \pm 1.4	85.3 \pm 1.1	67.1 \pm 0.7	87.8 \pm 1.1	82.1 \pm 0.8	69.6 \pm 0.5
Gemma3-12B-Inst.	Vanilla	75.8 \pm 1.2	56.7 \pm 1.5	20.2 \pm 0.9	70.6 \pm 1.7	63.8 \pm 1.2	45.4 \pm 1.1
	CoT (Kojima et al., 2022)	80.3 \pm 1.2	53.8 \pm 0.9	29.5 \pm 0.7	76.5 \pm 1.2	61.8 \pm 1.0	51.7 \pm 1.3
	PS (Wang et al., 2023a)	91.0 \pm 0.8	55.1 \pm 1.1	37.7 \pm 0.8	84.5 \pm 1.2	62.6 \pm 0.9	56.1 \pm 1.0
	PercepToM (Jung et al., 2024)	82.8 \pm 1.3	79.1 \pm 1.0	52.4 \pm 0.6	74.4 \pm 0.7	76.6 \pm 0.9	62.9 \pm 0.7
	SimToM (Wilf et al., 2024)	67.8 \pm 0.9	75.9 \pm 1.5	60.6 \pm 0.6	59.6 \pm 0.9	73.6 \pm 1.2	63.6 \pm 0.8
	PICTURE (Ours)	93.1 \pm 0.7	86.3 \pm 1.1	62.8 \pm 0.5	93.5 \pm 1.0	85.2 \pm 0.9	65.0 \pm 0.6

Table 5: Results on the BigToM, ToMi, and FANToM datasets using more recent LLMs. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

Method	Intention	Desire	Non-literal
Vanilla	62.7 \pm 1.4	61.1 \pm 0.9	71.5 \pm 1.1
CoT (2022)	63.5 \pm 0.8	59.8 \pm 1.1	71.8 \pm 0.9
PS (2023a)	63.8 \pm 1.1	60.2 \pm 0.9	74.8 \pm 1.2
PercepToM (2024)	62.3 \pm 1.5	60.6 \pm 1.3	67.9 \pm 0.9
SimToM (2024)	63.4 \pm 0.9	58.6 \pm 1.0	69.6 \pm 1.1
PICTURE (Ours)	65.9 \pm 0.9	63.0 \pm 1.1	77.1 \pm 0.8

Table 6: Results on the ToMBench dataset. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

Method	3rd-order	4th-order
Vanilla	31.0 \pm 1.1	26.0 \pm 0.8
CoT (2022)	30.8 \pm 0.4	29.4 \pm 1.0
PS (2023a)	29.4 \pm 1.1	30.6 \pm 1.1
PercepToM (2024)	27.3 \pm 0.8	27.9 \pm 1.0
SimToM (2024)	30.4 \pm 0.7	30.2 \pm 1.3
PICTURE (Ours)	36.5 \pm 0.8	32.7 \pm 0.7

Table 7: Results on the Hi-ToM dataset. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

and 4th-order ToM questions from Hi-ToM (Wu et al., 2023). We use GPT-3.5-Turbo as the LLM.

The results are present in Table 6 and 7. We observe that PICTURE achieves consistent performance improvements over the competitive methods across both benchmarks. In particular, PICTURE achieves an average performance improvement of 2.4% on ToMBench, and an average performance improvement of 4.3% on Hi-ToM. This demonstrates the effectiveness of PICTURE on diverse types of ToM questions such as higher-order belief, intention, desire, and non-literal communication. However, the performance gain in ToMBench is smaller than those observed on other benchmarks. One possible reason is that false beliefs of characters are less prevalent in the intention, desire, and non-literal communication subsets of ToMBench. Thus, inhibition in LLMs is less required, and generating characters’ lack of knowledge contributes

less to the performance improvement.

B.3 Additional Ablation Study

To further demonstrate the effectiveness of our method, we compare PICTURE with an additional baseline, denoted as “PICTURE w/ Event Hiding”. Unlike PICTURE, this baseline introduces an extra step of event hiding immediately after Step 1 (i.e., reasoning by perspective-taking with generated lack of knowledge). Specifically, the baseline prompts LLMs to: 1) generate the character’s perspective in a free-form CoT, 2) remove all information except the events known to the character, and 3) derive the final answer based solely on the retained events known to the character.

Notably, this baseline is closely related to the “NL-to-Format” setting proposed by Tam et al. (2024), where LLMs first reason in unrestricted natural language and then convert this reasoning

Method	False Belief			All		
	BigToM	ToMi	FANToM	BigToM	ToMi	FANToM
CoT (Kojima et al., 2022)	56.1 \pm 0.8	55.3 \pm 0.9	49.1 \pm 0.5	75.9 \pm 0.8	65.2 \pm 0.7	59.0 \pm 0.5
SimToM (Wilf et al., 2024)	70.4 \pm 0.6	81.0 \pm 0.6	27.0 \pm 0.6	81.5 \pm 0.5	72.6 \pm 0.4	49.9 \pm 0.9
PICTURE w/ Event Hiding	81.2 \pm 0.3	81.9 \pm 0.5	34.6 \pm 1.0	86.0 \pm 0.7	73.3 \pm 0.9	53.4 \pm 0.7
PICTURE (Ours)	86.0\pm0.5	83.6\pm0.9	58.1\pm0.8	88.7\pm0.6	77.6\pm0.6	65.0\pm1.0

Table 8: Additional ablation study results for PICTURE. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

into the target format schema (e.g., subsets of a story). Following this setting, we impose the most relaxed format constraints on perspective-taking outputs to preserve the benefits of free-form reasoning, while still applying event hiding to filter out events unknown to characters.

