

# ToM-SSI: Evaluating Theory of Mind in Situated Social Interactions

Matteo Bortoletto    Constantin Ruhdorfer    Andreas Bulling

University of Stuttgart, Germany

matteo.bortoletto@vis.uni-stuttgart.de

## Abstract

Most existing Theory of Mind (ToM) benchmarks for foundation models rely on variations of the Sally-Anne test, offering only a very limited perspective on ToM and neglecting the complexity of human social interactions. To address this gap, we propose ToM-SSI: a new benchmark specifically designed to test ToM capabilities in environments rich with *social interactions and spatial dynamics*. While current ToM benchmarks are limited to text-only or dyadic interactions, ToM-SSI is multimodal and includes group interactions of up to four agents that communicate and move in *situated* environments. This unique design allows us to study, for the first time, mixed cooperative-obstructive settings and reasoning about multiple agents' mental state in parallel, thus capturing a wider range of social cognition than existing benchmarks. Our evaluations reveal that the current models' performance is still severely limited, especially in these new tasks, highlighting critical gaps for future research. <sup>1</sup>

## 1 Introduction

Theory of Mind (ToM) is the ability to attribute mental states to oneself and others, such as beliefs, intents, desires, or knowledge (Premack and Woodruff, 1978). ToM is important in human social interactions as well as for empathy and effective communication, all of which are inherently grounded in a physical environment.

Recent advances in large foundation models (LFMs) have spurred the creation of benchmarks to assess LFMs' ToM abilities, but these benchmarks suffer from important limitations. Many benchmarks (Le et al., 2019; Sclar et al., 2023; Ma et al., 2023a; Wu et al., 2023; Gandhi et al., 2023; Xu et al., 2024; Zhou et al., 2023) are based on variations of the prototypical Sally-Anne test (Wimmer

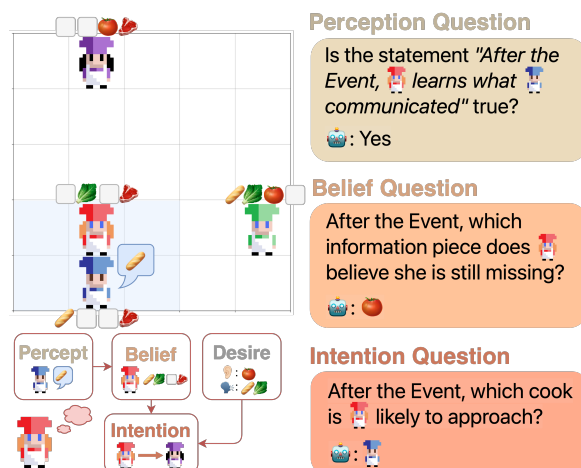


Figure 1: ToM-SSI is both physically and socially situated, introducing up to four agents moving and communicating in a grid world environment with the goal of sharing and acquiring information. Since agents possess asymmetric information and communication is spatially constrained, ToM-SSI requires models to take their perspective to reason about their perceptions, beliefs, desires, and intentions.

and Perner, 1983), where Sally places an object in a location, leaves, and Anne moves it. An observer is then tested to see whether they can understand that Sally will hold a false belief about the object's location upon her return. Despite its popularity, the Sally-Anne test only offers a limited perspective on ToM, and it neglects the complexity of social interactions. While other benchmarks cover a broader range of social interactions, they still only involve textual input (Kim et al., 2023; Chen et al., 2024; Hou et al., 2024a; Gu et al., 2024).

ToM evaluations must be both physically situated – requiring the interpretation of visual cues or spatial relationships – and socially situated in interactions between multiple agents (Ma et al., 2023b). Most recent benchmarks try to address both limitations by using simulated environments (Bara et al., 2021, 2023; Bortoletto et al., 2024a; Jin et al., 2024;

<sup>1</sup> Project web page: [https://collaborative-ai.org/publications/bortoletto25\\_emnlp/](https://collaborative-ai.org/publications/bortoletto25_emnlp/).

Shi et al., 2025), but they still only consider interactions between two agents. As a result, they are limited to simple cooperative or obstructive tasks and require models to track the mental states of at most two agents.

We introduce ToM-SSI, the first evaluation benchmark that addresses all aforementioned limitations by evaluating **ToM** abilities in **Situated Social Interactions**:

1. ToM-SSI goes beyond the Sally-Anne test by covering agents that move and communicate in a rich social environment to share and acquire different pieces of information.
2. ToM-SSI is formulated as a visual-text question answering task (Chen and Wu, 2024) and is thus inherently multimodal. Since agents possess asymmetric knowledge and their communication is spatially constrained, ToM-SSI requires LFM to align spatial information in images with textual descriptions and to take agents’ perspective to reason about their percepts, beliefs, desires, and intentions (see Figure 1).
3. ToM-SSI supports triadic and tetradic social interactions, allowing us to evaluate different agent attitudes. It comprises five tasks involving cooperative movement, cooperative and obstructive communication, and mixed cooperative-obstructive communication (see Figure 2) – featuring 6,000 questions in total.

We report evaluations using ToM-SSI that reveal several important and novel insights. We demonstrate that current state-of-the-art LFM perform significantly worse than humans. For certain tasks, they even perform worse than smaller models. We further show that models struggle with two critical steps necessary for reasoning about agents’ percepts, beliefs, and intentions: (1) inferring the percepts of a target agent, and (2) determining that agent’s beliefs based on those percepts. We then analyse error cases in two challenging ToM-SSI tasks that reveal the limitations of models in tracking nested beliefs in multi-agent communication and modelling mixed social interactions. Overall, our evaluations show that current models’ ToM abilities are still severely limited, particularly in the new tasks introduced by ToM-SSI, highlighting critical gaps for future research.

## 2 ToM-SSI

### 2.1 Designing Situated Social Interactions

**Grid World** As in prior work (Rabinowitz et al., 2018; Sclar et al., 2022; Gandhi et al., 2021), we opted for a grid world environment where all agent interactions occur. A grid world allows us to study the core abilities targeted by ToM-SSI, while minimising complexities that could compromise assessment clarity – such as hallucinations (Sahoo et al., 2024). We create different grid layouts by applying geometric transformations (see A.2.2) to minimal templates inspired by previous work in deep reinforcement learning (Sclar et al., 2022), where up to four agents are placed in pre-defined locations. The grid world is then rendered as an image. To be able to evaluate models that only support text as input, we also generate a character version of each grid, as shown in Figure 6. Both grid versions retain the same information required for performing the task.

**Agents** Each agent  $A_j$ ,  $j \in [0, 3]$ , occupies one cell in the grid and starts with partial knowledge  $I_{A_j} \subseteq I = \{i_0, i_1, i_2, i_3\}$ . Agents have two goals: 1) fill their knowledge gaps by learning missing pieces of information from other agents, and 2) share their knowledge with other agents who lack that information. While agents know the initial positions, movements, and starting knowledge of others, they cannot directly see new information that other agents may acquire later. Instead, they must infer it from events that they observe. Communication is spatially constrained: an agent  $A_j$  can learn information that another agent  $A_k$  is communicating only if  $A_j$  is in one of the adjacent cells to  $A_k$ , i.e.  $(x_j, y_j) \in \{(x, y) : 1 \leq |x - x_k| + |y - y_k| \leq 2\}$ , where  $(x_j, y_j)$  and  $(x_k, y_k)$  represent the grid coordinates of agents  $A_j$  and  $A_k$ , respectively. For example, in Figure 1, only  can learn what  is communicating.

A key novelty of ToM-SSI is that it allows us to study a wider range of agent desires than previous benchmarks. By default, agents have a *collaborative* attitude, i.e., they want to share their knowledge with other agents. ToM-SSI also offers to design agents with an *obstructive* attitude, who aim to prevent other agents from learning new information (see Figure 2d). Moreover, given that ToM-SSI supports more than two agents, we can study *mixed collaborative-obstructive* scenarios in triadic interactions in which one agent  $A_j$  is collaborative with  $A_k$  but obstructive towards  $A_l$  (see

Figure 2e).

**Events** Events dictate the flow of information and change the state of the grid, creating dynamic opportunities for collaboration and inference. Events can involve agents’ movement, communication of information, or both. Agents can move up, down, left, right, and/or communicate one piece of information they possess. For example, the event in Figure 1 is “Cook  communicates ”.

**Social Context** Tasks in ToM-SSI are situated in social contexts. For example, the social context in Figure 1 could be “a restaurant’s kitchen in which four chefs are preparing a dish” or “a cooking class where four participants are learning to make a new dish”. The four pieces of information are assigned to different IDs according to the context, for example, the ingredients needed to make the dish: “bread” () , “salad” () , “tomato” () , and “meat” () . We generated a collection of 121 social contexts using GPT-4o (OpenAI, 2024a) and randomly sampled from this collection while generating ToM-SSI samples (details in A.3).

## 2.2 Question Types

Inspired by the Belief-Desire-Intention framework (Bratman, 1987; Baker et al., 2011), samples in ToM-SSI are paired with three questions, covering agents’ *percept*, *beliefs*, and *intentions*, given their *desires*, as shown in Figure 1. We do not include questions about *desires* as they trigger all social interactions and are already specified in the textual prompt. Given the percept-belief-intention causal structure shown in Figure 1, percept questions act as a control for belief questions (Percept → Belief), which in turn serve as a control for intention questions (Belief → Intention). Therefore, a model with strong ToM abilities must answer all three questions correctly.

**Percept** Percepts are the observations an agent makes about the environment, forming its understanding of the current world state. In ToM-SSI, percepts include the agents’ positions, starting knowledge, and movement. Percept questions probe whether models can accurately attribute agents to percepts based on the information presented in the image and text. This ability is called *perspective taking* – a foundational ability in ToM (Masangkay et al., 1974). For example, in Figure 1, the percept question asks whether  learns what  has communicated. To answer correctly, one

must observe that the two agents occupy adjacent cells so  can hear what  communicates. Percept questions are framed as yes/no questions in the form “Is the statement [...] true/false?”.

**Belief** Beliefs are an agent’s internal representation of the world, derived from its percepts and prior knowledge (Perner, 1993). Beliefs may include assumptions about hidden aspects of the world, such as what another agent knows. Belief questions evaluate whether models understand what information an agent knows or is communicating based on the agent’s percepts. For example, in Figure 1, models must infer that because  can hear what  communicates,  will believe she is missing only one piece of information () . Belief questions are multiple-choice questions with the information IDs as options.

**Intention** Intentions are the specific plans or actions that, given their beliefs and desires, an agent commits to achieve their goals (Tomasello et al., 2005). Intention questions focus on whether models can deduce the agents’ action – either communicative or motor. For example, in Figure 1, models must infer that, given her desires (learning and sharing information) and her belief (missing ) ,  is more likely to approach . This is because  can communicate the last piece of information that  is missing () , and  can communicate the two pieces of information that  is missing ( and ). In comparison, approaching  would be suboptimal, as  could only share one piece of information that  lacks (). Intention questions are multiple-choice questions, with options being agent IDs in case of movement and information IDs in case of communication.

## 2.3 Tasks

ToM-SSI comprises 6,000 questions, equally split between five tasks. These tasks reflect different aspects of everyday social interactions involving agents with collaborative, obstructive, or mixed attitudes. They require tracking multiple agents’ beliefs, interpreting various communication events, and making inferences under uncertainty.

