HI-TOM : A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models

Yinghui He[*1] **Yufan Wu**[*1] **Yulong Chen**[2] **Naihao Deng**[1]

Abstract

Theory of Mind (ToM) is the ability to understand and reason about one's own and others' mental states, which plays a critical role in the development of intelligence, language understanding, and cognitive processes. While existing work has primarily focused on first and second-order ToM, we explore higher-order ToM, which involves recursive reasoning on others' beliefs. We introduce **HI-TOM** , a **Hi**gher Order **T**heory **o**f **M**ind benchmark, consisting of 1.8k Sally-Anne-like stories with multiple-choice questions. Our experimental evaluation using GPT-4 reveals a decline in performance on higher-order ToM tasks, indicating the limitations of current models. This highlights the challenges of reasoning in complex ToM scenarios and emphasizes the need for further advancements in large language models' higher-order ToM capabilities.

1. Introduction

Theory of Mind (ToM) refers to the ability to understand and reason the mental states, such as intentions and beliefs, of others and distinguish them from one's own (Premack & Woodruff, 1978). Such an ability has been considered as a crucial pointer in the development of intelligence functions (Premack & Woodruff, 1978; Bretherton & Beeghly, 1982; Frith & Frith, 2003), where researchers find that the ToM reasoning is highly related to linguistic and cognitive processes. Thus, the ToM has been widely used as a protocol to evaluate language understanding and reasoning ability of intelligence agents (Premack & Woodruff, 1978; Takano et al., 2006), such as young children (Osterhaus & Koerber, 2021).

Figure 1. A Sally-Anne-like story (Baron-Cohen et al., 1985). Sally, Anne, and Alex entered the room, and the target object milk is on the table (Scene 1). Then, Anne moved the milk (Scene 3) after Sally exited the room (Scene 2). The three questions at the bottom correspond to different orders of ToM reasoning concerning Scene 4. For demonstration purposes, we only show different orders for 3 agents as the proof-of-concept.

With the recent advance in large language models (LLMs), work has been done to evaluate the language skills of LLMs using ToM (Sap et al., 2022; Ullman, 2023). Most existing work is limited to first-order and second-order ToM, where LLMs are only asked to perform inference on others' belief of reality in one or two passes (the first and second-order questions in Figure 1). However, there has been evidence that higher-order ToM (third-order and beyond), which requires recursively reasoning on others' beliefs, is essential to communicate effectively in complicated scenarios, such as multi-party conversations (Liddle & Nettle, 2006; De Weerd et al., 2015; Ridinger & McBride, 2017; De Weerd et al., 2022). Such higher-order ToM is not well studied in the NLP community. One main reason is the paucity of such a benchmark that is carefully designed to systematically evaluate the ToM reasoning in LLMs.

Previous work constructs ToM benchmark using automatic

[*]Equal contribution [1]College of Engineering, University of Michigan, Ann Arbor, Michigan, United States [2]School of Engineering, Westlake University, Hangzhou, Zhejiang, China. Correspondence to: Yinghui He <huihui@umich.edu>, Yufan Wu <umwyf@umich.edu>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.

story generation scripts. Although simple and cheap, such a method cannot be directly extended to generating stories of higher-order ToMs because the generated stories contain insufficient information for raising a higher-order question. To address this issue, we theoretically show two necessities, namely *numbers of agents* and *key chapters*, of a story that contains higher-order ToM reasoning, which allows us to automatically generate high-quality and consistent stories without sophisticated designs.

Using such a protocol, we introduce **TOMH**, a multiple-choice question benchmark that is carefully designed for evaluating higher-order ToMs. TOMH consists of 1.8k Sally-Anne-like stories (Figure 1), with paired multiple-choice questions and answers. Different from previous datasets, TOMH contains stories from first-order to fourth-order ToMs. Further, with irregular structure and complex random distractors, TOMH is more robust against shortcuts (Le et al., 2019). We manually check the quality of constructed data, and empirically find TOMH is more challenging and diverse compared with previous datasets.

We experiment with GPT-4 (OpenAI, 2023) on TOMH under the zero-shot setting. Furthermore, we test chain-of-thought prompting (Wei et al., 2022) and conduct a thorough analysis. Experimental results show that though GPT-4 performs near perfectly on the first and second-order ToM, it suffers a significant performance drop on third-order and fourth-order ToM. We reveal that though chain-of-thought can help improve GPT-4's performance on third and fourth-order ToM, GPT-4 still fails in certain cases, and would prefer to take the "shortcut" instead of reasoning through the whole story. Our analysis shows that the claim of LLMs have the genuine ToM ability (Kosinski, 2023; Bubeck et al., 2023) is questionable, especially in the complicated scenarios of higher order ToM, where several rounds of recursive reasoning is required. To our knowledge, we are the first to introduce the benchmark for evaluating higher-order ToM reasoning.

