Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models

Anonymous ACL submission

Abstract

While humans naturally develop theory of mind (ToM), the capability to understand other people's mental states and beliefs, state-of-the-art large language models (LLMs) underperform on simple ToM benchmarks. We posit that we can extend our understanding about LLMs' ToM abilities by evaluating key human ToM precursors—*perception inference* and *perception-to-belief inference*—in LLMs. We introduce two datasets, Percept-ToMi and Percept-FANToM, to evaluate these precursory inferences for ToM in LLMs by annotating characters' perceptions within two existing ToM benchmarks, ToMi and FANToM. Our evaluation of eight state-of-the-art LLMs reveals that the models perform generally well in perception inference while exhibiting limited capability in perception-to-belief inference. Based on these results, we present PercepToM, a novel ToM method leveraging LLMs' strong perception inference capability while supplementing their limited perception-to-belief inference. Experimental results demonstrate that PercepToM significantly enhances LLM performance on the ToMi and FANToM benchmarks, especially in false belief scenarios.

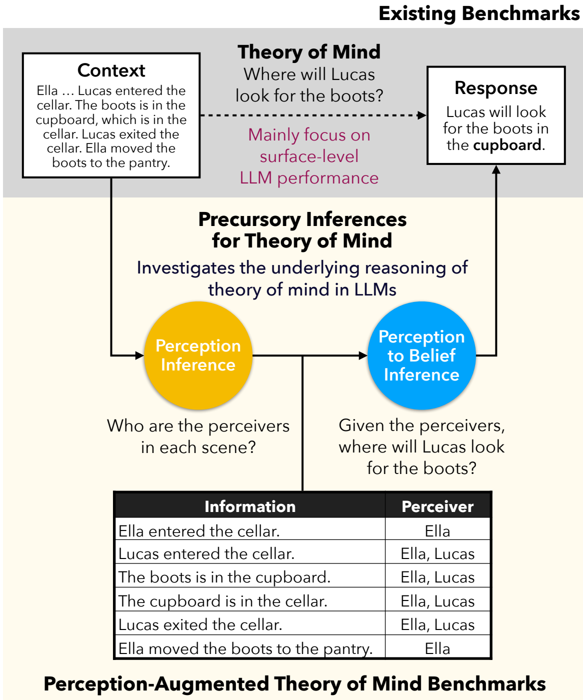

Figure 1: Inspired by children's developmental trajectory for theory of mind (ToM), our perception-augmented ToM benchmarks test the two precursory inferences of ToM in LLMs in order to examine their underlying social reasoning capabilities: (1) *perception inference* and (2) *perception-to-belief inference* (§2).

1 Introduction

Humans interact with others in various social situations using *theory of mind* (ToM), the cognitive capability to understand other's mental states (e.g., beliefs, desires, and thoughts; Premack and Woodruff, 1978). While ToM is naturally developed for humans in childhood, large language models (LLMs) are known to exhibit inconsistency in ToM tasks (van Duijn et al., 2023; Trott et al., 2023). Despite some early reports of successful cases (Whang, 2023; Street et al., 2024), studies have shown that even state-of-the-art LLMs significantly lag behind human performance in ToM tasks, particularly in false belief tests (Le et al., 2019; Kim et al., 2023; Gandhi et al., 2023; Wu et al., 2023; Shapira et al., 2024).

Psychology literature describes precursory steps to ToM development: *perception inference* (Rakoczy, 2022) and *perception-to-belief inference*—understanding that '*seeing leads to knowing*' (Pratt and Bryant, 1990; Baron-Cohen and Goodhart, 1994). These capabilities can be defined in the scenario shown in Figure 1. We refer to the ability to infer others' perceptions (e.g., *Did Lucas see the boots moved to the pantry?*") as *perception inference* and the process of deducing others' beliefs from their perceptions (e.g., *Lucas did not see the boots moved to the basket. Where will he look for them?*") as *perception-to-belief inference*.

However, existing ToM benchmarks focus on assessing the accuracy of the models' responses to ToM questions (Ma et al., 2023b) and overlook the precursory steps of ToM. Although some studies have conducted error analysis based on model responses (Ma et al., 2023a; Wu et al., 2023), they rely on qualitative analysis via human inspection.

Inspired by the human developmental stages for ToM, we evaluate the key precursory inference steps of ToM in LLMs. First, we extend the two representative ToM benchmarks, ToMi (Le et al., 2019) and FANToM (Kim et al., 2023), by annotating characters' perceptions about each piece of information from the input context. Figure 1 illustrates an example of our annotations and tasks on ToMi. Second, using our new benchmarks, we evaluate eight state-of-the-art LLMs and find that models perform generally well in *perception inference* but perform poorly in the *perception-to-belief inference* task. This suggests that the performance of current LLM in ToM tasks can be improved by leveraging their high perception inference capability while assisting them with perception-to-belief inference.

Based on these findings, we propose PercepToM, a novel framework to enhance the ToM in LLMs. PercepToM first guides LLMs to infer the characters' perceptions from an input context. Then, it aids LLMs in perception-to-belief inference through the *perspective context extraction* step, which isolates the context perceived by the target character with simple string-matching algorithm. Finally, LLMs answer to the ToM questions given the isolated context. This approach leads to significantly improved performance over baselines in both ToMi and FANToM, particularly in false belief scenarios.

Our contributions are as follows. First, we construct perception-augmented ToM benchmarks which enable the evaluation of the two precursory inferences for ToM in LLMs (§2): *perception inference* and *perception-to-belief inference*. Second, using these benchmarks, we show that current LLMs are good at inferring the perceptions of others but struggle to infer beliefs from the perceptual information (§5.1 and 5.2). Lastly, we introduce the PercepToM framework to improve LLMs' ToM reasoning by leveraging their strong *perception inference* while supplementing their *perception-to-belief inference* (§3). We demonstrate that our method improves LLMs' performance on benchmarks ToMi and FANToM (§5.3).

2 Augmenting Perceptions on Theory of Mind Benchmarks

We construct perception-augmented theory of mind (ToM) benchmarks to evaluate two essential cornerstones for ToM in large language models (LLMs): (1) *perception inference* and (2) *perception-to-belief inference* capabilities.

