ToMChallenges: A Principle-Guided Dataset and Diverse Evaluation Tasks for Exploring Theory of Mind

Xiaomeng Ma[1], Lingyu Gao[2], Qihui Xu[3]

[1]The Graduate Center, CUNY
[2]Toyota Technological Institute at Chicago
[3]Basque Center on Cognition, Brain and Language (BCBL)
xma3@gradcenter.cuny.edu, lygao@ttic.edu, q.xu@bcbl.eu

Abstract

Theory of Mind (ToM), the capacity to comprehend the mental states of distinct individuals, is essential for numerous practical applications. With the development of large language models (LLMs), there is a heated debate about whether they are able to perform ToM tasks. Previous studies have used different tasks and prompts to test the ToM on LLMs and the results are inconsistent: some studies asserted that these models are capable of exhibiting ToM, while others suggested the opposite. In this study, we present TOMCHALLENGES, a dataset for comprehensively evaluating the Theory of Mind based on the Sally-Anne and Smarties tests with a diverse set of tasks. In addition, we also propose an auto-grader to streamline the answer evaluation process. We tested three models: davinci, turbo, and gpt-4. Our evaluation results and error analyses show that LLMs have inconsistent behaviors across prompts and tasks. Performing the ToM tasks robustly remains a challenge for the LLMs. In addition, our paper wants to raise awareness in evaluating the ToM in LLMs and we want to invite more discussion on how to design the prompts and tasks for ToM tasks that can better assess the LLMs' ability. [1]

1 Introduction

With the recent advancement of large language models (LLMs; Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020), expectations for artificial intelligence systems to effectively interact with people have significantly increased. This may necessitate the development of human-like capabilities in these systems, such as reasoning not only about their own observations and beliefs but also understanding the mental states of others. This ability, termed as Theory of Mind (ToM), refers to the capacity to attribute mental states—such as beliefs, emotions, and intentions—to oneself and others (Wimmer and Perner, 1983; Gallese and Sinigaglia, 2011). In psychology, it is an essential milestone in the social development of a child. However, the challenges that persist are whether LLMs have already developed ToM capabilities and how to identify the appropriate tool to accurately assess these capabilities.

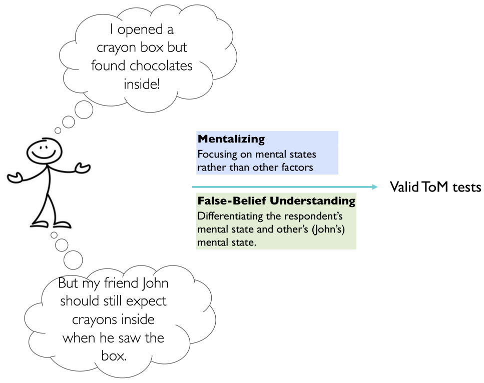


Figure 1: An example of Smarties test, as well as Mentalizing and False-Belief Understanding criteria.

Recent studies addressing those issues often draw inconsistent conclusions, some studies asserting that models exhibit ToM (Kosinski, 2023; Wu et al., 2023; Bubeck et al., 2023), some suggest the opposite (Le et al., 2019; Nematzadeh et al., 2018; Sap et al., 2022; Ullman, 2023a; Shapira et al., 2023), and others maintain caution and questions (Sileo and Lernould, 2023; Aru et al., 2023).

These varied results could be due to different evaluation methods. First, these studies have tested the models on different tasks, ranging from tasks of perspective-taking reasoning (i.e., does the other person know what I know; e.g., Kosinski, 2023) to intention ascription (i.e., what does a movie character intend to do at the end of an open-ended movie; e.g., Shapira et al., 2023). Additionally, the type of prompts varies across studies. For in-

[1]The data and code are available at https://github.com/xiaomeng-ma/ToMChallenges.

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 15–26
December 6–7, 2023. ©2023 Association for Computational Linguistics

stance, Le et al. (2019) and Sap et al. (2022) used question answering prompts, while Kosinski (2023) employed sentence completion prompts. This lack of clear principles in approaches poses challenges to the validity of ToM assessments for LLMs. If only specific prompts lead to high-performance results while others do not, it becomes questionable whether the correct responses truly reflect ToM or are simply the result of algorithmic shortcuts. Similarly, if some tasks are not valid for assessing ToM, the results cannot be interpreted in terms of models' ToM capability regardless of the conclusions drawn.

What is considered a valid ToM test? A valid test should be both theoretically grounded and methodologically validated to ensure it measures the intended subject, and the results are not skewed by other factors. From a theoretical standpoint, ToM theories in child development (Wellman et al., 2001; Quesque and Rossetti, 2020; Navarro, 2022) suggest that valid tests should focus on assessing the respondent's ability to a) represent mental states of one's own and others based on physical events (but not other factors such as emotions and intentions) (mentalizing), and b) differentiate one's own mental state and other's (false-belief understanding). Tasks not meeting these criteria might not be considered valid assessments because they either introduce confounding factors such as emotional or social ascription or fail to contrast the respondent's mental state and other's mental state.

