Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models

Ziqiao Ma **Jacob Sansom** **Run Peng** **Joyce Chai**
Computer Science and Engineering Division, University of Michigan
{marstin,jhsansom,roihn,chaijy}@umich.edu

Abstract

Large Language Models (LLMs) have generated considerable interest and debate regarding their potential emergence of Theory of Mind (ToM). Several recent inquiries reveal a lack of robust ToM in these models and pose a pressing demand to develop new benchmarks, as current ones primarily focus on different aspects of ToM and are prone to shortcuts and data leakage. In this position paper, we seek to answer two road-blocking questions: (1) How can we taxonomize a holistic landscape of machine ToM? (2) What is a more effective evaluation protocol for machine ToM? Following psychological studies, we taxonomize machine ToM into 7 mental state categories and delineate existing benchmarks to identify under-explored aspects of ToM. We argue for a holistic and situated evaluation of ToM to break ToM into individual components and treat LLMs as an agent who is physically situated in environments and socially situated in interactions with humans. Such situated evaluation provides a more comprehensive assessment of mental states and potentially mitigates the risk of shortcuts and data leakage. We further present a pilot study in a grid world setup as a proof of concept. We hope this position paper can facilitate future research to integrate ToM with LLMs and offer an intuitive means for researchers to better position their work in the landscape of ToM.

1 Introduction

The term *theory of mind* (ToM, sometimes also referred to as *mentalization* or *mindreading*) was first introduced by Premack and Woodruff (1978) as agents' ability to impute *mental states* to themselves and others. Many aspects of human cognition and social reasoning rely on ToM modeling of others' mental states (Gopnik and Wellman, 1992; Baron-Cohen, 1997; Gunning, 2018). This is crucial for understanding and predicting others' actions (Dennett, 1988), planning over others' beliefs and next actions (Ho et al., 2022), and various forms of reasoning and decision-making (Pereira et al., 2016; Rusch et al., 2020). Inspired by human ToM, AI researchers have made explicit and implicit efforts to develop a machine ToM for *social intelligence*: AI agents that engage in social interactions with humans (Krämer et al., 2012; Kennington, 2022) and other agents (Albrecht and Stone, 2018). A machine ToM enables an interactive paradigm of language processing (Wang et al., 2023), enhancing agents' capacity for interactions (Wang et al., 2021), explainable decision-making (Akula et al., 2022), dialogue communication (Qiu et al., 2022; Takmaz et al., 2023), and collaborative task planning (Bara et al., 2023).

Machine ToM has received an increasing amount of attention, especially as the field is reshaped by *large language models* (LLMs) such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023). This highlights an ongoing debate and discussion on whether a machine ToM has emerged in LLMs. While LLMs have demonstrated some capability of inferring communicative intentions, beliefs, and desires (Andreas, 2022; Kosinski, 2023; Bubeck et al., 2023), researchers also reported concerns regarding a lack of robust *agency* in LLMs for complex social and belief reasoning tasks (Sap et al., 2022; Shapira et al., 2023a) and in-context pragmatic communication (Ruis et al., 2022). Emerged or not emerged, that remains a question (or may not even be the central question to ask). In our view, existing evaluation protocols do not fully resolve this debate. Most current benchmarks focus only on a (few) aspect(s) of ToM, in the form of written stories, and are prone to data contamination, shortcuts, and spurious correlations (Trott et al., 2022; Aru et al., 2023; Shapira et al., 2023a). Prior to embarking on extensive data collection for new ToM benchmarks, it is crucial to address two key questions: (1) How can we taxonomize a holistic landscape of machine ToM? (2) What is a more effective evaluation protocol for machine ToM?

To embrace the transformation brought by LLMs and explore their full potential in understanding and modeling ToM, this position paper calls for a holistic investigation that taxonomizes ToM using the *Abilities in Theory of Mind Space* (ATOMS) framework (Beaudoin et al., 2020). After a review of existing benchmarks under this framework, we put forward a situated evaluation of ToM, one that treats LLMs as agents who are physically situated in environments and socially situated in interactions with humans. We hope this paper will offer an intuitive means to identify research priorities and to help gain a deeper understanding of, as well as to effectively utilize, LLMs in ToM modeling for AI agents in the future.

2 Large Language Models as Theory of Mind Agents

Since the advent of pre-trained language models, the research community has questioned whether they possess intrinsic mental states to represent the environment (Li et al., 2021; Storks et al., 2021; Hase et al., 2023) and comprehend the mental states of others (Sap et al., 2019; Zhang et al., 2021) through the textual description (observation) of behavioral cues. The relatively recent breakthroughs of LLMs have created many discussions and debates, primarily concerning the extent to which LLMs possess various capabilities required for a machine ToM. In this section, we first survey recent research presenting evidence and counter-evidence for the emergence of ToM in LLMs. We conclude the discussion with the limitations of current evaluation protocols.

2.1 Do Machine ToM Emerge in LLMs?

Evidence for emergent ToM in LLMs. Prior to the rise of large language models, there has been growing evidence and acknowledgment of a narrow and limited sense of agency in smaller language models. Andreas (2022) argues that language models have the capacity to predict relations between agents' observations, mental states, actions, and utterances, as they infer approximate representations of beliefs, desires, and intentions of agents mentioned in the context. These representations have a causal influence on the generated text, similar to an intentional agent's state influencing its communicative actions under a Belief-Desire-Intention (BDI) agent model (Bratman, 1987). Amidst the excitement surrounding the release of GPT-4 (OpenAI, 2023), researchers have searched for evidence of an emergent ToM in LLMs. Kosinski (2023) presents 20 case studies each of the unexpected contents task (Perner et al., 1987) and the unexpected transfer (Sally-Anne) task (Baron-Cohen et al., 1985). With direct comparisons to children's performance, the findings have been cited as potential evidence for a spontaneous emergence of ToM in LLMs. Bubeck et al. (2023) present a similar behavioral study with 10 cases of belief, emotion, and intention understanding, concluding that GPT-4 has an advanced level of ToM after qualitative comparison with predecessors. Other case studies have also shown aspects of machine ToM (Li et al., 2023; Holterman and van Deemter, 2023).

Limitations of ToM capabilities in LLMs. The above findings contradict the conclusions drawn in Sap et al. (2022)'s earlier study, which shows a clear lack of ToM in GPT-3 (Brown et al., 2020) on SocialIQA (Sap et al., 2019) and ToMi (Le et al., 2019) benchmarks. As a potential account, there has been criticism that the cognitive inquiries are anecdotal and inadequate for evaluating ToM in LLMs (Marcus and Davis, 2023; Mitchell and Krakauer, 2023; Shapira et al., 2023a). Following the same evaluation protocol, Ullman (2023) demonstrates that simple adversarial alternatives to Kosinski (2023) can fail LLMs. To further understand if the most recent variants of LLMs possess a robust ToM, Shapira et al. (2023a) present a comprehensive evaluation over 6 tasks and 3 probing methods, showing that a robust machine ToM is absent even in GPT-4 and that LLMs are prone to shortcuts and spurious correlations. Based on the ongoing debate, it can be concluded that, while LLMs exhibit some level of sensitivity at understanding others' mental states, this capability is limited and falls short of achieving robust human-level ToM (Trott et al., 2022; Shapira et al., 2023a).

2.2 Roadblocks in ToM Evaluation in LLMs

Given the pressing need for a robust machine ToM in LLMs and large-scale ToM benchmarks, researchers echo several difficulties in the evaluation protocol. Presently, ToM benchmarks suffer from three primary issues summarized as follows.

Limited aspects of ToM. The evaluation of machine ToM lacks consistency in the literature due to the ambiguity surrounding the specific mental states being targeted. Existing benchmarks often focus on limited numbers of mental states, such as the *intention* (Yoshida et al., 2008), *belief* (Grant

et al., 2017), *emotion* (Sap et al., 2019), and *knowledge* (Bara et al., 2021) of another agent. While all of these are necessary building blocks of machine ToM, we echo Shapira et al. (2023a)'s concern that the ToM capability of LLMs may have been overclaimed based on evaluations from only a specific aspect of ToM. To give a comprehensive assessment of a holistic machine ToM, a taxonomy is essential to enable researchers to effectively position their work with different focuses and priorities, which may be orthogonal to each other.

Data contamination. Data contamination refers to the lack of a verifiable train-test split that is typically established to test the ability of machine learning models to generalize (Magar and Schwartz, 2022). LLMs typically learn from internet-scale data, potentially giving them access during training to the data used to test them (Bubeck et al., 2023; Hagendorff, 2023). For ToM evaluation specifically, the training corpora of LLMs may contain research papers detailing these psychological studies. Many past studies used identical or slightly altered language prompts to test LLMs, leading to potential contamination issues (Ullman, 2023). To critically evaluate the performance of LLMs on ToM tasks, researchers must have access to the datasets used to train them (Dodge et al., 2021), which are unfortunately not available.