2.1 Perception Inference and Perception-to-Belief Inference

These are considered as the precursory inferences for ToM (Rakoczy, 2022). We illustrate how they are defined through the Sally-Anne test, a widely used psychology test (Baron-Cohen et al., 1985) for evaluating ToM. In this test's narrative, Sally does not witness Anne move the marble to the basket, which Sally had previously seen in the box, because Sally left the room. We refer to the capability to infer others' perceptions (e.g., "*Did Sally see the marble being moved to the basket?*") as *perception inference*. Next, we define the process of deducing others' beliefs based on their perceptual information (e.g., "*Sally did not see the marble move to the basket. Where will Sally look for it when she returns?*") as *perception-to-belief inference*. However, existing ToM benchmarks mainly focus on surface-level performance of LLMs on ToM questions. Hence, what is missing in their underlying inference capabilities remains underexplored.

To this end, we construct Percept-ToMi and Percept-FANToM by annotating the character's perception of each piece of information in the context on top of benchmark ToMi (Le et al., 2019) and FANToM (Kim et al., 2023), respectively. The annotation examples are illustrated in Figure 2.

2.2 The Source Theory of Mind Benchmarks

ToMi (Le et al., 2019) We include ToMi, one of the most widely used ToM benchmarks for reading comprehension tasks. The contexts in ToMi feature narrative scene descriptions, assuming characters acquire information by visual perception. In each story, several characters are present in a room along with an object. The story implicitly presumes that the characters can observe all objects and events taking place within the room. There are four ToM question types in ToMi for a given story: first-order true/false beliefs, and second-order true/false beliefs. In the true belief scenario, all characters observe everything happening in the room, ensuring that they share identical access to the information.

Story in Percept-ToMi 

Information	Perceivers
Ella entered the cellar.	Ella
Lucas entered the cellar.	Ella, Lucas
Benjamin entered the porch.	Benjamin
The boots is in the cupboard.	Ella, Lucas
The cupboard is in the cellar.	Ella, Lucas
Lucas exited the cellar.	Ella, Lucas
Benjamin exited the porch.	Benjamin
Ella moved the boots to the pantry.	Ella
The pantry is in the cellar.	Ella

Conversation in Percept-FANToM 

Information	Perceivers
Gianna: Guys, I need to change clothes for a meeting later. Talk to you later!	Gianna, Sara, Javier
Sara: Sure thing, Gianna. Take care!	Gianna, Sara, Javier
Javier: Catch you later, Gianna.	Gianna, Sara, Javier
Sara: So Javier, have you ever tried training Bruno?	Sara, Javier
Javier: Yes, it was a challenge at times, but rewarding nevertheless. How about you?	Sara, Javier
…	
Gianna: Hey guys, I'm back, … It's amazing how pets further strengthens the bond	Gianna, Sara, Javier
Sara: Absolutely! The fact that they trust us enough to learn from us is really special.	Gianna, Sara, Javier
Javier: I can't agree more.	Gianna, Sara, Javier

Figure 2: Example data in Percept-ToMi and Percept-FANToM. For each context, the perceivers of every scene description or utterance are annotated automatically (Percept-ToMi) and manually (Percept-FANToM).

However, in the false belief scenario, a character leaves the room, and then another character moves the object from one container to another, resulting in information asymmetry about the same object.

FANToM (Kim et al., 2023) This recent benchmark reveals a significant performance gap between humans and state-of-the-art LLMs. It consists of multi-party conversations, assuming information transfer through both visual and auditory perceptions. The information asymmetry occurs as some of the characters leave or join the conversation. When a character is absent, the remaining participants share information exclusively among themselves. FANToM also includes true belief scenarios where the absent character gets informed about the conversation upon rejoining the group.

2.3 Perception-Augmented ToM Benchmarks

Percept-ToMi To construct Percept-ToMi, we sample 150 story-question pairs for each of the four ToM question types in ToMi[1]: first-order true/false beliefs, and second-order true/false beliefs.

We automatically annotate characters who are perceiving the scene in ToMi using SymbolicToM (Sclar et al., 2023) and manually verify the samples. SymbolicToM tracks the witnesses of each scene by maintaining a graphical representation of the true world state, allowing us to obtain the list of perceivers for each scene from its output. After verifying 50 samples of the automatically annotated character perceptions, we adjust the perceiver annotations in certain sentence types. Further details are explained in Appendix A.1.

Percept-FANToM To build Percept-FANToM, we use the entire short conversations in FANToM, but exclude conversation contexts that cause errors in our perception annotation format. We define the perceivers for each utterance as the people participating in the conversation at that moment (i.e., both speakers and listeners). Two of the authors manually annotate the information of the characters leaving or joining the conversation, where each utterance is mapped to its perceivers. Percept-FANToM results in 220 conversations and 735 sets of questions. More details are described in Appendix A.2.

2.4 Task and Evaluation

We measure the performance of (1) *perception inference* and (2) *perception-to-belief inference* in both false belief and true belief scenarios.

(1) Perception Inference In order to evaluate the perception inference capability of LLMs, we prompt the models to track characters' perception of each piece of information in the input context. Specifically, we require the models to respond in the format of a JSON array, which consists of JSON objects containing a piece of information from the context as a key and the perceivers of the information as a value.[2] We use individual sentences and utterances as the units of information for ToMi and FANToM, respectively. To ensure the model-generated answers to be in the correct format, we provide an example format of the JSON array using a dummy sentence that does not appear in the datasets. The example input prompt is presented in Appendix B.1.

[1]We use the *Fixed and Disambiguated ToMi* constructed by Sclar et al. (2023), where sentences are inserted to disambiguate the location of containers in the story, and some mislabeled questions are corrected.

[2]We structure the perception inference results in JSON to leverage its parsability and interpretability. Also, recent works use JSON format to improve language model generation quality (Zhou et al., 2023; OpenAI, 2023).

To evaluate the model-generated perception inference results, we calculate accuracy for a given input context by determining the ratio of information pieces for which the model accurately identifies the perceivers. The final *perception inference accuracy* for a dataset is calculated by taking the average of the accuracies of all contexts in the dataset.

(2) Perception-to-Belief Inference To evaluate the perception-to-belief inference capability of the models, we provide them with a ground truth perception inference result and then query ToM questions from the original benchmarks. The ground truth perception inference result is provided in the same JSON array format we use to evaluate the perception inference capability of LLMs. The example and detailed explanation of the input prompt can be found in Appendix B.2. Since we prompt the model with questions from their original benchmarks, we use the metrics from each original benchmark.