From a methodological perspective, both psychology and NLP studies demand rigorous evaluation to ensure measurement validity. Unlike psychology studies where individual subjects can be randomly assigned to experimental and control conditions to yield reproducible results, LLMs like GPT-4, being a single 'subject', lack the capacity for reproducibility in the traditional sense. Therefore, any claims about an LLM possessing human-like capabilities must be substantiated after validation with a variety of prompts and tasks, provided these tasks align with the theoretical framework of the intended measurement.

Validity issues of current neural ToM tests Testing a few examples on a single format, as done by Kosinski (2023) and Bubeck et al. (2023), raises methodological questions and uncertainty about whether responses are shortcut-driven. In fact, Shapira et al. (2023) recently showed LLMs' inconsistent performance across ToM tasks, further indicating possible shortcuts and the idiosyncrasy of specific prompts. If relied upon singularly, these could lead to misinterpretations.

Meanwhile, several tasks from previous studies (e.g., Ullman, 2023b; Shapira et al., 2023) may not sufficiently adhere to Mentalizing and False-Belief Understanding criteria, casting doubt on whether these tasks genuinely reflect ToM or other capacity such as social ascription. In the study conducted by Ullman (2023b), adversarial variations such as transparent access and uninformative labels were used to evaluate the robustness of LLMs' ToM capability. For example, when the model is presented with a context where a transparent bag is filled with popcorn, but the label on the bag reads "chocolate," the model was likely to suggest that a person seeing the bag for the first time would believe it's full of chocolate, not popcorn, despite the bag's transparency. However, this variation might not be directly related to ToM. Successfully answering those questions may also require conceptual knowledge (e.g., what information can a transparent bag provide) and inferential biases (will the person trust the label or rely on their direct observation through the transparent bag?). Such issues could lead to evaluations straying from the Mentalizing and False-Belief Understanding criteria.

Likewise, certain tasks implemented in the Shapira et al. (2023) study, such as inferring another person's intention, did not distinguish between representations of self and others. Consequently, the model may depend on empathy (see Section 2 for differences between empathy and ToM) rather than ToM to accomplish the task, thereby failing to fulfill the Nonmerging criteria.

Auto-grader: Enabling diverse and large-scale evaluations One potential challenge to establishing a principle-guided yet diverse evaluation system is the intense human labor involved in evaluating models' responses. It may not be a significant issue when the task is in a constrained format such as true or false questions. However, when the diversity and the amount of tasks increase, which is necessary for a valid ToM test (e.g., ask models to provide reasoning so that one can better understand how the model reaches such a conclusion), a more efficient evaluation method becomes essential.

Present study To improve the validity of ToM tests, we propose a principle-guided dataset with a diverse set of tasks. In an effort to dissect the underpinnings of incorrect responses, we also conducted error analyses, particularly focusing on questions demanding reasoning. This approach offers a deeper insight into the cognitive process of the models when they arrive at incorrect conclusions. Finally, addressing the need for efficient evaluations, we have developed an autograder based on GPT-4 to streamline the evaluation process. This tool allows us to efficiently evaluate models' responses across a broader spectrum of tasks and on a larger scale, bringing a higher degree of accuracy and efficiency to the ToM testing process.

Our evaluations and error analyses show that current LLMs struggle to perform robustly on ToM tasks or reason in a manner characteristic of subjects possessing ToM. Moreover, we demonstrate that our auto-grader is highly proficient at automatically evaluating LLMs' responses across various tasks, paving the way for more efficient, larger-scale analyses for neural ToM.

2 Related Work

ToM in humans ToM in children significantly influences various facets of their development, including social competence, peer acceptance, and academic achievement (Carlson et al., 2013). Research has revealed substantial changes in children's understanding of mental states by the age of five (Wellman et al., 2001). Although ToM is often linked to cognitive abilities like empathy and visual-spatial attention, it's crucial to note that these are separate constructs involving distinct neurological and cognitive processes (Kanske et al., 2015; Schurz et al., 2021; Zaki and Ochsner, 2012). These abilities also yield largely divergent effects on other aspects of social and cognitive development (Happé et al., 2017). Take for instance an individual with ToM but not empathy. They have the intellectual ability to interpret and understand the thoughts, intentions, and beliefs of others. Nevertheless, when tasked with sharing or connecting with others' emotions, they may encounter difficulty.