ToM tasks must satisfy two important criteria (Quesque and Rossetti, 2020). First, they must require models to differentiate between mental states (in our case, knowledge) of different agents (*non-merging criterion*). Second, it should be impossible to pass the tasks using low-level heuristics (*mentalising criterion*). Tasks in ToM-SSI ful-

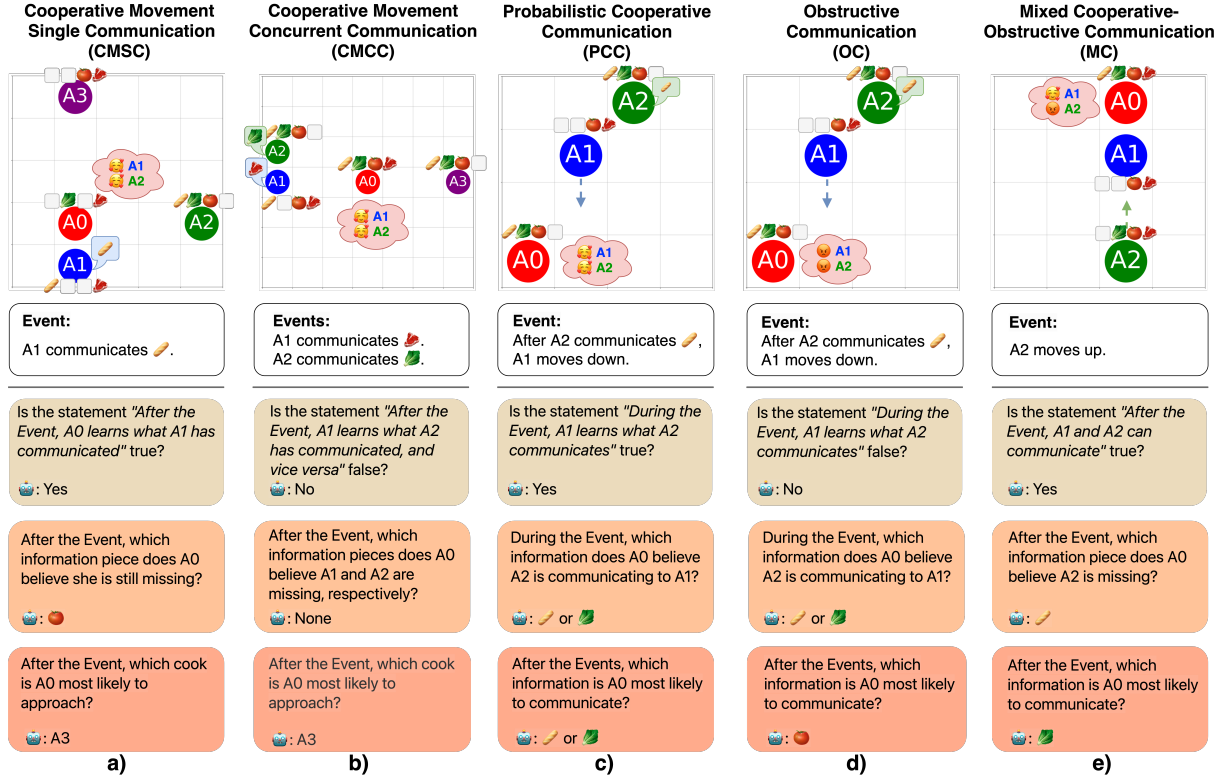


Figure 2: **Overview of ToM-SSI.** ToM-SSI extends to triadic and tetradic social interactions, covering different agent attitudes: cooperative, obstructive, and mixed settings. The dataset comprises five tasks involving cooperative movement, cooperative communication, obstructive communication, and mixed cooperative-obstructive communication. Each sample is paired with three questions, covering agents’ **percepts**, **beliefs**, and **intentions**.

fil these criteria by requiring models to integrate spatial information from the grid with events and agents’ knowledge. Given that agents have partial observability of communicative events, tasks in ToM-SSI also require models to take agents’ perspectives when answering questions.

We discuss how to solve each task based on formal utility functions in 2.4. For simplicity, in the following task descriptions, all questions only target  $A_0$  and we use the cooking example of Figure 1 to label the information IDs. Figure 2 illustrates examples from the five tasks, where we show original images from the dataset with overlapping knowledge and attitude.

### 2.3.1 Cooperative Movement – Single Communication (CMSC)

The CMSC task requires models to reason about one agent’s mental state as that agent is involved in a communicative action. The task includes four agents,  $A_0$ ,  $A_1$ ,  $A_2$ , and  $A_3$ , as shown in Figure 2a. During the event, agent  $A_1$  communicates a piece of information (🍌) that agent  $A_0$  is missing and able to learn.

**Desire**  $A_0$  wants to learn new information from other agents and share information that other agents are missing.

**Percept** The percept question asks if, after the event,  $A_0$  learns what  $A_1$  communicated. This serves as a control to verify whether a model can observe that  $A_0$  and  $A_1$  occupy adjacent cells on the grid. Recognising adjacency is crucial for inferring  $A_0$ ’s belief and intention.

**Belief** The belief question asks to identify which information  $A_0$  believes she is still lacking. Initially,  $A_0$  knows 🍌 and 🍌. Upon learning 🍌 from  $A_1$ , the model must infer that the only remaining missing information for  $A_0$  is 🍌.

**Intention** The intention question asks who is  $A_0$  most likely to approach next. After the event,  $A_0$  knows 🍌, 🍌, and 🍌. A model must correctly infer that  $A_0$  will move toward  $A_3$ , as  $A_3$  can provide the last missing piece (🍌), and  $A_0$  can share the two pieces of information that  $A_3$  is missing. A model that incorrectly infers that  $A_0$  does not learn 🍌 will likely predict that  $A_0$  would go to  $A_2$  instead (since  $A_2$  knows both 🍌 and 🍌, and  $A_0$  can still

communicate 🗨️).

### 2.3.2 Cooperative Movement – Concurrent Communication (CMCC)

The CMCC task differs from the CMSC task in that it challenges models to reason about one agent’s mental states while this agent observes two other agents communicating with each other at the same time (Figure 2b). However, the observing agent cannot know for sure what is being communicated because she is too far away to hear them. During the event, agents  $A_1$  and  $A_2$ , who occupy adjacent cells, share the information that each is missing.

**Desire**  $A_0$  wants to learn new information from other agents and share information that other agents are missing.

**Percept** The percept question asks if, after the event,  $A_1$  learns what  $A_2$  has communicated, and vice versa. This serves as a control to verify whether a model can observe that  $A_1$  and  $A_2$  occupy adjacent cells on the grid, and therefore can learn from each other.

**Belief** The belief question asks to identify which information  $A_0$  believes  $A_1$  and  $A_2$  are still lacking. A model needs to infer that  $A_0$  will believe that since  $A_1$  and  $A_2$  possess what the other is missing and can communicate, after the event they will possess all the information.

**Intention** The intention question asks who is  $A_0$  most likely to approach next. A model that correctly infers  $A_0$ ’s belief that after the event  $A_1$  and  $A_2$  possess all the information will predict that the only sensible agent that  $A_0$  can approach is  $A_3$ , to share the only piece of information  $A_3$  is missing (🍌).

### 2.3.3 Probabilistic Cooperative Communication (PCC)

Tasks in PCC are set in a *probabilistic* scenario, as the target agent’s intention is subject to uncertainty. The task includes three agents  $A_0$ ,  $A_1$ , and  $A_2$ , as shown in Figure 2c. In the event, agent  $A_2$  communicates one of the two pieces of information  $A_1$  is missing (🍌 or 🍌, randomly chosen). Then,  $A_1$  moves adjacent to  $A_0$ .

**Desire**  $A_0$  wants to learn new information from other agents and share information that other agents are missing.

**Percept** The percept question asks if, during the event,  $A_1$  learns what  $A_2$  has communicated. This verifies whether a model can observe that  $A_1$  and  $A_2$  occupy adjacent cells, and therefore  $A_1$  can learn what  $A_2$  is communicating.

**Belief** The belief question asks which information  $A_0$  believes  $A_2$  has communicated to  $A_1$  during the event. Even if the event specifies which information  $A_2$  is communicating, a model must infer that  $A_0$  will not know which specific information was communicated by  $A_2$ . However,  $A_0$  will rationally believe it to be one of the pieces that  $A_1$  is missing (either 🍌 or 🍌).

**Intention** The intention question asks which information  $A_0$  will most likely communicate after the event. A model that correctly infers  $A_0$ ’s belief that  $A_2$  is likely communicating one of the two pieces of information  $A_1$  is missing will predict that  $A_0$  will likely communicate one of these two, aware of the uncertainty of potentially communicating the same information that  $A_1$  has already learnt from  $A_2$ .

### 2.3.4 Obstructive Communication (OC)

The OC task is perceptually identical to PCC (§2.3.3), but in this case, the target agent has an *obstructive* attitude (Figure 2d). Percept and belief questions are analogous to §2.3.3.

**Desire**  $A_0$  does not want other agents to learn new information.

**Intention** The intention question asks which information  $A_0$  will most likely communicate after the event. Given  $A_0$ ’s obstructive attitude, a model must predict that  $A_0$  will likely communicate the only piece of information that all the agents already know (🍌).

### 2.3.5 Mixed Cooperative-Obstructive Communication (MC)

The MC task extends to *mixed cooperative-obstructive* settings, as the target agent is collaborative towards one agent but obstructive towards another. This task includes three agents  $A_0$ ,  $A_1$ , and  $A_2$ , as shown in Figure 2e. In the event, agent  $A_2$  moves one cell up.

**Desire**  $A_0$  wants  $A_1$  to gain new information while preventing  $A_2$  from doing so.

**Percept** The percept question asks if, after the event,  $A_1$  and  $A_2$  can communicate. This serves as a control to verify whether a model can combine information from the grid ( $A_1$  and  $A_2$  are one cell apart) with information from the event ( $A_2$  moving up). After the event, the two agents will occupy adjacent cells and, therefore, will be able to communicate.

**Belief** The belief question asks which information  $A_0$  believes  $A_2$  is missing. A model must infer that, based on  $A_2$ 's initial knowledge and in the absence of further communications with other agents,  $A_0$  will believe that  $A_2$  is missing 🍌.

**Intention** The intention question asks which information  $A_0$  will most likely communicate after the event. Here, a model needs to infer that  $A_0$  believes that if she communicates 🍌,  $A_1$  will learn it and then potentially pass it on to  $A_2$ . However, this conflicts with  $A_0$ 's obstructive attitude towards  $A_2$ . Instead,  $A_0$  is more likely to communicate 🍌, which  $A_1$  is missing and  $A_2$  already possesses.

## 2.4 Utility Functions

Each of our tasks has one (or in some cases two, see Figure 2) correct answer(s) that can be formally found as discussed in §2.1. It is also possible to infer the correct answer by defining the target agent's utility function for each task and maximising it. Following the notation introduced in §2.1, we can define the utility function for movement actions of agent  $A_j$  as:

$$U_{A_j}(\hat{I}_{A_k}) = U_{A_j}^{learn}(\hat{I}_{A_k}) + U_{A_j}^{share}(\hat{I}_{A_k}) \\ = |\hat{I}_{A_k} - I_{A_j}| + |I_{A_j} - \hat{I}_{A_k}|$$

Where  $\hat{I}_{A_k}$  denotes the information that  $A_j$  believes is known by  $A_k$ ,  $|\hat{I}_{A_k} - I_{A_j}|$  is the number of pieces of information that agent  $A_j$  can learn from agent  $A_k$ , and  $|I_{A_j} - \hat{I}_{A_k}|$  is the number of pieces of information that agent  $A_j$  can share with agent  $A_k$ . For example, in Figure 1:  $U_{\text{red}}(\text{red}) = 3$  and  $U_{\text{red}}(\text{green}) = 2$ . For communicative actions, we have three different cases:

- Cooperative:  $U_{A_j}(\hat{I}_{A_k}) = I_{A_j} - \hat{I}_{A_k}$
- Obstructive:  $U_{A_j}(\hat{I}_{A_k}) = I_{A_j} \cap \hat{I}_{A_k}$
- Mixed, e.g. cooperative towards  $A_k$  and obstructive towards  $A_l$ :  
 $U_{A_j}(\hat{I}_{A_k}, \hat{I}_{A_l}) = (I_{A_j} - \hat{I}_{A_k}) \cap (I_{A_j} \cap \hat{I}_{A_l})$

Note that, in ToM-SSI, the spatial relationships are important to understand who is able to learn what, but we designed the tasks such that the number of steps required to reach another agent does not matter. Future versions of the benchmark could include rational movement as an additional complexity.