		Short	Medium	Long
1st-order	2 agents	50	50	50
	3 agents	50	50	50
	4 agents	50	50	50
		Short	Medium	Long
2nd-order	2 agents	50	50	50
	3 agents	50	50	50
	4 agents	50	50	50
		Short	Medium	Long
3rd-order	3 agents	75	75	75
	4 agents	75	75	75
		Short	Medium	Long
4th-order	4 agents	150	150	150

Table 1. Number of HI-TOM story-question pairs in each setting in terms of the ToM order of the question, the context length of the story, and the agent number of the story. The division of context lengths is short (1 chapter, $5 \sim 15$ lines), medium (3 chapters, $15 \sim 25$ lines), and long (5 chapters, $25 \sim 30$ lines). The agent number of a story is always larger than or equal to the corresponding ToM order in our setting (proof in Theorem C.5 in Appendix C). We construct the same amount of story-question pairs for each ToM order.

2. Background

Theory of Mind. Most of the prior works focus on first or second-order ToM (Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2022), while higher-order ToM (third order and beyond [1]) remains under-explored. The concept of "orders" refers to the number of mental state attributions that are required to answer a particular question or reason about a particular scenario. For instance, a third-order ToM question can be "Where does Anne think that Sally thinks that Isabella searches for the milk?", where Sally's reasoning about Isabella is second-order, while Anna's reasoning on Sally's reasoning adds another order.

Higher-order ToM is useful in social interaction such as maintaining social networks (Liddle & Nettle, 2006), winning limited bidding (de Weerd & Verheij, 2011), efficiently coordinating cooperation (De Weerd et al., 2015; Ridinger & McBride, 2017), and winning unpredictable negotiations (De Weerd et al., 2022). Researchers from cognitive science have investigated second-order and higher-order ToM among young children via complex forms of false-belief tests, such as the Sally-Anne false-belief experiment (Baron-Cohen et al., 1985) in Figure 1.

Evaluating ToM in LLMs. Sap et al. (2022) find GPT-3's ToM ability on ToMi dataset (Le et al., 2019) is well below humans. Kosinski (2023); Bubeck et al. (2023) show the promising performance of recent LLMs such as GPT-3.5 and GPT-4 on ToM tasks. However, it is questionable whether LLMs have genuine ToM ability. Ullman (2023) find that for GPT-3.5, small variations that maintain the principles of ToM can cause a flip of the answer. Different from those works that only evaluate LLMs' ToM ability up to the second order, we take a step forward and evaluate LLM's ability in higher order ToM setting.

3. The TOMH Dataset

To evaluate models' ability on higher-order ToM [1], we build **HI-TOM**, a dataset that requires higher-order ToM reasoning to answer the questions correctly. All the TOMH stories, questions, and answers are automatically generated along with manual checking.

[1] We test LLMs' ToM ability up to fourth order in this paper.

3.1. Data Generation

We design our data generation scripts based on Nematzadeh et al. (2018), which are originally limited to first or second-order ToM stories. We extend the scripts to contain higher-order story-question pairs.

Story Overview. A HI-TOM story is composed of multiple chapters, with at least one of them being the key chapter. A *chapter* describes a scene where an object (e.g. *milk* in Figure 1) is moved from one container to another container (e.g. *box*) while one or more agents (e.g. *Anne* and *Alex*) are present at a room, while a *key chapter* is the scene where all the agents in the corresponding question are present together with the object in the question.

Generation Principles. HI-TOM stories follow two principles:

1. A story must contain at least one *key chapter* where all the agents in the corresponding question are present together with the object in the question.
2. The number of agents involved in a story should be greater than or equal to the ToM order in the corresponding question.

Intuitively, if principle 1 is violated, the agent in the question may never reason the object correctly. If principle 2 is violated, the answer at different order can overlap with each other, and the model may come up with the correct answer without conducting higher-order ToM. Appendix C articulate the intuition and provide a formal proof of these two principles.

