TIMETOM: Temporal Space is the Key to Unlocking the Door of Large Language Models' Theory-of-Mind

Anonymous ACL submission

Abstract

Theory of Mind (ToM)—the cognitive ability to reason about mental states of ourselves and others, is the foundation of social interaction. Although ToM comes naturally to humans, it poses a significant challenge to even the most advanced Large Language Models (LLMs). Due to the complex logical chains in ToM reasoning, especially in higher-order ToM questions, simply utilizing reasoning methods like Chain of Thought (CoT) will not improve the ToM capabilities of LLMs. We present TIMETOM, which constructs a temporal space and uses it as the foundation to improve the ToM capabilities of LLMs in multiple scenarios. Specifically, within the temporal space, we construct Temporal Belief State Chain (TBSC) for each character and inspired by the cognition perspective of the social world model, we divide TBSC into self-world beliefs and social world beliefs, aligning with first-order ToM (first-order beliefs) and higher-order ToM (higher-order beliefs) questions, respectively. Moreover, we design a novel tool-belief solver that, by considering belief communication between characters in temporal space, can transform a character's higher-order beliefs into another character's first-order beliefs under belief communication period. Experimental results indicate that TIMETOM can dramatically improve the reasoning performance of LLMs on ToM questions while taking a big step towards coherent and robust ToM reasoning.

1 Introduction

Humans continually try to reason about other people's mental states and understand how it might impact their actions (Frith and Frith, 2003). This capability, known as Theory of Mind (ToM) (Premack and Woodruff, 1978), is crucial for social interactions. With Large Language Models (LLMs) playing a growing role in our lives, developing LLMs with ToM could be better at teaching us, learning from us, communicating with us, collaborating with us, and understanding us (Gandhi et al., 2021, 2023; Rabinowitz et al., 2018; Shu et al., 2021).

Although ToM often comes naturally to humans, LLMs often make various errors in ToM reasoning (Figure.1C), such as ignoring the temporal order of events, generating outputs that violate commonsense, confusing the reasoning logic in higher-order ToM questions (He et al., 2023) and failing on "trivial" alternations to existing datasets (Kim et al., 2023; Ullman, 2023). Recently, various reasoning strategies like Chain of Thought (CoT) (Wei et al., 2022) and Tree of Thought (ToT) (Yao et al., 2023) have improved the reasoning abilities of LLMs in some tasks. However, these strategies are not suitable for ToM reasoning (Ma et al., 2023). Furthermore, Wilf et al. (2023) proposes the perspective-taking strategy to improve the ToM reasoning abilities of LLMs, but this strategy is not suitable for higher-order ToM reasoning. Currently, there is still a lack of effective reasoning strategies to improve the performance and robustness of LLMs in ToM reasoning tasks.

In this paper, we introduce TIMETOM, which initially constructs temporal space by adding timeline to the stories or dialogues. Within the temporal space, we construct Temporal Belief State Chain (TBSC) for each character based on the events they are aware of on the timeline. Meanwhile, inspired by a principle of modern cognitive science (Mitchell, 2023; Yue, 2022), which posits that humans construct abstract models of the social world and their self-world in their minds, we split the beliefs in TBSC into self-world beliefs and social world beliefs. We use self-world beliefs to answer first-order ToM questions and incorporate social world beliefs when answering higher-order ToM questions.

The reasoning difficulty of ToM questions significantly increases as the order of questions rises (Sclar et al., 2023), and currently, there is no effective reasoning strategy for solving higher-order

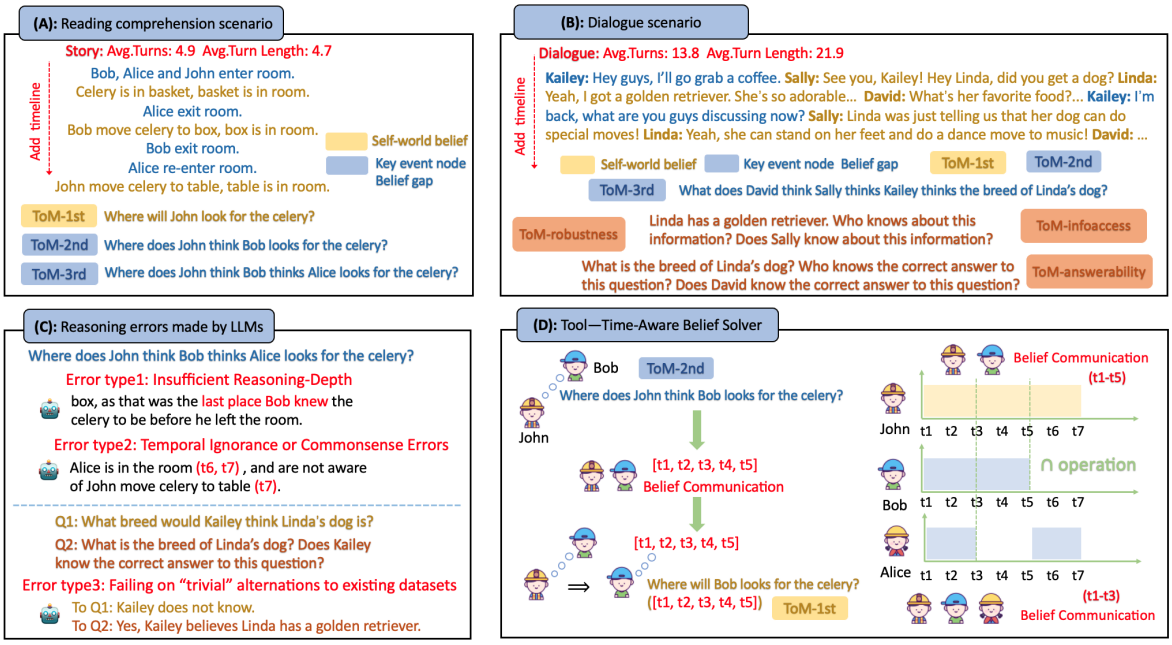

Figure 1: (A) and (B): The structure of story and dialogue, as well as ToM questions in reading comprehension and dialogue scenarios. (C): ToM reasoning errors made by LLMs. (D): Reasoning perspective of belief solver.