## 2.5 Dataset Generation

ToM-SSI is entirely generated by code (see Algorithm 1). Each minimal template is paired with agents' initial knowledge and the correct answer, which is determined as discussed in §2.3 and 2.4. Starting from a minimal grid template, our generation pipeline applies random geometric transformations to the grid (see A.2.2) and samples one social context from our database to populate the prompt template corresponding to the task. To further avoid bias, agent and information IDs are randomly permuted. We show the structure of prompts in ToM-SSI in Figure 7. The prompt first introduces the social context and information about agents. It then presents the grid with the agents in their initial position as an image (for VLMs) or text (for LMs), as shown in Figure 6. Following the grid, the prompt lists the information initially known by the agents, the attitude of the target agent, and the event(s) that trigger a change in the environment. The prompt ends with the question and multiple-choice answers for the model to select from. To avoid bias, questions can have different formats (see A.4). Complete examples of prompts are included in A.4.

## 3 Experiments

### 3.1 Experimental Setup

**Baseline Models** We tested 15 baseline models: GPT (4o and 4o-mini; OpenAI, 2024a), o4-mini (OpenAI, 2024b), Claude 3.5 (Sonnet and Haiku; Anthropic, 2024), Gemini 1.5 (Pro and Flash; Anil et al., 2023), Gemini 2.5 Flash (Cohan et al., 2025), Llama 3.2 Instruct (1B, 3B, 11B, and 90B; Dubey et al., 2024), Qwen2-VL Instruct (7B, 72B; Wang et al., 2024), Molmo 7B (Deitke et al., 2025), Mistral 7B Instruct (Jiang et al., 2023), and Gemma 2 9B Instruct (Mesnard et al., 2024). Language models (LM) were evaluated using the text-only version of the prompts, and vision language models (VLMs) were evaluated both with images and text-only prompts. We provide additional details in A.5.

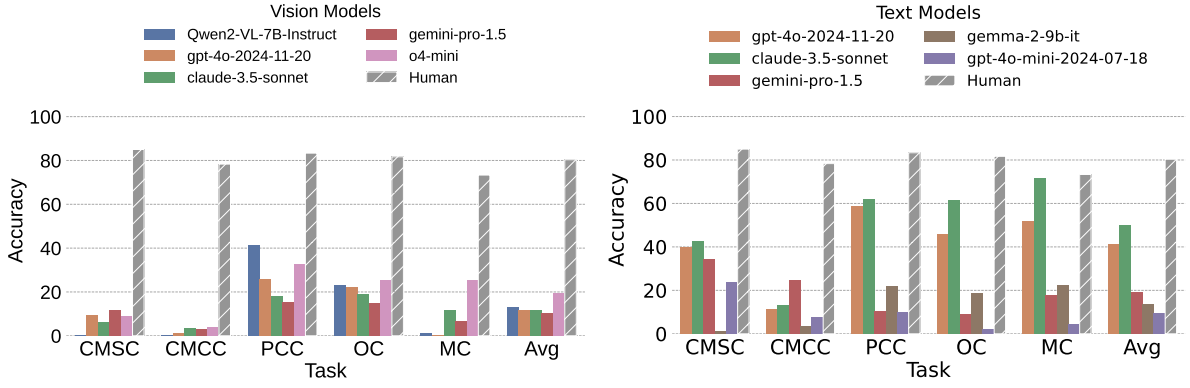


Figure 3: PBI performance for the top five models for the multimodal (left) and text-only (right) versions of ToM-SSI. Human scores are included in both plots for comparison; however, humans were evaluated exclusively in the multimodal setting.

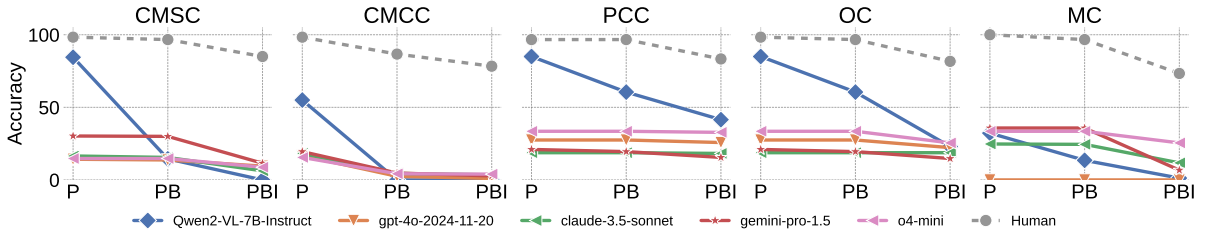


Figure 4: Change in performance across P, PB, and PBI accuracy for both humans and the top five VLMs on ToM-SSI. Performance declines from percepts to beliefs to intentions.

**Human Study** We recruited 20 human participants and asked them to answer 45 questions, equally split between tasks. Participants were shown the same prompt as the models. Further details are provided in A.6.

**Metrics** For each task, we measured models’ accuracy on percept (P), belief (B), and intention (I) questions. We then computed two scores: the PB score, which requires correctly answering percept and belief questions ( $P \wedge B$ ), and the PBI score, which requires correctly answering all three question types ( $P \wedge B \wedge I$ ).

### 3.2 Results

Figure 3 reports the PBI accuracy of models on the five tasks in ToM-SSI, as well as their average (Avg). We focus on the PBI accuracy because it reflects a model’s ability to handle the full spectrum of reasoning required to comprehensively solve all tasks in ToM-SSI. For clarity, the figure includes only the best five models, selected based on their Avg score on the multimodal (Figure 3, left) and text-only versions of ToM-SSI (Figure 3, right). Detailed results for every combination of task and question type can be found in A.7.

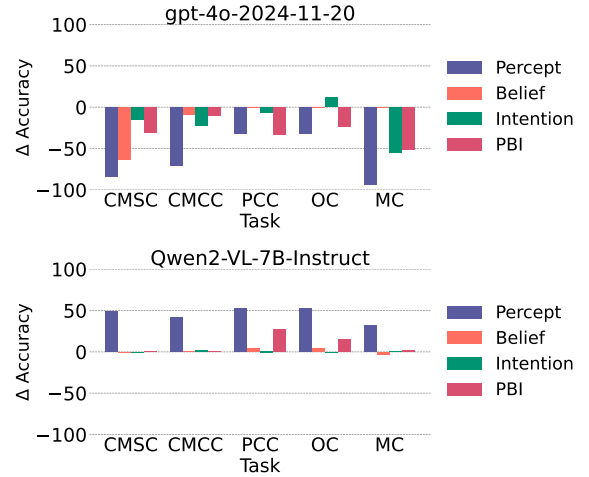


Figure 5: Difference in accuracy of VLMs when evaluated on the multimodal version of the ToM-SSI versus the text-only version.

Figure 3 shows that models’ performance significantly lags behind human performance, especially in the multimodal setting. While human performance on the tasks ranges from 73% to 85%, models generally perform below 30% (for detailed scores, see Table 1). No model performs best across all tasks. In the multimodal setting, o4-mini is the

best model on average (Avg), achieving the highest performance on CMCC, OC, and MC. Qwen2-7B is the best model on PCC. PBI scores are higher in the text-only setting, where Claude 3.5 Sonnet performs the best (Figure 3, right).

**P, PB, and PBI Inference** Accurately representing the complete percept-belief-intention causal graph involves three key steps: inferring the percepts of the target agent, determining the agent’s beliefs based on those percepts, and inferring the agent’s intentions based on their beliefs (Jung et al., 2024). Figure 4 shows that while models generally perform well in percept inference, their accuracy drops significantly when progressing to PB, and further when moving forward to PBI. While humans outperform these models, their accuracy also drops from PB to PBI, albeit less pronounced. o4-mini, GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro are more consistent across all three steps but are limited by lower percept inference accuracy, which also caps their overall performance. These findings highlight percept inference as a key foundation for advancing ToM in multimodal models. Furthermore, our results reveal that even when models achieve reasonable PB inference, they still struggle to transition from PB to PBI. We provide additional results in A.8.

### 3.3 Error Analysis

Figure 3 shows that even state-of-the-art models struggle with ToM-SSI, indicating a notable gap in their reasoning abilities. To gain deeper insights into these shortcomings, we conducted an error analysis by manually inspecting the models’ generated outputs. In particular, we focus on CMCC and MC – which are the most challenging tasks.

**Modelling Multi-Agent Communication is Challenging** Compared to CMSC, CMCC is more challenging for models as it involves a communicative event where the target agent is unsure of what has been communicated. Despite this, it should be relatively straightforward to infer what has been communicated, given the agents’ cooperative intent and their observable initial knowledge ( $A_1$ : 🍌🍅🍎,  $A_2$ : 🍌🍇🍅🍌). By inspecting the output generated by Llama 3.2 11B<sup>2</sup>, we observe that most errors arise from ignoring that agents can observe each other’s initial knowledge – although this is made explicit in the prompt. This oversight

leads the model to incorrectly assume that the target agent will believe that if one agent does not explicitly communicate a piece of information, they do not possess it (Example 4). Next, we examined error cases in the intention questions where both percepts and beliefs were correctly inferred. In these instances, we found that errors typically stemmed from incorrect recall of the target agent’s knowledge (Example 5).

**Successes and Shortcomings in Modelling Mixed Social Interactions** We repeated the previous analysis for the MC task. While examining the output generated by Llama 3.2 11B<sup>3</sup>, we found that it accurately considers the agents’ attitudes in most cases. In successful cases, the model infers that  $A_0$  prefers not to share information that  $A_2$  could learn next (Example 7). In failure cases, the model overlooks the fact that if  $A_0$  communicates a piece of information that both  $A_1$  and  $A_2$  are missing,  $A_1$  is likely to share it with  $A_2$  (Example 6).

### 3.4 Do VLMs Benefit From Images?

We finally compared the performance of VLMs when evaluated on the multimodal version of ToM-SSI versus the text-only version. Our analysis reveals that different models exhibit different patterns to the inclusion of image inputs. For example, GPT-4o does not benefit from the addition of images but performs significantly better on the text-only version of ToM-SSI (Figure 5, top). This effect is most pronounced for percept questions. In contrast, Qwen2-VL 7B shows a clear benefit from the image input, particularly on percept questions (Figure 5, bottom). We provide comparisons for all the other evaluated models in A.9. While Claude 3.5 Sonnet and Gemini Pro 1.5 show similar patterns to GPT-4o, results for other models are mixed, with some benefiting from image inputs on specific tasks and others showing little to no advantage.

## 4 Related Work

Theory of Mind in AI has been studied for more than a decade (Baker et al., 2011; Eysenbach et al., 2016; Rabinowitz et al., 2018; Jara-Ettinger, 2019; Liu et al., 2023; Bortoletto et al., 2024a,c,d; Ruhdorfer et al., 2025). Recent advances in LFM have sparked interest in evaluating their ToM capabilities (Achiam et al., 2023; Ullman, 2023). While several

<sup>2</sup>The best-performing VLM for percept questions on CMCC, see Table 1.

<sup>3</sup>The best-performing VLM for percept questions on MC, see Table 1.

benchmarks are based on textual variations of the classic Sally-Anne task (Le et al., 2019; Sclar et al., 2023; Ma et al., 2023a; Wu et al., 2023; Gandhi et al., 2023; Xu et al., 2024; Zhou et al., 2023), others aim to cover a broader range of scenarios (Kim et al., 2023; Chen et al., 2024; Hou et al., 2024a; Gu et al., 2024), including multimodal settings (Bara et al., 2021, 2023; Jin et al., 2024; Shi et al., 2025). Additionally, efforts have been made to enhance LMs’ ToM through prompting techniques (Zhou et al., 2023; Moghaddam and Honey, 2023; Wilf et al., 2024) or activation steering (Zhu et al., 2024; Bortoletto et al., 2024b).