Shortcuts and spurious correlations. The availability of shortcuts and spurious features has triggered many concerns that a model may leverage them to perform highly on a benchmark without robustly acquiring the desired skill (Sclar et al., 2023; Ullman, 2023; Shapira et al., 2023a). Recent findings suggest that LLMs tend to learn surface-level statistical correlations in compositional tasks, potentially leading to an illusion of systematic learning (Dziri et al., 2023). In all likelihood, LLMs are capable of learning ToM shortcuts in a similar manner.

3 Towards A Holistic Landscape of Machine Theory of Mind

3.1 Abilities in Theory of Mind Space (ATOMS) Framework

The evaluation of machine ToM lacks clarity and consistency across various literature, primarily due to the ambiguity surrounding the specific *mental states* being targeted. This ambiguity is not unique to the field of AI but is rooted in the complicated cognitive underpinnings of ToM. At the core of this ambiguity is the latent nature of *mental states*, the subject has privileged access to them while others can only infer the existence of these mental states based on observable behaviors or expressions (Dretske, 1979; Blakemore and Decety, 2001; Zaki et al., 2009). Thus, it is impossible to directly access and assess the mental states of a human, and ToM must be tested indirectly through humans' ability to understand the relationship between mental states and behaviors, especially by predicting how agents behave based on their mental states (Swettenham, 1996; Phillips et al., 2002).

While the exact definition of ToM remains a central debate, the AI community can benefit from looking at what psychologists have viewed as an initial step. In this paper, we follow Beaudoin et al. (2020)'s taxonomy of ToM sub-domains, *i.e.,* the Abilities in Theory of Mind Space (ATOMS). As shown in Figure 1, the space consists of 7 categories of mental states, including *beliefs*, *intentions*, *desires*, *emotions*, *knowledge*, *percepts*, and *non-literal communication*. We selected this taxonomy because it was derived from a comprehensive meta-analysis of ToM studies. The meta-analysis focused on young children aged 0-5 years at the early stage of cognitive development, such that the setups are simpler and more comparable, avoiding complicated physical and social engagements that cannot be trivially deployed on LLMs.

Beliefs. Beliefs are informational states that people judge to be true, usually decoupled from motivational states (Dennett, 1995; Eccles and Wigfield, 2002). Beliefs, the most studied mental states in the field of ToM, are usually tested in the form of false belief tasks, including the unexpected contents test (Perner et al., 1987), the unexpected transfer (Sally-Anne) Test (Baron-Cohen et al., 1985), the second-order false belief (Ice-cream Van) Test (Perner and Wimmer, 1985). Researchers also studied their connection to actions and emotions (Swettenham, 1996).

Intentions. Intentions are choices with commitment, usually associated with concrete actions towards a goal (Cohen and Levesque, 1990). As a critical component of ToM, Kennington (2022) has called for a more explicit treatment of intentions. Intentions have been extensively explored in psychology tests, e.g., behavioral re-enactment (Meltzoff, 1995), action prediction (Phillips et al., 2002), intention explanation (Smiley, 2001), and intention attribution to abstract figures (Castelli, 2006).

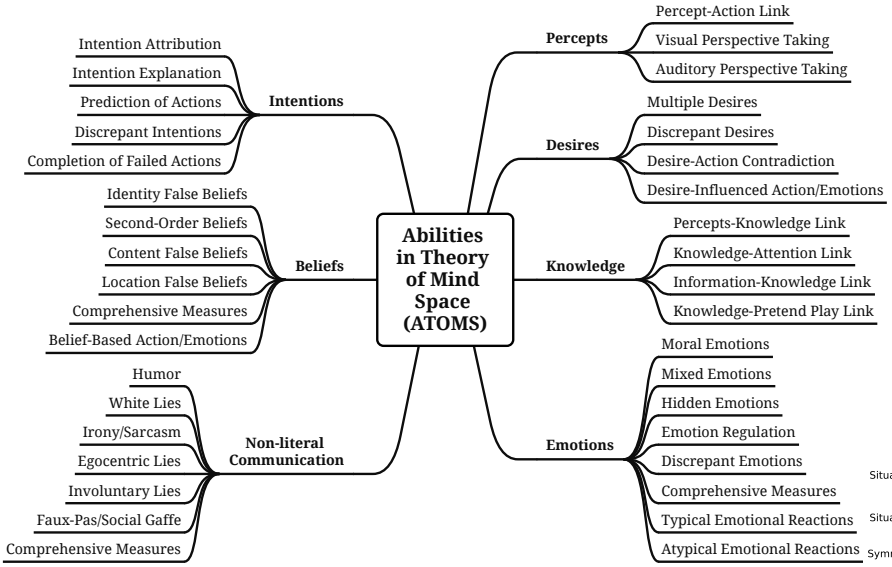


Figure 1: The ATOMS framework of Beaudoin et al. (2020), which identified 7 categories of mental states through meta-analysis of ToM studies for children.

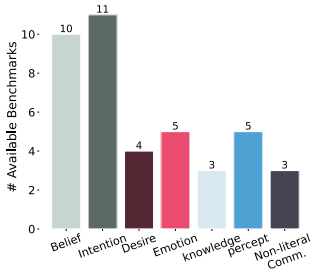


Figure 2: Number of available benchmarks for each mental state in ATOMS.

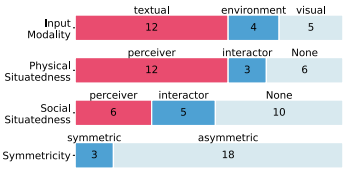


Figure 3: A comparison of benchmark settings.

Benchmarks and Task Formulations	Tested Agent			Situatedness				ATOMS Mental States									Sym.
	Task	Input Modality		Physical		Social		Belief		Intention		Des.	Emo.	Know.	Per.	NLC	
		Text	Nonling.	Per.	Int.	Per.	Int.	1st	2nd+	Act.	Com.						
EPISTEMIC REASONING (Cohen, 2021)	Infer	T	-					✓	✓								
TOMI (Nematzadeh et al., 2018)	QA	T	-	✓				✓	✓								
HI-TOM (He et al., 2023)	QA	T	-	✓				✓	✓								
MINDGAMES (Sileo and Lernould, 2023)	Infer	T	-	✓				✓	✓						✓		
ADV-CSFB (Shapira et al., 2023a)	QA	H	-	✓				✓									
CONVENTAIL (Zhang and Chai, 2010)	Infer	H	-			✓		✓			✓	✓					
SOCIALIQA (Sap et al., 2019)	QA	H	-			✓				✓			✓				
BEST (Tracey et al., 2022)	-	H	-			✓		✓					✓			✓	
FAUXPAS-EAI (Shapira et al., 2023b)	QA	H,AI	-			✓		✓								✓	
COKE (Wu et al., 2023)	NLG	AI	-			✓	✓			✓			✓				
TOM-IN-AMC (Yu et al., 2022)	Infer	H	-	✓		✓				✓	✓						
G4C (Zhou et al., 2023b)	NLG	H,AI	-	✓		✓	✓			✓	✓				✓		
VISUALBELIEFS (Eysenbach et al., 2016)	Infer	-	Cartoon	✓				✓								✓	
TRIANGLE COPA (Gordon, 2016)	QA	H	Cartoon	✓		✓				✓			✓				
MSED (Jia et al., 2022)	Infer	H	Images	✓								✓	✓				
BIB (Gandhi et al., 2021)	Infer	-	2D Grid	✓						✓		✓					
AGENT (Shu et al., 2021)	Infer	-	3D Sim.	✓						✓		✓			✓		
MTOM (Rabinowitz et al., 2018)	Infer	-	2D Grid	✓				✓		✓							
SYMMTOM (Sclar et al., 2022)	MARL	-	2D Grid	✓	✓	✓	✓							✓			✓
MINDCRAFT (Bara et al., 2021)	Infer	H	3D Sim.	✓	✓	✓	✓			✓				✓	✓		✓
CPA (Bara et al., 2023)	Infer	H	3D Sim.	✓	✓	✓	✓			✓	✓			✓	✓		✓

Table 1: A taxonomized review of existing benchmarks for machine ToM and their settings under ATOMS. We further break **beliefs** into first-order beliefs (1st) and second-order beliefs or beyond (2nd+); and break **intentions** into Action intentions and Communicative intentions. **Tasks** are divided into Inference, Question Answering, Natural Language Generation, and MultiAgent Reinforcement Learning. **Input** modalities consist of Text (Human, AI, or Template) and Nonlinguistic ones. The latter further breaks into Cartoon, Natural Images, 2D Grid World, and 3D Simulation. The **Situatedness** is divided into None, Passive Perceiver, and Active Interactor. **Symmetricity** refers to whether the tested agent is co-situated and engaged in mutual interactions with other ToM agents.