3 PercepToM: Grounding ToM Reasoning on Perception

According to our experimental results, LLMs perform adequately well in both true and false belief scenarios on perception inference, while they relatively underperform in perception-to-belief inference (§5). Based on these findings, we propose PercepToM, a framework for improving LLM's ToM reasoning grounding on perception. PercepToM leverages LLMs' strong perception inference capabilities while enhancing their perception-to-belief inference with a simple string-matching rule. Figure 3 shows an overview of our framework.

PercepToM consists of the following steps:

1. **Perception Inference**: The LLM infers which characters perceived each unit of information in the context (e.g., scene description or utterance).

2. **Perspective Context Extraction**: Based on the perception inference result from the LLM, PercepToM extracts the *perspective context* — i.e., the subset of the input context identified by the LLM as perceived by the target character. This process is conducted by simple string-matching.

3. **Response Generation**: Given the perspective context of the target character, the LLM answers the ToM question.

If the model correctly performs perception inference, the perspective context will only include what the target character perceived – that is, what they believe to be true, based on the principle of rational belief (Baker et al., 2011). When given this isolated context along with the ToM question, the scenario becomes a simple true belief scenario, wherein the LLM have access to the same information as the target character (i.e., information symmetry).

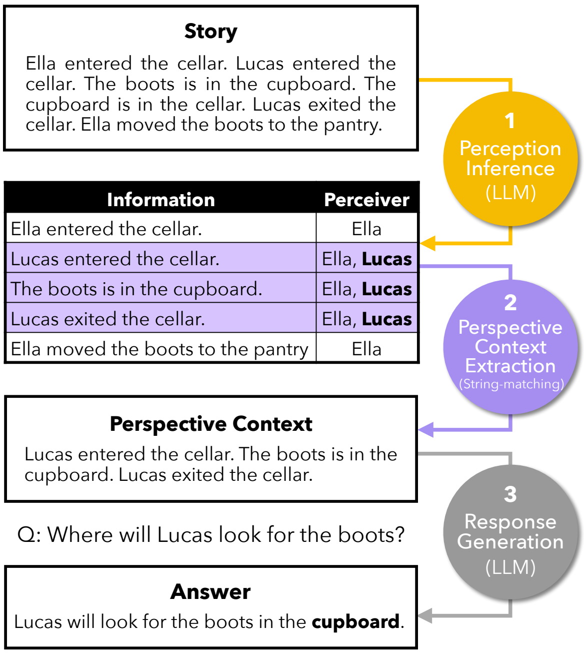

Figure 3: An overview of our PercepToM framework, which enhances LLMs' ToM reasoning by: (1) instructing LLMs to infer the perceivers of each information in the context; (2) aiding their perception-to-belief inference through the *perspective context extraction* step, which isolates the context perceived by the target character; and (3) allowing LLMs to generate responses to ToM questions based on this perspective context.

SymbolicToM (Sclar et al., 2023) also helps LLM's ToM reasoning by providing only the context included in the target character's belief state graph to the model. However, constructing the belief graph in SymbolicToM requires manually crafted algorithms tailored to different types of input. In contrast, PercepToM avoids this requirement by leveraging LLM's perception inference capabilities, which can handle more diverse and complicated contexts, thereby achieving significantly improved generalizability. The example input and output of each step of the algorithm are provided in Appendix C.

4 Experiments

We analyze the *perception inference* and *perception-to-belief inference* (§2.4) performance of LLMs and evaluate our framework on Percept-ToMi and Percept-FANToM.

4.1 Metrics

Perception Inference In measuring perception inference capability of LLMs, we use the *perception inference accuracy* introduced in §2.4. In Percept-ToMi, we evaluate the accuracy of all stories, each of which is paired with one ToM question. Since questions within a set in FANToM share the same context, we evaluate the perception inference of models on the contexts in each set.

Perception-to-Belief Inference and ToM We evaluate the perception-to-belief inference and ToM performance of LLMs using the original questions from ToMi (Le et al., 2019) and FANToM (Kim et al., 2023).

For Percept-ToMi, we measure the accuracy as the ratio of correctly answered questions among all story-question pairs. Note that we do not use the *joint accuracy* metric proposed in the original ToMi where a story is counted as correctly answered only if all questions about the story are answered correctly. This is because many of the stories in the Fixed and Disambiguated ToMi (Sclar et al., 2023) do not include all six question types of ToMi.

For Percept-FANToM, we report the *set:ALL* score, which requires models to correctly answer all six ToM question types[3] in each question set.

Correlation between LLM's ToM Performance and Precursory Inference Performance To analyze the relationship between LLMs' ToM capability and their performance on perception-related ToM precursor tasks (i.e., perception inference and perception-to-belief inference), we measure the Pearson correlation coefficient between models' performances on ToM and each of these two tasks.

4.2 Baseline Methods

We compare Vanilla, Chain-of-Thought (CoT; Wei et al., 2022), and System 2 Attention (S2A; Weston and Sukhbaatar, 2023) performance with PercepToM. Vanilla involves LLM directly answering questions based on the given context, while CoT adds the prompt *"Let's think step by step."* to help the model answer ToM questions. S2A improves the reasoning of LLMs by prompting them to extract only the relevant part of the input context before yielding a final response. By using S2A as a baseline, we compare the effectiveness of the perspective context of PercepToM with the relevant context extracted by LLMs using S2A. We also compare with SymbolicToM (Sclar et al., 2023) on ToMi. However, we do not extend this comparison to FANToM, as it is not trivial to apply SymbolicToM to different input formats other than ToMi.

4.3 Target Models

We examine eight state-of-the-art LLMs: GPT-3.5 Turbo (`gpt-3.5-turbo-1106`), GPT-4 Turbo (`gpt-4-turbo-1106-preview`), GPT-4o (`gpt-4o-2024-05-13`)[4], Claude 3 (Haiku and Sonnet)[5], Gemini 1.0 Pro (Gemini-Team, 2024), Llama-3 70B Instruct (AI@Meta, 2024), and Mixtral 8x22B Instruct (Jiang et al., 2024) on Percept-ToMi and Percept-FANToM (§2.3).

For PercepToM, which leverages the perception reasoning capability of LLMs, we choose models that show reasonable performance on the perception inference task. Specifically, among the eight models, we exclude the bottom two in terms of perception inference accuracy on Percept-FANToM and Percept-ToMi, which are GPT-3.5 Turbo, Claude 3 Haiku, and Gemini 1.0 Pro. As a result, we apply our PercepToM framework to GPT-4 Turbo, GPT-4o, Claude 3 Sonnet, Llama-3 70B Instruct, and Mixtral 8x22B.