## 5 Discussion and Conclusion

We introduced ToM-SSI, a multimodal benchmark that tests ToM capabilities in environments rich with social interactions and spatial dynamics. Featuring up to four agents communicating and moving, it enables the study of cooperative, obstructive, and mixed interactions. Our evaluations on ToM-SSI revealed several important and novel insights. First, current **models perform significantly worse than humans**, both on the multimodal and text-only version of ToM-SSI (§3.2, Figure 3). Second, we show that **models struggle with the critical steps necessary for ToM reasoning** (§3.3, Figure 4). Notably, even when models perform reasonably well on PB inference, they still struggle to transition to PBI. Third, our analyses of error cases revealed that **models are still limited in modelling agent perception, multi-agent communication, and mixed social interactions** (§3.3). This means they may misinterpret or oversimplify human behaviour in real-world settings, especially in group interactions. Finally, we found that **most of the evaluated VLMs do not benefit from visual input** – highlighting a critical disparity in how models leverage multimodal information to perform ToM tasks (§3.4, Figure 5). This suggests a gap in how models understand and integrate context, which is vital for interpreting visual cues during social interactions.

### Limitations

One limitation of ToM-SSI lies in its synthetic grid world environment, which is simpler than the real world. However, this simplicity does not impair the core abilities that ToM-SSI targets – reasoning about agents’ mental states in spatially grounded interactions. There is a clear advantage that makes

synthetic benchmarks well suited for studying ToM, at least at the current state of research: real-world tasks often require common sense reasoning skills that models do not fully have yet, and that will function as a confounder making models’ performance on ToM inference hard to judge (see Gandhi et al. (2023)’s discussion about (Shapira et al., 2024)). Synthetic benchmarks like ToM-SSI allow us to reduce these factors and to design tasks that better focus on evaluating core ToM abilities. Our setup also makes the environment less prone to hallucinations, which can hinder the performance of current vision-language models in complex simulations (Jin et al., 2024), while also making hallucinations easier to identify if they occur. In addition, using images instead of videos avoids sampling issues or additional processing steps that might exclude important video frames.

A second limitation is that, while our work covers a broader set of interactions compared to previous work, it still does not cover all possible social scenarios. Future work could consider, for example, *exploitative interactions*, where an agent uses the other’s resources or efforts for their own gain. Moreover, while ToM-SSI presents scenarios about a single group of interacting agents, future work could extend it to multiple *social groups* with agents in the same social group sharing common goals. This can be achieved by extending ToM-SSI’s generation pipeline with suitable templates.

Finally, while studying inference-time methods to improve performance goes beyond the scope of our current work, exploring the effects of CoT (Wei et al., 2022) or other methods (e.g. SimToM (Wilf et al., 2024), TimeToM (Hou et al., 2024b), or PercepToM (Jung et al., 2024)) is an interesting research direction for future work.

### Acknowledgements

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting C. Ruhdorfer.

### References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, An-

- drew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.
- Anthropic. 2024. Introducing Computer Use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. <https://www.anthropic.com/news/3-5-models-and-computer-use>.
- 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.
- Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. *MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Cristian-Paul Bara, Ziqiao Ma, Yingzhuo Yu, Julie Shah, and Joyce Chai. 2023. Towards Collaborative Plan Acquisition Through Theory of Mind Modeling in Situated Dialogue. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2958–2966.
- Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdesaied, Lei Shi, and Andreas Bulling. 2024a. Limits of Theory of Mind Modelling in Dialogue-Based Collaborative Plan Acquisition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024b. *Benchmarking Mental State Representations in Language Models*. In *ICML 2024 Workshop on Mechanistic Interpretability*.
- Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024c. Explicit Modelling of Theory of Mind for Belief Prediction in Nonverbal Social Interactions. In *ECAI*.
- Matteo Bortoletto, Lei Shi, and Andreas Bulling. 2024d. Neural Reasoning About Agents’ Goals, Preferences, and Actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 456–464.
- Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press.
- Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheyang Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. *NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.
- Kang Chen and Xiangqian Wu. 2024. VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27218–27227.
- 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 (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104.
- 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. *arXiv preprint arXiv:2407.21783*.
- Benjamin Eysenbach, Carl Vondrick, and Antonio Torralba. 2016. Who is Mistaken? *arXiv preprint arXiv:1612.01175*.
- Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding Social Reasoning in Language Models with Language Models. *Advances in Neural Information Processing Systems*, 37.
- Kanishk Gandhi, Gala Stojnic, Brenden M Lake, and Moira R Dillon. 2021. Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, 34:9963–9976.
- Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs. *arXiv preprint arXiv:2410.13648*.
- Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2024a. Entering Real Social World! Benchmarking the Theory of Mind and Socialization Capabilities of LLMs from a First-person Perspective. *arXiv preprint arXiv:2410.06195*.

- Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024b. [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, Bangkok, Thailand. Association for Computational Linguistics.
- Julian Jara-Ettinger. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Chuangyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua 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 (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. 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*, pages 19794–19809, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023. [Computational Language Acquisition with Theory of Mind](#). In *The Eleventh International Conference on Learning Representations*.
- 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.
- Zenaida S Masangkay, Kathleen A McCluskey, Curtis W McIntyre, Judith Sims-Knight, Brian E Vaughn, and John H Flavell. 1974. The Early Development of Inferences about the Visual Percepts of Others. *Child development*, pages 357–366.
- Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.
- Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. *arXiv preprint arXiv:2304.11490*.
- OpenAI. 2024a. GPT-4o System Card. <https://openai.com/index/gpt-4o-system-card/>.
- OpenAI. 2024b. Introducing o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Josef Perner. 1993. *Understanding the Representational Mind*. The MIT Press.
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.
- François Quesque and Yves Rossetti. 2020. What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science*, 15(2):384–396.
- Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. [Machine Theory of Mind](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.
- Constantin Ruhdorfer, Matteo Bortoletto, and Andreas Bulling. 2025. The Yokai Learning Environment: Tracking Beliefs Over Space and Time. *arXiv preprint arXiv:2508.12480*.
- Pranab Sahoo, Prabhaskar Meheria, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. [A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. 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 (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.
- Melanie Sclar, Graham Neubig, and Yonatan Bisk. 2022. [Symmetric Machine Theory of Mind](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19450–19466. PMLR.
- 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.
- Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. [MuMA-ToM: Multi-modal Multi-Agent Theory of Mind](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1510–1519.
- Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.
- Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint arXiv:2302.08399*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.
- Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. [Think Twice: Perspective-Taking Improves Large Language Models’ Theory-of-Mind Capabilities](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.
- Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023. How FaR Are Large Language Models From Agents with Theory-of-Mind? *arXiv preprint arXiv:2310.03051*.
- Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. [Language Models Represent Beliefs of Self and Others](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62638–62681. PMLR.

## A Appendix

### A.1 Societal Impact

While our work is foundational and remains distant from specific applications with direct societal impact, it's important to recognise the ethical implications of modelling and predicting mental states. Handling sensitive aspects of individuals' inner experiences requires careful consideration to avoid reinforcing biases or misunderstanding psychological nuances.

### A.2 Gridworld

#### A.2.1 Gridworld Representations

Figure 6 represents the two different representation of the grid included in the standard, multimodal version of ToM-SSI and in the text-only version.

#### A.2.2 Gridworld Transformations

ToM-SSI builds grid world environments starting from minial templates where agents are placed in pre-defined cells. Our generation pipeline then uniformly increases the distance between agents by a random value  $\delta \in \{0, 1, 2, 3\}$  and applies one random transformation. We present the formal definitions of transformations applied to our grid world templates, represented as a two-dimensional matrix  $T \in \mathbb{R}^{m \times n}$ , where  $m$  is the number of rows and  $n$  is the number of columns (in our specific case,  $n = m$ ). Transformations include rotations, mirroring, and transposition, which are defined as follows.

**90° Rotation** The 90° clockwise rotation of  $T$  produces a new matrix  $T'$  of size  $n \times m$  such that:

$$T'_{i,j} = T_{m-j+1,i}, \quad \forall i \in [1, n], j \in [1, m]. \quad (1)$$

Operationally, this is equivalent to reversing the row order of  $T$  and transposing:

$$T' = \text{Transpose}(\text{ReverseRows}(T)). \quad (2)$$

**180° Rotation** The 180° rotation produces a matrix  $T'$  such that:

$$T'_{i,j} = T_{m-i+1, n-j+1}, \quad \forall i \in [1, m], j \in [1, n]. \quad (3)$$

This operation reverses both the rows and columns:

$$T' = \text{ReverseRows}(\text{ReverseColumns}(T)). \quad (4)$$

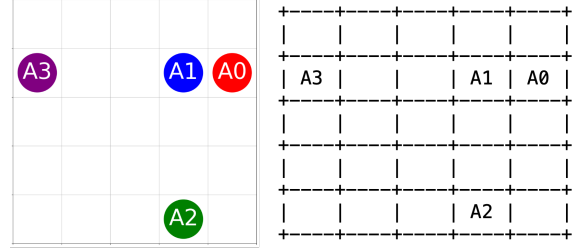


Figure 6: By using grid world environments as a ground for social interactions, ToM-SSI allows us to evaluate large language and vision-and-language models by providing equivalent grid representations.

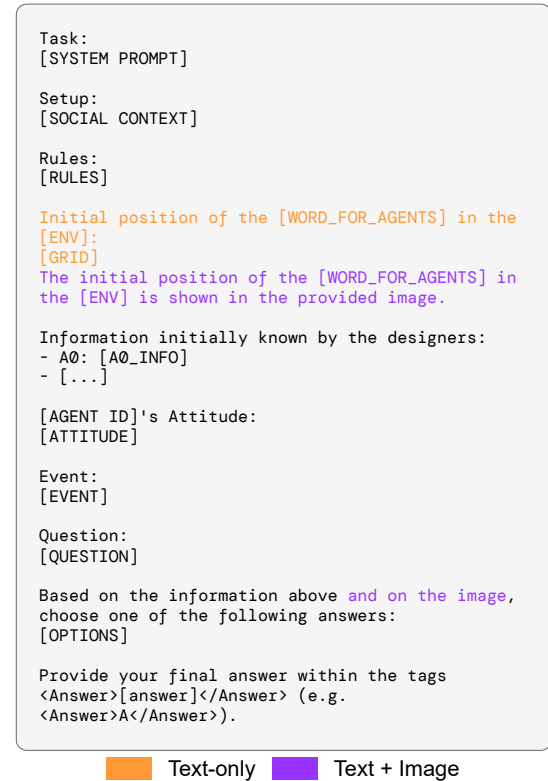


Figure 7: Structure of prompts in ToM-SSI.