ToM tasks Quesque and Rossetti (2020) reviewed tasks frequently employed to assess ToM. Among these, the *False Belief* task, one of the most widely utilized tasks in human and language model studies, fulfills the criteria. This task requires participants to infer the belief of a character who holds a false belief about a particular scenario, which contrasts with the participants' updated belief of the same scenario. The Smarties and the Sally-Ann tests are the two most frequently employed *False Belief* tasks. For instance, in the Smarties Test, a child is shown a box labeled as 'candies'. After revealing that the box indeed contains crayons rather than candies, the child is asked what another person, unaware of the box's contents, would guess is inside. Younger children often answer 'crayons', while older children, understanding others would base their belief on the box's label, answer 'candies' (Gopnik and Astington, 1988).

On the other hand, several tasks either do not demand the distinction between one's own mental state and that of others or they actually measure processes not directly related to ToM. The tasks in Shapira et al. (2023) - *Intention Ascription* (included in the SOCIAL IQA dataset; Sap et al., 2019) and *Animated Shapes* - fall under this category. These tasks often foster shared representations between self and others, rather than creating a distinction (Brass et al., 2009). For example, in the *Animated Shapes* task, participants watch short animated films featuring geometrical shapes, and they are then asked to interpret the thoughts or feelings of these shapes. However, this task probes more into empathy rather than ToM.

Evaluations of ToM in LLMs ToM evaluations in LLMs vary greatly in terms of tasks and prompts. Nematzadeh et al. (2018) was the first work for evaluating ToM in LLMs, finding all models unsuccessful. In 2019, Le et al. (2019) found that the question-answer benchmarks of the time were prone to data biases, allowing models to develop corner-cutting heuristics due to a rigid event sequence template for each task type. To mitigate this, they introduced new evaluation methods along with a novel dataset. Sap et al. (2022) later evaluated GPT-3 (Brown et al., 2020) on this dataset, reporting only 55 - 60% accuracy, even after few-shot fine-tuning with GPT-3-Davinci.

Recent two studies tested GPT-4 on a few *False Belief* examples using sentence completion Kosinski (2023) and question-answer prompts Bubeck et al. (2023). Both studies reported GPT-4 achieving $\geq$ 90% accuracy, leading to suggestions of spontaneous ToM emergence in LLMs. However, this claim was disputed by subsequent research (Ullman, 2023a; Shapira et al., 2023). As noted in Section 1, Ullman (2023a) introduced adversarial variations to the false belief questions used in

Kosinski (2023), which resulted in a significant decrease in LLMs' performance. Shapira et al. (2023) evaluated LLMs across a range of tasks ToM, finding that current LLMs, including GPT-4, struggled to perform consistently. The tasks included the *False Belief* task from Kosinski (2023), the *False Belief* task with adversarial variations (Ullman, 2023a), the *Animated Shapes* task adapted from Heider and Simmel (1944), and a set of common sense reasoning tasks including the *Intention Ascription* task (Sap et al., 2019). Their findings indicated that current LLMs struggle to consistently perform well on these tasks. The high performance of GPT-4 observed in the initial studies (Kosinski, 2023; Bubeck et al., 2023) may reflect shallow heuristics, not robust ToM capabilities.

3 ToMChallenges and Tasks

We aim to build a corpus based on two types of tests: *Sally–Anne Test* and *Smarties Test*, which fit the ToM test criteria. Below we describe how we construct ToMChallenges data, and how we design our evaluation tasks.

3.1 Dataset Construction

While Le et al. (2019) proposed the inclusion of distractors to prevent models from adopting corner-cutting heuristics, it is important to note that distractors are more relevant for fine-tuning rather than zero-shot probing. Given the ongoing discussions surrounding the zero-shot performance of models in recent studies (Kosinski, 2023; Ullman, 2023b) and we care more about the model's inherent capabilities, we introduce a dataset without distractors as below to maintain our focus, with examples displayed in Tables 1 and 2. We created 30 variations of each test (e.g., changing the person's name, location, and items), and the details of the tests and variables are described as follows.

Sally-Anne Test The Sally-Anne Test was first introduced by Baron-Cohen et al. (1985) and has been widely used in psychology studies. The test typically involves two characters, Sally and Anne, where Anne hides an object while Sally's away. The children were usually asked where would Sally look for the object when she returns. The narrative consists of the following components: (1) a location L, where the event takes place, (2) two agents, A and B, where A moved the object while B one is away (3) an object O, whose position changed in the narrative, and (4) two containers, C1 and C2, representing the object's initial and updated positions, respectively. Using these components, we construct narratives as shown in Table 1.[2]

For each narrative, we create 6 questions following Le et al. (2019) to comprehensively evaluate the model's understanding of the narrative and the

Variables	L: attic, A: Neila, B: Juanita, O: towel, C1: closet, C2: cabinet
Narrative $\mathcal{N}$	*Neila* and *Juanita* were hanging out in the *attic*. They saw a *closet* and a *cabinet*. They found a *towel* in the *closet*. *Juanita* left the *attic*. *Neila* moved the *towel* to the *cabinet*.
REALITY	Where is the *towel* currently? Answer: The cabinet.
BELIEF	Where was the *towel* previously? Answer: The closet.
After *Juanita* came back to the *attic*, †	
1STA	where would *Neila* look for the *towel*? Answer: The closet.
1STB	where would *Juanita* look for the *towel*? Answer: The cabinet.
2NDA	where would *Neila* think *Juanita* would look for the *towel*? Answer: The cabinet.
2NDB	where would *Juanita* think *Neila* would look for the *towel*? Answer: The cabinet.