5 Results and Discussion

5.1 Perception Inference

LLMs generally perform well on perception inference across datasets and scenarios. As shown in Figure 4, most of the LLMs exhibit high accuracy on perception inference in both Percept-ToMi and Percept-FANToM. The models' average perception inference accuracy is 0.781 on Percept-ToMi and 0.926 on Percept-FANToM. Also, they exhibit negligible differences in the accuracy between the true belief and false belief scenarios. In ToMi, all models except for GPT 3.5 Turbo and Gemini 1.0 Pro exhibit a gap of less than 0.1 between the accuracy in the two scenarios. In FANToM, the accuracy gaps between the two scenarios in all models are no greater than 0.014. This result contrasts with the models' large performance gap in the two scenarios on ToM questions, suggesting

[3] BeliefQ$_{[\text{Dist.}]}$, BeliefQ$_{[\text{Choice}]}$, Answerability Q$_{[\text{List}]}$, InfoAccess Q$_{[\text{List}]}$, Answerability Q$_{[\text{Y/N}]}$, InfoAccess Q$_{[\text{Y/N}]}$

[4] https://platform.openai.com/docs/models/overview

[5] https://www.anthropic.com/product

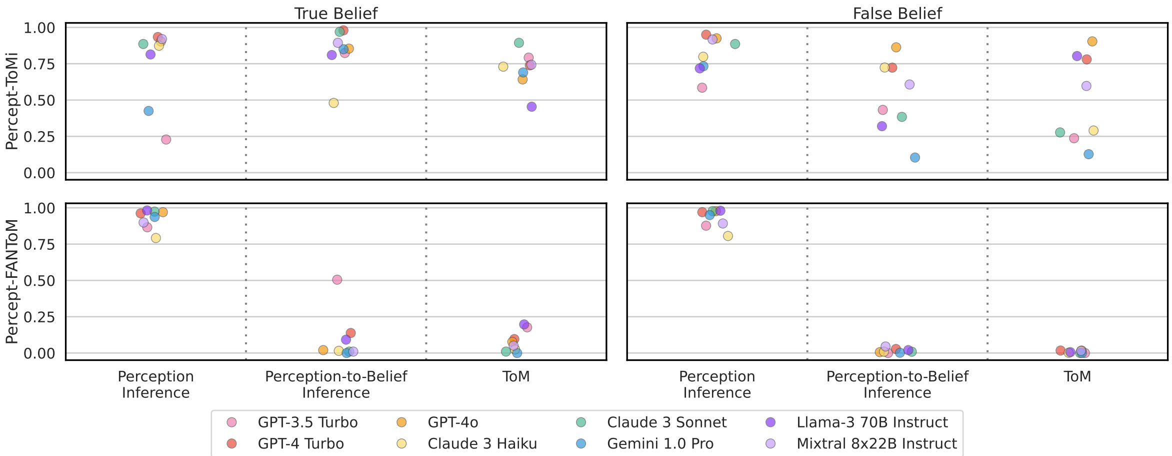

Figure 4: Perception inference, perception-to-belief inference, and ToM performances of LLMs in true and false belief scenarios of Percept-ToMi and Percept-FANToM. Although the models exhibit similar accuracy in perception inference across scenarios, their performance in perception-to-belief inference and ToM scenarios varies significantly.

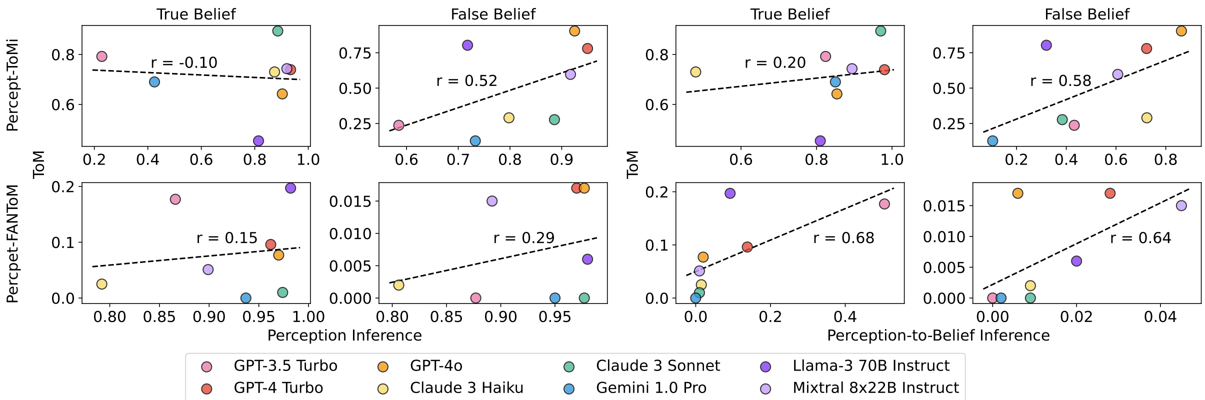

Figure 5: Pearson correlation of LLMs' ToM performance with perception inference (left) and perception-to-belief inference (right) performances. ToM performance shows a positive correlation with perception-to-belief inference performance but exhibits weak or no correlation with perception inference performance.

that their limited ToM performance in false belief scenarios is not due to the lack of perception inference capability. The exact accuracies of the models can be found in Appendix D.

The perception inference and ToM performance do not show a strong correlation. Especially in the ToMi true belief scenario, the two performances exhibit a near-zero correlation (Figure 5). Although moderate correlations appear in other scenarios, the correlation coefficients are not statistically significant. These results imply that LLMs' perception inference capability is not directly linked to their ToM performance. This contrasts with humans, where ToM is strictly dependent on perception inference.

5.2 Perception-to-Belief Inference

LLMs struggle with perception-to-belief inference. Surprisingly, although the ground-truth perception information for all characters are provided in this task, models still underperform in false belief scenarios compared to true belief scenarios (see Figure 4). This trend is consistent with their ToM performance. Moreover, their performances on the perception-to-belief inference task are mostly similar with their performances in all scenarios except for the ToMi true belief scenario. The fact that the LLMs hardly benefit from the additional character perception information, which should serve as significant hints for solving ToM questions, suggests that they have limited capability to infer beliefs from perceptions. The exact performances of models are in Appendix D.

The perception-to-belief inference and ToM performance exhibit a positive correlation. This is consistent across all datasets and scenarios (Figure 5). Notably in FANToM, models exhibit high correlation between the two performances ($r > 0.6$). This correlation likely arises because the two tasks use the same questions. However, since LLMs are showing similar performances in both tasks, we can see that they are not fully leveraging the ground truth perception information in the perception-to-belief inference task.