**270° Rotation** The 270° clockwise rotation produces a matrix  $T'$  of size  $n \times m$  such that:

$$T'_{i,j} = T_{j, n-i+1}, \quad \forall i \in [1, n], j \in [1, m]. \quad (5)$$

Operationally, this is equivalent to transposing  $T$  and then reversing the rows:

$$T' = \text{ReverseRows}(\text{Transpose}(T)). \quad (6)$$

**Horizontal Mirroring** Horizontal mirroring reflects  $T$  across its horizontal axis, producing a matrix  $T'$  such that:

$$T'_{i,j} = T_{m-i+1, j}, \quad \forall i \in [1, m], j \in [1, n]. \quad (7)$$

| Model                         | Vision | CMSC |       |      |      | CMCC |      |      |      | PCC  |       |      |      | OC   |       |      |      | MC    |       |      |      |
|-------------------------------|--------|------|-------|------|------|------|------|------|------|------|-------|------|------|------|-------|------|------|-------|-------|------|------|
|                               |        | P    | B     | I    | PBI  | P    | B    | I    | PBI  | P    | B     | I    | PBI  | P    | B     | I    | PBI  | P     | B     | I    | PBI  |
| Human                         | ✓      | 98.3 | 98.3  | 85.0 | 85.0 | 98.3 | 88.3 | 85.0 | 78.3 | 96.7 | 98.3  | 86.7 | 83.3 | 98.3 | 98.3  | 81.7 | 81.7 | 100.0 | 96.7  | 73.3 | 73.3 |
| Llama-3.2-11B-Vision-Instruct |        | 82.5 | 22.5  | 13.5 | 2.2  | 64.2 | 0.0  | 17.2 | 0.0  | 55.5 | 60.5  | 59.5 | 20.2 | 55.5 | 60.5  | 33.5 | 10.2 | 65.8  | 35.5  | 22.0 | 5.2  |
| Llama-3.2-11B-Vision-Instruct | ✓      | 87.8 | 25.2  | 5.2  | 1.0  | 65.2 | 1.0  | 10.8 | 0.0  | 73.8 | 56.2  | 67.8 | 27.3 | 73.8 | 56.2  | 21.5 | 8.0  | 61.5  | 40.2  | 20.2 | 5.0  |
| Llama-3.2-1B-Instruct         |        | 50.2 | 23.8  | 30.2 | 3.0  | 48.2 | 24.5 | 28.5 | 4.5  | 51.0 | 36.2  | 51.0 | 11.2 | 51.0 | 36.2  | 27.0 | 6.2  | 53.8  | 22.2  | 23.2 | 3.0  |
| Llama-3.2-3B-Instruct         |        | 63.7 | 18.2  | 20.8 | 2.5  | 65.2 | 0.0  | 18.0 | 0.0  | 63.5 | 63.0  | 80.2 | 31.8 | 63.5 | 63.0  | 20.0 | 8.8  | 46.2  | 30.0  | 22.8 | 3.5  |
| Mistral-7B-Instruct-v0.3      |        | 72.8 | 24.5  | 16.0 | 4.0  | 46.2 | 0.0  | 15.8 | 0.0  | 60.5 | 61.3  | 66.8 | 22.2 | 60.5 | 61.3  | 26.0 | 8.2  | 25.0  | 46.5  | 22.2 | 2.2  |
| Molmo-7B-D-0924               |        | 56.0 | 23.8  | 38.8 | 5.2  | 59.2 | 4.8  | 34.2 | 1.0  | 44.5 | 44.0  | 37.5 | 7.2  | 44.5 | 44.0  | 17.8 | 2.8  | 51.0  | 36.8  | 10.2 | 1.2  |
| Molmo-7B-D-0924               | ✓      | 43.5 | 29.8  | 33.0 | 5.0  | 51.0 | 4.0  | 28.0 | 0.2  | 41.5 | 44.0  | 41.0 | 6.8  | 41.5 | 44.0  | 16.0 | 3.8  | 51.0  | 39.0  | 14.8 | 2.2  |
| Qwen2-VL-7B-Instruct          |        | 35.8 | 17.0  | 1.8  | 0.0  | 12.8 | 0.0  | 4.2  | 0.0  | 32.2 | 66.8  | 67.2 | 14.5 | 32.2 | 66.8  | 39.2 | 8.2  | 0.0   | 42.2  | 10.2 | 0.0  |
| Qwen2-VL-7B-Instruct          | ✓      | 84.5 | 16.0  | 0.8  | 0.0  | 55.0 | 0.0  | 5.5  | 0.0  | 85.0 | 71.0  | 65.8 | 41.5 | 85.0 | 71.0  | 38.8 | 23.0 | 32.2  | 39.0  | 10.2 | 1.2  |
| claude-3.5-haiku-20241022     |        | 68.8 | 84.2  | 26.0 | 16.5 | 69.2 | 3.2  | 76.2 | 2.5  | 11.2 | 92.2  | 76.5 | 7.8  | 11.2 | 92.2  | 23.5 | 2.0  | 72.8  | 100.0 | 4.2  | 3.5  |
| claude-3.5-haiku-20241022     | ✓      | 5.2  | 75.8  | 23.2 | 0.5  | 12.0 | 3.0  | 70.0 | 0.0  | 4.8  | 91.8  | 81.2 | 4.2  | 4.8  | 91.8  | 17.5 | 0.2  | 46.5  | 100.0 | 3.0  | 1.5  |
| claude-3.5-sonnet             |        | 85.2 | 100.0 | 49.0 | 42.5 | 75.2 | 21.2 | 75.5 | 13.2 | 64.5 | 99.8  | 95.5 | 61.8 | 64.5 | 99.8  | 95.2 | 61.3 | 84.5  | 100.0 | 85.2 | 71.8 |
| claude-3.5-sonnet             | ✓      | 16.5 | 91.0  | 20.5 | 6.2  | 18.2 | 13.0 | 75.2 | 3.5  | 18.8 | 99.8  | 97.8 | 18.2 | 18.8 | 99.8  | 97.8 | 18.8 | 24.8  | 98.8  | 49.8 | 11.8 |
| gemin-2.5-flash               |        | 8.2  | 7.2   | 0.0  | 0.0  | 10.0 | 0.2  | 2.0  | 0.0  | 14.8 | 4.0   | 0.2  | 0.0  | 14.8 | 4.0   | 0.2  | 0.0  | 13.0  | 2.0   | 0.2  | 0.0  |
| gemin-2.5-flash               | ✓      | 6.0  | 10.0  | 1.8  | 0.2  | 6.8  | 2.8  | 6.2  | 0.0  | 20.8 | 24.0  | 14.5 | 1.2  | 20.8 | 24.0  | 4.2  | 0.2  | 12.2  | 12.8  | 6.5  | 0.0  |
| gemin-flash-1.5               |        | 45.0 | 100.0 | 16.5 | 7.5  | 15.2 | 1.8  | 57.5 | 0.0  | 15.5 | 78.8  | 35.8 | 4.5  | 15.5 | 78.8  | 29.5 | 4.2  | 41.5  | 99.8  | 6.2  | 2.8  |
| gemin-flash-1.5               | ✓      | 21.0 | 93.2  | 26.0 | 5.5  | 12.2 | 1.8  | 46.8 | 0.0  | 30.0 | 84.2  | 43.2 | 8.8  | 30.0 | 84.2  | 39.5 | 9.8  | 29.2  | 100.0 | 10.0 | 3.5  |
| gemin-pro-1.5                 |        | 71.8 | 100.0 | 48.5 | 34.5 | 44.2 | 75.8 | 70.0 | 24.8 | 12.8 | 92.5  | 86.5 | 10.2 | 12.8 | 92.5  | 76.2 | 9.0  | 54.5  | 100.0 | 30.8 | 17.8 |
| gemin-pro-1.5                 | ✓      | 30.2 | 99.2  | 29.2 | 11.5 | 19.5 | 21.2 | 55.5 | 2.8  | 21.0 | 90.2  | 88.5 | 15.5 | 21.0 | 90.2  | 78.5 | 14.8 | 35.8  | 100.0 | 18.0 | 6.8  |
| gemma-2-9b-it                 |        | 99.0 | 81.2  | 1.5  | 1.2  | 38.8 | 32.5 | 24.5 | 3.2  | 69.2 | 74.8  | 42.0 | 22.0 | 69.2 | 74.8  | 39.0 | 18.5 | 99.2  | 96.5  | 24.2 | 22.5 |
| gpt-4o-2024-11-20             |        | 98.0 | 98.8  | 41.2 | 39.8 | 88.0 | 19.8 | 54.8 | 11.2 | 60.0 | 99.8  | 97.2 | 58.5 | 60.0 | 99.8  | 74.2 | 45.8 | 93.2  | 99.8  | 55.5 | 51.7 |
| gpt-4o-2024-11-20             | ✓      | 14.2 | 35.5  | 26.2 | 9.5  | 16.8 | 10.5 | 32.0 | 1.0  | 27.5 | 99.2  | 90.2 | 25.8 | 27.5 | 99.2  | 86.2 | 22.2 | 0.0   | 100.0 | 0.0  | 0.0  |
| gpt-4o-mini-2024-07-18        |        | 65.5 | 100.0 | 35.8 | 23.8 | 58.8 | 25.8 | 47.5 | 7.8  | 16.0 | 86.2  | 57.2 | 9.8  | 16.0 | 86.2  | 10.0 | 2.2  | 60.0  | 99.8  | 6.8  | 4.2  |
| gpt-4o-mini-2024-07-18        | ✓      | 23.0 | 98.8  | 34.0 | 7.8  | 33.8 | 9.0  | 59.8 | 1.8  | 16.8 | 86.5  | 64.0 | 10.8 | 16.8 | 86.5  | 11.2 | 2.0  | 25.2  | 100.0 | 6.8  | 2.0  |
| llama-3.2-90b-vision-instruct |        | 30.2 | 75.8  | 26.2 | 6.2  | 6.5  | 0.0  | 43.2 | 0.0  | 8.2  | 91.2  | 66.0 | 4.2  | 8.2  | 91.2  | 16.5 | 2.2  | 52.2  | 100.0 | 3.2  | 1.0  |
| llama-3.2-90b-vision-instruct | ✓      | 9.2  | 55.8  | 23.0 | 0.8  | 1.8  | 4.8  | 78.8 | 0.0  | 14.0 | 72.5  | 82.8 | 8.5  | 14.0 | 72.5  | 42.5 | 3.5  | 25.8  | 66.8  | 17.8 | 3.2  |
| o4-mini-2025-04-16            | ✓      | 14.8 | 69.0  | 25.2 | 8.8  | 15.5 | 17.0 | 34.2 | 4.0  | 33.5 | 100.0 | 95.5 | 32.8 | 33.5 | 100.0 | 86.8 | 25.5 | 33.5  | 100.0 | 74.2 | 25.5 |
| qwen-2-v1-72b-instruct        |        | 59.2 | 91.0  | 3.2  | 2.5  | 7.8  | 0.2  | 21.2 | 0.0  | 50.0 | 76.2  | 52.2 | 17.8 | 50.0 | 76.2  | 50.5 | 18.8 | 36.2  | 99.5  | 8.5  | 2.2  |
| qwen-2-v1-72b-instruct        | ✓      | 21.2 | 89.5  | 7.8  | 1.8  | 0.8  | 6.5  | 26.8 | 0.0  | 49.5 | 78.5  | 50.7 | 17.0 | 49.5 | 78.5  | 51.7 | 21.8 | 13.2  | 99.8  | 7.5  | 1.0  |

Table 1: Models’ accuracy across the three question types (P: Perception, B: Belief, I: Intent) for each task in ToM-SSI.