(reasoning depth $m \geq 2$) ToM questions. We consider that the key to higher-order ToM reasoning lies in capturing the belief communication between characters. We design a novel tool—belief solver, which first parses each character's perceptible time set based on their TBSC and then calculates the intersections of the time set of different characters to determine at which times they achieve belief communication. Furthermore, as illustrated in Figure.1D, since John's understanding of Bob's belief only occurs during the belief communication period, the higher-order ToM question of John's estimation of Bob's beliefs can be transformed into a first-order ToM question of what Bob's belief state is like during the belief communication period. In reasoning about higher-order ToM questions, LLMs generate an initial reasoning process based on the character's TBSC, and the belief solver transforms the higher-order ToM questions into first-order ToM questions under belief communication period, which serves as feedback to inspire LLMs to refine their initial reasoning process on higher-order ToM questions.

Experimental results Experimental results on the ToMI (Le et al., 2019), BigToM (Gandhi et al., 2023), and FanToM (Kim et al., 2023) benchmarks indicate that TIMETOM dramatically improves the reasoning performance of LLMs on ToM questions in multiple scenarios (Figure.1A and Figure.1B), while taking a big step towards coherent and robust ToM reasoning. It's noteworthy that TIMETOM is well-suited for higher-order ToM questions, demonstrating good performance in third-order ToM questions. Furthermore, TIMETOM can be applied to situations involving agent communication which is commonly occurring in the real-world.

The main contributions of this work are as follows: (1) We construct temporal space and use it as the foundation. (2) Within the temporal space, we construct TBSC, which is a comprehensive representation of a character's beliefs, including the temporal evolution of thing states and clear temporal cognition of key social events that create belief gaps. From the social world model's cognitive perspective, TBSC is split into self-world and social world beliefs, aligning with first-order ToM (first-order beliefs) and higher-order ToM (higher-order beliefs) questions respectively. (3) Within the temporal space, we design a novel tool—belief solver, which serves as feedback to inspire LLMs to refine their initial reasoning process in higher-order ToM questions. (4) Temporal space is the key to unlocking the door of LLMs' Theory of Mind. Extensive experimental results indicate that TIMETOM dramatically improves the reasoning performance and robustness of LLMs on ToM questions.

2 TIMETOM Overview

Figure.1C illustrates the errors LLMs made in ToM reasoning, often ignoring the temporal order of

events and confusing the reasoning logic in higher-order ToM questions. Concurrently, explicitly representing the timeline not only allows LLM to have a clearer temporal understanding of the events in the story and dialogue, but we also observe the association between higher-order questions (higher-order beliefs) and first-order questions (first-order beliefs) on the timeline. Building upon this insight, we introduce TIMETOM, improving the ToM capabilities of LLMs in interactive dialogue and reading comprehension scenarios. The overall procedure of TIMETOM is shown in Figure 2.

2.1 Constructing Temporal Space

In reading comprehension scenarios, each sentence in the story corresponds to a specific time point. Similarly, in interactive dialogue scenarios, each utterance in the dialogue corresponds to a specific time point. Illustrating with the case of reading comprehension scenarios, given the input story x: *Sentence1, Sentence2,..., SentenceN*, prompt p_{cts}, and model $\mathcal{M}$, TIMETOM adds a complete timeline for input story x to form x_t:

$$x_t = \mathcal{M}(p_{cts}||x). \tag{1}$$

For example, as illustrated in Figure 2, the model explicitly adds time points before each sentence for the given input story. Here, $||$ denotes concatation and p_{cts} is shown in Appendix C.2.

2.2 Constructing Temporal Belief State Chain for Each Character

ToM questions focus on the beliefs of characters, including their own beliefs as well as their estimations of others' beliefs. Given story x_t within temporal space, prompt p_{tbsc} and model $\mathcal{M}$, TIMETOM construct TBSC c_{tbsc} for each character, based on the events they are aware of on the timeline:

$$c_{tbsc} = \mathcal{M}(p_{tbsc}||x_t) \tag{2}$$

For example, as illustrated in Figure 2, Alice is aware of events between **t1** to **t3**, but since she leaves the room at **t3** and re-enters the room at **t6**, he cannot aware of events between **t4** to **t5**. TBSC is a comprehensive representation of a character's beliefs, which includes the temporal evolution of object states, such as celery is in the basket at **t2**, in the box at **t4**, and on the table at **t7** as well as key social events that create belief gaps with clear temporal logic, such as Alice exits the room at **t3**, Bob exits the room at **t5**, and Alice re-enters the room at **t6**. Here, $||$ denotes concatation and p_{tbsc} is shown in Appendix C.2.

2.3 Time-Aware ToM-Question Answer from Social World Model Perspective

Mitchell (2023) posits that humans construct abstract models of the social world and their self-world in their minds. Inspired by this cognitive perspective, we divide the belief in TBSC into self-world belief and social world belief. We consider self-world belief as the perception of the state and information of things around oneself, while social world belief is the perception of other characters' actions that may lead to a belief gap[1]. Given character's TBSC c_{tbsc}, prompt p_{self} and model $\mathcal{M}$, TIMETOM generates self-world belief $c_{tbsc-self}$ via belief compression, i.e., focusing on events in TBSC about the states and information of things:

$$c_{tbsc-self} = \mathcal{M}(p_{self}||c_{tbsc}). \tag{3}$$

Given the character's self-world belief $c_{tbsc-self}$ and the comprehensive belief c_{tbsc} after incorporating social world beliefs, prompt p_{qa}, and model $\mathcal{M}$, TIMETOM use self-world belief $c_{tbsc-self}$ to answer first-order ToM questions and comprehensive belief c_{tbsc} when answering higher-order ToM questions:

$$\begin{aligned} y_{first\_order} &= \mathcal{M}(p_{qa}||c_{tbsc-self}) \\ y_{higher\_order} &= \mathcal{M}(p_{qa}||c_{tbsc}). \end{aligned} \tag{4}$$