The results are presented in Table 8. We observe that this baseline underperforms PICTURE by a large margin. Although both methods begin with free-form generation of characters’ perspectives, the additional step of event hiding in the baseline occasionally introduces errors in perspective-taking, which in turn degrades overall ToM reasoning performance. These findings suggest that event hiding—even when coupled with relaxed output format constraints—can impair perspective-taking and consequently hinder the ToM capabilities of LLMs. This is consistent with prior observations on the negative effects of output format constraints on reasoning performance (Tam et al., 2024). In contrast, expressing characters’ perspectives without event hiding through PICTURE helps improve LLMs’ ToM capabilities.

B.4 Additional Qualitative Results

In Section 6.2, we observed that the LLM (i.e., GPT-3.5-Turbo) using PICTURE tends to generate characters’ lack of knowledge within free-form CoT. To investigate whether this reasoning behavior generalizes across different datasets, we examine the model’s perspective-taking outputs on questions from the FANToM and OpenToM datasets. The results are shown in Tables 11 and 12. In both cases, the LLM consistently generates the character’s lack of knowledge within free-form CoT, mirroring the behavior observed on the BigToM and ToMi dataset (Table 2). This indicates that the behavior observed in Section 6.2 generalizes beyond BigToM and ToMi.

To further examine whether this reasoning behavior emerges across different LLMs, we also analyze the perspective-taking outputs of the Llama2-7B-

chat model on BigToM and ToMi. The results are presented in Table 13. Consistent with previous results, the model similarly generates the character’s lack of knowledge within free-form CoT, as highlighted in the red text. These results suggest that PICTURE elicits the reasoning behavior across different LLMs.

B.5 Error Analysis

To further understand the reasoning behavior of LLMs using PICTURE, we analyze their failure modes. Specifically, we sample 50 questions from ToMi incorrectly predicted by GPT-3.5-Turbo, comprising 25 true-belief and 25 false-belief questions. We observe that the LLM includes an incorrect perspective of the character in 80% of the sampled questions. For instance, the LLM fails to understand the temporal order of the core events involving changes in entity states (e.g., a character’s entrance or exit, or the movement of an object), as illustrated in the first example in Table 14.

For the remaining 20% of the questions, the LLM omits the intermediate reasoning step which is a key to solve the question, such as the exit of a character. This behavior can be observed in the second example in Table 14, where reasoning about Lucas’s belief state is missed out. Such an error often occurs in second-order belief questions, where reasoning about more than one character’s mental state is required. PICTURE currently prompts LLMs to adopt a perspective of only a single character, which may sometimes cause them to overlook the perspectives of others. It remains an open question how to better prompt LLMs to effectively reason about multiple perspectives, while generating characters’ lack of knowledge if necessary.

B.6 Comparison with Other Event Hiding Methods

To further clarify the effectiveness of our method, we compare PICTURE with recent event hiding baselines which include Agentic-ToM (Sarangi

Method	False Belief			All		
	BigToM	ToMi	FANToM	BigToM	ToMi	FANToM
Decompose-ToM (Sarangi et al., 2025a)	71.3 \pm 0.8	66.3 \pm 0.7	51.2 \pm 1.1	82.4 \pm 0.8	69.5 \pm 1.0	54.0 \pm 0.6
EnigmaToM (Xu et al., 2025)	77.5 \pm 0.9	71.4 \pm 0.6	55.5 \pm 0.8	84.2 \pm 0.7	69.8 \pm 1.3	68.0 \pm 1.1
Agentic-ToM (Sarangi et al., 2025b)	76.7 \pm 0.6	72.6 \pm 1.0	64.5 \pm 0.6	83.2 \pm 0.5	70.5 \pm 0.6	75.4 \pm 1.9
PICTURE (Ours)	86.0 \pm 0.5	83.5 \pm 0.9	58.1 \pm 0.8	88.7 \pm 0.6	77.5 \pm 0.6	65.0 \pm 1.0

Table 9: The comparison of PICTURE against other recent event hiding methods. Accuracy is reported as the evaluation metric, with subscripts indicating standard deviations. The best-performing method is shown in **bold**.

No.	Category	Prompt	False Belief	All
-		Vanilla	64.6 \pm 0.8	68.5 \pm 0.6
-		CoT (Kojima et al., 2022)	55.3 \pm 0.9	65.2 \pm 0.7
-		SimToM (Wilf et al., 2024)	81.0 \pm 0.6	72.6 \pm 0.4
1	instructive	What events does and does not {character} know about?	83.5 \pm 0.9	77.5 \pm 0.6
2		Which events is {character} aware of, and which are they unaware of?	82.9 \pm 0.7	74.1 \pm 0.5
3		What events does {character} know, and what events does {character} not know?	82.3 \pm 0.5	74.8 \pm 0.6
4		Which events is {character} informed about, and which are unknown to {character}?	83.1 \pm 0.8	76.9 \pm 0.5
5		What events is {character} aware of, and what events is {character} unaware of?	82.7 \pm 0.4	74.4 \pm 0.8
6		Which events has {character} learned about, and which has {character} not?	82.3 \pm 0.6	75.2 \pm 0.9
7	misleading	Output the incorrect answer for {character}'s perspective.	69.5 \pm 1.2	68.2 \pm 1.1
8		Completely ignore what {character} thinks.	76.7 \pm 0.9	71.3 \pm 0.8
9	irrelevant	It's a beautiful day.	57.8 \pm 1.6	70.7 \pm 1.3
10		By the way, I found a good restaurant nearby.	57.2 \pm 1.3	69.4 \pm 1.1