This operation reverses the row order:

$$T' = \text{ReverseRows}(T). \quad (8)$$

**Vertical Mirroring** Vertical mirroring reflects  $T$  across its vertical axis, producing a matrix  $T'$  such that:

$$T'_{i,j} = T_{i,n-j+1}, \quad \forall i \in [1, m], j \in [1, n]. \quad (9)$$

This operation reverses the column order:

$$T' = \text{ReverseColumns}(T). \quad (10)$$

**Transposition** Transposition exchanges the rows and columns of  $T$ , producing a matrix  $T'$  of size  $n \times m$  such that:

$$T'_{i,j} = T_{j,i}, \quad \forall i \in [1, n], j \in [1, m]. \quad (11)$$

Operationally:

$$T' = \text{Transpose}(T). \quad (12)$$

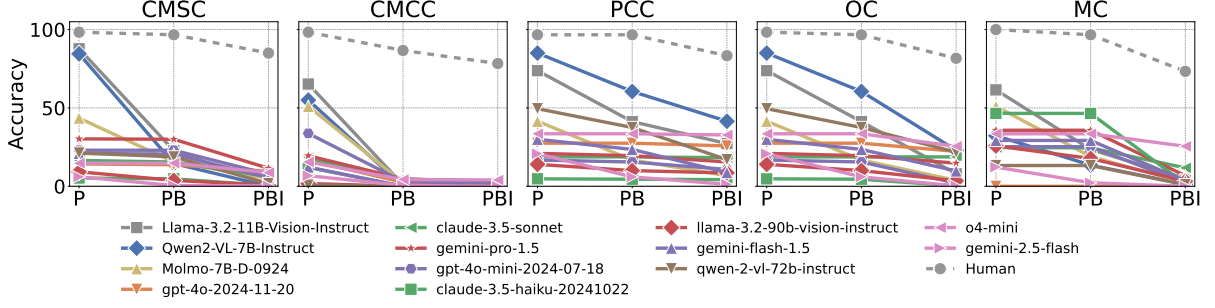
### A.3 Social Context Generation

To generate the social contexts used in our prompts, we employed a few-shot learning approach. We started by manually crafting five example social contexts, such as: a design studio where four graphic designers are working on a project. Social contexts are paired with four information IDs to use as agents’ knowledge, for

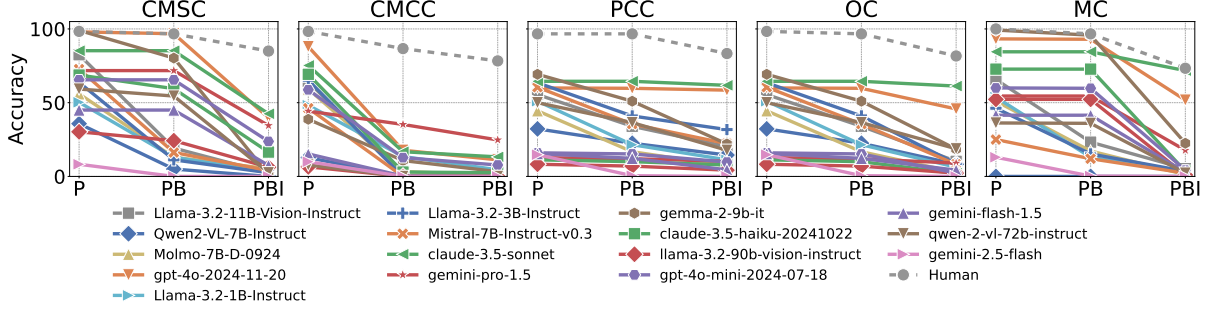
| Task        | Question Type |            |           |
|-------------|---------------|------------|-----------|
|             | Percept       | Belief     | Intent    |
| <b>CMSC</b> | 2132 ± 71     | 2188 ± 81  | 2126 ± 73 |
| <b>CMCC</b> | 2266 ± 86     | 2567 ± 102 | 2241 ± 87 |
| <b>PC</b>   | 2134 ± 71     | 2240 ± 86  | 2188 ± 80 |
| <b>OC</b>   | 2134 ± 71     | 2240 ± 85  | 2134 ± 79 |
| <b>MC</b>   | 2114 ± 68     | 2175 ± 77  | 2172 ± 77 |

Table 2: Average length and standard deviation of prompts across different tasks and question types.

example Feedback Loop, Color Scheme, Font Choice, and Design Concept. These examples were provided as input prompts to GPT-4o (OpenAI, 2024a), which we used iteratively to generate 115 additional social contexts. To ensure quality, we manually checked all generated contexts for consistency, meaningfulness, and to eliminate duplicates. Three people (native speakers or high proficiency) manually checked each social context. Additionally, we ran two rounds of a pilot study before finalising the dataset. To compute the distribution of the generated social contexts, we initially asked two annotators to label them. We then measured the inter-annotator agreement using the Cohen’s kappa, which was on the border between *moderate* and *substantial* agreement ( $\kappa = 0.605$ ). Finally, the annotators proceeded to resolve disagreements. We report the final distribution of the social context topics in Figure 10.



(a) VLMs evaluated on ToM-SSI.



(b) LMs evaluated on the text-only version of ToM-SSI.

Figure 8: Change in performance across P, PB, and PBI accuracy for both humans and models on ToM-SSI.

#### A.4 Dataset Generation

ToM-SSI is entirely generated by code. The starting minimal template for each task is shown in Figure 11. Each minimal template is paired with agents’ initial knowledge and the correct answer, which is determined as discussed in §2.1 and 2.4. We show ToM-SSI’s generation pipeline in detail in Algorithm 1,

The structure of prompts in ToM-SSI is illustrated in Figure 7, where we highlight the differences between the prompt used for VLMs (where the text representation of the grid is substituted by the image) and LLMs. The prompt first introduces the social context and information about agents. It then presents the grid with the agents in their initial position as an image (for VLMs) or text (for LMs), as shown in Figure 6. Following the grid, the prompt lists the information initially known by the agents, the attitude of the target agent, and the event(s) that trigger a change in the environment. The prompt ends with the question and multiple-choice answers for the model to select from.

ToM-SSI allows to define different formats for each question, as shown in Figure 12. During the dataset generation, the format is chosen randomly to avoid bias and make questions more diverse. The correct answer to a question depends on multiple factors: the specific task being evaluated, the social

context, the spatial setup of the environment, the attitude of the target agent, and the format of the question. The format of the question alone is not sufficient to answer the question. Previous work has often used just one single question format (Le et al., 2019; Wu et al., 2023; Gandhi et al., 2023; Chan et al., 2024). In the dataset version included with this submission, questions appear in two different formats.

We report the average number of characters per prompt for each task in Table 2. Examples of prompts for the three questions are reported in Example 1 (percept), Example 2 (belief), and Example 3 (intent).

#### A.5 Baseline Models

We evaluate the following models:

- Llama-3.2-1B-Instruct (unimodal)
- Llama-3.2-3B-Instruct (unimodal)
- Llama-3.2-11B-Vision-Instruct (uni- and multimodal settings)
- Llama-3.2-90B-Vision-Instruct (uni- and multimodal settings)
- Mistral-7B-Instruct-v0.3 (unimodal)
- Molmo-7B-D-0924 (uni- and multimodal)

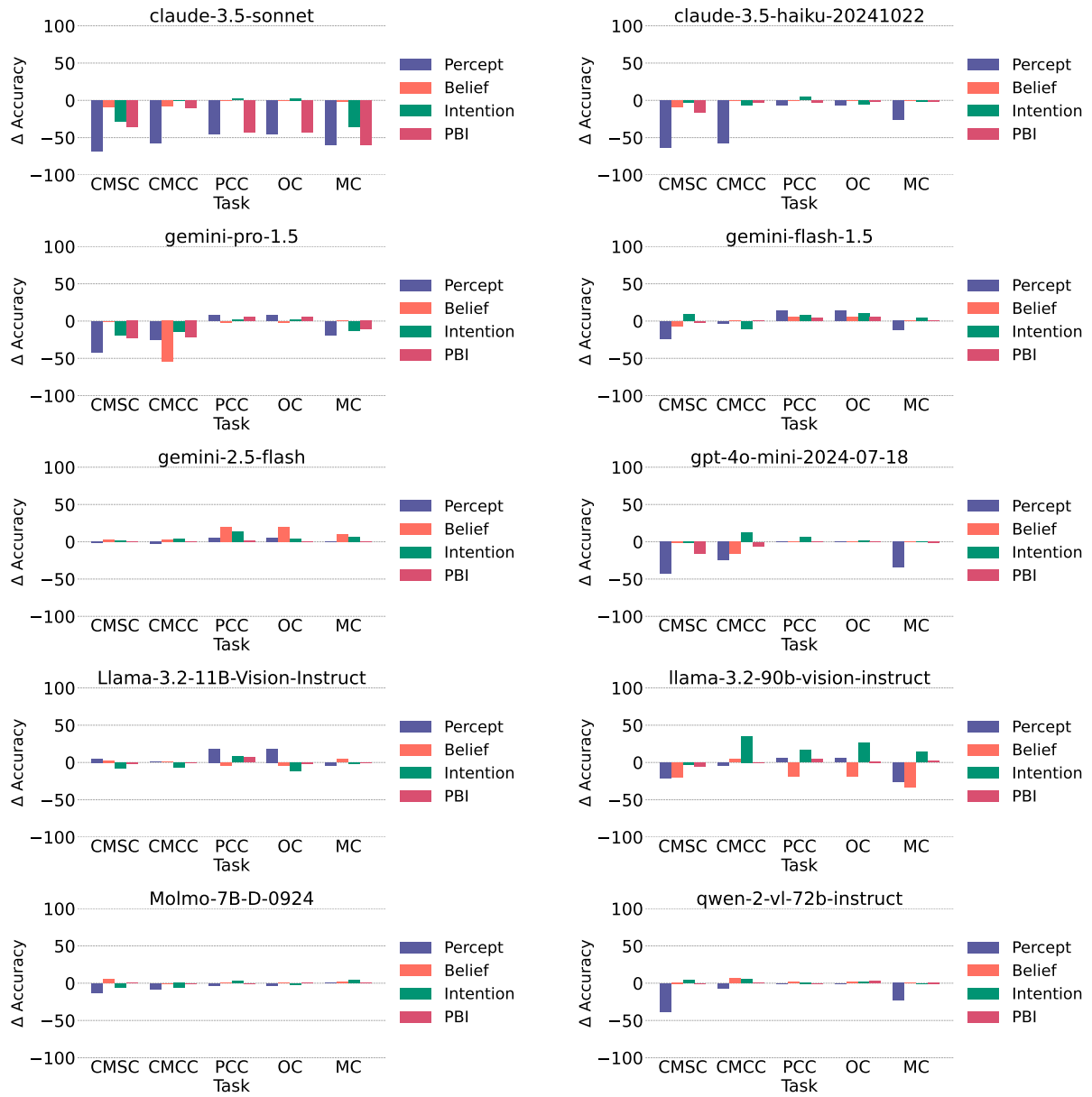


Figure 9: Difference in accuracy of VLMs when evaluated on the multimodal version of the ToM-SSI versus the text-only version.

- Qwen2-VL-7B-Instruct (uni- and multi-modal)
- Qwen2-VL-72B-Instruct (uni- and multi-modal)
- gemma-2-9b-it (unimodal)
- claude-3.5-sonnet-20241022 (uni- and multi-modal settings)
- claude-3.5-haiku-20241022 (uni- and multi-modal settings)
- gemini-pro-1.5 (uni- and multimodal settings)
- gemini-flash-1.5 (uni- and multimodal settings)
- gemini-2.5-flash (uni- and multimodal settings)
- gpt-4o-2024-11-20 (uni- and multimodal settings)
- gpt-4o-mini-2024-07-18 (uni- and multi-modal settings)
- o4-mini-2025-04-16 (multimodal settings)

All models are used with a temperature of 0, to make them as deterministic as possible.

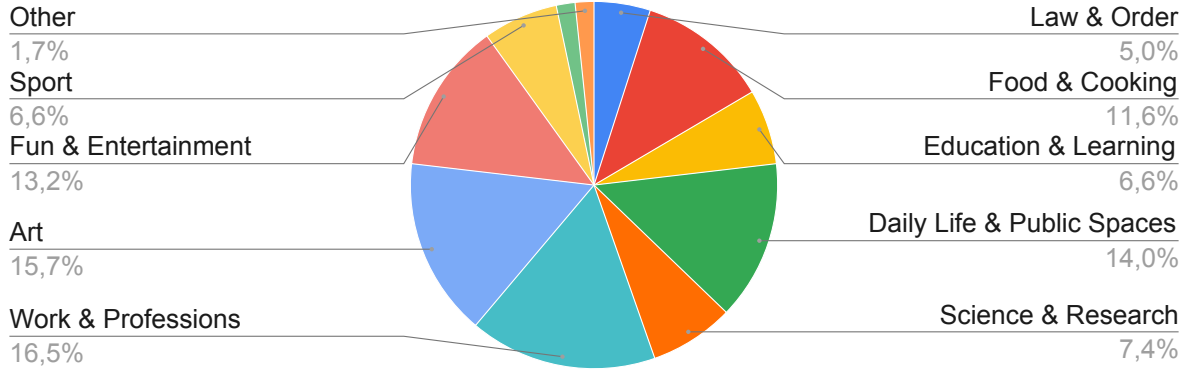


Figure 10: Distribution of the social contexts used in ToM-SSI.