Chapter Generation. In each chapter, the number of agents varies from one (A1) to four (A4). Agents possess a true belief (TB) of the object location if the object movement happens before he leaves the room, in which the agent knows about the movement. Or the agent possesses a false belief (FB) of the object location if the object movement happens after he leaves the room, in which the agent does not know about the movement (eg. Sally in Figure 1). The possible combinations of agent numbers and belief types yield 7 different chapter types, where A1-FB does not exist because a single agent in a chapter moves the object himself and thus always holds a true belief. Table 5 in Appendix A shows example chapters of the seven types.

```
Input: ToM order: n ∈ {1, 2, 3, 4}
       Agent number: k ∈ {2, 3, 4}
       Number of chapters: ℓ ∈ {1, 3, 5}
       Story components: Names, Rooms,
                         Containers, Objects
       Actual object locations: Map
       Agents' beliefs: Belief
Output: story, question, answer
function STORY(n, k, ℓ, Names, Rooms, Objects)
    if n > k then
        return None
    obj_pool ← [ ]
    key ← RANDOM(1 : ℓ, s = 1)
    ns ← RANDOM(Names, s = k)
    for i ← 0 to l do
        r ← RANDOM(Rooms, s = 1)
        if i = key then
            chap, obj ← KEYCHAP(ns, r)
        else
            chap, obj ← CHAP(ns, r)
        add chap into story
        add obj into obj_pool
        update Map and Belief
    objs ← RANDOM(obj_pool, s = 1)
    question = Q_GEN(ns, objs)
    answer = A_GEN(Map, Belief)
    return story, question, answer
```

Algorithm 1: Map records which container contains which objects. $Belief$ records agents' belief of others' thoughts (from first to fourth order). CHAP and KEYCHAP generate a chapter or a key chapter by populating the input story components into some pre-defined templates. Similarly, Q_GEN populates the names of agents and objects into question templates and returns the generated $question$. A_GEN selects the corresponding answer given the agents' belief states and the actual location of each object. RANDOM($list$, s) is a function that randomly outputs s items from $list$.

In terms of the key chapter, its type depends on the ToM order of the story-question pair, which is the number of agents involved in the question. As the true belief (TB) does not pose any unalignment between the mental belief and the state of the physical world (everyone still holds a true belief), we select FB in the multi-agent setting to make the reasoning process more challenging. Therefore, we have four types of key chapters: A1-TB for the first order, A2-FB for the second order, A3-FB for the third order, and A4-FB for the fourth order. Note that we have A1-TB for the first-order because A1-FB does not exist.

Generation Scripts. Algorithm 1 provides the pseudocode for the generation process of the story, question and answer pairs in HI-TOM. Our program takes a list of story components for story generation, two dictionaries Map and $Belief$ that record actual object location and agents' belief, as well as three inputs n, k, and ℓ corresponding to the three attributes of a story-question pair: *ToM order* of the question, *agent number* of the story, and *context length* (short, medium or long) of the story. The context length depends on the number of chapters in the story: 1 chapter for a short context, 3 for a medium one, and 5 for a long one.

The Hi-ToM generator program incorporates 7 generators for different chapter types, which we generally represent as Chap. Especially, we utilize the function KeyChap to generate a key chapter. For the generation of each chapter, we randomly choose the story components and populate them into the chapter template in Chap. Complying with principle 1, we randomly choose one chapter to be a key chapter during the generation. In the end, we use the names and objects appearing in the story to generate the question and refer to Map and $Belief$ to generate the answer.

Assumptions. We assume that all the containers where objects are placed are transparent, following Nematzadeh et al. (2018). Once an agent enters a location, they become aware of the container that contains each object in this location.

Refinement. Following Le et al. (2019), we adapt the generation scripts from *ToM* in order to avoid shortcuts in answering. For instance, in *ToM*, the object in the question is always the one that is last moved, which may provide shortcuts for LLMs to achieve decent performance without understanding the story. To address this issue, we randomize the distribution of each move so that the objects in the question are not necessarily the last to be moved.

In addition, we incorporate random distractors in our generation. Specifically, we involve the movement of an irrelevant character at an irrelevant location in the story arbitrarily, trying to distract models and posing challenges to models' reasoning processes. With the additional distractor characters, the distribution of agent numbers may slightly deviate from those in Table 1. We still use the original agent number to describe a refined story for conciseness.

To increase the amount of information that LLMs need to capture and reason on, long stories are designed to have at least one agent re-entering a location, as exemplified in Table 2 line 11.

Also, we design the questions to be multiple-choice questions to avoid LLMs from generating vague answers like "the answer may either be the box or the table".

Table 2 shows an example third-order story question pair, with the refinement highlighted.

	Story
1	**Alexander, Jackson, and Benjamin entered the garage.** ♢
2	The apple is in the green_drawer.
3	Jackson exited the garage.
4	**Alexander moved the apple to the blue_box.** ♡
5	Alexander and Benjamin exited the garage.
6	Alexander and Benjamin entered the playroom.
7	The banana is in the red_treasure_chest.
8	Benjamin exited the playroom.
9	**Benjamin entered the TV room.** ♣
10	Alexander moved the banana to the red_basket.
11	**Alexander, Jackson, and Benjamin entered the garage.** ♠
12	The pumpkin is in the blue_box.
13	Jackson moved the pumpkin to the green_drawer.
	Question: Where does Alexander think that Jackson thinks that Benjamin searches for the apple? A. blue_box, B. green_drawer, C. red_treasure_chest **Answer: A. blue_box.**

Table 2. An example third-order story-question pair in ToMh along with the answer. The question is designed to be multiple-choice. We omit some irrelevant sentences and reformat the question and answer to ease the reading. We embolden sentences relevant to illustrate the key chapter (♢ and ♠), where all agents gather at a location at the same time. Random distractor (♣), which is inserted to distract the model and pose challenges to its reasoning process. Important sentence where the target object blue_box appears in the story (♡). The answer to the question is blue_box, since all the three agents re-entered the garage in line 11 and all three are aware that the apple is in the blue_box.