For example, as illustrated in Figure 2, John uses only self-world belief (celery is in the basket at **t2**, in the box at **t4**, and on the table at **t7**) to answer the first-order ToM question "*Where will John look for the celery?*" and incorporates social world belief (Alice exits the room at **t3**, Bob exits the room at **t5**, and Alice re-enters the room at **t6**) when answering higher-order ToM questions "*where does John think Bob looks for the celery?*". Here, $||$ denotes concatation, p_{self} and p_{qa} is shown in Appendix C.2.

2.4 Time-Aware Belief Solver

We achieve a more comprehensive and clearer representation of characters' beliefs by establishing a TBSC for each character, which improve the performance of LLMs in answering first-order and higher-order ToM questions. However, as the order

[1]Considering self-world belief and social world belief in this way aligns well with first-order and higher-order ToM questions, making it very suitable for ToM scenarios.

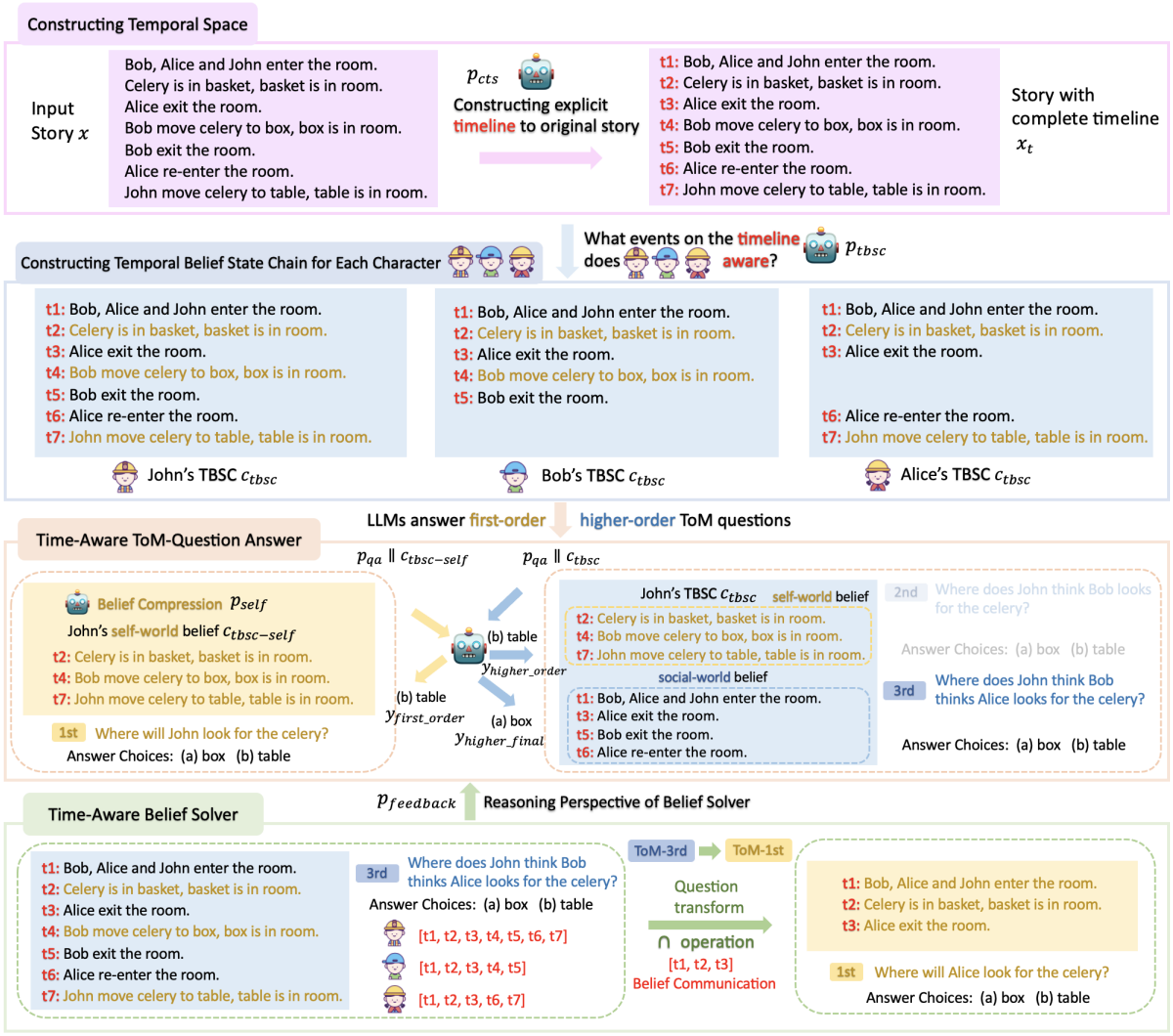

Figure 2: Pipeline overview of TIMETOM, which constructs a temporal space and uses it as the foundation to improve the ToM capabilities of LLMs. TIMETOM does not require training, it works in a zero-shot setting.

of ToM questions increases, the depth of reasoning required becomes deeper, relying solely on the character's TBSC still leads to logical errors in reasoning. Through in-depth consideration of belief communication between characters, we design an external tool—belief solver, which provides a novel reasoning perspective that effectively reduces the depth of reasoning, and serves as a feedback to inspire LLMs to refine their initial reasoning process on higher-order ToM questions.