Desires. Desires are motivational states that do not necessarily imply commitment, though they are usually emotionally charged and affect actions (Malle and Knobe, 2001; Kavanagh et al., 2005). Typical studies along this line include the Yummy-Yucky Task (Repacholi and Gopnik, 1997) for discrepant preferences from different individuals, the multiple desires within one individual (Bennett and Galpert, 1993), and the relationship between desires and emotions/actions (Wellman and Woolley, 1990; Colonnesi et al., 2008).

Emotions. Emotions are mental states associated with an individual's feelings and affective experiences, which could impact beliefs and behaviors (Frijda et al., 1986; Damasio, 2004). Most ToM studies on emotions focus on typical (Knafo et al., 2009) and atypical (Denham, 1986) emotional reactions to situations. Other studies also encompass affective perspective taking (Borke, 1971), understanding hidden emotions (Harris et al., 1986), and morally related emotions (Pons and Harris, 2000).

Knowledge. Many controversies revolve around the definition of knowledge as justified true beliefs (Gettier, 2000). In the context of AI, knowledge typically consists of information and organized representations of the world, which can be used to simplify understanding and address intricate reasoning and planning (Schank and Abelson, 2013). ToM studies usually involve understanding the absence of knowledge (Aronson and Golomb, 1999) as well as the connection between knowledge and perception (Ruffman and Olson, 1989) and attention (Moll et al., 2006).

Percepts. Humans are situated in the physical and social environments. To enable AI agents to operate in the world and communicate with humans, the sensory and social aspects of perception are crucial in a machine ToM. Along this line, psychological studies have investigated the perceptual perspective taking (Masangkay et al., 1974) and understanding the influence of limited perception on actions (Hadwin et al., 1997).

Non-literal communications. Being able to understand non-literal and figurative communication helps humans to perform pragmatic inference and reason about hidden words behind their written meanings (Giora, 2003). Non-literal communication has been recognized as an advanced ToM capability, spanning a wide spectrum of humor and deceptions (Happé, 1994), sarcasm (Sullivan et al., 1995), and faux-pas (social gaffe) situations (Baron-Cohen et al., 1999).

3.2 A Taxonomized Review of Benchmarks

The ATOMS framework can serve as an intuitive reference for researchers to identify their research priorities and situate their work better in the landscape of literature. We further take the initiative to provide a systematic review of existing benchmarks for machine ToM under the umbrella of ATOMS.[1] Although there are independent research initiatives on certain ToM facets like intention classification, emotion modeling, and aspects of non-literal communications, we primarily focus on those that explicitly target ToM or inferences of latent mental states. Besides the ToM dimensions in ATOMS, we further characterize the benchmarks on their task formulation, input modalities, physical and social situatedness, and symmetricity (whether the tested agent is co-situated and engaged in mutual interactions with other ToM agents). We summarize our review in Table 1 and discuss our observations and under-explored aspects of ToM evaluation.

Many aspects of ToM are under-explored. As shown in Figure 2, we notice an overwhelming research focus on the intention and belief aspects of machine ToM. Several other aspects of ToM have not received enough attention. While the field of NLP has thoroughly explored different facets of emotion and non-literal communication, e.g., in the context of dialogue systems, ToM has rarely been explicitly mentioned as motivation. More connections and integrative efforts are clearly needed.

Lack of clear targeted mental states. Explicitly mentioning the Sally-Anne Test (Baron-Cohen et al., 1985) as inspiration, Grant et al. (2017) developed the predecessor of ToMi (Le et al., 2019). Similarly, Nematzadeh et al. (2018) cited the Ice-cream Van Test (Perner and Wimmer, 1985) as motivation and the FauxPas-EAI (Shapira et al., 2023b) benchmark followed the study of Baron-Cohen et al. (1999). While these benchmarks are cognitively grounded and target one particular aspect of ToM, the majority often incorporate multiple mental states without clear descriptions, which could make it challenging to measure the actual progress (Raji et al., 2021).

Lack of situatedness in a physical and social environment. Figure 3 illustrates the configurations of benchmarks. Each bar in the chart represents a distinct benchmark characteristic, and each segment within the bar illustrates the proportion of benchmarks with one specific setting. An immediate observation is a noticeable lack of benchmarks that encompass both physical and social environments, which highlights an existing research disparity in the field. We notice that many existing benchmarks are story-based, which verbalize the agent's perception of the environment and the behaviors of other agents in the form of story episodes, usually with language templates. The semantics of the environment are given by high-level events (e.g., Sally entered the kitchen). Many aspects of physical and social situatedness are overlooked in these benchmarks, e.g., spatial relations, the task and motivation of agents, and their action trajectories.

Lack of engagement in environment. We point out that existing benchmarks primarily adopt a passive observer role to test language agents. Yet the crucial aspects of interaction and engagement between the agent and other entities involved have

[1] We maintain a repository for relevant literature at https://github.com/Mars-tin/awesome-theory-of-mind.

been overlooked. Among all the benchmarks we reviewed, only three of them treat the tested model as an active agent, one that perceives the physical and social context, reasons about others' mental states, communicates with other agents, and interacts with the environment to complete pre-defined tasks (Sclar et al., 2022; Bara et al., 2021, 2023).

4 Towards A Situated Theory of Mind

4.1 Why A Situated ToM?

There have been concerns that cognitive inquiries are inadequate for gaining insight into understanding ToM for LLMs (Mitchell and Krakauer, 2023; Shapira et al., 2023a). However, we believe that the primary problem lies in using story-based probing as proxies for psychological tests, which situate human subjects in specific physical or social environments and record their responses to various cues. We, therefore, call for a situated evaluation of ToM, in which the tested LLMs are treated like agents who are physically situated in environments and socially situated in interactions with others.

Situated evaluation covers more aspects of ToM. Although it is possible to frame the situations as narratives and cover all mental states using text-only benchmarks, certain aspects of ToM can only be effectively studied within specific physical or social environment (Carruthers, 2015). This is because humans have the ability to infer the mental states of others through various modalities such as visual perception, actions, attention (gazes or gestures), and speech (Stack et al., 2022). For instance, studying perceptual disparities can be challenging with text-only datasets, as they often reduce complex scenarios to rule-based manipulations over negations in the prompts (Sileo and Lernould, 2023). Benchmarks that are not situated also face challenges when it comes to implementing coordination between agents, e.g., aligning intentions towards joint actions (Jain et al., 2019) and pragmatic generation (Zhu et al., 2021a; Bao et al., 2022).

Situated evaluation mitigates data contamination. A situated ToM evaluation can mitigate data contamination, as researchers can design scenarios in simulated settings that are unlikely to be part of the LLM's training data. Carefully designed benchmarks can also incorporate seen and unseen environments to assess generalization to new tasks and new environments, fundamentally addressing the issue of data contamination (Gandhi et al., 2021).

Situated evaluation mitigates shortcuts. By employing situated evaluation, the risk of taking shortcuts can be mitigated. Many of the existing ToM benchmarks are either limited in scale or adopt text templates to verbalize a (few) predefined scenario(s) and prompt LLMs for answers, giving answers away from syntactic structures and positional information (Le et al., 2019; Sclar et al., 2023). In a situated setting, on the contrary, we rely on simulated environments to manipulate evaluation data at scale, so that the environment, the states, and the action traces in the environment can be randomized to avoid the statistical spurious correlations. While situated evaluation can mitigate shortcuts, it does not eliminate the issue completely. For example, Aru et al. (2023) have reported that shortcuts can emerge in grid world setups if the design is not careful enough and randomness is limited. We emphasize that careful design and consideration are still required to curate any ToM benchmark.

4.2 A Preliminary Exploration in Grid World

In this section, we present a proof-of-concept study on a situated evaluation of ToM on LLMs. We choose to conduct our pilot study in MiniGrid (Chevalier-Boisvert et al., 2018), a simple and commonly used environment for ToM studies in the machine learning community (Rabinowitz et al., 2018; Sclar et al., 2022). Through basic grid world representation, we can create tasks to challenge LLMs to reason about many aspects of physical and social situatedness, e.g., spatial relations, partial observability, agent's action trajectories, and from there, their beliefs, intent, emotions, etc. This is in stark contrast to existing story-based ToM benchmarks, which only contain high-level event episodes. We demonstrate that a diverse range of challenging ToM tests, covering all mental states from ATOMS, can be effectively created in a situated manner using a simple 2D grid world.