The initial prompt with † is applied to 1STA, 1STB, 2NDA, and 2NDB.

Table 1: An example for Sally-Anne Test.

Variables	L: attic, A: Neila, B: Juanita, C: bag, O1: plate, O2: vest
Narrative $\mathcal{N}$	*Neila* found a *bag* in the *attic*. The label on the *bag* says *plate*. *Neila* couldn't see what was inside the *bag*. *Neila* opened the *bag* and found a *vest*. There is no *plate* in the *bag*. *Neila* closed the *bag* and put it back. *Juanita* entered the *attic* and saw the *bag*.
REALITY	What was in the *bag*? Answer: A vest.
BELIEF	What was supposed to be in the *bag*? Answer: A plate.
After *Juanita* opened the *bag*, †	
1STA	what would *Neila* expect to find in the *bag*? Answer: A vest.
1STB	what would *Juanita* expect to find in the *bag*? Answer: A plate.
2NDA	what would *Neila* think *Juanita* would expect to find in the *bag*? Answer: A plate.
2NDB	what would *Juanita* think *Neila* would expect to find in the *bag*? Answer: A plate.

The initial prompt with † is applied to 1STA, 1STB, 2NDA, and 2NDB.

Table 2: An example for Smarties Test.

[2]The agents' names were selected from CMU Name Corpus. All the names are female names. We manually crafted L, O, C1, and C2.

Narrative $\mathcal{N}$	*Neila* and *Juanita* were hanging out in the *attic*. They saw a *closet* and a *cabinet*. They found a *towel* in the *closet*. *Juanita* left the *attic*. *Neila* moved the *towel* to the *cabinet*.
Fill-in-the-Blank	Fill in the blank (<>): $\mathcal{N}$ After *Juanita* came back to the *attic*, *Neila* would think *Juanita* would look for the *towel* in the < >. Answer:
Multiple Choice	Choose the correct answer from A or B for the following question: Question: $\mathcal{N}$ After *Juanita* came back to the *attic*, where would *Neila* think *Juanita* would look for the *towel*? A. *cabinet* B. *closet*
True/False	Given the context, judge True or False of the given statements A and B respectively: $\mathcal{N}$ Statements: A. Juanita would look for the towel in the cabinet. B. Juanita would look for the towel in the closet.
CoT True/False	Given the context, reason through statements A and B step by step and provide a True or False judgment based on the reasoning: $\mathcal{N}$ Statements: A. Juanita would look for the towel in the cabinet. B. Juanita would look for the towel in the closet.
Q&A	Answer the question based on the context: Context: $\mathcal{N}$ Questions: After Juanita came back to the attic, where would Neila think Juanita would look for the towel? Answer:
Text Completion	Complete the following paragraph: $\mathcal{N}$ After *Juanita* came back to the *attic*, *Neila* would think *Juanita* would look for the *towel* in

Table 3: An illustrative example for different task templates of the Sally-Anne Test using 2NDA question as an example, ignoring line breaks in templates for space saving.

agents' mental states: REALITY focuses on the updated/current position of O, and BELIEF focuses on the initial/previous position. The first-order belief (1STA and 1STB) questions ask the agents' beliefs, and the second-order belief (2NDA and 2NDB) questions ask one agent's belief regarding the other agent's mental state.

Smarties Test The Smarties Test was first introduced by Gopnik and Astington (1988) and has also been widely adopted in psychology studies. In a typical Smarties test, the child is presented with a 'Smarties' box that actually contains something else. The child is then asked what they think another person, who has not seen the contents of the box, would believe is inside. The narrative consists of the following components: (1) two agents, A and B, where A saw the contents and B didn't, (2) one container C that holds the object, and (3) two objects, O1 and O2, where O1 is the labeled content and O2 is the actual content. Using these components, we construct narratives for the Smarties Test as shown in Table 2.

The questions of the Smarties Test narrative are similar in nature to those of the Sally-Anne Test, but the REALITY question focuses on the actual object in the container, and the BELIEF question focuses on the container's label.

3.2 Task Formulation

Previous studies have used a single task (e.g. question-answering task or sentence completion) task to evaluate the model's performance. In order to test the robustness of the model's performance, it is necessary to adapt the questions into a variety of tasks. We construct different prompts to create 6 task formats, as demonstrated in Table 3. These tasks can be categorized into three groups based on the level of freedom in generation:

Fully-Constrained Fully-constrained generation limits the model's output to specific predefined structures or responses. In this group, we design 3 tasks, i.e., Fill-in-the-Blank, Multiple Choice, and True or False questions.