Time Set Parsing We parse each character's perceptible time set based on their TBSC. For example, as illustrated in Figure 2, the set of times that John, Bob, and Alice can perceive are as follows:

$$\begin{aligned} T_{John} &= [t_1, t_2, t_3, t_4, t_5, t_6, t_7] \\ T_{Bob} &= [t_1, t_2, t_3, t_4, t_5] \\ T_{Alice} &= [t_1, t_2, t_3, t_6, t_7] \end{aligned} \tag{5}$$

Belief Communication between Different Characters To determine at which times belief communication occurs between different characters, we perform intersection operations on the sets of times perceived by each character, as parsed in the previous step:

$$\begin{aligned} BC_{John,Bob} &= T_{John} \cap T_{Bob} \\ BC_{John,Bob,Alice} &= T_{John} \cap T_{Bob} \cap T_{Alice} \end{aligned} \tag{6}$$

where $BC_{John,Bob} = [t_1, t_2, t_3, t_4, t_5]$ represents the set of times for belief communication between John and Bob, the same applies to $BC_{John,Bob,Alice} = [t_1, t_2, t_3]$.

Transforming Higher-order ToM problems into First-order ToM problems Consider second-order ToM question *"Where does John think Bob looks for the celery?"*, since John's understanding of Bob's belief only occurs during $BC_{John,Bob}$

period, this question can be transformed into a first-order ToM question *"Where does Bob look for the celery?"* under $BC_{John,Bob}$. It is worth noting that this transformation is also applicable to third-order and higher-order ToM problems.

Inspiring LLMs to Reason on Higher-order ToM Questions Through in-depth consideration of belief communication between characters, we observe that higher-order ToM questions can be transformed into first-order ToM questions under belief communication periods. Given this reasoning process as feedback $p_{feedback}$, LLM's initial reasoning process y_{higher_order}, and model $\mathcal{M}$, TIMETOM lets the LLM reason again:

$$y_{higher\_final} = \mathcal{M}(p_{feedback}||y_{higher\_order}). \quad (7)$$

We hope that through this method, the LLM can pay attention to the belief communication between characters as well as the connection between higher-order beliefs and first-order beliefs to refine their initial reasoning outputs on higher-order ToM questions. Here, $||$ denotes concatation, $p_{feedback}$ is shown in Appendix C.2.

3 Experiments

3.1 Settings

Benchmarks We evaluate TIMETOM within reading comprehension and interactive dialogue scenarios, using ToMI, BigToM, and FanToM benchmarks. Compared to stories in reading comprehension scenarios, dialogues are more aligned with real-world scenarios requiring ToM reasoning. Furthermore, dialogues in FanToM are significantly longer, with a larger number of subtopics and characters per dialogue. This poses a greater challenge for LLMs, as it demands the LLMs' capability to comprehend the complete dialogue utterance and reason about each character's beliefs. Detailed explanations for story (or dialogue) structure, each type of question in ToMI, BigToM, and FanToM benchmarks, and evaluation metrics can be found in Appendix B.

Baselines We employ five widely utilized LLMs: three open source – Llama2-7b, 13b, and 70b chat (Touvron et al., 2023) – and two closed source: GPT-3.5-Turbo-0613 and GPT-4-0613 to evaluate TIMETOM. To highlight the effectiveness of TIMETOM, we evaluate LLMs 0-shot on ToM benchmarks with and without our TIMETOM. Additionally, we compare TIMETOM with the CoT (Wei et al., 2022) and SimToM (Wilf et al., 2023), where SimToM is a recently proposed prompting framework specifically designed to improve the reasoning performance of LLMs on ToM questions, achieving state-of-the-art results. To make a fair comparison, we uniformly set the temperature to 0 (GPT-series models) or 0.3 (Llama2-series models) and top_p to 0.95 for all experiments. We reproduce our TIMETOM prompts in Appendix C.

3.2 Main Results

In Tables 1 and 2, we report the reasoning performance of LLMs for ToM questions in reading comprehension scenarios and dialogue scenarios.

Substantial Improvement across Different LLMs and Scenarios From widely utilized commercial LLMs (GPT-series) to open-source models (Llama2-series), and from reading comprehension to interactive dialogue scenarios, TIMETOM leads to substantial performance improvement. Specifically, in the reading comprehension scenario, we achieve an average absolute accuracy improvement of **+19.43%** and **+9.63%**, as well as **+13.38%** and **+10.81%** over the 0-shot and SIMTOM baselines for the ToMI and BigTOM benchmark, respectively. A larger improvement is observed in the interactive dialogue scenario, where TIMETOM achieves an average absolute accuracy improvement of **+44.7%** and **+13.6%** over the 0-shot and SIMTOM baselines for the FanTOM benchmark.

Well-suited for Higher-order ToM Reasoning On the challenging higher-order ToM questions, TIMETOM yields **+29.00%** and **+18.75%** absolute accuracy improvement over the 0-shot and SIMTOM GPT-4 baselines, as well as **+16.8%** and **+17.9%** absolute accuracy improvement over the 0-shot and 0-shot-CoT GPT-4 baselines for the ToMI and FanTOM benchmark, respectively. An equally impressive result is observed across other model types. Furthermore, based on the GPT-4 model, we test the performance of baselines and TIMETOM on the third-order ToM problems of ToMI-Extend[2] benchmark. As shown in Figure 3, compared with SIMTOM and 0-shot-CoT, the performance of TIMETOM does not degrade as the order of the ToM question increases, indicating its suitability for higher-order ToM reasoning. Meanwhile, TIMETOM exhibits the most outstanding

[2]Sclar et al. (2023) construct third-order ToM questions by making simple modifications to the story structure of the original ToMI benchmark