Environment and Task Setups We introduced 9 different ToM evaluation tasks for each mental state under ATOMS, and 1 reality-checking task to test LLMs' understanding of the world. It is important to acknowledge that our experiment serves as a proof of concept and does not aim to cover the entire spectrum of machine ToM, as our case studies are far from being exhaustive or systematic.

- **Reality Check**: Given the sequence of actions, predict the closest object at the end of the trajectory. The task is designed to test LLMs' under-

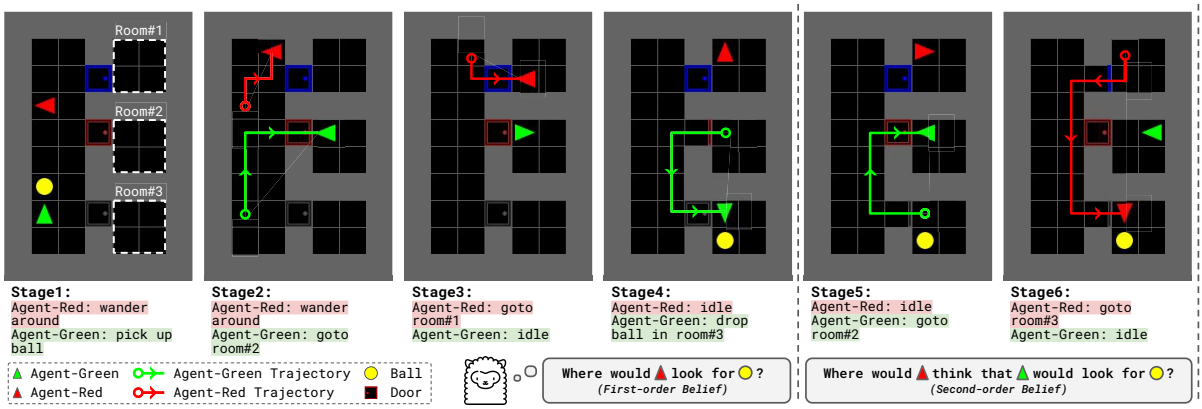


Figure 4: An overview of the first and second order false belief task illustrated in a grid world setup. We simulate the unexpected transfer scenarios with two agents, and verbalize the environment and action traces to test if LLMs hold a correct understanding of the agents' false beliefs.

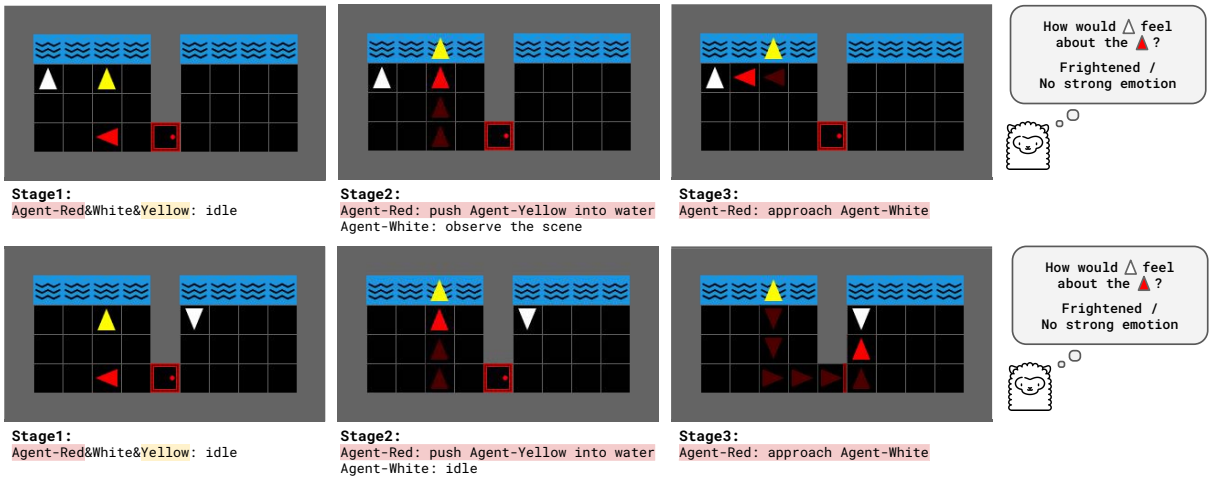


Figure 5: An overview of the morally related emotional reaction tasks illustrated in a grid world setup. We simulate scenarios where an agent either directly witnesses or is ignorant of a morally related event, and verbalize the environment and action traces to test if LLMs hold a correct prediction of the agent's emotional reaction.

standing of relocations in the grid world.

- **Short-term Intention**: Given an incomplete trajectory and a goal, predict the next action.
- **Long-term Intention**: Given an incomplete trajectory and a list of subgoals, predict the next subgoal that the agent is planning to achieve.
- **Desire**: Given a complete trajectory, predict if the agent demonstrates a preference for objects.
- **Percepts**: Given a complete trajectory, predict if the agent has a partial or full observation.
- **Belief**: The classic unexpected transfer task with possible first and second order false belief.
- **Non-literal Communication**: Given a trajectory and a statement from the agent, judge whether the agent is being deceptive.
- **Knowledge**: Given a trajectory, predict the object whose location is unknown to the agent.
- **Emotion**: The classic perception-emotion link test, where emotions are evoked in response to witnessing an emotionally stimulating situation.

We detail two case studies and leave examples of each task in Appendix A.

Case Study 1: Beliefs. Our belief experiments emulate the classic unexpected transfer tasks (Baron-Cohen et al., 1985; Perner and Wimmer, 1985). As is shown in Figure 4, we simulate this disparity of belief state and world state in MiniGrid. The first-order belief task features a main room with three connected side rooms, two agents named `Red` and `Green`, and a ball. Each instance of the belief experiment begins with `Green` placing the ball in `Room#2` while `Red` watches. `Red` then enters a separate `Room#1` and shuts the door. While `Red` is inside of this closed room, `Green` transfers the ball to `Room#3`. `Red` presumably holds a *false belief* about the location of the ball, believing it is in `Room#2` though it is now in `Room#3`. Similarly, we implement the second-order belief task to test an incorrect belief that one agent holds about the belief of another. After `Green` has finished transferring the ball, it navigates to the room originally

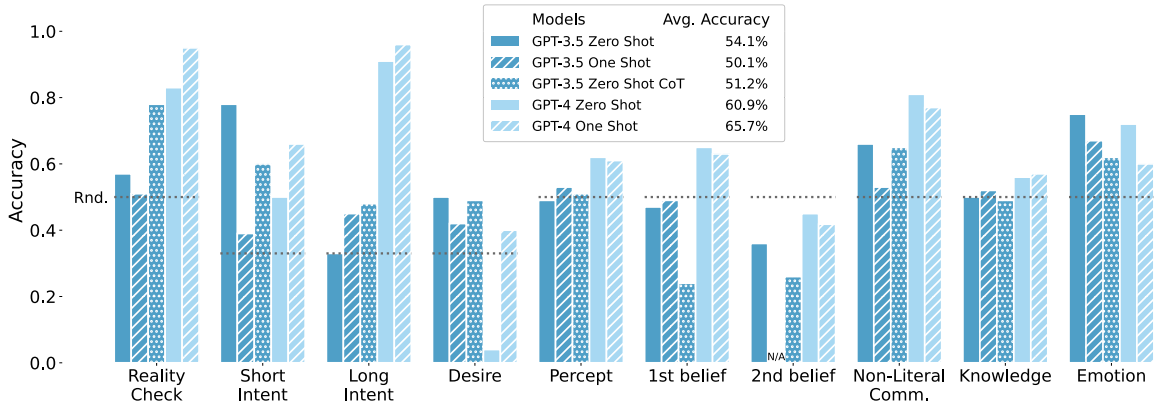


Figure 6: The LLMs' performance across the 10 tasks is illustrated. Each bar shows how one LLM performed with a specific prompting method. Overall, the tasks are tough for all LLMs tested. The effectiveness of one-shot and CoT prompting is not consistent across the board. Some results are N/A as the prompt went out of the context window.

containing the ball and shuts the door. `Red` then navigates to the room now containing the ball and sees the true location of the ball. Still, `Green` presumably possesses a false belief about `Red`'s belief. In both tasks, LLMs are queried with two versions of the world: a false one with the ball in the original room, and a true one with the ball in the third room (its actual location). LLMs must correctly respond that the agents hold a false belief.

Case Study 2: Emotions. While the belief tasks highlight the importance of physical situatedness, we further demonstrate that social interactions can be simulated in the grid world. As is shown in Figure 5, We design morally related events that stimulate emotions (e.g., fear, appreciation). In this task, LLMs are queried to predict the emotional response of `Agent-White`, who either directly witnesses or is ignorant of this event. LLMs must correctly respond that the agent holds an emotional reaction only if it observes the event.