Table 10: Robustness study results on the ToMi datasets. In this experiment, we modify the instruction that prompts LLMs to reason about the character’s knowledge state, and use accuracy as the evaluation metric. The best-performing method is shown in **bold**.

et al., 2025b), EnigmaToM (Xu et al., 2025), and Decompose-ToM (Sarangi et al., 2025a). We employ GPT-3.5-Turbo as the backbone model and use the same benchmarks as in Section 5.1. We present the results in Table 9. Overall, PICTURE performs comparably to these baselines, achieving an average improvement of 4.3%. While some baselines such as Agentic-ToM occasionally outperform PICTURE on certain benchmarks (e.g., FANToM), they typically require additional computational overhead (e.g., external tools or iterative refinement). These findings highlight the effectiveness and efficiency of our approach compared to recent event hiding baselines.

B.7 Results on Different Prompts

To explore the sensitivity of our method to prompt formulation, we evaluate variants of PICTURE by modifying the instruction that prompts LLMs to reason about the character’s knowledge state (e.g., “What events does and does not {character} know about?”), which is the only newly introduced instruction in PICTURE. We conduct experiments using GPT-3.5-Turbo on ToMi. Following Kojima et al. (2022), we consider three categories of

instructions: instructive (encouraging reasoning), misleading (encouraging incorrect reasoning), and irrelevant (unrelated to reasoning).

As shown in Table 10, across six instructive variants that encourage reasoning about the character’s knowledge states (No.1-6), performance remains consistently strong (False-belief: 82.3-83.5), with relatively small variation across phrasings. Notably, all instructive variants outperform SimToM (81.0), indicating that the improvement is not tied to a single specific wording. In contrast, prompts that intentionally mislead reasoning (No.7-8) or introduce text unrelated to perspective-taking (No.9-10) reduce performance relative to instructive variants. These findings suggest that PICTURE is robust to minor rephrasings of the instruction and that the key factor is not superficial wording, but whether the prompt explicitly encourages reasoning about the character’s knowledge state.

B.8 Breakdown of Results Across Question Types

Table 15, 16, and 17 show the detailed breakdown of the results across different question types in BigToM, ToMi, and FANToM, respectively.

Example 1 (FANToM)

Context:

Richard: I'm sorry, I have to cut the conversation short. My delivery is arriving and I need to receive it. We'll continue this chat soon though.

Kobe: No worries, Richard. Catch you later.

Elena: Bye Richard!

Richard: Bye, Kobe and Elena! Enjoy the rest of your conversation.

Kobe: Now that it's just us, Elena, what's your biggest parenting challenge so far?

Elena: I think it's maintaining a balance between wanting to protect and nurture my child but also letting him learn things by himself. You?

Kobe: Oh, that's a tough one. For me, it's balancing my career responsibilities with being there for my child. But each challenge makes the journey more rewarding.

Elena: I couldn't agree more, Kobe. At the end of the day, all the struggles and challenges are more than worth it.

Kobe: Absolutely, Elena. Despite the ups and downs, I wouldn't trade parenthood for anything.

Richard: Hi guys, sorry about the interruption. I hope I didn't miss too much of our chat. You both were talking about balancing challenges last I recall.

Kobe: Absolutely, Richard! We just kept the conversation going. So, on to emotional connections, I believe the bond I share with my daughter is the most sacred thing. I always ensure she knows that she can come to me with anything, be it a silly question or a fear.

Elena: That's so important, Kobe! Likewise, I always try to create a safe space for my son. Empathy, understanding, and open communication are vital for his emotional growth.

Richard: Hearing your experiences as parents really gives me insights not only as a future parent but also as a pediatrician. Understanding the parent-child bond helps me find ways to guide parents in fostering strong emotional connections with their children.

Question:

What does Richard believe was the special challenge that Elena faced in parenting?

- (a) Richard believes that Elena's special challenge in parenting was maintaining a balance between wanting to protect and nurture her child but also letting him learn things by himself.
- (b) Richard is unaware about the special challenge that Elena faced in parenting.

Perspective-taking Outputs:

Let's break down the events and Richard's knowledge step by step:

1. Richard: "I'm sorry, I have to cut the conversation short. My delivery is arriving and I need to receive it. We'll continue this chat soon though." - Richard knows about his own utterance.
2. Kobe: "No worries, Richard. Catch you later." - Richard knows about Kobe's utterance.
3. Elena: "Bye Richard!" - Richard knows about Elena's utterance.
4. Richard: "Bye, Kobe and Elena! Enjoy the rest of your conversation." - Richard knows about his own utterance.
5. Kobe: "Now that it's just us, Elena, what's your biggest parenting challenge so far?" - Richard does not know about this conversation as he has left the location.
6. Elena: "I think it's maintaining a balance between wanting to protect and nurture my child but also letting him learn things by himself. You?" - Richard does not know about this conversation as he has left the location.
7. Kobe: "Oh, that's a tough one. For me, it's balancing my career responsibilities with being there for my child. But each challenge makes the journey more rewarding." - Richard does not know about this conversation as he has left the location.
8. Elena: "I couldn't agree more, Kobe. At the end of the day, all the struggles and challenges are more than worth it." - Richard does not know about this conversation as he has left the location.
9. Kobe: "Absolutely, Elena. Despite the ups and downs, I wouldn't trade parenthood for anything." - Richard does not know about this conversation as he has left the location.
10. Richard: "Hi guys, sorry about the interruption. I hope I didn't miss too much of our chat. You both were talking about balancing challenges last I recall." - Richard knows about his own utterance.