Model	ToMI				BigTOM	
	Overall	False-Belief	First-Order	Second-Order	Overall	False-Belief
0-Shot						
Llama2-7b-chat	44.50	28.25	39.00	40.00	52.50	53.50
Llama2-13b-chat	51.00	39.25	54.75	34.75	55.25	46.50
GPT-3.5-turbo	68.60	67.25	68.75	52.75	78.50	69.50
GPT-4	66.50	25.50	50.75	65.50	97.50	99.00
0-Shot-CoT						
Llama2-7b-chat	43.70	24.00	45.00	37.75	50.50	39.50
Llama2-13b-chat	45.00	16.50	43.00	37.00	57.25	52.50
GPT-3.5-turbo	64.10	34.00	58.50	53.00	80.75	71.50
GPT-4	74.40	74.25	73.75	62.25	97.75	99.00
SIMTOM						
Llama2-7b-chat	48.10	40.00	47.25	39.25	56.25	75.00
Llama2-13b-chat	61.10	35.50	53.75	53.75	57.75	62.50
GPT-3.5-turbo	72.80	81.00	74.75	57.25	84.00	78.00
GPT-4	87.80	87.75	93.75	75.75	96.00	98.00
TIMETOM						
Llama2-7b-chat	64.30 **(+19.80,+16.20)**	47.25 **(+19.00, +7.25)**	56.50 **(+17.50, +9.25)**	57.75 **(+17.75, +18.50)**	68.75 **(+16.25, +12.50)**	84.50 **(+31.00, +9.50)**
Llama2-13b-chat	67.20 **(+16.20, +6.10)**	44.75 **(+5.50, +9.25)**	61.25 **(+6.50, +7.50)**	57.00 **(+22.25, +3.25)**	77.75 **(+22.50, +20.00)**	89.50 **(+43.00, +27.00)**
GPT-3.5-turbo	80.80 **(+12.20, +8.00)**	82.00 **(+14.75, +1.00)**	80.50 **(+11.75, +5.75)**	71.50 **(+18.75, +14.25)**	93.75 **(+15.25, +9.75)**	96.00 **(+26.50, +18.00)**
GPT-4	96.00 **(+29.50, +8.20)**	98.75 **(+73.25, +11.00)**	95.50 **(+44.75, +1.75)**	94.50 **(+29.00, +18.75)**	97.00 **(-0.50, +1.00)**	99.00 **(+0.00, +1.00)**
Avg	**(+19.43, +9.63)**	**(+28.13, +7.13)**	**(+20.13, +6.06)**	**(+21.94, +13.69)**	**(+13.38, +10.81)**	**(+25.13, +13.88)**

Table 1: TIMETOM results on ToMI across False-Belief, First-Order, Second-Order, and All question types. Since the BigToM benchmark contains only First-Order ToM questions, we only report results across False-Belief and All question types. We present **absolute accuracy difference** between TIMETOM and the baselines (0-shot and SIMTOM). Results for True-Belief and Mem-Real question types can be found in Appendix D.2.

performance on both first-order and higher-order ToM questions.

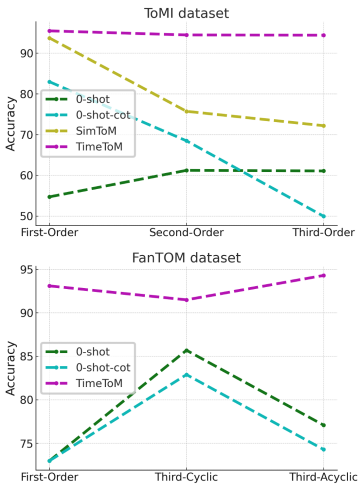

Figure 3: Performance comparison of TIMETOM and baselines on first-order and higher-order ToM questions.

Better ToM Reasoning Robustness We use ALL* and All score from Table 2 to evaluate the ToM reasoning robustness of baselines and TIMETOM. We achieve **+33.0% (× 4.8)** and **+31.3% (× 4.0)** absolute accuracy improvement over the 0-shot and 0-shot-CoT GPT-4 baselines for ALL* score, which requires correct answers to all five types of ToM questions (Belief, Answerability[List, Y/N], and Infoaccess[List, Y/N]). For All scores under the Answerability question and Infoaccess question, which require correct answers to both list-type and Y/N-type questions, we achieve **+27.8% (× 2.2)** and **+26.1% (× 2.0)**, as well as **+28.7% (× 2.2)** and **+17.3% (× 1.5)** absolute accuracy improvement over the 0-shot and 0-shot-CoT GPT-4 baselines, respectively. A similar improvement is noticeable in the llama2-70b-chat model, although the degree of improvement is not as large.

4 The Effect of Key Components

Given TIMETOM's strong performance, we analyze its key components: (1) Foundation: constructing temporal space. (2) From the perspective of first-order and higher-order ToM questions, considering temporal belief construction and compression as well as time-aware belief solver.

Constructing Temporal Space The construction of temporal space provides LLMs with a clearer un-

Model	ALL* Question Types	Belief Questions				Answerability Questions	infoaccess Questions
		Overall	First-order	Third-acyc	Third-cyc	All	All
0-Shot							
Llama2-70b-chat	0.0	6.5	8.7	0.0	5.7	4.3	8.7
GPT-4	8.7	76.2	73.0	77.1	85.7	23.5	23.5
0-Shot-CoT							
Llama2-70b-chat	3.5	69.7	64.3	77.1	80.0	11.3	13.9
GPT-4	10.4	75.1	73.0	74.3	82.9	25.2	34.8
TIMETOM							
Llama2-70b-chat	6.1 **(+6.1, +2.6)** **(× ∞, × 1.7)**	79.0 **(+72.5, +9.3)**	75.7 **(+67.0, +11.4)**	80.0 **(+80.0, +2.9)**	88.6 **(+82.9, +8.6)**	17.4 **(+13.1, +6.1)** **(× 4.0, × 1.5)**	15.7 **(+7.0, +1.8)** **(× 1.8, × 1.1)**
GPT-4	41.7 **(+33.0, +31.3)** **(× 4.8, × 4.0)**	93.0 **(+16.8, +17.9)**	93.1 **(+20.1, +20.1)**	94.3 **(+17.2, +20.0)**	91.5 **(+5.8, +8.6)**	51.3 **(+27.8, +26.1)** **(× 2.2, × 2.0)**	52.2 **(+28.7, +17.3)** **(× 2.2, × 1.5)**
Avg	**(+19.6, +17.0)**	**(+44.7, +13.6)**	**(+43.6, +15.8)**	**(+48.6, +11.5)**	**(+44.4, +8.6)**	**(+20.5, +16.1)**	**(+17.9, +9.6)**