Experiment Setups. For each task, we create 100 instances following a prompt template that consists of `[environment description]`, `[agent description]`, `[observability statement]`, `[task statement]`, `[actions sequences]`, `[QA]`. We select GPT-4 (`gpt-4-0314`) and ChatGPT (`gpt-3.5-turbo-0613`) for evaluation on the 9 tasks.[2] Following prior work (Hu et al., 2022; Shapira et al., 2023a), we adopt MC-probing for LLMs that don't produce probabilities, which directly instructs LLMs to generate only the letter corresponding to the answer. Besides zero-shot evaluation, we also explored one-shot learning and Chain-of-Thought (CoT) prompting (Wei et al., 2022). More details are available in Appendix B.

[2]We use the `ChatCompletion.create` function from `openai` package.

Results and Discussion. We observe that LLMs exhibit some level of sensitivity for some mental states. Especially, GPT-4 scores up to 91% zero-shot accuracy and 96% one-shot accuracy in the long-term intention task. However, we also highlight the shortcomings of LLMs in some mental states of ATOMS to varying degrees, especially, in terms of predicting preferences, perception limitations, missing knowledge, and higher-order beliefs. These findings align with previous research (Sap et al., 2022; Trott et al., 2022; Shapira et al., 2023a), further confirming that LLMs are not yet reliable and comprehensive ToM agents. From the reality-checking task, we observe that GPT-3.5 reaches 78% accuracy with CoT prompting and GPT-4 significantly surpasses its predecessors with 83% zero-shot accuracy and 95% one-shot accuracy. Solving this reality check by no means implies that LLMs have a general perception ability of the real world, but that as a proof of concept, they demonstrate a certain (but still limited) level of situated awareness within the context of a basic abstract grid world. This implies that researchers can begin utilizing them as powerful building blocks for situated agents in complex ToM tasks. We note that it is always possible to come up with more challenging reality-checking questions to expose the limitations of LLMs, or to provide more guided prompts to assist LLMs in successfully completing ToM tasks. Undoubtedly, further research is required along this exciting yet challenging trajectory to advance ToM in LLMs and AI agents built upon LLMs.

5 Discussions and Action Items

5.1 The Scope of Machine Theory of Mind

Be specific about the mental states studied. Existing benchmarks often lack a clear target mental

state, making it challenging to interpret the results and measure the actual progress. To mitigate the risk of overestimating LLMs' ToM capabilities, it is recommended that future benchmark developers provide specific details regarding the targeted mental state(s) they intend to assess.

Broaden the Scope of Machine ToM. A breadth of mental states and their sub-domains have already been covered by AI benchmarks (Table 1). We observed an overwhelming emphasis on the benchmarks and modeling of *beliefs* and *intentions*, while other aspects have received insufficient attention. Still, there are considerably many blank spaces in the landscape of machine ToM, especially for more complicated forms of knowledge, desires, perspective-tasking, and emotional experiences beyond typical social situations.

5.2 Design New Theory of Mind Benchmarks

Avoid shortcuts and spurious correlations. The evaluation of LLMs itself presents significant challenges, not only in the case of ToM. Existing benchmarks suffer from issues such as data leakage and spurious correlations. Especially, shortcut solutions have been consistently reported in recent years (Le et al., 2019; Shapira et al., 2023a; Aru et al., 2023). We are in pressing need of new benchmarks with scalable sizes, high-quality human annotations, and privately held-out sets for evaluation.

Avoid unfair evaluations from prompting. Previous work has shown that CoT prompting can improve the performance of LLMs in ToM tasks (Li et al., 2023; Moghaddam and Honey, 2023; Shapira et al., 2023a). Various recent prompting mechanisms have also been developed to improve LLM's capability on ToM tasks (Zhou et al., 2023a; Leer et al., 2023). In the evaluation of LLMs' ToM capabilities, we recommend the careful documentation of prompts used and the avoidance of implicit human guidance to ensure a fair comparison.

Move on to a situated ToM. We call for a situated evaluation of ToM, in which the tested LLMs are treated like agents who are physically situated in environments and socially situated in interactions with others. A situated setup covers a wider range of ToM aspects. With carefully designed benchmarks with diverse environments and unseen test sets, a situated setup can help address data contamination issues and assess generalization to new tasks and environments. Furthermore, a situated setup allows for more complicated evaluation protocols than simple inference and QA tasks.

Consider a mutual and symmetric ToM. ToM is symmetric and mutual in nature, as it originally imputes the mental states of self and others. Prior research is largely limited to passive observer roles (Grant et al., 2017; Nematzadeh et al., 2018; Le et al., 2019; Rabinowitz et al., 2018) or speaker in a speaker-listener relationship (Zhu et al., 2021b; Zhou et al., 2023b). We encourage more studies on how humans and agents build and maintain common ground with a human ToM and a machine ToM through situated communication (Bara et al., 2021; Sclar et al., 2022). Besides, more research is needed to understand if LLMs possess early forms of intrinsic mental states given observation cues of the world. While we need to develop machines that impute the mental states of humans, humans should also develop a theory of AI's mind (ToAIM) (Chandrasekaran et al., 2017) by understanding the strengths, weaknesses, beliefs, and quirks of these black box language models.

5.3 Neural Language Acquisition and ToM

Both psychological studies (Bloom, 2002; Tomasello, 2005) and computational simulations (Liu et al., 2023) have demonstrated the effectiveness of ToM, especially intention, in language acquisition. Instead of concentrating on eliciting ToM in LLMs, we should contemplate whether certain ToM elements should be inherently present in LLMs or perhaps introduced alongside language pretraining. More research is needed to understand the connection between neural word acquisition and ToM development in machines.

6 Conclusion

In this position paper, we survey and summarize the ongoing debate regarding the presence of a machine ToM within LLMs, and identify the inadequate evaluation protocols as the roadblock. Many benchmarks focus only on a few aspects of ToM, and are prone to shortcuts. To mediate this issue, we follow the ATOMS framework to offer a holistic review of existing benchmarks and identify underexplored aspects of ToM. We further call for a situated evaluation of ToM, one that is physically situated in environments and socially situated in interactions with humans. We hope this work can facilitate future research towards LLMs as ToM agents, and offer an intuitive means for researchers to position their work in the landscape of ToM.

Ethical Statement

The dataset created in this study includes instances that are synthetically generated from planners and RL algorithms, as well as ones created by humans. Human subjects research is approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (IRB-HSBS) under eResearch ID `HUM00234647`. The text generated by LLMs could potentially contain harmful, toxic, or offensive content. The authors have ensured that the data does not contain personally identifiable information or offensive content.

Limitations

Our current benchmark only covers 100 instances for each task, adding up to only 1000 instances. Our experiment serves as a proof of concept and does not aim to cover the entire spectrum of machine ToM, as our case studies are far from being exhaustive or systematic. In the future, we plan to create a more systematic benchmark with a larger scale and various forms of evaluation. Additionally, it is worth noting that the ATOMS framework is derived from human ToM studies conducted with children under the age of 5. Consequently, this framework primarily focuses on the early developmental stages of ToM, capturing the naive and potentially rudimentary aspects of ToM. For more advanced ToM capability, we point to some recent frameworks proposed by Osterhaus and Bosacki (2022) and Stack et al. (2022).

Acknowledgements

This work was supported in part by NSF IIS-1949634, NSF SES-2128623, and by the Automotive Research Center at the University of Michigan. Without implying any agreement with the contents as presented in this work, the authors extend their appreciation to Susan Gelman for her valuable feedback. The authors would like to thank all anonymous reviewers for their valuable feedback.

References

Arjun R Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. 2022. Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *Iscience*, 25(1):103581.

Stefano V Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James N Aronson and Claire Golomb. 1999. Preschoolers' understanding of pretense and presumption of congruity between action and representation. *Developmental Psychology*, 35(6):1414.

Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. Mind the gap: Challenges of deep learning approaches to theory of mind. *Artificial Intelligence Review*, pages 1–16.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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, IJCAI-23*, pages 2958–2966. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Cristian-Paul Bara, CH-Wang Sky, 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.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

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

Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29:407–418.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.

Mark Bennett and Linda Galpert. 1993. Children's understanding of multiple desires. *International Journal of Behavioral Development*, 16(1):15–33.

Sarah-Jayne Blakemore and Jean Decety. 2001. From the perception of action to the understanding of intention. *Nature reviews neuroscience*, 2(8):561–567.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Helene Borke. 1971. Interpersonal perception of young children: Egocentrism or empathy? *Developmental psychology*, 5(2):263.

Michael Bratman. 1987. *Intention, plans, and practical reason*. The University of Chicago Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Peter Carruthers. 2015. Perceiving mental states. *Consciousness and cognition*, 36:498–507.