Semi-Constrained Semi-constrained generation involves partial guidance by specific rules or structures, while still allowing some flexibility in the model's responses. This group consists of 2 tasks, i.e., Chain-of-Thought (CoT) True or False questions and Question Answering (Q&A) tasks.

Open-Ended Open-ended generation enables the model to generate responses without being restricted by predefined rules or structures, leading to more diverse and varied outputs. An example of this group is Text Completion.

3.3 Experimental Setup

We evaluate the zero-shot performance of three models: text-davinci-003 and gpt-3.5-turbo-0301, and gpt-4-0613 (OpenAI, 2022). For the hyperparameters of all models, we set the temperature as 0, top_p as 1, and both frequency penalty and presence penalty as 0. Due to the different natures of our task design, we choose different maximum token limits for each task as follows: 10 tokens for Fill-in-the-Blank, 2 tokens for Multiple Choice, 20 tokens for True or False, 100 tokens for CoT True or False, and 50 tokens for both Question Answering and Text Completion.

3.4 Answer Evaluation and Auto-grader

For the fully-constrained tasks, the models' answers can be graded easily since there are standard answers. We first apply a python function to grade these answers, and the results are double checked by human annotators. For the semi-constrained and open-ended tasks, the answers don't necessarily follow a standard form and are graded by human annotators. The rubrics to grade these answers include: 1) the answer is correct; 2) the answer doesn't contain any information that can not be inferred from the narrative.

In order to improve the efficiency of grading, we develop an auto-grader based on the gpt-4-0613 model with a grading prompt. The grading prompt consists of a general template of the narrative and guidelines of how to construct gold answers for the 6 questions. The model then grades the generated answers based on the gold answers. In addition, an example of a generated answer and grading pair was also included in the prompt for in-context learning. An example of the grading prompt is included in Appendix A. The output of the auto-grader consists of two parts: the reasoning part, where it outputs the gold answers to 6 questions; and the grade part, where it grades the generated answer. An example of the auto-grader's output is shown in Table 4.

We apply the auto-grader to evaluate the answers in two tasks: Q&A and Text Completion. First, we evaluate the gold answers output by the auto-grader. The auto-grader achieved 100% accuracy on all Sally-Anne and Smarties narratives, showing it can effectively produce gold answers for the 6 questions. Then we evaluated the grading results by comparing them to the human annotated results. The auto-grader achieved 100% accuracy on Q&A task and over 90% accuracy on Text Completion task. These results demonstrated that the auto-grader could be an effective tool in evaluating more freely generated answers.

4 Results and Analyses

In this section, we present the results of our evaluation for all models on Sally-Anne and Smarties tests. As we create 30 variations of the narrative for each test, and each narrative comes with 6 questions (REALITY, BELIEF, 1STA, 1STB, 2NDA, 2NDB), and each question is tested on 6 tasks, an idealized model that is capable to solve Theory of Mind tasks should be able to achieve high accuracy on all questions across different tasks and in most of the narratives.

4.1 Accuracy by Question and Task

The accuracy of each question type is calculated by averaging the accuracy over 30 narratives (e.g., an accuracy of 50% for 1STA question means that the model answered correctly for 15 out of the 30 narratives). Figure 2 and 3 show the average accuracy of 6 types of questions in different prompts for Sally-Anne Test and Smarties Test.

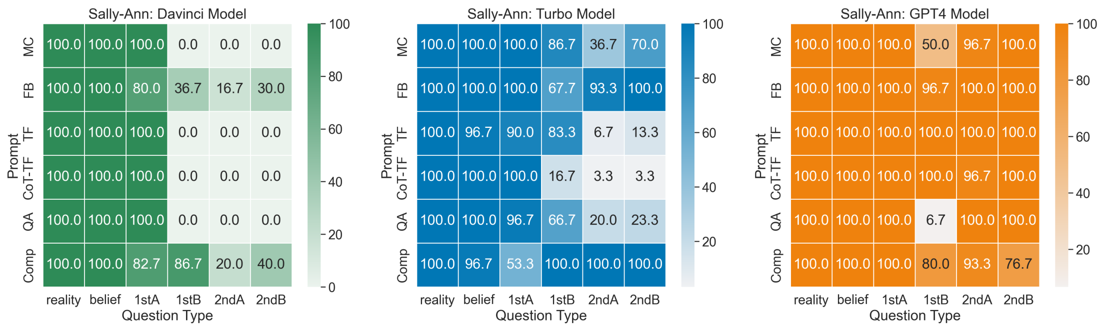
For the Sally-Anne tests, all three models are able to achieve near-perfect accuracy on REALITY, BELIEF, and 1STA questions for all prompts, indicating that the models can reason based on facts. For 1STB question that requires reasoning both the belief of A and B, the gpt-4 model achieved better performance than the gpt-3.5 models (davinci and turbo). For 2NDA and 2NDB questions, gpt-3.5 models struggled to understand one person's belief about another person's belief, while gpt-4 answered most of the narratives correctly. For different tasks, the models behaved differently. All three models achieved the best overall performance with the Text Completion task, followed by the Fill-in-the-Blank task. In addition, introducing Chain-of-Thought did not improve the model's performance on True/False task.