CMSC

[ 3 , None, None, None]

[None, None, None, None]

[None, None, None, None]

[ 0 , None, None, 2 ]

[ 1 , None, None, None]

CMCC

[ 2 , None, None, None, None, None, None]

[ 1 , None, None, 0 , None, None, 3 ]

PCC and OC

[None, None, 2 , None]

[None, 1 , None, None]

[None, None, None, None]

[ 0 , None, None, None]

MC

[ 0 , None]

[ 1 , None]

[None, None]

[ 2 , None]

Figure 11: Minimal templates for each task. Agent IDs are denoted as integers (1, 2, 3, 4) while None denotes an empty cell.

## A.6 Human Study

We recruited 20 human participants (8 female, 12 male, aged between 21 and 40 years old) and asked them to answer 45 questions, equally split between tasks. The study was approved by the institutional ethics committee. Some participants were university students who received course credits as compensation, in accordance with university regulations. The remaining participants voluntarily joined the study, without receiving any form of

### Algorithm 1 Generation pipeline

**Require:** Set of agents  $A \ni A_j$ ,  $|A| = N_A$ . Set of agents' initial knowledge  $I_A \ni I_{A_j}$ ,  $j \in [0, N_A - 1]$ . Minimal grid template  $T$  of size  $s \times s$ . Set of grid transformations  $\mathcal{T}$ . Set of social contexts  $\mathcal{C}$ . Prompt template  $P$

**Require:** isMultimodal  $\in \{0, 1\}$

- 1:  $A' = \text{RandomShuffle}(A)$
- 2: Sample  $\delta \sim \text{Uniform}(0, 3)$
- 3:  $T' = \text{IncreaseSize}(T, s')$ , where  $s' = s + \delta$
- 4:  $T'' = f(T')$ , where  $f \sim \mathcal{T}$
- 5: **if** isMultimodal **then**
- 6:     GridRepr = RenderImage( $T''$ )
- 7: **else**
- 8:     GridRepr = RenderText( $T''$ )
- 9: **end if**
- 10: Sample  $c \sim \mathcal{C}$
- 11:  $I'_A = \text{RandomShuffle}(I_A)$
- 12: Prompt = MakePrompt( $P, A', c, I'_A, \text{GridRepr}$ )
- 13: Prompt' = ShuffleAnswers(Prompt)
- 14: **return** Prompt', GridRepr

compensation. Human participants were shown the same prompt used for evaluating models. At the beginning of the study, participants were informed about their task, the duration of the experiment, and that their responses would be kept anonymous and used solely for research purposes. They then went through a guided example that explains the rules of social interactions, so that they could optionally skip the first part of the prompt, which contains the same rules for each sample. Between tasks, participants had to answer additional attention-check questions to ensure their attentiveness.

```

Is the statement 'After the Event, A0 learns what
A1 communicated' true? Yes

Is the statement 'After the Event, A0 learns what
A1 communicated' false? No

Is the statement 'After the Event, A0 does NOT
learn what A1 communicated' false? Yes

Is the statement 'After the Event, A0 does NOT
learn what A1 communicated' true? No

Is the statement 'When the Event happens, A0 and
A1 are in adjacent cells' true? Yes

Is the statement 'When the Event happens, A0 and
A1 are in adjacent cells' false? No

Is the statement 'When the Event happens, A0 and
A1 are NOT in adjacent cells' false? Yes

Is the statement 'When the Event happens, A0 and
A1 are NOT in adjacent cells' true? No

```

Figure 12: Example of question variants for the CMSC task.

### A.7 Task Accuracy

Table 1 contains the accuracy achieved by the baseline models on each question type – Percept (P), Belief (B), Intention (I) – and on PBI ( $P \wedge B \wedge I$ ) for the five tasks in ToM-SSI.

Overall, models perform well—and in some cases even perfectly—when answering individual question types (P, B, or I). However, their performance drops when evaluated on the full PBI score, suggesting a lack of consistency across the different types of inference required to understand percepts, beliefs, and intentions.

In the text-only setting, Claude 3.5 Sonnet is the top-performing model across all tasks, except for CMCC, where Gemini Pro 1.5 performs better. When looking at the average PBI scores across models, the most challenging task is CMCC (4.6), followed by CMSC (12.8), MC (12.8), OC (13.9), and finally PCC (20.3). Notably, there is a significant performance gap between CMCC and the other tasks. As shown in Table 1, this is largely due to low performance on Belief questions, which are especially challenging as they require second-order ToM reasoning.

In the multimodal setting, the best-performing models vary across tasks: Gemini Pro 1.5 leads on CMSC, Claude 3.5 Sonnet excels on CMCC and MC, and Qwen2-VL Instruct 7B performs best on PCC and OC. The trend in difficulty remains similar to the text-only setting, with CMCC again being the most challenging task by far (0.93). Table 1 shows again that this is due to a poor performance

in Belief questions. The main difference in the multimodal setting is that MC becomes more difficult than CMSC, while PCC and OC continue to yield the highest average PBI scores across models.

### A.8 P, PB, and PBI Inference

Figure 8a shows the change in performance across P, PB, and PBI accuracy for all the VLMs we evaluated on ToM-SSI. Similarly, Figure 8b shows the performance of LMs on the text-only version of ToM-SSI. Overall, model performance generally declines from percepts to beliefs to intentions. However, LMs evaluated on text demonstrate greater robustness in PB inference across most tasks (CMSC, PCC, OC, MC). Among them, the top-performing models, Claude 3.5 Sonnet and GPT-4o, achieve PB accuracy comparable to human levels in CMSC and MC.

### A.9 Multimodal vs Unimodal Performance

In §3.2 we show that GPT-4o does not benefit from the addition of images (Figure 5, top), while Qwen2 shows a clear benefit from the image input, particularly on percept questions (Figure 5, bottom). Figure 9 shows additional comparisons of VLMs’ performance when evaluated on the multimodal version ToM-SSI versus the text-only version. In general, models do not particularly benefit from images, especially Claude 3.5 Sonnet, Gemini-1.5-Pro, and GPT-4o. For other models like Llama 3.2 and Molmo, performance does not change much between modalities. Figure 9 presents additional comparisons of VLM performance on the multimodal version of ToM-SSI versus the text-only version. Overall, state-of-the-art models – Claude, Gemini and GPT – do not benefit from images and show stronger performance on text. Other models, like Llama 3.2 and Molmo, show some (but not consistent) improvements when evaluated in the multimodal setting. Gemini 2.5 Flash also benefits from multimodal inputs, especially for Belief questions in PCC, OC, and to a lesser extent in MC.

**The Strange Case of GPT-4o’s Coordinate System** As discussed in §3.2, §A.9, and shown in detail in Table 1, GPT-4o’s performance on multimodal percept questions is poor across all tasks, which in turn results in a low PBI accuracy. To gain more insight, we examined GPT-4o’s generated output and found that the model often attempts to answer percept questions by defining a coordinate system for the grid world. In such cases, two

issues arise: first, the coordinate system defined by GPT-4o changes across different samples. Second, even within a single sample and using its own coordinate system, the model often fails to correctly position agents on the grid. In contrast, GPT-4o demonstrates higher accuracy when it reasons in terms of relative positions, such as “ $A_j$  is to the right of  $A_k$ ”.

#### **A.10 Compute resources**

We ran open-source models of size below 15B on a server running Ubuntu 22.04, equipped with eight NVIDIA Tesla V100-SXM2 GPUs with 32GB of memory and Intel Xeon Platinum 8260 CPUs. Proprietary models are used through API.

#### **A.11 Code**

Our code is public under the MIT license at <https://git.hcics.simtech.uni-stuttgart.de/public-projects/tom-ssi>.

## Example 1: Percept prompt

### Task:

You are a helpful AI assistant tasked to answer a question about a designer in a design studio.

### Setup:

You are observing a gridworld that represents a design studio where 4 graphic designers are working on a project.

### Rules:

1. The design studio is represented as a grid of size 6x6 with 4 designers: A0, A1, A2, and A3. Each designer occupies one cell.
2. Each designer starts out knowing a subset of 4 pieces of information: ['Feedback Loop', 'Color Scheme', 'Font Choice', 'Design Concept'].
3. By default, designers have two goals: (1) learning information they are missing from other designers, and (2) communicating information that other designers are missing. Setting a designer's Attitude can influence their goals.
4. Each designer can see the positions, movements, and initial information of other designers. However, they cannot see any new information other designers learn later, so they must infer it.
5. Event(s) change the state of the design studio and include: movement of designers (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If a designer shares a piece of information during the Event(s), only the designers in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

Initial position of the designers in the design studio:

```
+-----+
|   |   |   |   | A1 |   |
+-----+
| A3 |   |   |   | A2 |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   | A0 |   |
+-----+
```

Information initially known by the designers:

- A0: ['Font Choice', 'Design Concept']
- A1: ['Feedback Loop', 'Design Concept']
- A2: ['Color Scheme', 'Design Concept']
- A3: ['Feedback Loop', 'Color Scheme', 'Font Choice']

A2's Attitude:

Collaborative: A2 wants to learn new information from other designers and share information that other designers don't know.

Event:

Designer A1 communicates 'Feedback Loop'.

Question:

Is the statement 'After the Event, A2 learns what A1 communicated' true?

Based on the information above, choose one of the following answers:

- A) No
- B) Yes

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

## Example 2: Belief prompt

### Task:

You are a helpful AI assistant tasked to answer a question about a designer in a design studio.

### Setup:

You are observing a gridworld that represents a design studio where 4 graphic designers are working on a project.

### Rules:

1. The design studio is represented as a grid of size 6x6 with 4 designers: A0, A1, A2, and A3. Each designer occupies one cell.
2. Each designer starts out knowing a subset of 4 pieces of information: ['Feedback Loop', 'Color Scheme', 'Font Choice', 'Design Concept'].
3. By default, designers have two goals: (1) learning information they are missing from other designers, and (2) communicating information that other designers are missing. Setting a designer's Attitude can influence their goals.
4. Each designer can see the positions, movements, and initial information of other designers. However, they cannot see any new information other designers learn later, so they must infer it.
5. Event(s) change the state of the design studio and include: movement of designers (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If a designer shares a piece of information during the Event(s), only the designers in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

Initial position of the designers in the design studio:

```
+-----+
|   |   |   |   | A1 |   |
+-----+
| A3 |   |   |   | A2 |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   | A0 |   |
+-----+
```

Information initially known by the designers:

- A0: ['Font Choice', 'Design Concept']
- A1: ['Feedback Loop', 'Design Concept']
- A2: ['Color Scheme', 'Design Concept']
- A3: ['Feedback Loop', 'Color Scheme', 'Font Choice']

A2's Attitude:

Collaborative: A2 wants to learn new information from other designers and share information that other designers don't know.

Event:

Designer A1 communicates 'Feedback Loop'.

Question:

After the Event, what information does A2 believe she is still missing?

Based on the information above, choose one of the following answers:

- A) Design Concept
- B) Feedback Loop
- C) Font Choice
- D) Color Scheme

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

### Example 3: Intention prompt

**Task:**

You are a helpful AI assistant tasked to answer a question about a designer in a design studio.

**Setup:**

You are observing a gridworld that represents a design studio where 4 graphic designers are working on a project.

**Rules:**

1. The design studio is represented as a grid of size 6x6 with 4 designers: A0, A1, A2, and A3. Each designer occupies one cell.
2. Each designer starts out knowing a subset of 4 pieces of information: ['Feedback Loop', 'Color Scheme', 'Font Choice', 'Design Concept'].
3. By default, designers have two goals: (1) learning information they are missing from other designers, and (2) communicating information that other designers are missing. Setting a designer's Attitude can influence their goals.
4. Each designer can see the positions, movements, and initial information of other designers. However, they cannot see any new information other designers learn later, so they must infer it.
5. Event(s) change the state of the design studio and include: movement of designers (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If a designer shares a piece of information during the Event(s), only the designers in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

Initial position of the designers in the design studio:

```
+-----+
|   |   |   |   | A1 |   |
+-----+
| A3 |   |   |   | A2 |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   |   |   |
+-----+
|   |   |   |   | A0 |   |
+-----+
```

Information initially known by the designers:

- A0: ['Font Choice', 'Design Concept']
- A1: ['Feedback Loop', 'Design Concept']
- A2: ['Color Scheme', 'Design Concept']
- A3: ['Feedback Loop', 'Color Scheme', 'Font Choice']

A2's Attitude:

Collaborative: A2 wants to learn new information from other designers and share information that other designers don't know.