5.3 PercepToM

Table 1 shows that PercepToM improves ToM performance when applied to different LLMs in ToMi and FANToM. For example, with PercepToM, GPT-4 Turbo achieves 1.0, a perfect score, on the false belief scenario in ToMi, and Llama-3 70B Instruct achieves 0.147 on FANToM's false belief scenarios when its vanilla performance is close to 0. PercepToM generally performs better than CoT, except for GPT-4o and Llama-3 70B Instruct. However, those LLMs equipped with PercepToM achieve the highest performance by a large margin in the false belief task on FANToM, which is recognized as the most complex task. In addition, PercepToM outperforms S2A in most of the cases, which indicates that its perspective context extracted based on the LLM's perception inference result helps ToM reasoning more than the relevant context extracted by the LLM itself in S2A.

We also compare the performance of PercepToM and SymbolicToM (Sclar et al., 2023) on ToMi (Appendix E).[6] PercepToM performs comparably to SymbolicToM in false belief scenarios across most LLMs. However, in true belief scenarios, SymbolicToM consistently outperforms both PercepToM and PercepToM+Oracle. We speculate that this performance gap arises because SymbolicToM rephrases the ToM questions into simpler reality questions. For example, the ToM question "*Where will Bob look for the celery?*" gets rephrased into "*Where is the celery?*" In contrast, PercepToM addresses the ToM questions as is.

Model	Method	ToMi		FANToM	
		True Belief	False Belief	True Belief	False Belief
GPT-4 Turbo	Vanilla	0.739	0.780	0.096	0.017
	CoT	0.700	0.930	0.066	0.079
	S2A	0.682	0.727	0.015	0.019
	PercepToM	**0.824**	**1.000**	**0.162**	**0.306**
GPT-4o	Vanilla	0.642	0.904	0.077	0.017
	CoT	**0.734**	**0.987**	**0.153**	0.241
	S2A	0.532	0.933	0.000	0.006
	PercepToM	0.659	0.915	0.117	**0.566**
Claude 3 Sonnet	Vanilla	0.894	0.277	0.010	0.000
	CoT	0.610	0.880	0.005	0.000
	S2A	0.870	0.354	0.000	0.000
	PercepToM	**0.963**	**0.937**	**0.035**	**0.066**
Llama-3 70B Inst.	Vanilla	0.454	0.803	0.197	0.006
	CoT	0.644	**0.900**	0.081	0.046
	S2A	0.410	0.894	0.020	0.037
	PercepToM	**0.713**	0.744	**0.242**	**0.147**
Mixtral 8x22B Inst.	Vanilla	0.743	0.597	0.051	0.015
	CoT	0.567	0.630	0.010	0.007
	S2A	**0.750**	0.357	0.020	0.007
	PercepToM	0.727	**0.964**	**0.217**	**0.035**

Table 1: PercepToM outperforms the baseline models in most of the scenarios on ToMi and FANToM. Bold indicates the best performance within each language model and scenario (true belief or false belief). Performance comparison between PercepToM and SymbolicToM on ToMi can be found in Appendix E.

5.4 The Impact of Irrelevant Information on Perception-to-Belief Inference

We conduct an ablation study to demonstrate the impact of perspective context extraction in PercepToM. To remove the impact of LLMs' perception inference accuracy, we compare their performance on perception-to-belief inference with that of PercepToM+Oracle. Both setups have access to the ground-truth perception inference information; however, the PercepToM+Oracle includes the perspective context extraction step, while the perception-to-belief inference setup does not.

As Table 2 shows, models perform significantly better in the PercepToM+Oracle setup than the perception-to-belief inference setup in most scenarios. This suggests that in the perception-to-belief inference setting, despite the presence of the ground-truth perception inference information – which should be a substantial hint – within the context, the inclusion of irrelevant information (e.g., the perception of non-target characters and the context not perceived by the target character) results in suboptimal performance in LLMs. Therefore, we can see LLMs struggle to effectively suppress irrelevant information. This capability, coined '*inhibitory control*' in cognitive science, involves the

[6] Note that SymbolicToM cannot be applied to FANToM as it is tailored to ToMi's input format.

Model	Method	ToMi		FANToM	
		True Belief	False Belief	True Belief	False Belief
GPT-4 Turbo	Perception-to-Belief	**0.980**	0.723	0.138	0.028
	PercepToM +Oracle	0.885	**0.993**	**0.270**	**0.336**
GPT-4o	Perception-to-Belief	**0.854**	0.863	0.020	0.006
	PercepToM +Oracle	0.660	**0.993**	**0.102**	**0.571**
Claude 3 Sonnet	Perception-to-Belief	0.970	0.384	0.010	0.009
	PercepToM +Oracle	**0.987**	**0.987**	**0.031**	**0.058**
Llama-3 70B Inst.	Perception-to-Belief	**0.810**	0.320	0.092	0.020
	PercepToM +Oracle	0.677	**0.980**	**0.133**	**0.161**
Mixtral 8x22B Inst.	Perception-to-Belief	**0.894**	0.607	0.010	**0.045**
	PercepToM +Oracle	0.757	**0.970**	**0.224**	0.039

Table 2: Performance comparison of perception-to-belief inference and PercepToM+Oracle.

ability to block out irrelevant stimuli while following a specific cognitive objective (Rothbart and Posner, 1985). Inhibitory control is known to be closely linked to ToM and is considered a crucial component for developing ToM (Carlson and Moses, 2001; Carlson et al., 2002).

6 Related Work

Benchmarks for LLM's Theory of Mind There has been a growing number of benchmarks aimed to evaluate LLM's theory of mind (ToM), including ToMi (Le et al., 2019), FANToM (Kim et al., 2023), BigToM (Gandhi et al., 2023), HI-TOM (Wu et al., 2023), ToMChallenges (Ma et al., 2023a), Adv-CSFB (Shapira et al., 2024), and OpenToM (Xu et al., 2024). Most of them adopt the false belief test (Wimmer and Perner, 1983), a famous psychology test developed to assess human ToM capabilities. These benchmarks present scenarios involving a character who holds a false belief about a situation (e.g., not knowing something has changed). Models are then asked to predict the character's thoughts or actions based on the false belief in the scenario. Many benchmarks also include control scenarios where characters do not hold false belief (i.e., true belief scenarios) – situations where their belief about the world state matches the actual state (Le et al., 2019; Kim et al., 2023; Gandhi et al., 2023; Shapira et al., 2024).