The Smarties test showed a different accuracy pattern from the Sally-Anne test. All three models showed some difficulties answering the fact-based questions, REALITY, BELIEF, and 1STA questions correctly. For the 1STB questions, all three models had worse performance than the 1STB questions in the Sally-Anne test. For 2NDA and 2NDB questions, gpt-4 model and turbo model had similar performance as in the Sally-Anne test, while the davinci model achieved better performance than the Sally-Anne test. We observe that the davinci model performed best on the Text Completion task, the turbo model performed best on the Multiple-Choice task, and gpt-4 model performed best on the Fill-in-the-Blank task.

By comparing the different tests, questions, and tasks, it is clear that the models can not reliably perform ToM tasks well. The models are generally good at fact-based questions and struggle more

Reasoning:
This is a Sally-Anne story. Person A is Sabra and Person B is Hillary. The item is a mitten. Location X is the briefcase and Location Y is the bucket.
Gold Answers:
1. reality: The mitten is currently in the bucket because Sabra moved it there.
2. belief: The mitten was in the briefcase before.
3. 1st order belief of Sabra: Sabra would look for the mitten in the bucket because she moved it there.
4. 1st order belief of Hillary: Hillary would look for the mitten in the briefcase because she didn't know that Sabra moved it.
5. 2nd order belief of Sabra: Sabra would think that Hillary looks for the mitten in the briefcase because she knew that Hillary didn't see her move the mitten to the bucket.
6. 2nd order belief of Hillary: Hillary would think that Sabra looks for the mitten in the briefcase because she didn't know that Sabra already moved it.
The response to the prompts suggests that Hillary would look for the mitten in the bucket. The answer is incorrect because it is contradicting gold answer 4.
Grade: Incorrect.

Table 4: An example of Auto-grader output on answer 1STB question.



MC = Multiple Choice, FB = Fill-in-the-Blank, TF = True/False, CoT-TF = Chain-of-Thought True/False, QA = Question Answering, Comp = Text Completion

Figure 2: The average accuracy for questions in Sally-Anne test for different prompts.

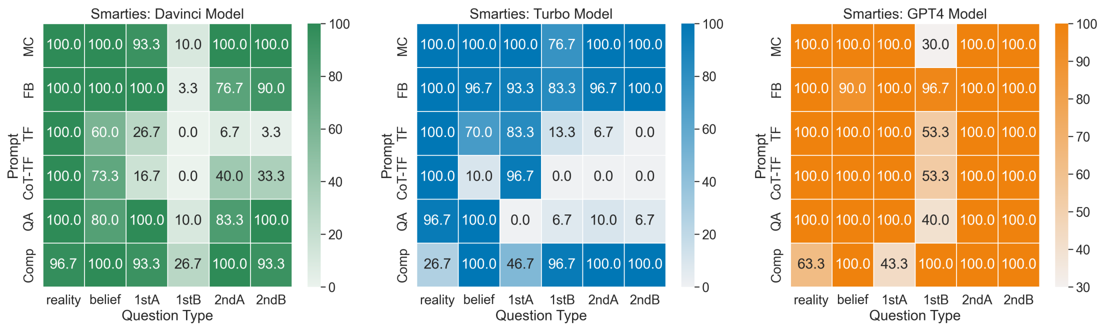


Figure 3: The average accuracy of questions in Smarties test for different prompts.

	Sally-Anne	Smarties
Gold Answers	100	100
Q&A	100	100
Text Completion	95.7	91.4

Table 5: The accuracy of auto-grader on Gold Answer, Q&A and Text Completion.

with questions that require reasoning through different agents' beliefs. The models are also sensitive to the prompts and framing the question into different tasks could significantly affect the model's performance.

4.2 Accuracy by Narratives

The accuracy of each narrative is calculated as the average accuracy over six question types. Although the narratives are generated through the same template, the models produced different answers. For example, for some narratives, the model is able to answer all the questions correctly, while for others the model's accuracy drops. Table 6 and Table 7 show the average accuracy of Sally-Anne and Smarties tests across narratives. For both tests, the gpt-4 model has the best and most stable performance, which has the highest average accuracy and lowest standard deviation.