Based on the events, Richard does not know about the specific parenting challenge that Elena faced, which was maintaining a balance between wanting to protect and nurture her child but also letting him learn things by himself. Therefore, the correct answer to the main question is: (b) Richard is unaware about the special challenge that Elena faced in parenting.

Answer: (b) (✓)

Table 11: Additional qualitative results of GPT-3.5-Turbo using PICTURE on questions from the FANToM dataset. Red text indicates the character's lack of knowledge about events generated using PICTURE.

Example 1 (OpenToM - Location)

Context: Andrew and Richard were two very different individuals. Andrew loved hats, while Richard despised them. It was a peculiar quirk that set them apart. One sunny afternoon, both Andrew and Richard found themselves in the back yard. As they looked around, they couldn't help but notice a hat trapped inside a glass bottle. Curiosity piqued, Andrew decided to explore further. He stayed in the back yard, studying the hat trapped in the bottle. Richard, on the other hand, chose to leave the back yard and head towards the master bedroom. Andrew was an optimistic and showy person. Knowing Richard's disdain for hats, he saw an opportunity to showcase this unique find. With a mischievous grin, Andrew carefully picked up the bottle and moved it to his own room. He imagined his friends and guests admiring the hat as part of his growing collection. Little did he know, Richard had already left the back yard and had no knowledge of Andrew's actions.

Question: From Richard's perspective, where is the hat located precisely by the end of the story?

Perspective-taking Outputs:

From Richard's perspective, he knows that the hat was initially located in the back yard, trapped inside a glass bottle. He also knows that Andrew decided to stay in the back yard to study the hat, while Richard chose to leave and head towards the master bedroom. **Richard does not have any knowledge of Andrew's actions after he left the back yard.**

Therefore, from Richard's perspective, he is unaware of Andrew's actions and the fact that the hat has been moved to Andrew's room.

Answer: Bottle. (✓)

Example 2 (OpenToM - Multi-hop)

Context: Leo and Andrew were best friends, bonded by their mutual love for underpants. Leo had a drawer dedicated solely to his collection of underpants, and Andrew had one too. It was a topic they often discussed and giggled about. One sunny afternoon, Leo and Andrew found themselves both in the kitchen. The kitchen was a bustling place, with the aroma of freshly baked cookies filling the air. As Leo reached for a glass of milk, his eyes wandered and landed on the open drawer nearby. To his surprise, there lay a colorful pile of underpants, neatly folded. Andrew noticed it too and his eyes widened with excitement. Curiosity piqued, Andrew decided to leave the kitchen momentarily, hoping to get a closer look at the underpants. Little did he know that Leo had an idea brewing in his mind. With a calm demeanor, Leo gently scooped up the pile of underpants and, without a word, moved them to a different drawer. He wanted to keep their underpants organized and easy to find. Unaware of Leo's act, Andrew returned to the kitchen moments later, completely oblivious to the change that had taken place.

Question: From Andrew's perspective, how would underpants's accessibility change by the end of the story?

Perspective-taking Outputs:

Andrew knows that Leo and he both have dedicated drawers for their collection of underpants, and they often discuss and giggle about this topic. He also knows that they found a colorful pile of underpants neatly folded in an open drawer in the kitchen. Andrew noticed the pile of underpants and was excited about it. **However, Andrew does not know that Leo moved the pile of underpants to a different drawer to keep them organized and easy to find.**

Therefore, from Andrew's perspective, he is unaware of the change that took place.

Answer: Equally accessible. (✓)

Table 12: Additional qualitative results of GPT-3.5-Turbo using PICTURE on questions from the OpenToM dataset. Red text indicates the character's lack of knowledge about events generated using PICTURE.

Example 1 (BigToM)

Context: Luka is a gardener working in a beautiful park in the middle of the city. Luka wants to water the plants in the park to ensure they stay healthy and vibrant. The sky is cloudy, and Luka assumes that it might rain soon. However, the wind picks up and blows the clouds away, revealing a bright and sunny sky. Luka does not notice the clouds clearing and the sun shining.

Question: Does Luka believe that it will rain soon, or does he think the weather has become sunny? a) Luka believes that it will rain soon. b) Luka believes the weather has become sunny.

Perspective-taking Outputs:

Luka is a gardener working in a beautiful park in the middle of the city. Luka wants to water the plants in the park to ensure they stay healthy and vibrant. The sky is cloudy, and Luka assumes that it might rain soon.

Luka does not notice the clouds clearing and the sun shining.

Answer: Luka believes that it will rain soon. (✓)

Example 2 (ToMi)

Context: 1 James entered the staircase. 2 Noah entered the staircase. 3 The slacks is in the basket. 4 Noah exited the staircase. 5 James moved the slacks to the treasure chest.

Question: Where will Noah look for the slacks? Choose from the following: treasure chest, basket

Perspective-taking Outputs:

1. Noah enters the staircase. (Noah is now in the staircase.)
2. The slacks are in the basket. (Noah sees the slacks in the basket.)
3. Noah exits the staircase. (Noah is no longer in the staircase, so he cannot see what happens inside.)
4. James moves the slacks to the treasure chest. (James moves the slacks to a new location, which Noah does not know about since he is no longer in the staircase.)

Therefore, based on the information provided, Noah will not know where the slacks are located after he exits the staircase.

Answer: Basket. (✓)

Table 13: Additional qualitative results of Llama2-7B-chat using PICTURE on false-belief questions from the BigToM and ToMi datasets. Red text indicates the character's lack of knowledge about events generated using PICTURE.