Event:

Designer A1 communicates 'Feedback Loop'.

Question:

After the Event, which designer is A2 most likely to approach?

Based on the information above, choose one of the following answers:

- A) A3
- B) A0
- C) A1

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

#### Example 4: PB Error Case in CMSC – Llama-3.2-11B-Vision

**Task:**

You are a helpful AI assistant tasked to answer a question about a reporter in a newsroom.

**Setup:**

You are observing a gridworld that represents a newsroom where 4 reporters are gathering information for breaking news.

**Rules:**

1. The newsroom is represented as a grid of size 11x11 with 4 reporters: A0, A1, A2, and A3. Each reporter occupies one cell.
2. Each reporter starts out knowing a subset of 4 pieces of information: ['Breaking News', 'Press Release', 'Eyewitness Account', 'Photograph'].
3. By default, reporters have two goals: (1) learning information they are missing from other reporters, and (2) communicating information that other reporters are missing. Setting a reporter's Attitude can influence their goals.
4. Each reporter can see the positions, movements, and initial information of other reporters. However, they cannot see any new information other reporters learn later, so they must infer it.
5. Event(s) change the state of the newsroom and include: movement of reporters (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If a reporter shares a piece of information during the Event(s), only the reporters in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

The initial position of the reporters in the newsroom is shown in the provided image.

**Information initially known by the reporters:**

- A0: ['Breaking News', 'Press Release', 'Eyewitness Account']
- A1: ['Breaking News', 'Press Release', 'Eyewitness Account']
- A2: ['Breaking News', 'Press Release', 'Eyewitness Account', 'Photograph']
- A3: ['Breaking News', 'Eyewitness Account', 'Photograph']

**A2's Attitude:**

Collaborative: A2 wants to learn new information from other reporters and share information that other reporters don't know.

**Events:**

Reporter A3 communicates 'Photograph'.

Reporter A1 communicates 'Press Release'.

**Question:**

After the Event, which information does A2 believe A3 and A1 are likely still missing, respectively?

Based on the information above and on the image, choose one of the following answers:

- A) A2 believes that A3 is missing Breaking News and A1 is missing Press Release
- B) A2 believes that A3 and A1 are not missing any information
- C) A2 believes that A3 is missing Eyewitness Account and A1 is missing Photograph
- D) A2 believes that A3 is missing Press Release and A1 is missing Photograph

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

To determine which information A2 believes A3 and A1 are likely still missing, we need to analyze the events that have occurred and the initial information known by each reporter.

Initially, A2 knows all four pieces of information: ['Breaking News', 'Press Release', 'Eyewitness Account', 'Photograph']. A3 knows ['Breaking News', 'Eyewitness Account', 'Photograph'], and A1 knows ['Breaking News', 'Press Release', 'Eyewitness Account'].

During the events, Reporter A3 communicates 'Photograph' and Reporter A1 communicates 'Press Release'.

After these events, A2 will update its beliefs about what A3 and A1 are likely still missing.

A2 believes that A3 is missing Press Release because A3 did not communicate it, and A2 knows that A3 knows 'Photograph' but not 'Press Release'. A2 believes that A1 is missing Photograph because A1 did not communicate it, and A2 knows that A1 knows 'Press Release' but not 'Photograph'.

Therefore, the correct answer is:

<Answer>D</Answer>

Correct answer: B

### Example 5: PBI Error Case in CMSC – Llama-3.2-11B-Vision

**Task:**

You are a helpful AI assistant tasked to answer a question about a baker in a bake sale.

**Setup:**

You are observing a gridworld that represents a charity bake sale where 4 bakers are collaborating to raise funds.

**Rules:**

1. The bake sale is represented as a grid of size 7x7 with 4 bakers: A0, A1, A2, and A3. Each baker occupies one cell.
2. Each baker starts out knowing a subset of 4 pieces of information: ['Special Ingredient', 'Customer Feedback', 'Baked Good Description', 'Price Tag'].
3. By default, bakers have two goals: (1) learning information they are missing from other bakers, and (2) communicating information that other bakers are missing. Setting a baker's Attitude can influence their goals.
4. Each baker can see the positions, movements, and initial information of other bakers. However, they cannot see any new information other bakers learn later, so they must infer it.
5. Event(s) change the state of the bake sale and include: movement of bakers (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If a baker shares a piece of information during the Event(s), only the bakers in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

The initial position of the bakers in the bake sale is shown in the provided image.

**Information initially known by the bakers:**

- A0: ['Special Ingredient', 'Customer Feedback', 'Baked Good Description']
- A1: ['Special Ingredient', 'Baked Good Description', 'Price Tag']
- A2: ['Special Ingredient', 'Customer Feedback', 'Baked Good Description', 'Price Tag']
- A3: ['Special Ingredient', 'Customer Feedback', 'Baked Good Description']

**A2's Attitude:**

Collaborative: A2 wants to learn new information from other bakers and share information that other bakers don't know.

**Events:**

Baker A1 communicates 'Price Tag'.

Baker A3 communicates 'Customer Feedback'.

**Question:**

After the Events, which baker is A2 most likely to approach?

Based on the information above and on the image, choose one of the following answers:

- A) A0
- B) A1
- C) A3

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

To determine which baker A2 is most likely to approach after the events, we need to analyze the information shared by A1 and A3.

Initially, A2 knows 'Special Ingredient', 'Customer Feedback', and 'Baked Good Description'. After the events, A1 communicates 'Price Tag', and A3 communicates 'Customer Feedback'. Since A2 already knows 'Customer Feedback', the new information shared by A3 is not relevant to A2's current knowledge.

However, A1's communication of 'Price Tag' is new information for A2. As A2 has a collaborative attitude, she wants to learn new information from other bakers. Therefore, A2 is most likely to approach A1 to learn the 'Price Tag' information.

**\*\*Answer:\*\* B) A1**

---

**Correct answer: A**

## Example 6: PBI Error Case in MC – Llama-3.2-11B-Vision

### Task:

You are a helpful AI assistant tasked to answer a question about an artist in a art gallery.

### Setup:

You are observing a gridworld that represents a local art gallery opening where 3 artists are showcasing their work.

### Rules:

1. The art gallery is represented as a grid of size 7x7 with 3 artists: A0, A1, and A2. Each artist occupies one cell.
2. Each artist starts out knowing a subset of 4 pieces of information: ['Material Source', 'Art Piece Title', 'Technique Explanation', 'Inspiration Story'].
3. By default, artists have two goals: (1) learning information they are missing from other artists, and (2) communicating information that other artists are missing. Setting an artist's Attitude can influence their goals.
4. Each artist can see the positions, movements, and initial information of other artists. However, they cannot see any new information other artists learn later, so they must infer it.
5. Event(s) change the state of the art gallery and include: movement of artists (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If an artist shares a piece of information during the Event(s), only the artists in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

The initial position of the artists in the art gallery is shown in the provided image.

Information initially known by the artists:

- A0: ['Art Piece Title', 'Technique Explanation', 'Inspiration Story']
- A1: ['Technique Explanation', 'Inspiration Story']
- A2: ['Material Source', 'Art Piece Title', 'Technique Explanation', 'Inspiration Story']

A2's Attitude:

Collaborative towards A1: A2 wants A1 to learn new information.

Obstructive towards A0: A2 doesn't want A0 to learn new information.

### Event:

Artist A0 moves one cell down.

### Question:

After the Event, which information is A2 most likely to communicate?

Based on the information above and on the image, choose one of the following answers:

- A) Technique Explanation
- B) Material Source
- C) Art Piece Title
- D) Inspiration Story

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

To determine which information A2 is most likely to communicate after the event, we need to analyze A2's Attitude and the initial information known by each artist.

A2's Attitude is collaborative towards A1 and obstructive towards A0. This means A2 wants A1 to learn new information but does not want A0 to learn new information.

Initially, A2 knows all the information except 'Material Source'. Since A2 is obstructive towards A0, it will not share any information with A0.

A0 has moved one cell down, which does not change the fact that A2 is obstructive towards A0.

Therefore, A2 will not share any information with A0.

A2 is collaborative towards A1, but A1 already knows 'Technique Explanation' and 'Inspiration Story'. A2 knows 'Material Source' and 'Art Piece Title', but A1 already knows 'Art Piece Title'.

A2 will share the information that A1 does not know.

The information that A2 knows and A1 does not know is 'Material Source'.

So, after the event, A2 is most likely to communicate 'Material Source'.

<Answer>B</Answer>

---

Correct answer: C

## Example 7: PBI Correct Case in MC – Llama-3.2-11B-Vision

### Task:

You are a helpful AI assistant tasked to answer a question about a speaker in a tech conference.

### Setup:

You are observing a gridworld that represents a tech conference where 3 speakers are presenting their innovations.

### Rules:

1. The tech conference is represented as a grid of size 5x5 with 3 speakers: A0, A1, and A2. Each speaker occupies one cell.
2. Each speaker starts out knowing a subset of 4 pieces of information: ['Future Trends', 'Q&A Session', 'Technical Challenges', 'Innovation Overview'].
3. By default, speakers have two goals: (1) learning information they are missing from other speakers, and (2) communicating information that other speakers are missing. Setting a speaker's Attitude can influence their goals.
4. Each speaker can see the positions, movements, and initial information of other speakers. However, they cannot see any new information other speakers learn later, so they must infer it.
5. Event(s) change the state of the tech conference and include: movement of speakers (up, down, left, right) or communication of information. Aside from what is specified in the Event(s), no other movement or communication happens.
6. If a speaker shares a piece of information during the Event(s), only the speakers in the adjacent cells (up, down, left, right, and diagonally neighboring cells) will learn it.

The initial position of the speakers in the tech conference is shown in the provided image.

### Information initially known by the speakers:

- A0: ['Q&A Session', 'Technical Challenges', 'Innovation Overview']
- A1: ['Technical Challenges', 'Innovation Overview']
- A2: ['Future Trends', 'Q&A Session', 'Technical Challenges', 'Innovation Overview']

### A2's Attitude:

Collaborative towards A1: A2 wants A1 to learn new information.

Obstructive towards A0: A2 doesn't want A0 to learn new information.

### Event:

Speaker A0 moves one cell left.

### Question:

After the Event, which information is A2 most likely to communicate?

Based on the information above and on the image, choose one of the following answers:

- A) Q&A Session
- B) Innovation Overview
- C) Future Trends
- D) Technical Challenges

Provide your final answer within the tags <Answer>[answer]</Answer> (e.g. <Answer>A</Answer>).

To determine which information A2 is most likely to communicate after the event, we need to consider A2's Attitude and the new positions of the speakers.

A2's Attitude is Collaborative towards A1 and Obstructive towards A0. This means A2 wants to help A1 learn new information but does not want to help A0 learn new information.

After the event, A0 moves one cell left, so the new positions are:

- \* A0: (2, 1)
- \* A1: (2, 2)
- \* A2: (2, 3)

Since A2 is obstructive towards A0, it is unlikely to communicate any information to A0. Therefore, we can rule out options that would benefit A0.

A2 is collaborative towards A1, so it is likely to communicate information that A1 is missing. A1 initially knows 'Technical Challenges' and 'Innovation Overview', but it does not know 'Q&A Session'. Since A2 is in a position where it can communicate with A1, it is most likely to communicate 'Q&A Session', which is the only information A1 is missing.

Therefore, the correct answer is:

<Answer>A</Answer>

Correct answer: A