3.2. Dataset Information

Dataset Characteristics. Table 1 shows the number of story-question pairs under each setting. Table 3 shows a comparison between Hi-ToM and the other ToM datasets.

Compared to prior datasets, Hi-ToM supports the evaluation of ToM ability in the third and fourth order, while the prior datasets are limited to the first and second order. Moreover, the average length per story (# L / S) in Hi-ToM is 19.78, surpassing 15.05 for *ToM* / *ToM-easy* and 8.86 for ToMi. The average agent number in Hi-ToM is 3.38, also larger than 3.22 for *ToM* / *ToM-easy* and 2.75 for ToMi. The longer length and larger agent number suggest the intricate reasoning and logic behind each story.

Quality Control. For all story-question pairs in Hi-ToM, we manually verify that each story is coherent. Even with the presence of random distractors, one agent should appear at only one location at a time, and one object should only be bonded with one location. Then, we verify that all the stories introduce the correct amount of agents or movements, all the questions are solvable, and all the answers are correct.

4. Multiple-Choice (MC) Prompting

Following Sap et al. (2022), we prompt the model in a multiple-choice (MC) fashion, where the model chooses an answer between the given choices. We input a subset of our dataset consisting of 540 story-question pairs. Table 6 in Appendix B shows an example of the MC prompt.

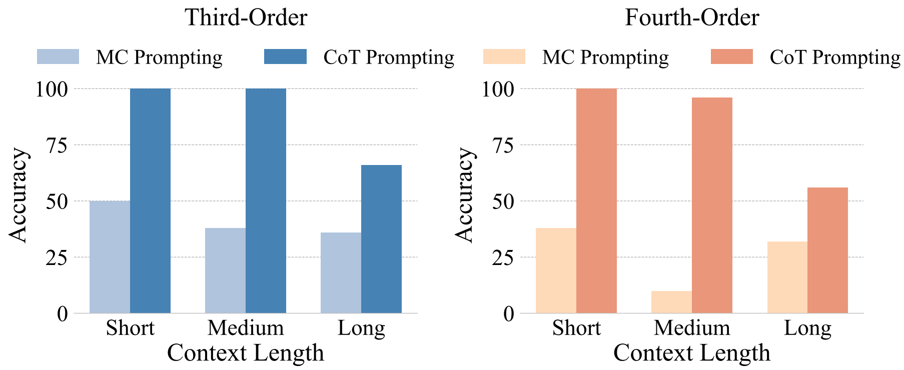

Figure 2. Comparison of accuracy of GPT-4 on ToMH with multiple-choice (MC) prompting and CoT prompting. The story-question pairs used in the CoT prompts are the same as those in the MC prompts.

Datasets	*ToM / ToM-easy*	ToMI	ToMH
1st	✓	✓	✓
2nd	✓	✓	✓
3rd	✕	✕	✓
4th	✕	✕	✓
#Story	96k	6k	1.8k
#Agentc	3.22	2.75	3.38
#Line	15.05	8.86	19.48
#Token	106.86	41.3	102.78

Table 3. Comparison between Hi-ToM and other datasets. 1st, 2nd, 3rd, and 4th refer to whether a dataset contains story-question pairs of a specific ToM order. "#Story" represents the number of stories. "#Agents", "#Line", and "#Token" represent average numbers per story.

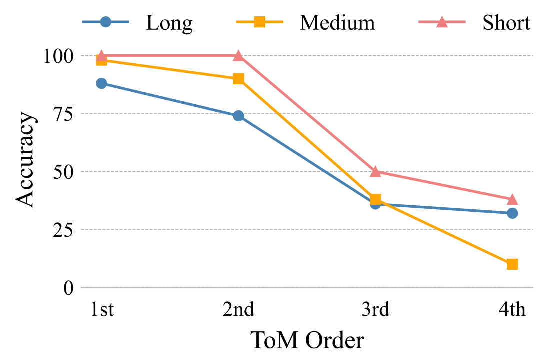

Figure 3. Accuracy of GPT-4 on ToMH by question orders, split by different context lengths.