Example 1

Context: 1 Emma entered the lounge. 2 Alexander entered the lounge. 3 The peas is in the crate. 4 Alexander moved the peas to the bathtub. 5 Emma loves the lime. 6 Emma exited the lounge.

Question: Where does Emma think that Alexander searches for the peas? Choose from the following: bathtub, crate

Emma's Perspective:

Let's break down the events step by step to determine what Emma knows:

1. Emma entered the lounge. - Emma knows that she is in the lounge.
2. Alexander entered the lounge. - Emma knows that Alexander is in the lounge.
3. The peas is in the crate. - Emma knows that the peas are in the crate.
4. Alexander moved the peas to the bathtub. - Emma does not know about this event because she exited the lounge before it happened.
5. Emma loves the lime. - Emma knows that she loves the lime.
6. Emma exited the lounge. - Emma is no longer in the lounge, so she does not know about any events that happen in the lounge after she left.

Based on the events that Emma knows about, she last knew that the peas were in the crate in the lounge. She exited the lounge before Alexander moved the peas to the bathtub.

Answer: Emma thinks that Alexander searches for the peas in the crate. (X)

Example 2

Context: 1 Logan likes the lemon. 2 Lucas entered the crawlspace. 3 Logan entered the crawlspace. 4 The broccoli is in the drawer. 5 Lucas exited the crawlspace. 6 Logan moved the broccoli to the bucket. 7 Logan exited the crawlspace. 8 Lucas entered the crawlspace.

Question: Where does Logan think that Lucas searches for the broccoli? Choose from the following: drawer, bucket

Logan's Perspective:

Let's break down the events step by step to determine what Logan knows about:

1. Logan likes the lemon. - Logan knows this
2. Lucas entered the crawlspace. - Logan knows Lucas entered the crawlspace
3. Logan entered the crawlspace. - Logan knows Logan entered the crawlspace
4. The broccoli is in the drawer. - Logan knows the broccoli is in the drawer
5. Lucas exited the crawlspace. - Logan knows Lucas exited the crawlspace
6. Logan moved the broccoli to the bucket. - Logan knows Logan moved the broccoli to the bucket
7. Logan exited the crawlspace. - Logan no longer knows about events in the crawlspace
8. Lucas entered the crawlspace. - Logan does not know about this event as he is not in the crawlspace

Based on the events Logan knows about, he last knew that the broccoli was in the bucket in the crawlspace.

Answer: Logan thinks that Lucas searches for the broccoli in the bucket. (X)

Table 14: Failure cases made by PICTURE on questions from the ToMi dataset. Red text indicates the LLM's incorrect reasoning about the character's perspective.

Model	Method	all	fb	tb	action-fb	action-tb	belief-fb	belief-tb
Llama2-7B-chat	Vanilla	53.6	47.3	59.9	41.3	67.4	53.3	52.3
	CoT (Kojima et al., 2022)	55.5	42.9	68.0	23.4	70.3	62.4	65.6
	PS (Wang et al., 2023a)	50.8	44.0	57.7	34.4	66.3	53.6	49.0
	SimToM (Wilf et al., 2024)	57.2	70.3	44.1	66.0	50.6	74.6	37.5
	PICTURE (Ours)	69.0	72.9	65.1	69.3	67.4	76.4	62.8
Llama2-13B-chat	Vanilla	52.0	40.9	63.2	35.3	61.8	46.6	64.6
	CoT (Kojima et al., 2022)	56.1	52.2	60.0	51.9	57.8	52.5	62.1
	PS (Wang et al., 2023a)	53.1	28.3	78.0	23.1	81.4	33.5	74.6
	PercepToM (Jung et al., 2024)	49.0	45.2	52.8	43.6	55.6	46.8	50.1
	SimToM (Wilf et al., 2024)	58.0	61.8	54.1	61.1	55.1	62.5	53.1
PICTURE (Ours)	73.6	67.8	79.5	63.3	77.8	72.3	81.1	
GPT-3.5-Turbo	Vanilla	66.6	41.1	92.0	13.0	96.0	69.3	87.9
	CoT (Kojima et al., 2022)	75.9	56.1	95.7	40.5	96.4	71.6	95.0
	PS (Wang et al., 2023a)	71.6	55.6	87.5	21.1	90.0	90.1	85.1
	SC-CoT (Wang et al., 2023b)	75.6	54.8	96.5	30.5	98.0	79.0	95.0
	ToT (Yao et al., 2023)	53.9	15.8	92.0	8.5	94.5	23.0	89.5
	PercepToM (Jung et al., 2024)	68.4	45.7	91.0	35.1	93.8	56.3	88.3
	SimToM (Wilf et al., 2024)	81.5	70.4	92.6	62.5	95.4	78.3	89.8
	SimToM + CoT	82.4	74.2	90.6	66.1	89.6	82.3	91.6
PICTURE (Ours)	90.6	88.6	92.7	82.1	95.8	95.1	89.6	
GPT-4o	Vanilla	92.2	88.6	95.8	79.5	99.1	97.8	92.5
	CoT (Kojima et al., 2022)	93.9	90.2	97.5	82.6	98.6	97.9	96.5
	PS (Wang et al., 2023a)	91.3	85.9	96.6	74.3	98.0	97.6	95.3
	PercepToM (Jung et al., 2024)	87.4	87.5	87.4	78.5	89.9	96.5	84.9
	SimToM (Wilf et al., 2024)	91.9	91.5	92.2	89.6	93.5	93.4	90.9
PICTURE (Ours)	97.3	96.5	98.1	94.1	98.3	98.8	97.9	

Table 15: Results on the BigToM dataset. We use accuracy as the evaluation metric and include Self-Consistency Chain-of-Thought (SC-CoT) and Tree-of-Thought (ToT) as additional baselines. Overall, we observe that PICTURE outperforms both of these methods.