Fulvia Castelli. 2006. The valley task: Understanding intention from goal-directed motion in typical development and autism. *British journal of developmental psychology*, 24(4):655–668.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. 2018. Minimalistic gridworld environment for gymnasium.

Michael Cohen. 2021. Exploring roberta's theory of mind through textual entailment.

Philip R Cohen and Hector J Levesque. 1990. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261.

Cristina Colonnesi, Carolien Rieffe, Willem Koops, and Paola Perucchini. 2008. Precursors of a theory of mind: A longitudinal study. *British Journal of Developmental Psychology*, 26(4):561–577.

Antonio R Damasio. 2004. Emotions and feelings. In *Feelings and emotions: The Amsterdam symposium*, volume 5, pages 49–57. Cambridge University Press Cambridge.

Susanne A Denham. 1986. Social cognition, prosocial behavior, and emotion in preschoolers: Contextual validation. *Child development*, pages 194–201.

Daniel Dennett. 1995. Do animals have beliefs. *Comparative approaches to cognitive science*, 111.

Daniel C Dennett. 1988. Précis of the intentional stance. *Behavioral and brain sciences*, 11(3):495–505.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Fred I Dretske. 1979. Simple seeing. In *Body, mind, and method*, pages 1–15. Springer.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Jacquelynne S Eccles and Allan Wigfield. 2002. Motivational beliefs, values, and goals. *Annual review of psychology*, 53(1):109–132.

Benjamin Eysenbach, Carl Vondrick, and Antonio Torralba. 2016. Who is mistaken? *arXiv preprint arXiv:1612.01175*.

Nico H Frijda et al. 1986. *The emotions*. Cambridge University Press.

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.

Edmund L Gettier. 2000. Is justified true belief knowledge? *Causal Theories of Mind*, page 135.

Rachel Giora. 2003. *On our mind: Salience, context, and figurative language*. Oxford University Press.

Alison Gopnik and Henry M Wellman. 1992. Why the child's theory of mind really is a theory. *Mind & Language*, 7(1–2):145–171.

Andrew Gordon. 2016. Commonsense interpretation of triangle behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*.

David Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

Julie Hadwin, Simon Baron-Cohen, Patricia Howlin, and Katie Hill. 1997. Does teaching theory of mind have an effect on the ability to develop conversation in children with autism? *Journal of autism and developmental disorders*, 27:519–537.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

Paul L Harris, Kara Donnelly, Gabrielle R Guz, and Rosemary Pitt-Watson. 1986. Children's understanding of the distinction between real and apparent emotion. *Child development*, pages 895–909.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2706–2723.

Yinghui He, Yufan Wu, 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*.

Mark K Ho, Rebecca Saxe, and Fiery Cushman. 2022. Planning with theory of mind. *Trends in Cognitive Sciences*.

Bart Holterman and Kees van Deemter. 2023. Does chatgpt have theory of mind? *arXiv preprint arXiv:2305.14020*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Unnat Jain, Luca Weihs, Eric Kolve, Mohammad Rastegari, Svetlana Lazebnik, Ali Farhadi, Alexander G Schwing, and Aniruddha Kembhavi. 2019. Two body problem: Collaborative visual task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6699.

Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. 2022. Beyond emotion: A multi-modal dataset for human desire understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1512–1522.

David J Kavanagh, Jackie Andrade, and Jon May. 2005. Imaginary relish and exquisite torture: the elaborated intrusion theory of desire. *Psychological review*, 112(2):446.

Casey Kennington. 2022. Understanding intention for machine theory of mind: a position paper. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 450–453. IEEE.

Ariel Knafo, Carolyn Zahn-Waxler, Maayan Davidov, Carol Van Hulle, JoAnn L Robinson, and Soo Hyun Rhee. 2009. Empathy in early childhood: Genetic, environmental, and affective contributions. *Annals of the New York Academy of Sciences*, 1167(1):103–114.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. 2012. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective*, pages 215–240. Springer.

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.

Courtland Leer, Vincent Trost, and Vineeth Voruganti. 2023. Violation of expectation via metacognitive prompting reduces theory of mind prediction error in large language models. *arXiv preprint arXiv:2310.06983*.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. 2022. Pretrained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212.

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*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165.

Bertram F Malle and Joshua Knobe. 2001. The distinction between desire and intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, 45:67.

Gary Marcus and Ernest Davis. 2023. How not to test gpt-3.

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.

Andrew N Meltzoff. 1995. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental psychology*, 31(5):838.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Henrike Moll, Cornelia Koring, Malinda Carpenter, and Michael Tomasello. 2006. Infants determine others' focus of attention by pragmatics and exclusion. *Journal of Cognition and Development*, 7(3):411–430.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Christopher Osterhaus and Sandra L Bosacki. 2022. Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, 64:101021.

Gonçalo Pereira, Rui Prada, and Pedro A Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.

Josef Perner, Susan R Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.

Josef Perner and Heinz Wimmer. 1985. "john thinks that mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.

Ann T Phillips, Henry M Wellman, and Elizabeth S Spelke. 2002. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, 85(1):53–78.

F. Pons and P. Harris. 2000. *Test of Emotion Comprehension: TEC*. University of Oxford.

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

Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Betty M Repacholi and Alison Gopnik. 1997. Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, 33(1):12.

Ted K Ruffman and David R Olson. 1989. Children's ascriptions of knowledge to others. *Developmental Psychology*, 25(4):601.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*.

Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. 2020. Theory of mind and decision science: towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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.

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, pages 19450–19466.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023a. Clever hans or neural theory of mind? stress testing social reasoning in large language models.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023b. How well do large language models perform on faux pas tests. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. 2021. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pages 9614–9625. PMLR.

Damien Sileo and Antoine Lernould. 2023. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*.

Patricia A Smiley. 2001. Intention understanding and partner-sensitive behaviors in young children's peer interactions. *Social Development*, 10(3):330–354.

Caoimhe Harrington Stack, Effat Farhana, Xinyu Shen, Simeng Zhao, and Angela Maliakal. 2022. Framework for a multi-dimensional test of theory of mind for humans and ai systems. In *The Tenth Annual Conference on Advances in Cognitive Systems*.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4902–4918.

Kate Sullivan, Ellen Winner, and Natalie Hopfield. 1995. How children tell a lie from a joke: The role of second-order mental state attributions. *British journal of developmental psychology*, 13(2):191–204.

J Swettenham. 1996. Can children be taught to understand false belief using computers? child psychology & psychiatry & allied disciplines, 37 (2), 157–165.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernández. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Jennifer Tracey, Owen Rambow, Claire Cardie, Adam Dalton, Hoa Trang Dang, Mona Diab, Bonnie Dorr, Louise Guthrie, Magdalena Markowska, Smaranda Muresan, et al. 2022. Best: The belief and sentiment corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2460–2467.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2022. Do large language models know what humans know? *arXiv preprint arXiv:2209.01515*.

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

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. 2023. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*.

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

Henry M Wellman and Jacqueline D Woolley. 1990. From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35(3):245–275.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, and Minlie Huang. 2023. Coke: A cognitive knowledge graph for machine theory of mind. *arXiv preprint arXiv:2305.05390*.

Wako Yoshida, Ray J Dolan, and Karl J Friston. 2008. Game theory of mind. *PLoS computational biology*, 4(12):e1000254.

Mo Yu, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Jing Li, Yue Yu, and Jie Zhou. 2022. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. *arXiv preprint arXiv:2211.04684*.

Jamil Zaki, Niall Bolger, and Kevin Ochsner. 2009. Unpacking the informational bases of empathic accuracy. *Emotion*, 9(4):478.

Chen Zhang and Joyce Y Chai. 2010. Towards conversation entailment: An empirical investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 756–766.

Haode Zhang, Yuwei Zhang, Li Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao Ming Wu, and Albert YS Lam. 2021. Effectiveness of pre-training for few-shot intent classification. In *2021 Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pages 1114–1120. Association for Computational Linguistics (ACL).

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

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023b. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021a. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pages 12901–12911. PMLR.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021b. Few-shot language coordination by modeling theory of mind. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12901–12911. PMLR.

A Task Settings and Data Collection

In this section, we provide an in-depth explanation of the ten tasks outlined in section 4.2. Task 0 serves as a "reality check" to assess LLMs' grasp of the physical world, particularly relocations within the grid world. Tasks 1 through 9 each emphasize distinct facets of ToM. All these tasks utilize MiniGrid, a streamlined 2D grid-world environment. (Chevalier-Boisvert et al., 2018).