Sally-Anne	davinci	turbo	gpt-4
MC	0.50±0	0.82±0.17	0.91±0.10
FB	0.61±0.13	0.93±0.09	0.99±0.03
TF	0.5±0	0.65±0.10	1±0
CoT-TF	0.5±0	0.57±0.12	0.99±0.03
QA	0.5±0	0.68±0.17	0.84±0.04
Comp	0.72±0.15	0.92±0.10	0.92±0.12

Table 6: The average accuracy and standard deviation for narratives in the Sally-Anne test for different prompts.

Smarties	davinci	turbo	gpt-4
MC	0.84±0.03	0.95±0.07	0.88±0.08
FB	0.78±0.12	0.96±0.10	0.88±0.10
TF	0.33±0.11	0.46±0.12	0.92±0.08
CoT-TF	0.44±0.15	0.34±0.06	0.92±0.08
QA	0.79±0.12	0.37±0.10	0.90±0.08
Comp	0.85±0.09	0.78±0.13	0.84±0.13

Table 7: The average accuracy for stories in the Smarties test for different prompts.

4.3 Error Analysis

We further looked into the errors the models made, especially for the questions that the models had low accuracy. We focused our error analysis on the Q&A and Text Completion tasks, since the output of these two tasks contains more information to analyze. The errors can be divided into three major types:[3] a) True Failure of ToM, b) Overly conservative, c) Hallucination. The summary of the error counts of each type of error in Q&A and Text Completion tasks is shown in Table 8.

The errors of True Failure are similar to the errors the younger children would make, where the model assumed that an agent knew something they shouldn't know. An example of the wrong answer is '*Hillary would most likely look in the bucket where Sabra moved the mitten*.' This type of error is more common in the davinci and turbo models, and more frequently occurs in Sally-Anne's narrative than the Smarties narrative.

Overly conservation errors happen when the model is being too conservative and refuses to make inferences about the agent's belief. This type of error is common in the turbo and the gpt-4 models, where the model produces answers like '*The context does not provide information on where Juanita would look for the towel when she returns.*'. In addition, this error is more common in the Smarties narrative than in the Sally-Anne narrative.

[3]There are also miscellaneous answers, such as *'Neila would expect to find a surprise inside'*. These answers are not considered in error analysis.

	True Failure		Conservative		Hallucination	
	SA	Sm	SA	Sm	SA	Sm
davinci	136	58	0	6	4	1
turbo	66	0	3	114	14	38
gpt-4	15	18	28	17	0	11

SA = Sally-Anne, Sm = Smarties

Table 8: The total error counts of 6 questions in Q&A and Text Completion tasks for 3 models.

Hallucination error is identified when the answer includes information that can not be inferred from the narrative, or the answer contains contradicting information than the narrative. An example error would be: '*In the backpack, there was a note that said, "This backpack belongs to Norina"*.', where '*note*' was not mentioned in the narrative at all. This type of error is more frequently found in the turbo model.

The error analyses showed that the models failed on the ToM tasks not only because they could not reason about reality and people's beliefs, but also because of the inherent limitation of LLMs. For example, the hallucination errors and the overly conservative errors are related to the inference process of the LLMs, which has always been a challenging part of the NLP field.

5 Conclusions

In this study, we proposed ToMCHALLENGES to comprehensively test the ToM on LLMs. The dataset is constructed based on the Sally-Anne and Smarties tests. For each test, we created a template to generate variations of the test. In addition, we incorporated 6 types of questions to examine the model's understanding of reality, belief, 1st order belief, and 2nd order belief. We also included 6 tasks with different prompts for evaluation, considering the impact of prompts on model performance. This evaluation method serves a dual purpose: it not only measures whether the model has ToM capacity, but also measures the robustness of the model in performing the ToM tasks. In addition, we also create an effective auto-grader that achieved high accuracy in evaluating the more free-formed answers of the ToM tasks.

Using 30 variations of Sally-Anne and Smarties tests, we found that the GPT-based models can not reliably perform the ToM tasks, since the accuracy varies across different tasks, questions, and narratives. Through error analysis, we found that the failure of the models can not be simply concluded as they lack the ability to reason reality and beliefs. Instead, many of the errors reflect the inherent limitations of the LLMs, which might not be related to the cognitive explanations of the ToM.

We hope our study could invite more discussions on evaluating the ToM in LLMs. Further studies could investigate how and why different prompt types would affect the model's performance, how to improve the robustness of the model's performance, how to interpret the results more appropriately, etc. In addition, one limitation of our study is that we did not conduct human experiments. It is unclear if human adults are able to achieve (almost) perfect performance across different stories, prompts and tasks. Future studies could also compare human results and the LLMs' behaviors side by side for evaluation. We hope that future discussions could not only focus on whether LLMs exhibit the ToM, but also attend to the significance and implications of the ToM in LLMs.

References

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.

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

Marcel Brass, Perrine Ruby, and Stephanie Spengler. 2009. Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528):2359–2367.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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*.