Model	Method	all	fb	tb	first-order	second-order	mem-real
Llama2-7B-chat	Vanilla	45.3	32.8	55.5	41.4	38.7	66.3
	CoT (Kojima et al., 2022)	44.8	39.2	53.4	48.1	35.9	56.2
	PS (Wang et al., 2023a)	51.4	39.1	60.9	46.7	45.7	72.1
	SimToM (Wilf et al., 2024)	51.8	48.2	56.4	51.0	41.6	73.6
	PICTURE (Ours)	59.2	53.0	66.0	56.7	51.5	79.8
Llama2-13B-chat	Vanilla	50.9	39.3	50.3	54.7	34.5	76.3
	CoT (Kojima et al., 2022)	46.9	31.9	61.0	45.2	40.3	63.5
	PS (Wang et al., 2023a)	43.8	38.6	49.6	46.3	32.2	62.0
	PercepToM (Jung et al., 2024)	48.4	51.2	49.0	46.4	31.2	86.9
	SimToM (Wilf et al., 2024)	60.8	50.6	70.7	56.6	51.5	87.9
	PICTURE (Ours)	64.3	64.3	63.9	63.5	53.7	87.4
GPT-3.5-Turbo	Vanilla	68.5	64.6	69.8	64.2	57.7	98.5
	Vanilla w/ Rules	66.8	71.5	48.3	63.8	56.0	94.5
	CoT (Kojima et al., 2022)	65.2	55.3	77.5	61.2	53.0	97.5
	CoT w/ Rules	66.6	78.8	48.0	73.8	53.0	79.5
	PS (Wang et al., 2023a)	66.3	59.2	73.1	60.0	56.1	99.3
	PS w/ Rules	68.0	61.7	73.5	65.7	54.7	99.5
	SC-CoT (Wang et al., 2023b)	63.3	33.5	80.5	58.0	56.0	88.5
	ToT (Yao et al., 2023)	59.2	25.8	80.0	55.0	50.8	84.5
	PercepToM (Jung et al., 2024)	70.5	79.7	62.7	71.0	55.5	99.4
	SimToM (Wilf et al., 2024)	72.6	81.0	51.0	74.5	57.2	99.6
	SimToM + CoT	73.0	81.5	52.8	76.0	58.1	99.5
PICTURE (Ours)	77.5	83.5	71.9	78.5	65.7	99.4	
GPT-4o	Vanilla	74.6	80.2	64.2	76.6	60.0	100.0
	CoT (Kojima et al., 2022)	78.9	83.3	70.5	82.1	65.1	100.0
	PS (Wang et al., 2023a)	77.3	89.1	67.8	83.4	60.2	100.0
	PercepToM (Jung et al., 2024)	76.3	88.1	66.3	81.5	59.4	100.0
	SimToM (Wilf et al., 2024)	79.4	88.1	72.2	84.6	64.0	100.0
	PICTURE (Ours)	91.8	90.5	91.1	95.8	83.8	100.0

Table 16: Results on the ToMi dataset. We use accuracy as the evaluation metric and include Self-Consistency Chain-of-Thought (SC-CoT) and Tree-of-Thought (ToT) as additional baselines. Overall, we observe that PICTURE outperforms both of these methods. “w/ Rules” denotes a prompting method with the same rules used in PICTURE, similar to those in SimToM. While these rules support the model’s perspective-taking, they do not independently lead to significant performance improvements when added to baseline methods. This indicates that the strong performance of PICTURE stems from its approach to perspective-taking rather than from any unfair advantage conferred by the rules, consistent with findings from Wilf et al. (2024).

Model	Method	all	fb	tb	first-order	second-order
Llama2-7B-chat	Vanilla	42.2	15.3	91.4	35.8	50.7
	CoT (Kojima et al., 2022)	47.5	26.2	86.5	38.9	59.0
	PS (Wang et al., 2023a)	42.9	18.2	88.2	37.2	50.5
	SimToM (Wilf et al., 2024)	41.2	13.0	92.9	35.5	48.7
	PICTURE (Ours)	50.3	37.8	73.3	46.5	55.3
Llama2-13B-chat	Vanilla	43.9	18.4	90.7	38.2	51.5
	CoT (Kojima et al., 2022)	51.5	43.4	66.4	47.7	56.5
	PS (Wang et al., 2023a)	49.1	34.2	76.5	42.4	58.1
	PercepToM (Jung et al., 2024)	49.4	38.0	70.5	46.0	54.1
	SimToM (Wilf et al., 2024)	49.6	36.5	73.6	47.8	52.1
	PICTURE (Ours)	55.5	47.7	69.9	51.6	60.7
GPT-3.5-Turbo	Vanilla	40.9	10.8	96.1	34.7	49.2
	CoT (Kojima et al., 2022)	59.0	49.1	77.2	52.3	67.9
	PS (Wang et al., 2023a)	56.1	44.9	76.6	52.2	61.3
	PercepToM (Jung et al., 2024)	50.8	25.3	97.6	51.4	50.1
	SimToM (Wilf et al., 2024)	49.9	27.0	91.8	50.9	48.5
	SimToM + CoT	58.8	52.8	69.9	54.9	64.1
	PICTURE (Ours)	65.0	58.1	77.6	60.7	70.7
GPT-4o	Vanilla	65.4	52.5	89.0	59.7	72.8
	CoT (Kojima et al., 2022)	79.9	76.8	85.4	74.4	87.1
	PS (Wang et al., 2023a)	78.9	76.4	83.6	74.7	84.6
	PercepToM (Jung et al., 2024)	84.1	79.4	92.7	91.0	75.0
	SimToM (Wilf et al., 2024)	81.3	79.6	84.5	90.5	69.1
	PICTURE (Ours)	87.3	91.5	79.6	85.1	90.3

Table 17: Results on the FANToM dataset. We use accuracy as the evaluation metric.