4.1. Set-Ups

We do ToM evaluations on GPT-4 (OpenAI, 2023), a strong language model that exhibits state-of-the-art performance on various professional and academic benchmarks (Bubeck et al., 2023). We evaluate the ToM ability of GPT-4 on three dimensions: the context length of the stories, the number of agents in the stories, and the ToM order in the questions. The different combinations of these dimensions yield 27 different settings (Table 1). Following Kojima et al. (2022); Moghaddam & Honey (2023), we adopt the zero-shot setting in our experiments, where we prompt the model with the story, question, and several possible answers without any prior demonstration.

4.2. Results and Discussion

Significant Performance Drop between Second and Third Order. Figure 3 indicates that the performance of GPT-4 gets worse as we increase the ToM order of the story-question pair for each context length. The performance degradation from first to second order is less than 10%. However, we observe a sharp drop of $40\% \sim 52\%$ from second-order to third-order ToM regardless of the context length. For "short" tasks, the story structure and question formulation in the third order is almost the same as in the second order, except for the additional layer of the reasoning process in the question that makes the question a third order instead of a second order. Therefore, the model experiences a significant performance drop because of the extra order we add to the question.

Model Takes Shortcuts. One observation is that, in some cases, GPT-4 consistently chooses the final location of the object despite the variation in the story structure and the true answer. We suspect that instead of reasoning about agents' mental states throughout the whole story, the model takes a "shortcut" and only chooses the last location (the true location) of the object in the story. This situation occurs frequently on higher ToM orders and on long context of lower orders as well. But since the majority of answers are actually the initial locations of the objects because the key chapters have false-belief types, taking shortcuts turns out to lower the model performance. This might have led to the performance decline on higher-order stories and lower-order

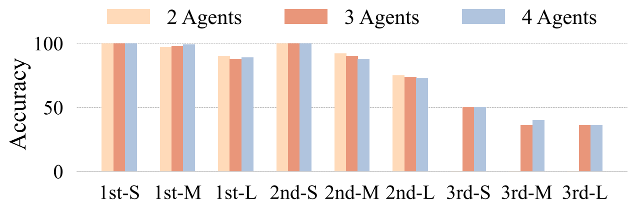

Figure 4. Accuracy of GPT-4 on TOMH versus the number of agents in different settings. 1st, 2nd, and 3rd represents the ToM order. S, M, and L respectively stand for short, medium, and long context length.

long stories.

Accuracy for Medium vs. Long. In general, the model performs worse on TOMH as the context length increases, as shown in Figure 3. A longer context implies a longer tracking process on agents' mental states and the appearance of more containers and locations, which may distract GPT-4 and result in lower accuracy. However, for fourth-order questions, we notice that the accuracy for medium context is lower than that for long context.

The reason we suspect is also related to the model shortcuts. As GPT-4 keeps choosing the final location, it fails on most of the medium story-question pairs whose answers are very likely to be the initial location because of the false-belief key chapter. On the other hand, long story-question pairs contain at least one re-entry chapter, hence producing a larger answer pool. The chance of the final location being the true answer increases with the re-entry of characters, as demonstrated in Table 2.

Agent Number vs. Accuracy. Figure 4 indicates that the accuracy scores across different agent numbers are almost the same under each setting on the x-axis. This implies that the number of agents in the stories does not pose a significant influence on the model accuracy as long as the context length and the order number stay the same. One possible explanation is that GPT-4 is good at extracting relevant information from the story based on the question, and irrelevant agents barely distract the model from tracking the mental states of relevant agents.

5. Chain-of-Thought (CoT) Prompting

Recent works have found that CoT prompting significantly improves LLMs' performance on various reasoning tasks (Kojima et al., 2022; Zhang et al., 2023; Liu et al., 2023). In particular, a zero-shot setting with a simple CoT prompt like "Let's think step by step" is able to significantly increase the answer accuracy of GPT-4 on lower-order ToM tasks (Moghaddam & Honey, 2023).

5.1. Set-Ups

Section 4.2 shows that GPT-4 performs poorly on higher-order ToM tasks. Here we conduct follow-up experiments to test whether CoT prompting improves the model performance on higher-order ToM tasks. Following Kojima et al. (2022); Moghaddam & Honey (2023), we prompt GPT-4 with a beginning sentence "Read the following story and answer the multiple choice question. *Think step by step*" followed by a story-question pair in HI-TOM. The statistics of our test sets on the third or fourth-order story-question pairs conform with Table 1. See Table 6 in Appendix B for an example prompt.