Stephanie M Carlson, Melissa A Koenig, and Madeline B Harms. 2013. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):391–402.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vittorio Gallese and Corrado Sinigaglia. 2011. What is so special about embodied simulation? *Trends in cognitive sciences*, 15(11):512–519.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pages 26–37.

Francesca Happé, Jennifer L Cook, and Geoffrey Bird. 2017. The structure of social cognition: In (ter) dependence of sociocognitive processes. *Annual review of psychology*, 68:243–267.

Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.

Philipp Kanske, Anne Böckler, Fynn-Mathis Trautwein, and Tania Singer. 2015. Dissecting the social brain: Introducing the empatom to reveal distinct neural networks and brain–behavior relations for empathy and theory of mind. *NeuroImage*, 122:6–19.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.

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.

Ester Navarro. 2022. What is theory of mind? a psychometric study of theory of mind and intelligence. *Cognitive Psychology*, 136:101495.

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. Openai api. https://api.openai.com/.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin trivial. 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.

Matthias Schurz, Joaquim Radua, Matthias G Tholen, Lara Maliske, Daniel S Margulies, Rogier B Mars, Jerome Sallet, and Philipp Kanske. 2021. Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, 147(3):293.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

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*.

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

Tomer D. Ullman. 2023b. Large language models fail on trivial alterations to theory-of-mind tasks. *CoRR*, abs/2302.08399.

Henry M Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684.

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.

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*.

Jamil Zaki and Kevin N Ochsner. 2012. The neuroscience of empathy: progress, pitfalls and promise. *Nature neuroscience*, 15(5):675–680.

A Appendix

You are a helpful grader. You will be given a series of Prompt and Response pairs about Theory of Mind. The core of grading rubric is to see if the answer can correctly reason the reality and the knowledge of the people in the story. You first need to reason about the context in the Prompt and figure out the key facts.
There are two types of story in the context.
The Sally-Ann story is about two people (Person A and Person B) found an item in location X. Person B left the room and Person A moved the item to location Y.
Key facts for Sally-Ann story include:
1. reality: the current location of the item, which is location Y.
2. belief: the previous location of the item, which is location X.
3. 1st order belief of Person B: what Person B thinks the current location of the item, which is location X.
4. 1st order belief of Person A: what Person A thinks the current location of the item, whic is location Y.
5. 2nd order belief of Person B: what Person B thinks that Person A thinks the current location of the item, which is location X.
6. 2nd order belief of Person A: what Person A thinks that Person B thinks the current location of the item, which is location X.
The Smarties story is about Person A found a container. The label on the container says item X, but there's no item X in the container. Person A opened the container and found item Y. Later, Person B entered the room and saw the container.
Key facts for Smarties story include:
1. reality: what's inside the container, which is item Y.
2. anti-reality: what supposed to be in the container according to the label, which is item X.
3. 1st order belief of Person B: what Person B expects to find in the container, which is item X .
4. 1st order belief of Person A: what Person B expects to find in the container, which is item Y (because Person A already opened the bag and found Y).
5. 2nd order belief of Person B: what Person B thinks that Person A expects to find in the container, which is item X.
6. 2nd order belief of Person A: what Person A thinks that Person B thinks is in the container, which is item X.
After identifying the key facts of the context, you can take a look at the response to see if that is align with the key facts. If the response is align with the key facts and does not contain any information that cannot be inferred from the context, it can be graded as 'Correct'. Otherwise, it is 'Incorrect'.

Here is an example:
user:
Propmt: Complete the following paragraph: Neila found a bag in the attic. The label on the bag says plate. Neila couldn't see what was inside the bag. Neila opened the bag and found a vest. There is no plate in the bag. Neila closed the bag and put it back. Juanita entered the attic and saw the bag. In the bag, there was
Response: a vest.
assistant:
Reasoning:
This is a Smarties story. Person A is Neila and Person B is Juanita. The container is a bag. Item X is plate and item Y is vest.
Key facts:
1. reality: A vest is in the bag because Neila found a vest in the bag.
2. beleif: There supposed to be a plate in the bag because the label says plate.
3. 1st order belief of Juanita: Juanita expects fo find a plate in the bag because the label says plate and doesn't know there's actually a vest in the bag.
4. 1st order belief of Neila: Neila expects to find a vest in the bag because she already opened the bag and found a vest.
5. 2nd order belief of Juanita: Juanita thinks that Neila expects to find a plate in the bag, because she doesn't know that Neila already opened the bag and found a vest.
6. 2nd order belief of Neila: Neila thinks that Juanita expects to find a plate in the bag, because she knows that Juanita doesn't know there's actually a vest in the bag.
The response to the prompt suggests that there was a vest in the bag, which is align with reality. Therefore the response is correct.
Grade: Correct.

Appendix: Example Prompt for Auto-grader.