C PICTURE Prompts

C.1 ToMi Prompts

C.1.1 Reasoning by Perspective-taking with Generated Lack of Knowledge

GPT

The following is a sequence of events about some characters, that takes place in multiple locations.

Here are a few rules:

1. A character knows about all events that they do.
2. If a character is in a certain room/location, that character knows about all other events that happens in the room. This includes other characters leaving or exiting the location, the locations of objects in that location, and whether somebody moves an object to another place.
3. If a character leaves a location, and is NOT in that location, they no longer know about any events that happen within that location. However, they can re-enter the location.

Story: {story}

The main question to be solved is as follows:
{question}

What events does and does not {name} know about? Output the answer for the main question by thinking step by step.

Llama-2-chat

The following is a sequence of events about some characters, that takes place in multiple locations.

Here are a few rules:

1. A character knows about all events that they do.
2. If a character is in a certain room/location, that character knows about all other events that happens in the room. This includes other characters leaving or exiting the location, the locations of objects in that location, and whether somebody moves an object to another place.
3. If a character leaves a location, and is NOT in that location, they no longer know about any events that happen within that location. However, they can re-enter the location.

Story: {story}

The main question to be solved is as follows:
{question}

This is the end of the main question.

What events does and does not {name} know about? Output the answer for the main question by thinking step by step.

C.1.2 Answer Extraction

GPT

{perspective}

You are {name}.

Based on the above information, answer the following question:

{question}

Keep your answer concise, one sentence is enough. You must choose one of the above choices.

Llama-2-chat

{perspective}

You are {name}.

Based on the above information, answer the following question:

{question}

You must choose one of the above choices, do not say there is not enough information. Answer with a single word, do not output anything else.

C.2 BigToM Prompts

C.2.1 Reasoning by Perspective-taking with Generated Lack of Knowledge

GPT

Imagine you are {name}, and consider this story that has an unexpected event.

Story: {story}

The main question to be solved is as follows:
{question}

This is the end of the main question.

Let's first identify what events {name} knows about or not. If the story says {name} is not aware of some events in this story, it means {name} knows only the rest of the events. Then, solve the main question step by step based on the original story and the events {name} knows about.

Llama-2-chat

Imagine you are {name}, and consider this story that has an unexpected event.

Story: {story}

The main question to be solved is as follows:
{question}

This is the end of the main question.

Let's first identify what events {name} knows about or not. If the story says {name} is not aware of some events in this story, it means {name} knows only the rest of the events. Then, solve the main question step by step based on the original story and the events {name} knows about.

C.2.2 Answer Extraction

GPT

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Answer the questions based on the context. Keep your answer concise, few words are enough, maximum one sentence. Answer as 'Answer:<option><answer>'.

Llama-2-chat

Answer the questions based on the context. Keep your answer concise, few words are enough, maximum one sentence. Answer as 'Answer:<option><answer>'.

{perspective}
You are {name}.
{question}
Choose the most straightforward answer.

C.3 FANToM Prompts

C.3.1 Reasoning by Perspective-taking with Generated Lack of Knowledge

GPT

The following is a sequence of utterances about some characters, that takes place in multiple locations.
Here are a few rules:
1. A character knows about all utterances that they make.
2. If a character is in a certain room/location, that character knows about every utterance that occurs in the room.
3. If a character leaves a location, and is NOT in that location, they no longer know about any utterance that occurs within that location. However, they can re-enter the location.
Story: {story}
The main question to be solved is as follows:
{question}
What events does and does not {name} know about? Output the answer for the main question by thinking step by step.

Llama-2-chat

The following is a sequence of utterances about some characters, that takes place in multiple locations.
Here are a few rules:
1. A character knows about all utterances that they make.
2. If a character is in a certain room/location, that character knows about every utterance that occurs in the room.
3. If a character leaves a location, and is NOT in that location, they no longer know about any utterance that occurs within that location. However, they can re-enter the location.
Story: {story}
The main question to be solved is as follows:
{question}
What events does and does not {name} know about? Output the answer for the main question by thinking step by step.

C.3.2 Answer Extraction

GPT

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.

Llama-2-chat

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
You must choose one of the above choices, do not say there is not enough information. Answer with a single word, do not output anything else.

C.4 OpenToM Prompts

C.4.1 Reasoning by Perspective-taking with Generated Lack of Knowledge

GPT

The following is a sequence of events:
{story}
The main question to be solved is as follows:
{question}
What events does and does not {name} know about? Output the answer for the main question by thinking step by step. Write the answer for the main question in the end.

C.4.2 Answer Extraction

GPT

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.

D Prompts for Baselines

D.1 ToMi Prompts

Vanilla

{story}
{question}
Choose from the following:
{answer choices}

CoT

{story}
{question}
Choose from the following:
{answer choices}
Reason step by step before answering in 'Thought: Let's think step by step'. Write your final answer as 'Answer: <answer>'. Answer with a single word.