5.2. Results and Discussion

Think Step-by-Step is Effective. As shown in Figure 2, on third and fourth-order story-question pairs, CoT prompting raises the accuracy by 16% to 84% compared to MC prompting. Under CoT prompting, we get almost perfect accuracy and mostly correct step-by-step thinking processes in short and medium contexts. Table 4 (a) shows a completely correct step-by-step answer of GPT-4 to a fourth-order short story-question pair. In the long context, the model's performance also increases from 34% to 64%, and from 32% to 54% for the third and fourth order respectively. Yet, there is still room for improvement on the accuracy on long context even under CoT prompting.

Deficiency in Recursive Reasoning. Although the model performance has improved significantly, the detailed thinking process of GPT-4 still exposes its deficiency in higher-order ToM reasoning. Table 4 (b) shows an example answer which is correct on the first and second-order reasoning (step 1 and 2) but becomes both logically and factually incorrect on the third and fourth (step 3 and 4). In step 3, for example, the model focuses on what Ava and William "both" think, which is irrelevant to how Ava thinks William thinks of Isabella's mental state. A correct third step is *"Ava would believe that William also thinks that Isabella thinks the pineapple is still in the red_basket, as she doesn't know that William moved the pineapple and Isabella saw the movement"*.

Failure to Capture Key Information in Recursive Reasoning. Another reason why GPT-4 fails on higher-order ToM tasks is that it ignores key information in recursive reasoning. Table 4 (c) shows an example in which GPT-4 completely neglects the reappearance of the location "pantry". As everyone gathers again in the pantry and sees everything including the cabbage in the red_basket, all of the agents need to update their perception of each others' mental states. However, GPT-4 neglects such key information and chooses the answer based on only the first chapter of the story.

	Type	Story & Question	GPT-4 Answer
(a)	4TH-A4-SHORT	Noah, Liam, Isla and Hannah entered the den. The orange is in the **red_treasure_chest**. Noah exited the den. Liam moved the orange to the **green_pantry**. *Q: Where does Noah think that Liam thinks that Isla thinks that Hannah searches for the orange?*	The orange was initially in the red_treasure_chest, and all the people saw that ✓. After Noah exited the den, Liam moved the orange to the green_pantry. Therefore, Hannah, Isla, and Liam would be aware of the new location, but Noah would not be ✓. So, Noah thinks that Liam thinks that Isla thinks that Hannah searches for the orange in the last location he knew it to be before he left the den ✓. That location is the **red_treasure_chest (Correct)**.
(b)	4TH-A4-MEDIUM	Ava, William, Isabella and Evelyn entered the closet. The pineapple is in the **red_basket**. Ava exited the closet. William moved the pineapple to the **blue_treasure_chest**. *Q: Where does Ava think that William thinks that Isabella thinks that Evelyn searches for the pineapple?*	Let's walk through the story: 1. Ava last saw the pineapple in the red_basket and does not know it has been moved ✓. 2. Ava thinks that William also believes the pineapple is in the red_basket, as she didn't see him move it ✓. 3. Ava and William would both think that Isabella thinks the pineapple is still in the red_basket because that's where it was the last time she saw it ✗. 4. Finally, Ava, William, and Isabella would all believe that Evelyn is looking for the pineapple in the blue_treasure_chest because that's where it was when Evelyn last saw it ✗. Therefore, the answer is **blue_treasure_chest (Incorrect)**.
(c)	4TH-A4-LONG	James, Liam, Amelia, and Isla entered the **pantry** and saw everything. The cabbage is in the **green_suitcase**. James exited the pantry. Liam moved the cabbage to the **red_basket**. *<Irrelevant chapters>* **James entered the pantry.** Liam, Amelia, and Isla are in the pantry. The orange is in the red_basket. James exited the pantry. Liam moved the orange to the green_suitcase. *Q: Where does James think that Liam thinks that Amelia thinks that Isla searches for the cabbage?*	The last time the cabbage's location was updated, Liam moved the cabbage to the red_basket ✓. James was not present when this move happened ✓. The last time James saw the cabbage, it was in the green_suitcase ✗. Therefore, James thinks that Liam thinks the cabbage is in the green_suitcase ✗. Amelia was also not present during this update ✗. The last time Amelia saw the cabbage, it was in the green_suitcase ✗. So, Liam thinks that Amelia thinks the cabbage is still in the green_suitcase ✗. Finally, Isla was not present during either of these updates ✗. The last time Isla saw the cabbage, it was still in the original location, the green_suitcase ✗. Therefore, Amelia thinks that Isla is still searching for the cabbage in the green_suitcase ✗. So, James thinks that Liam thinks that Amelia thinks that Isla searches for the cabbage in the **green_suitcase (Incorrect)**.

Table 4. Example answers of GPT-4 to fourth-order story-question pairs under CoT prompting. The type encodes the order, agent numbers, and context length respectively (e.g., 4TH-A4-SHORT means a four-agent short story with a fourth-order question). For demonstration purposes, we only present the main task of a story here and omit the irrelevant chapters in the story, and merge successive "enter" actions in one sentence.