Task 0: Reality Check

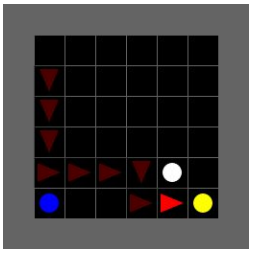


Figure 7: Illustration for *Reality Check* task

In this scenario, the agent is tasked with visiting 0-2 of the objects in the grid world. The agent has full observation in the world for efficient navigation. At least 2 objects (12 maximum) are placed in the environment. To test an LLM's ability to understand the physical actions taken by the agent, we ask it about the distance between the agent and various objects after it accomplishes a number of actions. The action planner is either a shortest path planner towards specified objects or a random action generator.

After the agent has finished 10 random actions, or right after it has visited one object with an optimal action planner, the task-related question will be generated with the following format:

`After having taken these actions, which item is the agent closer to?`

A. `<object1.color> <object1.name>`

B. `<object2.color> <object2.name>`

Here `object1` and `object2` both exist in the environment, and one of them is guaranteed to be the target object if there is such a goal. There are always two options for this task.

Along with the task-related question, the prompt includes a description of the grid world environment, the action space of the agent (only going forward, turn left / right in this task), a board-like depiction of the initial state, the list of actions taken by the agent, and the agent's location and face direction for each step. For more details on prompting, please refer to section B.

The data for Task 0 are autonomously generated using seeds and a shortest-path planner.

Task 1: Short Term Intention

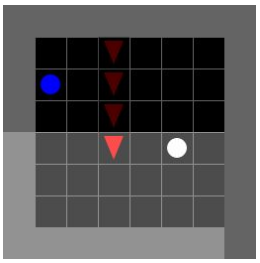


Figure 8: Illustration for *Short Term Intention* task

In this scenario, the agent is tasked with visiting either of the two objects in the grid world. The LLM is not provided with the goal object. Rather, it must determine the goal object by examining and understanding the agent's trajectory. In this task, the agent has full observation in the world, and there are exactly two objects placed. The object types and colors are randomly generated. The size of the room can be randomly sampled from the range 6 by 6 to 12 by 12.

To test an LLM's understanding of short-term intention, we ask it to predict the next action of the agent given its previous trajectory.

The agent's trajectory halts at a random step, with the exception of the precise moment when the optimal paths to the two objects diverge. This "cutting point" is set as an exception and also serves as the mean for the normal distribution from which the stopping point is sampled.

By restricting the cutting point in this manner, we guarantee the trajectory included in the prompt to be optimal for reaching the potential goal objects. This reduces the ambiguities of our experiments, and thus improves the significance of our results.

The task-related question has the following format:

`Which action will the agent take next?`

A. `left`

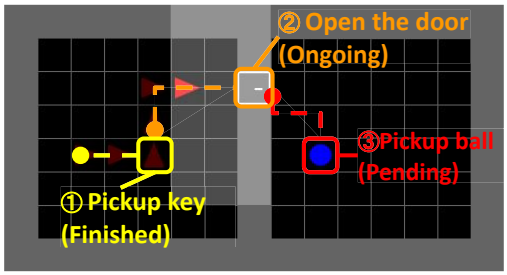
B. `right`

C. `forward`

The LLM should be able to choose which action would be next were the agent to continue its optimal trajectory to the goal object. The data for Task 1 are

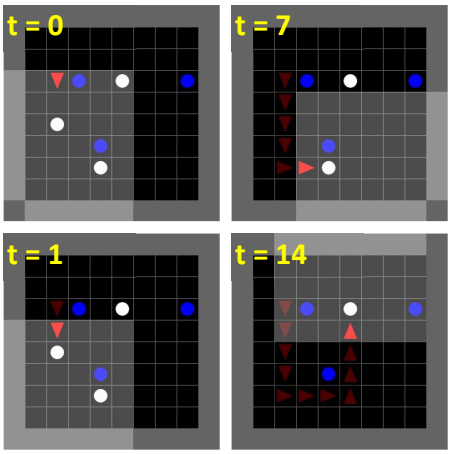
autonomously generated using seeds and a shortest-path planner.

Task 2: Long Term Intention



Task setting: complete three subgoals in as few steps as possible
Q: which subgoal is the agent currently trying to complete?
A. Locate and pick up a key
B. Locate and go through a door
C. Navigate to the object in the new room

Figure 9: Illustration for *Long term intention* task



Task setting: pickup objects in the environment. It may or may not have its own preference over some objects.
Q: Which object (if any) does the agent prefer?
A. Blue ball
B. White ball
C. No preference

Figure 10: Illustration for *Desire* task

In this scenario, the agent needs to complete the following subgoals in as few steps as possible: 1) Locate and pick up a key; 2) Locate and go through a door; 3) Navigate to the object in the new room. There are two rooms in this setting, which are connected by a locked door. The key of the door is always in the room in which the agent is initially located. The object is always placed in the other room.

We provide an LLM with a subset of the agent's trajectory and ask it which subgoal the agent is currently trying to complete.

The task-related question has the following format:

Based on the agent's trajectory thus far, which subgoal is the agent currently trying to complete?

A. Locate and pick up a key
B. Locate and go through a door
C. Navigate to the object in the new room

The data for Task 2 are autonomously generated using seeds and a shortest-path planner.

Task 3: Desire

In this scenario, the agent is required to pick up three objects as soon as possible. There are 2 types of objects in the world (e.g., blue balls and white balls), 3 of each. The agent may or may not have a preference for one object type (we stratify the data such that in 50% of the episodes the agent has a preferred object type and in the other 50% the agent lacks a preference). We also deduct from the final reward given to the agent when it takes a large number of steps to finish picking up three objects.

This task tests whether an LLM is able to determine the desire of the agent (for one object type or the other) by examining its trajectory. We prepared the following question for this task:

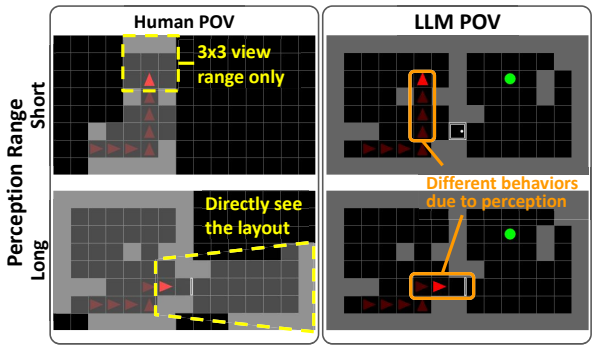
Which object does the agent prefer?

A. white ball
B. blue ball
C. no preference

The data for Task 3 are autonomously generated using seeds and Reinforcement Learning. We use the PPO algorithm (Schulman et al., 2017) to train the model. In the scenario wherein a preference is present, the preferred object type yields 10 times more reward than the non-preferred one. In the scenario wherein the preference is absent, both object types yield identical rewards.

Task 4: Percept

In this scenario, the agent is instructed to navigate in two rooms and reach the goal in the other room as fast as possible. In contrast with previous task settings, the agent has either a very limited visual range (3 x 3 grid in the front), or an "infinitely" large visual range (for practical purposes, the visual range is actually a 101 x 101 grid). Naturally, an agent with a smaller viewing range will



Task setting: navigate to the green ball with given perception range.
Q: Given only the environment layout and agent's trajectory, what is the most likely perceptual range that the agent possesses?
A. 3 blocks
B. Infinitely many blocks
*The perceptual range refers to the number of cells that an agent can see in front of it

Figure 11: Illustration for *Percept* task

make more mistakes (e.g., navigating to a dead end) while trying to reach the goal object. Obstacles are randomly placed in each room to block the agent's view.

LLMs are expected to only look at the trajectory of the agent in an environment, and determine whether the agent has a limited view range or a nearly full view range. The question format is as follows:

Based on the agent's actions, what is the most likely perceptual range that the agent possesses? The perceptual range refers to the number of cells that an agent can see in front of it.

A. 3 blocks

B. infinitely many blocks

We manually collected 100 trials in total for both situations.

Task 5 & 6: First and Second Order Belief

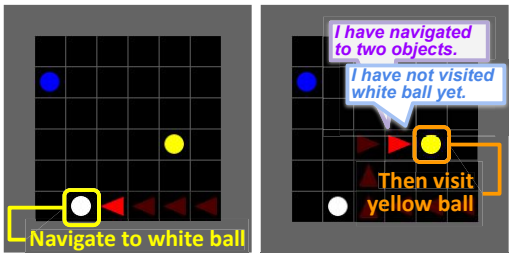
In this scenario, there are two agents in the environment. Both of them are initially in the main room (on the left side of the whole grid world; see Figure 4). On the right side, there are three small rooms. Agents can freely go in and out of each room. They can see everything inside the current room and can see the other rooms through the door if it is open.