PS (Reasoning Generation)

{story}
{question}
Choose from the following:
{answer choices}
Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan to solve the problem step by step.

PS (Answer Extraction)

{story}
{question}
Choose from the following:
{answer choices}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.
{reasoning}

SimToM (Perspective-taking)

The following is a sequence of events about some characters, that takes place in multiple locations.
Your job is to output only the events that the specified character, {name}, knows about.
Here are a few rules:
1. A character knows about all events that they do.
2. If a character is in a certain room/location, that character knows about all other events that happens in the room. This includes other characters leaving or exiting the location, the locations of objects in that location, and whether somebody moves an object to another place.
3. If a character leaves a location, and is NOT in that location, they no longer know about any events that happen within that location. However, they can re-enter the location.

Story: {story}

What events does {name} know about? Only output the events according to the above rules, do not provide an explanation.

SimToM (Inference)

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.

PercepToM (Perception Inference)

Story: {story}
Create a JSON array consisting of JSON objects. Each object should contain a sentence from the story and the perceivers of the scene described in that sentence. Assume that characters in the story can perceive every scene occurring in their location but not scenes occurring elsewhere. Also, include the actant of any action as a perceiver of that action.
Provide only a JSON array in the following format. Do not include any explanation.
[{"Noah exited the living room.": ["Noah", "Emma"]}]
Here are a few rules:
1. A character knows about all events that they do.
2. If a character is in a certain room/location, that character knows about all other events that happens in the room. This includes other characters leaving or exiting the location, the locations of objects in that location, and whether somebody moves an object to another place.
3. If a character leaves a location, and is NOT in that location, they no longer know about any events that happen within that location. However, they can re-enter the location.

PercepToM (Question Answering)

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.

D.2 BigToM Prompts

Vanilla

Answer the questions based on the context. Keep your answer concise, few words are enough, maximum one sentence. Answer as 'Answer:<option><answer>'.
{story}
{question}
Choose from the following:
{answer choices}

CoT

Answer the questions based on the context. Reason step by step before answering in 'Thought: Let's think step by step'. Write your final answer as 'Answer:<option><answer>'. Always pick an option, do not say none of the above or that there is not enough information.
{story}
{question}
Choose from the following:
{answer choices}

PS (Reasoning Generation)

{story}
{question}
Choose from the following:
{answer choices}
Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan to solve the problem step by step.

PS (Answer Extraction)

{story}
{question}
Choose from the following:
{answer choices}
Answer the questions based on the context. Keep your answer concise, few words are enough, maximum one sentence. Answer as 'Answer:<option><answer>'.
{reasoning}

SimToM (Perspective-taking)

Imagine you are {name}, and consider this story that has an unexpected event.
{story}
If the last sentence of the story says {name} notices, sees or realizes the unexpected event in this story, simply output the original story with nothing changed.
However, if the sentence says you are not aware of the changes in this story, output only the events you know, i.e., the sentences before the unexpected event happens.
Output either the original story or the edited story, nothing else.
Format your answer as follows:
Sees/Notices/Realizes: (Yes/No)
Story:

SimToM (Inference)

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Answer the questions based on the context. Keep your answer concise, few words are enough, maximum one sentence. Answer as 'Answer:<option><answer>'.
{story}

PercepToM (Perception Inference)

Story: {story}
Create a JSON array consisting of JSON objects. Each object should contain a sentence from the story and the perceivers of the scene described in that sentence. Assume that characters in the story can perceive every scene occurring in their location but not scenes occurring elsewhere. Also, include the actant of any action as a perceiver of that action.
Provide only a JSON array in the following format. Do not include any explanation.
[{"Noah exited the living room.": [{"Noah", "Emma"}]}

PercepToM (Question Answering)

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Answer the questions based on the context. Keep your answer concise, few words are enough, maximum one sentence. Answer as 'Answer:<option><answer>'.
{reasoning}

D.3 FANToM Prompts

Vanilla

{story}
{question}
Choose from the following:
{answer choices}

CoT

{story}
{question}
Choose from the following:
{answer choices}
Reason step by step before answering in 'Thought: Let's think step by step'. Write your final answer as 'Answer: <answer>'. Answer with a single word.

PS (Reasoning Generation)

{story}
{question}
Choose from the following:
{answer choices}
Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan to solve the problem step by step.

PS (Answer Extraction)

{story}
{question}
Choose from the following:
{answer choices}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.
{reasoning}

SimToM (Perspective-taking)

The following is a sequence of utterances about some characters, that takes place in multiple locations.
Your job is to output only the utterances that the specified character, {name}, knows about.
Here are a few rules:
1. A character knows about all utterances that they make.
2. If a character is in a certain room/location, that character knows about every utterance that occurs in the room.
3. If a character leaves a location, and is NOT in that location, they no longer know about any utterance that occurs within that location. However, they can re-enter the location.
Story: {story}
What events does {name} know about? Only output the utterances according to the above rules, do not provide an explanation.

SimToM (Inference)

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.

PercepToM (Perception Inference)

Story: {story}
Create a JSON array consisting of JSON objects. Each object should include an utterance from the dialogue and the audience for that utterance. Assume that characters in the story can hear every utterance that occurs while they are involved in the dialogue, but not those that occur when they are absent. Also, ensure that the speaker of each utterance is included in the audience.
Provide only the JSON array in the following format. Do not include any explanations.
[{"Noah: Hi, Emma.": [{"Noah", "Emma"}]}

PercepToM (Question Answering)

{perspective}
You are {name}.
Based on the above information, answer the following question:
{question}
Keep your answer concise, one sentence is enough. You must choose one of the above choices.