A. ToMh Details

Table 5 shows examples of the seven chapter types that compose the ToMh 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 ToMh 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.

MC Prompt	CoT Prompt
Read the following story and answer the multiple-choice question (without explanations): 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 Lucas think that Jacob thinks that Carter searches for the strawberry? 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 Lucas think that Jacob thinks that Carter searches for the strawberry? 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 ToMh 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 definition, theorem, 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 $i \leq j$):

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

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. □