This task reproduces the unexpected transfer (Sally-Anne) test (Baron-Cohen et al., 1985; Perner and Wimmer, 1985), with both first-order and second-order belief checking. In the first-order belief task, one agent does not see the second agent transfer a ball from one room to another. Presumably, therefore, the agent falsely believes the ball to be in its previous location. The second-order belief task extends the first-order belief task by enabling the agent with the false belief to see the ball in its new location. The other agent, however, does not witness the first agent rectifying its false belief, so it presumably holds a second-order false belief (about the belief state of the first agent). By varying the observations that each agent makes (e.g., whether or not each agent sees the transfer of the ball from one room to the another) as well as the order of these events, this task setting allows us to check an LLM's first-order and second-order belief capabilities.

Rather than asking the LLM directly where to find the object, we provide two board-like belief states as options. We then query the LLM about which board-like state the agent is more likely to believe.

Data for tasks 5 & 6 are collected via rule-based planners with several scenarios.

Task 7: Non-Literal Communication



Task setting: visiting all the objects, and report its progress with texts.
Q: The agent claims that it at one point navigated to blue ball. Based on the agent's actions, is it telling the truth?
A. Yes
B. No

Figure 12: Illustration for *Non-literal communication* task

This task focuses specifically on one form of non-literal communication: lying. Within this task, the LLM is told explicitly that there is an agent tasked with navigating to all of the objects within the grid world. In each instance of the task, however, the agent only visits a subset of the objects in the environment. The LLM is subsequently told that the agent has claimed success in visiting a particular object. This object is randomly selected so that sometimes it is an object that has actually been visited and other times it is not.

The question format is as follows:

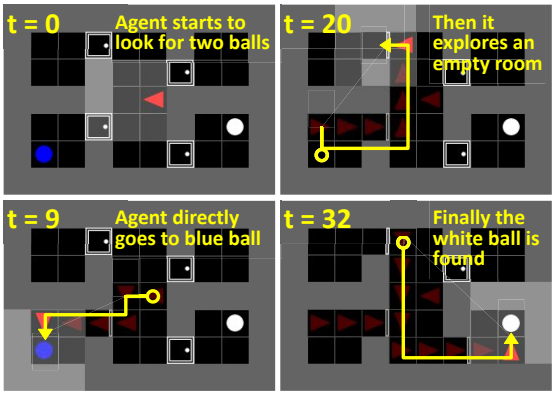
Based on the agent's actions, is it telling the truth?

A. yes

B. no

To successfully complete this task, the LLM must combine its knowledge about the physical occurrences taking place within the grid world with its knowledge about lying, a vital component in Beaudoin et al. (2020)'s category of non-literal communication.

Task 8: Knowledge



Task setting: pick up both two objects in the environment with limited visibility. One of their location is known. The other is unknown. LLM will only see the environment layout and agent's trajectory.
Q: Based on the agent's actions, does it know the position of blue ball before?
A. Yes
B. No

Figure 13: Illustration for *Knowledge* task

This task requires the agent to pick up two objects that exist in the environment. Both two objects are known to be placed separately in two of the four rooms.

In this scenario, the agent is informed of the location of one object, while the location of the second object remains unknown. The agent is instructed to collect the objects in a specific sequence. Ideally, if the agent knows an object's location, it should proceed directly to the appropriate room. Otherwise, it will have to search for the object in the yet-to-be-explored rooms.

We include the agent's entire trajectory in the prompt. We then ask the LLM to determine if the agent knows the position of one object before. The question format is presented as follows:

Based on the agent's actions, does it know the position of <color> <name> before?

A. Yes

B. No

The data for Task 8 are autonomously generated using seeds and a rule-based planner.

Task 9: Emotions

This task (see Figure 5) requires LLMs to infer the emotions of agents in a situated context from their physical behaviors. Specifically, it involves variations on a theme involving a small lake of water. In every variation, one agent pushes another agent into the lake. In some of these variations, an observer is privy to the situation. The prompt then asks the LLM about this observer's feelings towards both the victim and the perpetrator. Presumably, the observer should experience sympathy for the victim and anger (or a similarly negative emotion) for the perpetrator.

How would <observer.name> most likely feel about <pusher.name>?

A. no strong emotion

B. angry

In other variations, a helper comes along and pulls the victim out of the lake. Presumably, the observer should feel positive emotions (e.g. respect, gratitude) for this helper. The question format is as follows:

How would <observer.name> most likely feel about <helper.name>?

A. no strong emotion

B. respectful

B Prompting and Reproducibility

In this pilot study, our data curation follows a uniform structure across all tasks similar to prior work (Li et al., 2022), deviating only slightly to account for task-specific circumstances.[3]

Environment Description Each prompt begins with a description of the two-dimensional world wherein the task will take place. Specifically, our prompting code provides LLMs with the dimensions of the game board and a method to reference specific cells (column-first Cartesian coordinates).

[3]The data for this pilot study is available at https://huggingface.co/datasets/sled-umich/2D-ATOMS.

Each prompt subsequently itemizes the various objects that are situated in the world, along with their coordinates and attributes. Although this prompting structure could be easily adapted to handle many different types of attributes, we focused only on color for the sake of simplicity. Additionally, this section assigns each object a "label": a single letter that the prompt uses to represent the object in a printed grid representation.

Agent Description, Observability, and Task The next section of each prompt is a detailed description of the agent(s) occupying the grid world. In the example below, which was taken from task 1, only one agent occupies the grid world. In multi-agent tasks, this section details the various properties of all agents. Whether single or multi-agent, however, the basic properties are the same. The prompt first details the position and direction of the agent. It is located in a specific cell, and it always faces up, down, left, or right. The prompt then specifies the various actions that an agent is capable of taking (e.g. "forward", "left", and "open"). The agent's labels are then specified. Within the printed grid representation, the agent is always represented as a V-like shape depending on the direction that it is facing. For instance, the prompting tool uses < to represent an agent that is facing left. Finally, the prompt specifies two key attributes of the agent: (1) its level of observability (e.g. whether or not it can see into adjacent rooms) and (2) its goal. Sometimes these two descriptors are heavily modified, restricted, or removed altogether so that they do not interfere with the task. For instance, when testing LLMs for percepts, the prompt does not specify the visual range of the agent.

Action Sequences Following the agent description section, the prompt prints out a board representation, a multi-line sequence of plain text that appears two-dimensional when printed out. Here, the various objects and agents are depicted in position by their associated labels. Additionally, the configuration of the walls is specified by a perimeter of W's. Next, the prompt specifies a sequence of actions that take place over the course of the episode being considered by the LLM. In the episode depicted below, the agent navigates part of the way to a red box. These actions always specify the new position and orientation of the agent.

Questions and Answer Candidates Finally, each prompt contains a question that the LLM must answer. These questions are the most task-specific portion of the prompt, so their contents vary, however, they are always multiple-choice. Additionally, they always contain a set of instructions below heeding the LLM to return only a single letter in its response.

Sample Prompt for Task 1

```
This is a grid-like 2D world
The grid world consists of 6 rows and
6 columns, 0-based
We use (i,j) to represent the i-th
column (from left to right) and j-th
row (from top to bottom).

The following is a list of objects in
this world. Each line starts with the
object's position and is followed by
its attributes
(2, 3): key, grey; represented by this
label: G
(4, 4): box, red; represented by this
label: H

Walls are depicted using the symbol W

There is an agent at (2, 2) facing
left

The agent can take the following
actions:
- left: makes the agent face left of
where it is currently facing
- right: makes the agent face right
of where it is currently facing
- forward: makes the agent move one
step in the direction it is currently
facing
- open: makes the agent open a door
that it is in front of
- pickup: makes the agent pick up
the object that it is in front of
- drop: makes the agent drop an item
that it is holding
- stay: makes the agent stay where it
currently is for a timestep

The agent is represented by the
following labels depending on which
direction it is facing:
- Facing left: <
- Facing up: ^
- Facing right: >
- Facing down: v

The agent has full observability,
meaning it can see the entire world

The agent has been instructed to
navigate to one of the two objects
in the environment, although you do
not know which

This is the starting state of the
board:
```

```
    0 1 2 3 4 5
0 | W W W W W W
1 | W O O O O W
2 | W O < O O W
3 | W O G O O W
4 | W O O O H W
5 | W W W W W W
```

```
This list contains a sequence of
actions taken by the agent
(Step 1) The agent took action left
and is now at (2, 2) facing down
(Step 2) The agent took action left
and is now at (2, 2) facing right
(Step 3) The agent took action forward
and is now at (3, 2) facing right
(Step 4) The agent took action forward
and is now at (4, 2) facing right
(Step 5) The agent took action right
and is now at (4, 2) facing down

Which action will the agent take next?
A: left
B: right
C: forward

Please ONLY respond using the letter
corresponding to your answer
Do not generate any text other
than the letter
```