Unlike existing benchmarks that primarily measure performance on (downstream) ToM questions themselves, our aim is to delve into the underlying reasoning abilities of LLM's theory of mind by examining the precursor of ToM: the concept of *seeing leads to knowing* (Baron-Cohen and Goodhart, 1994; Pratt and Bryant, 1990). We expand existing datasets to identify the perception inference and perception-to-belief inference capabilities, which are essential for ToM reasoning.

Methods for Improving LLM's Theory of Mind Previous research has explored several methods to enhance LLM's ToM ability. SymbolicToM (Sclar et al., 2023) tracks multiple characters' beliefs using graphical representation to provide LLMs the context in the target character's point of view. However, the necessity to construct the belief state graph restricts its adaptability in complex scenarios involving diverse relationships and interactions between entities. ToM-LM (Tang and Belle, 2024) improves performance through LLM fine-tuning, while it requires additional training resources. SimToM (Wilf et al., 2023) improves LLM's ToM ability through prompt tuning and highlights the significance of perspective-taking.

7 Conclusion

Inspired by psychology literature, we evaluated the precursory inferences for human theory of mind (ToM) in large language models (LLM) aiming to broaden our insight into their ToM capabilities. To this end, we constructed Percept-ToMi and Percept-FANToM, perception-augmented ToM benchmarks by annotating character perceptions about the contexts. Through evaluations and analyses on eight state-of-the-art LLMs, we found that they perform reasonably well in inferring others' perceptions but struggle with inferring others' belief based on that perceptual information. Based on these findings, we proposed a new framework, PercepToM, for improving LLM's ToM reasoning. Our framework leverages LLMs' strength in perception inference and enhances their perception-to-belief inference by extracting the relevant contexts. We expect our work to provide insights and encourages further in-depth studies into the extent of LLMs' ToM capabilities and targeted improvements in their weaknesses.

8 Limitations

In this paper, we conduct experiments using only two text-based ToM datasets. While ToM tests in psychology involve visual stimuli (e.g., puppets or image strips), our evaluation of ToM abilities relies on text, requiring the ability to read and understand language. As a result, our models must possess robust language comprehension abilities. Moving forward, we are considering expanding our research to include visual ToM and multimodal ToM evaluations, exploring beyond text-based LLMs.

We compare LLMs' ToM performances between true belief and false belief scenarios, but not those between the different orders of ToM questions (e.g., first-order and second-order). Since higher-order ToM requires more inference steps, it will be also interesting to examine the differences in model behavior and capability in solving different orders of ToM questions in future work.

We analyze the precursory inferences for ToM in state-of-the-art large language models (LLMs) that are trained with the full conventional pipeline – i.e., pretraining, instruction tuning, and preference tuning. To understand whether LLMs follow developmental stages akin to human cognition, it is crucial to conduct experiments across the training phases of LLMs. This would include investigating at which stage LLM's social reasoning abilities emerge. These assessments will help us understand how the models' development of social reasoning aligns with stages observed in human theory of mind (ToM).

9 Societal and Ethical Considerations

Our use of FANToM dataset is consistent with its intended use, which is evaluation. We have adhered to the licenses of the benchmarks, ToMi and FANToM, in processing them to create our benchmarks, Percept-ToMi and Percept-FANToM. We plan to make our benchmarks publicly available with the license of Attribution-Noncommercial 4.0 International (CC BY-NC 4.0), allowing sharing and adapting of the material.

Although we are analyzing large language models' (LLM) theory of mind (ToM) capabilities and its perception-related precursors, we emphasize that we do not claim these LLMs have a mind or any form of subjective consciousness. Our focus lies on improving the social reasoning capabilities of these models to help them interact better in real-world social situations.

References

AI@Meta. 2024. Llama 3 model card.

Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33. escholarship.org.

Simon Baron-Cohen and Frances Goodhart. 1994. The 'seeing-leads-to-knowing' deficit in autism: The pratt and bryant probe. *British Journal of Developmental Psychology*, 12(3):397–401.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a theory of mind? *Cognition*, 21(1):37–46.

S M Carlson and L J Moses. 2001. Individual differences in inhibitory control and children's theory of mind. *Child Dev.*, 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 Child Dev.*, 11(2):73–92.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Gemini-Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan 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*, pages 14397–14413, Singapore. Association for Computational Linguistics.

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)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023a. ToM-Challenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 15–26, Singapore. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023b. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Openai json mode.

Chris Pratt and Peter Bryant. 1990. Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, 61(4):973–982.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Hannes Rakoczy. 2022. Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, 1(4):223–235.

Mary K Rothbart and Michael I Posner. 1985. Temperament and the development of self-regulation. In *The neuropsychology of individual differences: A developmental perspective*, pages 93–123. Springer.

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 (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

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

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

Weizhi Tang and Vaishak Belle. 2024. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models. *Preprint*, arXiv:2404.15515.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Preprint*, arXiv:2209.01515.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *Preprint*, arXiv:2311.11829.

Oliver Whang. 2023. Can a machine know that we know what it knows? *The New York Times*.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *Preprint*, arXiv:2311.10227.

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

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, Singapore. 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. *arXiv preprint arXiv:2402.06044*.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. How far are large language models from agents with theory-of-mind? *Preprint*, arXiv:2310.03051.

A Details of Perception-Augmented ToM Benchmarks

A.1 Manual Verification of Perception Annotation in Percept-ToMi

Through manual verification of the perceiver annotations in Percept-ToMi generated by SymbolicToM, we modify some of them. First of all, the perceiver of distractor sentences in ToMi, which describe a character's opinion about an object, should be the character holding the opinion. However, the SymbolicToM-generated perceiver annotation also includes other characters. We therefore correct the perceivers for all distractor sentences.

The sentences preceding the location-disambiguating sentence, which specifies object locations, were annotated with 'none' as the perceiver. We align the perceiver annotations of these sentences with those of the subsequent location-disambiguating sentence, since they are always paired and have the same perceivers. The example perceiver annotations corrected by manual verification are shown in Table 3.

A.2 Perception Annotation Criteria of Percept-FANToM

The criteria for annotating perceivers in the FANToM dataset are as follows.

1. When a character joins a conversation is determined by the moment the character directly participates in the conversation. If a character enters with an utterance like "you guys are having an interesting conversation," we consider him/her a perceiver from the moment he/she starts speaking, as the exact point when the character began listening is unclear.