Table 2: TIMETOM results on FanToM. We present **absolute accuracy difference** between TIMETOM and the baselines (0-shot and 0-shot-CoT), and **green** will appear on metrics related to the **robustness** of ToM reasoning. Results for list-type and binary-type questions can be found in Appendix D.2.

derstanding of object states and character actions, especially for those models with weaker cognitive abilities. We conduct experiments on the Llama2 series models in 0-shot and 0-shot with timeline settings, results show that the construction of the temporal space has led to significant performance improvements in true-belief and mem-real questions, which are associated with the real-world state. Moreover, the clear cognition brought by the temporal space is also helpful for reasoning about false belief questions. Detailed results data can be found in Appendix D.1. Case 1 in Figure 4 vividly illustrates the benefits of constructing temporal space.

Temporal Belief Construction and Compression Within the temporal space, we construct TBSC for the characters and utilize self-world belief, obtained through belief compression, to answer first-order ToM questions. By comparing the performance of the Llama2-7b-chat and Llama2-13b-chat models on first-order ToM questions in Tables 1 and 4, an observed improvement of **+3.75%** and **+2.75%** brought by belief construction and compression is noted. A larger improvement **+10.50%** and **+29.50%** appears in the GPT-3.5-turbo and GPT-4 models, as they inherently possess stronger language comprehension abilities, construct better TBSCs, perform more effective belief compression.

Tool—Time-Aware Belief Solver Within the temporal space, we design a novel tool—belief solver to inspire the reasoning process of LLMs on higher-order ToM questions. As shown in Table 3, the incorporation of belief solver results in a performance improvement of **+7.0%** and **+16.0%** for the GPT-3.5-turbo and GPT-4 models, respectively.

Model	Response	Result	Relation
GPT-3.5-turbo	Initial	65.0	Collaboration
	Tool	69.0	
	Final	72.0	+7.0
GPT-4	Initial	79.0	Tool Dominates
	Tool	96.0	
	Final	95.0	+16.0

Table 3: Comparison of initial, tool (belief solver as prompt), and final (belief solver as feedback) response performance of GPT series models on higher-order ToM questions under ToMI benchmark.

5 Analysis

In this section, we analyze belief solver as prompt vs feedback and the extension of TIMETOM to situations encompassing agent communication.

Belief Solver as Prompt vs Feedback Given that the belief solver can transform higher-order ToM questions into first-order ToM questions under belief communication periods, why don't we use it directly as a prompt when answering higher-order ToM questions? There are two reasons: (1) Given the GPT-4 model's exceptional language comprehension capabilities, it can construct accurate TBSC for each character. Consequently, it can accurately determine periods of belief communication through intersection operations of TBSC between characters. But as the model's ability decreases, e.g., for the GPT-3.5-turbo model, the

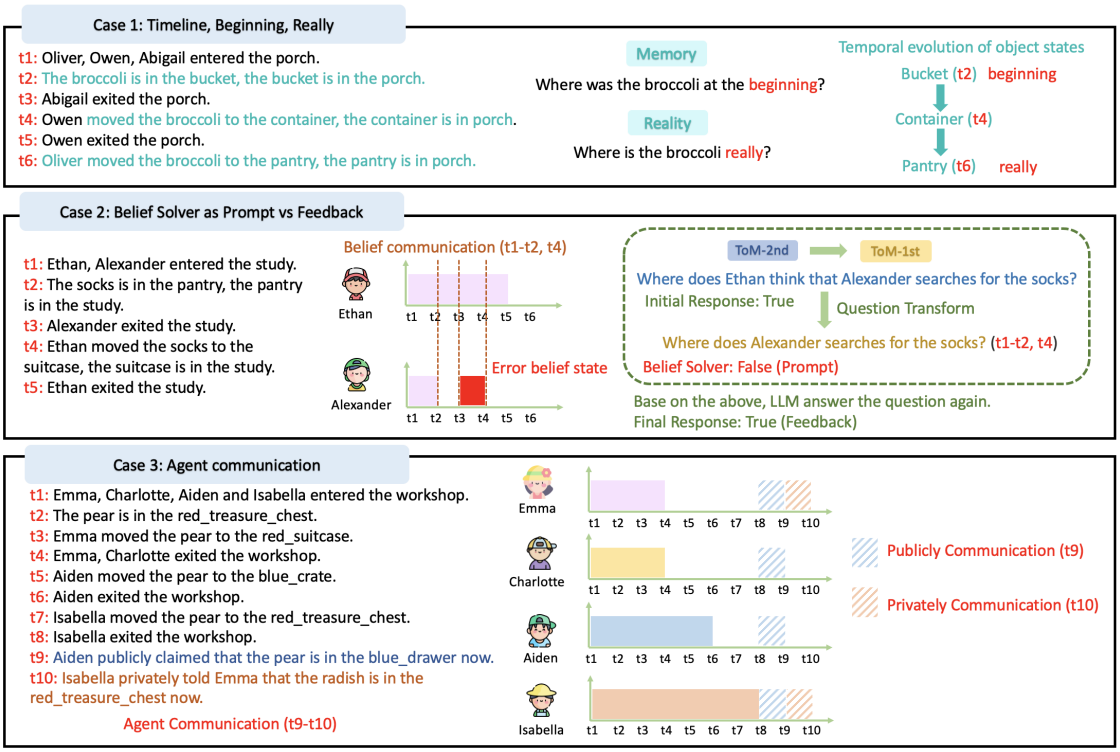

Figure 4: Case 1: The benefit of constructing temporal space. Case 2: The comparison between the belief solver as prompt and as feedback. Case 3: The application of TIMETOM in situations involving agent communication.