6. Conclusion

In this paper, we introduce HI-TOM, the first ToM benchmark that contains higher-order ToM tasks. We reveal that LLMs' performance suffers significantly on ToM tasks from the second to the third order. Although CoT prompting can improve the LLMs' performance on higher-order ToM tasks, we show that flaws exist in LLMs' intermediate reasoning steps. By proposing HI-TOM, we hope to address the challenges of ToM in complicated scenarios and spark further research on enhancing the reasoning ability of LLMs.

Future efforts may focus on constructing benchmarks that evaluate a wider range of higher-order ToM reasoning abilities, particularly in the context of creating intricate and realistic scenarios other than some simple actions. Also, it is crucial to pay attention to preventing LLMs from taking shortcuts to arrive at correct answers, as this would lead to overestimations of their true reasoning capabilities.

References

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

Bretherton, I. and Beeghly, M. Talking about internal states: The acquisition of an explicit theory of mind. *Developmental psychology*, 18(6):906, 1982.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

de Weerd, H. and Verheij, B. The advantage of higher-order theory of mind in the game of limited bidding. In *Proceedings Workshop 'Reasoning about other Minds', CEUR Workshop Proceedings*, volume 751, pp. 149–164, 2011.

De Weerd, H., Verbrugge, R., and Verheij, B. Higher-order theory of mind in the tacit communication game. *Biologically Inspired Cognitive Architectures*, 11:10–21, 2015.

De Weerd, H., Verbrugge, R., and Verheij, B. Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, 36(2):30, 2022.

Frith, U. and Frith, C. D. Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358 (1431):459–473, 2003.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Kosinski, M. Theory of mind may have spontaneously emerged in large language models, 2023.

Le, M., Boureau, Y.-L., and Nickel, M. 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)*, pp. 5872–5877, 2019.

Liddle, B. and Nettle, D. Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, 4(3-4):231–244, 2006.

Liu, X., Pang, T., and Fan, C. Federated prompting and chain-of-thought reasoning for improving llms answering. *arXiv preprint arXiv:2304.13911*, 2023.

Moghaddam, S. R. and Honey, C. J. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*, 2023.

Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., and Griffiths, T. L. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*, 2018.

OpenAI. Gpt-4 technical report, 2023.

Osterhaus, C. and Koerber, S. The development of advanced theory of mind in middle childhood: A longitudinal study from age 5 to 10 years. *Child Development*, 92(5):1872–1888, 2021.

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

Ridinger, G. and McBride, M. Theory of mind ability and cooperation. *Manuscript, Univ. California, Irvine*, 2017.

Sap, M., LeBras, R., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

Takano, M., Arita, T., et al. Asymmetry between even and odd levels of recursion in a theory of mind. *Proceedings of ALife X*, pp. 405–411, 2006.

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

A. HI-TOM Details

Table 5 shows examples of the seven chapter types that compose the HI-TOM stories: A1-TB, A2-TB, A3-TB, A4-TB, A2-FB, A3-FB, and A4-FB.

A1-TB	A2-TB	A3-TB	A4-TB
Sally entered the kitchen. The milk is on the table. Sally moved the mlik to the box.	Sally entered the kitchen. Anne entered the kitchen. The milk is on the table. Sally moved the milk to the box.	Sally entered the kitchen. Anne entered the kitchen. Alex entered the kitchen. The milk is on the table. Sally moved the milk to the box.	Sally entered the kitchen. Anne entered the kitchen. Alex entered the kitchen. Sam entered the kitchen. The milk is on the table. Sally moved the milk to the box.
A2-FB	**A3-FB**	**A4-FB**	
Sally entered the kitchen. Anne entered the kitchen. The milk is on the table. *Anne exited the kitchen.* Sally moved the milk to the box.	Sally entered the kitchen. Anne entered the kitchen. Alex entered the kitchen. The milk is on the table. *Anne exited the kitchen.* Sally moved the milk to the box.	Sally entered the kitchen. Anne entered the kitchen. Alex entered the kitchen. Sam entered the kitchen. The milk is on the table. *Anne exited the kitchen.* Sally moved the milk to the box.	

Table 5. Examples of the seven chapter types that compose the HI-TOM stories. A2, A3, and A4 respectively represent 2, 3, and 4 agents appearing in the chapter. FB represents false-belief, indicating that the agent that exits in the middle has a false belief on the final location of the object. On the contrary, TB or true-belief indicates that all agents involved have a true belief on the object's final location.