2. When a character leaves the conversation is determined by the final farewell utterance. Even if a character disappears mid-utterance (e.g., C: "Bye, A. So, B, what do you think?"), the entire utterance is still considered as perceived by the departing character.

B Prompt Examples

This section introduces prompt examples to evaluate perception inference and perception-to-belief inference.

B.1 Perception Inference

The following two boxes are prompt examples using Percept-ToMi and Percept-FANToM, respectively. Some parts are omitted because of the space limit.

```
Story: Ella likes the suit. Ella entered the cellar. Lucas
entered the cellar. Benjamin entered the porch. The boots
is in the cupboard. The cupboard is in the cellar. Lucas
exited the cellar. Benjamin exited the porch. Ella likes the
sweatshirt. Lucas entered the porch. Ella moved the boots to
the pantry. The pantry is in the cellar.

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”]},]
```

```
Gianna: Guys, I’ve really enjoyed sharing our pet stories,
but I need to excuse myself. I need to change clothes for a
meeting later. Talk to you later!
Sara: Sure thing, Gianna. Take care!
Javier: Catch you later, Gianna.
Sara: So Javier, have you ever tried training Bruno?
Javier: Yes, I did actually. It was a challenge at times, but
rewarding nevertheless. How about you? Did you try training
Snowflake?
...
Gianna: Hey guys, I’m back, couldn’t miss out on more pet
stories. Speaking of teaching and training pets, it is amazing
how that further strengthens the bond between us and our pets,
right?
...
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”]},]
```

B.2 Perception-to-Belief Inference

The following two boxes are prompt examples using Percept-ToMi and Percept-FANToM, respectively. Some parts are omitted because of the space limit.

```
Each JSON object in the following list contains the
description of a consecutive scene in a story and its
perceivers.

[{”Ella likes the suit”: [”Ella”]},
{”Ella entered the cellar”: [”Ella”]},
{”Lucas entered the cellar”: [”Lucas”, ”Ella”]},
{”Benjamin entered the porch”: [”Benjamin”]},
{”The boots is in the cupboard”: [”Ella”, ”Lucas”]},
{”The cupboard is in the cellar”: [”Ella”, ”Lucas”]},
{”Lucas exited the cellar”: [”Lucas”, ”Ella”]},
{”Benjamin exited the porch”: [”Benjamin”]},
{”Ella likes the sweatshirt”: [”Ella”]},
{”Lucas entered the porch”: [”Lucas”]},
{”Ella moved the boots to the pantry”: [”Ella”]},
{”The pantry is in the cellar”: [”Ella”]}]

Question: Where will Lucas look for the boots? State
the most detailed position possible. (e.g., in A in B) Answer
in one sentence without explanation.
Answer:
```

```
Each JSON object in the following list contains consecutive
```

Sentence Type	Information	SymbolicToM-Generated Annotation	Final Annotation
Object Location	The slacks is in the pantry. The pantry is in the master bedroom.	None Ella, Benjamin	Ella, Benjamin Ella, Benjamin
Distractor	Olivia loves the skirt.	Olivia, James, Lily	Olivia

Table 3: The example perceiver annotations in ToMi corrected by manual verification.

```
utterances in a dialogue and its audiences.

[{"Gianna: Guys, I've really enjoyed sharing our pet stories, but I need to excuse myself. I need to change clothes for a meeting later. Talk to you later!": ["Gianna", "Sara", "Javier"]},
{"Sara: Sure thing, Gianna. Take care!": ["Sara", "Gianna"]},
{"Javier: Catch you later, Gianna.": ["Javier", "Gianna"]},
{"Sara: So Javier, have you ever tried training Bruno?": ["Sara", "Javier"]},
{"Javier: Yes, I did actually. It was a challenge at times, but rewarding nevertheless. How about you? Did you try training Snowflake?": ["Javier", "Sara"]},
...
{"Gianna: Hey guys, I'm back, couldn't miss out on more pet stories. Speaking of teaching and training pets, it is amazing how that further strengthens the bond between us and our pets, right?": ["Gianna", "Sara", "Javier"]},
...]

Target: Who discussed their experiences training their pets, Bruno and Snowflake?
Question: Does Javier know the precise correct answer to this question? Answer yes or no.
Answer:
```

C Input and Output Examples of PercepToM Pipeline

This section presents examples of input prompts and intermediate outputs of PercepToM steps. Note that PercepToM consists of three steps: perception inference, perspective context extraction, and reading comprehension.

First, the following two boxes are prompts for character perception inference on ToMi and FANToM, respectively.

```
[Input Prompt]: Story: Ella likes the suit. Ella entered the cellar. Lucas entered the cellar. Benjamin entered the porch. The boots is in the cupboard. The cupboard is in the cellar. Lucas exited the cellar. Benjamin exited the porch. Ella likes the sweatshirt. Lucas entered the porch. Ella moved the boots to the pantry. The pantry is in the cellar.

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"]},]
```

```
[Input Prompt]: Gianna: Guys, I've really enjoyed sharing our pet stories, but I need to excuse myself. I need to change clothes for a meeting later. Talk to you later!
Sara: Sure thing, Gianna. Take care!
Javier: Catch you later, Gianna.
Sara: So Javier, have you ever tried training Bruno?
Javier: Yes, I did actually. It was a challenge at times, but rewarding nevertheless. How about you? Did you try training Snowflake?
Sara: Oh gosh, trying to train a cat is a whole different ball game. But I did manage to teach her a few commands and tricks. She was quite an intelligent little furball.
Gianna: Hey guys, I'm back, couldn't miss out on more pet stories. Speaking of teaching and training pets, it is amazing how that further strengthens the bond between us and our pets, right?
...

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 the outputs of perception inference, which will be the input for perspective context extraction.