TBSC of the characters it builds will have a certain probability of error, and then the probability of obtaining an incorrect belief communication periods is greatly increased when performing the TBSC intersection operation between characters. (2) It is more effective to use the belief solver as feedback to inspire the reasoning process of LLMs on higher-order ToM questions. LLMs will consider the initial reasoning perspective and the reasoning perspective provided by the feedback to form a final response, and when the reasoning perspective provided by the feedback is accurate and effective, the LLMs also acknowledge this perspective, which corresponds to the GPT-4 model case. Conversely, if the reasoning perspective offered by the feedback is erroneous, the LLMs have a certain probability of recognizing this error and realizing the integration of useful information from both perspectives to achieve better performance, which corresponds to the GPT-3.5-turbo model case. Case 2 in Figure 4 and Table 3 offers qualitative and quantitative analysis support for the above two reasons.

Applicable to Situations Involving Communication between Agents. In real-world interactions, people engage in sharing their innermost thoughts with each other, including both their perceptions of situations and observations of others' behaviors. He et al. (2023) recently propose a benchmark encompassing agent communication, considering TIMETOM in this situation, which can model agent communication as a belief communication between agents at a specific time point. As shown in case 3 of Fig 4, TIMETOM has good applicability to situations involving agent communication.

6 Conclusion

In this paper, we propose TIMETOM to improve the ToM capabilities of LLMs in reading comprehension and interactive dialogue scenarios. Specifically, we first construct temporal space which serves as the foundation. Building on this, we develop several key components: character's belief state chain construction, social world model cognition-inspired belief compression, and tool—belief solver. Extensive experimental results show that TIMETOM substantially improves the reasoning performance of LLMs on ToM questions while making a significant advance towards coherent and robust ToM reasoning. The temporal space, serving as a key, unlocks the door to the LLMs' Theory of Mind. Furthermore, we find that TIMETOM can also be extended to situations involving agent communication which is commonly occurring in the real-world.

Limitations

There are two major limitations in TIMETOM. Firstly, the belief solver relies on constructing an accurate TBSC for characters. We conduct experiments on models with a parameter scale of 7B or larger. For models with less than 7B parameters, due to their relatively weaker instruction understanding ability, the error rate in constructing TBSCs is higher, which in turn affects the effectiveness of the belief solver. However, with the continuous development of LLMs, this limitation can be solved easily. Secondly, we focus on ToM reasoning for textual modality, it is also important to perform effective multimodal ToM reasoning, which we treat as future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akshatha Arodi and Jackie Chi Kit Cheung. 2021. Textual time travel: A temporally informed approach to theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4162–4172.

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

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2023. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*.

Kanishk Gandhi, Gala Stojnic, Brenden M Lake, and Moira R Dillon. 2021. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, 34:9963–9976.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Joy He-Yueya, Gabriel Poesia, Rose E Wang, and Noah D Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *arXiv preprint arXiv:2304.09102*.

Zhiting Hu and Tianmin Shu. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

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.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. *arXiv preprint arXiv:2310.19619*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Melanie Mitchell. 2023. Ai's challenge of understanding the world. *Science*, 382(6671):eadm8175.

Deepak Nathani, David Wang, Liangming Pan, and William Yang Wang. 2023. Maf: Multi-aspect feedback for improving reasoning in large language models. *arXiv preprint arXiv:2310.12426*.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. 2018. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*.

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

François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

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

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*.

Melanie Sclar, Graham Neubig, and Yonatan Bisk. 2022. Symmetric machine theory of mind. In *International Conference on Machine Learning*, pages 19450–19466. PMLR.

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*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Yutao Yue. 2022. A world-self model towards understanding intelligence. *IEEE Access*, 10:63034–63048.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Appendix

A Background and Related Work

Existing ToM Benchmarks Previous evaluations for the ToM of LLMs are primarily focused on testing models using situation descriptions (i.e., narratives) (Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2022; Shapira et al., 2023), also referred as reading comprehension scenarios. Recently, considering ToM capabilities play an even more important role in understanding dynamic social interactions, Kim et al. (2023) introduce FanToM, which tests models using interactive dialogues, also referred as dialogue scenarios.

LLMs Lack of ToM Capabilities Several studies (Gandhi et al., 2023; Sap et al., 2022; Kim et al., 2023; Wilf et al., 2023; Ullman, 2023) have shown that LLMs have poor reasoning performance and robustness on ToM tasks in a zero-shot setting, even with the current state-of-the-art GPT-4 (Achiam et al., 2023) model. With LLMs becoming increasingly integrated into our everyday lives, developing LLMs with ToM is very necessary.

Enhancing LLMs Reasoning Capabilities Recent prompt-based methods enhance the reasoning abilities of LLMs by guiding them to produce intermediate reasoning steps. For example, CoT (Wei et al., 2022) guides LLMs to generate step-by-step derivations before producing the final answer, LtM (Zhou et al., 2022) decomposes a target question into a series of subquestions, ToT (Yao et al., 2023) using tree-structured search to find better reasoning chains, Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) adds self-verification steps for rectifying reasoning errors. RAP (Hao et al., 2023; Hu and Shu, 2023) repurposes an LLM as a world model by prompting the LLM to predict the next state s_{t+1} of reasoning after applying a reasoning step a_t to the current state s_t. Apart from prompt-based methods, Declarative (He-Yueya et al., 2023) uses external symbolic solver to solve the equations in reasoning steps. MAF (Nathani et al., 2023) uses multiple external tools such as calculator, programming syntax to generate feedback to refine initial reasoning output.