B. Experiment Details

Table 6 shows example prompts we use on GPT-4, in MC and CoT fashion respectively.

```
Read the following story and answer the
multiple-choice question. Provide the
answer without explanation.
Story:
1 Lucas entered the kitchen.
2 Jacob entered the kitchen.
3 Carter entered the kitchen.
4 The strawberry is in the red box.
5 Jacob moved the strawberry to the
green crate.
6 Jacob exited the kitchen.
7 Lucas moved the strawberry to the blue
bottle.
8 Lucas exited the kitchen.
9 Carter exited the kitchen.
Question: Where does Charlotte think
Jack thinks Hannah thinks William thinks
the carrot is?
Choices: A. red box, B. green crate, C.
blue bottle.
```

```
Read the following story and answer the
multiple-choice question. Think step-
by-step.
Story:
1 Lucas entered the kitchen.
2 Jacob entered the kitchen.
3 Carter entered the kitchen.
4 The strawberry is in the red box.
5 Jacob moved the strawberry to the
green crate.
6 Jacob exited the kitchen.
7 Lucas moved the strawberry to the blue
bottle.
8 Lucas exited the kitchen.
9 Carter exited the kitchen.
Question: Where does Charlotte think
Jack thinks Hannah thinks William thinks
the carrot is?
Choices: A. red box, B. green crate, C.
blue bottle.
```

Table 6. Example MC prompt and CoT prompt. The correct answer should be "B. green crate", since Jacob exited the kitchen earlier than Lucas moving the strawberry to the blue bottle.

C. Supplementary Proof

Two important ideas in the construction of HI-TOM are: (1) The number of agents involved in a story should be greater than the ToM order in the corresponding question; (2) A story must contain at least one multi-agent chapter where all the agents mentioned in the question are present.

Here, we formalize these two ideas by introducing a sequence of definitions, theorems, and proof.

Definition C.1. We use the notation K to denote the set of natural numbers from 1 to n, where n is the number of moves of the object in the question. For instance, in the example in Table 2, $K = \{1\}$.

Definition C.2. We use the notation f to denote a function such that given a positive integer $k \in K$, $f(k)$ returns the container of the object in the question after its k-th move. For instance, in the example in Table 2, $f(1) =$ blue box.

Definition C.3. The answer to question "Where does $\mathbf{A}_1$ thinks that $\mathbf{A}_2$ thinks that ... $\mathbf{A}_n$ searches for the object **O**" is:

$$Ans = f(max(T_{A_1} \cap T_{A_2} \cap \dots T_{A_n}))$$

where T_{A_i} is the set of moves of the object in the question, observed by agent $\mathbf{A}_i$.

Remark C.4. $max(T_{A_1} \cap T_{A_2} \cap \dots T_{A_n})$ represents the index of the last move of the object in the question during their common observation. So the formula above reflects that A_1's inference of others' belief is essentially the last known container in their witness.

Theorem C.5. *The number of agents involved in a story should be greater than or equal to the ToM order in the corresponding question.*

Proof. Let (s, q) be a story-question pair. Suppose the k is the number of agents in s. We will prove that if the ToM order of q is larger than n, then the answer to q is the same as the answer to a k-th order question.

For $T_{A_1} \cap T_{A_2} \cap \dots T_{A_n}$, we first consider the case that $n = k + 1$. Then $\exists i, j \in 1, 2, \dots n$ such that $T_{A_i} = T_{A_j}$. Further we get (suppose the $i \leq j$):

$$\begin{aligned} T_{A_1} \cap T_{A_2} \cap \dots T_{A_n} &= T_{A_1} \cap T_{A_2} \cap \dots T_{A_i} \dots T_{A_{j-1}} \cap T_{A_j} \cap T_{A_{j+1}} \dots \cap T_{A_n} \\ &= T_{A_1} \cap T_{A_2} \cap \dots T_{A_i} \dots T_{A_{j-1}} \cap T_{A_i} \cap T_{A_{j+1}} \dots \cap T_{A_n} \\ &= T_{A_1} \cap T_{A_2} \cap \dots T_{A_i} \dots T_{A_{j-1}} \cap T_{A_{j+1}} \dots \cap T_{A_n} \end{aligned}$$

We see that the extra terms due to larger ToM order are eliminated after simplification. The final answer still corresponds to a k-th order question rather than n-th order. Applying the same logic, for any $n > k$, there will be $(n - k)$ pairs of identical terms. After discarding the extra term in each pair, we finally get the answer to the k-th order question.

Consequently, it is redundant to analyze questions with unmatched ToM order and number of agents as their answers are totally identical to those in proper questions. For simplicity, we thus require that number of agents should be greater than the ToM order in the question. □

Theorem C.6. *A story must contain at least one chapter where all the agents in the corresponding question are present together with the object in the question.*

Proof. To ensure that $T_{A_1} \cap T_{A_2} \cap \dots T_{A_n}$ is not empty (otherwise function f will receive no input), all agents have to gather together at least once. □