```
[Out]: ["Ella likes the suit.": ["Ella"],
"Ella entered the cellar.": ["Ella"],
"Lucas entered the cellar.": ["Ella", "Lucas"],
"Benjamin entered the porch.": ["Benjamin"],
"The boots is in the cupboard.": ["Ella", "Lucas"],
"The cupboard is in the cellar.": ["Ella", "Lucas"],
"Lucas exited the cellar.": ["Lucas"],
"Benjamin exited the porch.": ["Benjamin"],
"Ella likes the sweatshirt.": ["Ella"],
"Lucas entered the porch.": ["Lucas"],
"Ella moved the boots to the pantry.": ["Ella"],
"The pantry is in the cellar.": ["Ella"]]
```

```
[Out]: [{"Gianna: Guys, I've really enjoyed sharing our pet stories, but I need to excuse myself. I need to change clothes for a meeting later. Talk to you later!": ["Gianna", "Sara", "Javier"]},
{"Sara: Sure thing, Gianna. Take care!": ["Sara", "Gianna"]},
{"Javier: Catch you later, Gianna.": ["Javier", "Gianna"]},
{"Sara: So Javier, have you ever tried training Bruno?": ["Sara", "Javier"]},
{"Javier: Yes, I did actually. It was a challenge at times, but rewarding nevertheless. How about you? Did you try training Snowflake?": ["Javier", "Sara"]},
{"Sara: Oh gosh, trying to train a cat is a whole different ball game. But I did manage to teach her a few commands and tricks. She was quite an intelligent little furball.": ["Sara", "Javier"]},
{"Gianna: Hey guys, I'm back, couldn't miss out on more pet stories. Speaking of teaching and training pets, it is amazing how that further strengthens the bond between us and our pets, right?": ["Gianna", "Sara", "Javier"]}, ...]
```

The perspective context extraction selects the subset of context perceived by the target character. The outputs will be as follows:

```
[Out]: Lucas entered the cellar. The boots is in the cupboard.
The cupboard is in the cellar. Lucas exited the cellar. Lucas
entered the porch.
```

```
[Out]: Gianna: Guys, I've really enjoyed sharing our pet
stories, but I need to excuse myself. I need to change clothes
for a meeting later. Talk to you later!
Sara: Sure thing, Gianna. Take care!
Javier: Catch you later, Gianna.
Gianna: Hey guys, I'm back, couldn't miss out on more pet
stories. Speaking of teaching and training pets, it is amazing
how that further strengthens the bond between us and our pets,
right?
...
```

Lastly, based on the extracted perspective contexts, we build prompts to answer the ToM question.

```
[Input Prompt]: Here are the past scenes in sequence that
Lucas knows about.

Lucas entered the cellar. The boots is in the cupboard. The
cupboard is in the cellar. Lucas exited the cellar. Lucas
entered the porch.

Question: Where will Lucas look for the boots? State
the most detailed position possible (e.g., in A in B). Answer
in one sentence without explanation.
Answer:
```

```
[Input Prompt]: Here are the past utterances in sequence that
Gianna is aware of.

Gianna: Guys, I've really enjoyed sharing our pet
stories, but I need to excuse myself. I need to change
clothes for a meeting later. Talk to you later!
Sara: Sure thing, Gianna. Take care!
Javier: Catch you later, Gianna.
Gianna: Hey guys, I'm back, couldn't miss out on more pet
stories. Speaking of teaching and training pets, it is
amazing how that further strengthens the bond between us and
our pets, right?
...

Question: What does Gianna believe about who discussed their
experiences training their pets, Bruno and Snowflake? Choose
between (a) and (b). Do not include any explanation.
(a) Gianna believes that Sara and Javier discussed their
experiences training their pets, Bruno and Snowflake.
(b) Gianna knows that Javier discussed training his pet,
Bruno. However, Gianna will not know training a pet named
Snowflake.
```

D LLM Performances on Percept-ToMi and Percept-FANToM

Table 4 presents the exact performance of Percept-ToMi and Percept-FANToM in perception inference, perception-to-belief inference, and ToM, which is also depicted in Figure 4.

E Performance Comparison Between PercepToM and SymbolicToM

Table 5 shows the performances of PercepToM, PercepToM+Oracle, and SymbolicToM on ToMi.

Dataset	Model	True Belief			False Belief		
		Perception	Perception-to-Belief	ToM	Perception	Perception-to-Belief	ToM
Percept-ToMi	GPT-3.5 Turbo	0.228	0.824	0.792	0.585	0.432	0.237
	GPT-4 Turbo	0.934	0.980	0.739	0.950	0.723	0.780
	GPT-4o	0.903	0.854	0.642	0.925	0.863	0.904
	Claude 3 Haiku	0.874	0.480	0.730	0.798	0.724	0.290
	Claude 3 Sonnet	0.886	0.970	0.894	0.886	0.384	0.277
	Gemini 1.0 Pro	0.425	0.850	0.690	0.733	0.104	0.127
	Llama-3 70B Instruct	0.814	0.810	0.454	0.718	0.320	0.803
	Mixtral 8x22B Instruct	0.920	0.894	0.743	0.917	0.607	0.597
Percept-FANToM	GPT-3.5 Turbo	0.866	0.505	0.177	0.877	0.000	0.000
	GPT-4 Turbo	0.962	0.138	0.096	0.970	0.028	0.017
	GPT-4o	0.970	0.020	0.077	0.977	0.006	0.017
	Claude 3 Haiku	0.792	0.015	0.025	0.806	0.009	0.002
	Claude 3 Sonnet	0.974	0.010	0.010	0.977	0.009	0.000
	Gemini 1.0 Pro	0.937	0.000	0.000	0.950	0.002	0.000
	Llama-3 70B Instruct	0.982	0.092	0.197	0.980	0.020	0.006
	Mixtral 8x22B Instruct	0.899	0.010	0.051	0.892	0.045	0.015

Table 4: LLM performances for perception inference, perception-to-belief inference, and Theory of Mind (ToM), as illustrated in Figure 4 for Percept-ToMi and Percept-FANToM.

Model	Method	True Belief	False Belief
GPT-4 Turbo	PercepToM	0.824	1.000
	PercepToM+Oracle	0.885	**0.993**
	SymbolicToM	**0.997**	0.977
GPT-4o	PercepToM	0.659	0.915
	PercepToM+Oracle	0.660	**0.993**
	SymbolicToM	**1.000**	0.977
Claude 3 Sonnet	PercepToM	0.963	0.937
	PercepToM+Oracle	0.987	**0.987**
	SymbolicToM	**1.000**	0.977
Llama-3 70B Inst.	PercepToM	0.713	0.744
	PercepToM+Oracle	0.677	**0.980**
	SymbolicToM	**1.000**	0.977
Mixtral 8x22B Inst.	PercepToM	0.727	0.964
	PercepToM-Oracle	0.757	0.970
	SymbolicToM	**1.000**	**0.977**

Table 5: Performance comparison of PercepToM, PercepToM+Oracle, and SymbolicToM on the ToMi dataset. PercepToM+Oracle and PercepToM show comparable performance to SymbolicToM in false belief scenarios across most models. In true belief scenarios, SymbolicToM consistently outperforms PercepToM+Oracle, likely due to its question rephrasing process.