Although many methods have been introduced to enhance the reasoning ability of LLMs, they are not suitable for ToM reasoning. Wilf et al. (2023) adopts the perspective-taking strategy to enhance the ToM reasoning abilities of LLMs, but this strategy falls short in addressing higher-order ToM reasoning. Sclar et al. (2023)[3] tracks each entity's beliefs and their estimation of other entities' beliefs, through graphical representations. However, it requires a substantial amount of external memory as well as being difficult to apply to context-rich ToM scenarios. Furthermore, there are currently no tools specifically dedicated to ToM reasoning.

B Benchmark Details and Evaluation Metrics

B.1 ToMI

ToMI (Le et al., 2019) is a benchmark in the reading comprehension scenarios, strictly imitating the Sally-Anne test, including the story, questions, and answer choices. The structure of the story is as follows:

1 Characters Sally and Anne are in a room.
2 Sally moves the celery from basket to box.
3 Anne can choose to leave the room before Sally moves the celery, or can choose to stay.
4 Anne will know the new location of the celery only if she is in the room while the celery is being moved.

Figure 5: Story structure of ToMI.

Five types of ToM questions are proposed: first-order or second-order, exploring characters' true or false beliefs (i.e., beliefs that are consistent or inconsistent with reality) as well as questions exploring reality and memory (zeroth-order ToM, (Sclar et al., 2022)). The formatted description for each type of question is as follows:

Reality: Where is celery really?

Memory: Where was celery at the beginning?

First-Order Belief Sally: Where will Sally look for celery?

First-Order Belief Anne: Where will Anne look for celery?

Second-Order Belief Sally: Where does Sally believe Anne will look for celery?

Second-Order Belief Anne: Where does Anne believe Sally will look for celery?

[3]This method works by explicitly memorizing beliefs of each character, rather than utilizing LLMs for ToM reasoning

Meanwhile, in **first-order belief** and **second-order belief** questions, both **false belief** and **true belief** are involved. For example: Sally moves the celery from basket to box without Anne observing this action. A **first-order belief** question: "Where will Anne look for celery?" Since Sally has moved the celery, Anne's belief will be incorrect – this type of question is called **false belief** and has its counterpart in **true belief** questions, where Anne's belief about the world is correct.

An updated version of ToMI proposed from (Arodi and Cheung (2021); Sap et al. (2022)) that has relabelled mislabelled second-order questions and disambiguated the location of containers after their reference. Sclar et al. (2023) expands the story structure by introducing more characters and containers.

All questions have two possible answers: the **original O location**, and the **final O location**. ToMI is a binary multiple-choice task, with random accuracy being 50%. In our experiments, we select the exact same test dataset used in SIMTOM (i.e., the updated version of ToMI) to ensure a fair comparison.

B.2 BigToM

BigToM (Gandhi et al., 2023) is another benchmark in the reading comprehension scenario and also follows the Sally-Anne test format. It is generated by GPT-4. Unlike ToMI, BigToM tells stories in more natural language and is not limited to changes in object locations. In our experiments, we focus on the "Forward belief" questions rather than "Backward belief" to align more closely with the structure of ToMI questions. The definitions of **true belief** and **false belief** questions in BigToM are the same as in ToMI. BigToM is also a binary multiple-choice task, with random accuracy being 50%.

B.3 FanToM

FanToM (Kim et al., 2023) is a benchmark in interactive dialogue scenarios. Dialogues involve information asymmetry (Quesque and Rossetti, 2020), with characters joining and leaving the dialogues while it continues, to simulate distinct mental states. The structure of dialogue is illustrated in Figure 6. The authors of FanToM formulate multiple types of questions (belief questions, answerability questions[list, binary], and infoaccess questions[list, binary]) that demand the same underlying reasoning to identify the illusory or false sense of ToM capabilities in LLMs:

First-Order Belief: What does Gianna believe about who discussed their experiences training their pets, Bruno and Snowflake?

Third-Order Acyclic Belief: What does Sara believe about Javier's belief regarding Gianna believe about who discussed their experiences training their pets, Bruno and Snowflake?

Third-Order Cyclic Belief: What does Sara believe about Javier's belief regarding Sara believe about who discussed their experiences training their pets, Bruno and Snowflake?

Answerability_List: Who discussed their experiences training their pets, Bruno and Snowflake? List all the characters who know the precise correct answer to this question.

Answerability_Binary Who discussed their experiences training their pets, Bruno and Snowflake? Does Gianna (Javier, Sara) know the precise correct answer to this question?

Infoaccess_List: Sara and Javier discussed their experiences training their pets, Bruno and Snowflake. List all the characters who know this information.

Infoaccess_Binary: Sara and Javier discussed their experiences training their pets, Bruno and Snowflake. Does Gianna (Javier, Sara) know this information?

In our experiments, we focus not only on the performance of our method in answering individual question types but also on its ability to reason coherently and robustly across multiple question types. All questions have two possible answers, with random accuracy being 50%.

B.4 Evaluation Metrics

Following Wilf et al. (2023), we report accuracy for all questions under ToMI and BigToM. Following Kim et al. (2023), for FanToM, we report accuracy for belief, answerability[list] and infoaccess[list] questions. The weighted F1 scores are reported for answerability[binary] and infoaccess[binary] questions. To evaluate the reasoning robustness of LLMs on ToM questions, we report the All score for answerability and infoaccess questions requiring models to be correct on both list-type and binary-type questions, the ALL* score which

1 Gianna: Guys, I've really enjoyed sharing our pet stories, but I need to excuse myself. I need to change clothes for a meeting later. Talk to you later!
2 Sara: Sure thing, Gianna. Take care!
3 Javier: Catch you later, Gianna.
4 Sara: So Javier, have you ever tried training Bruno?
5 Javier: Yes, I did actually. It was a challenge at times, but rewarding nevertheless. How about you? Did you try training Snowflake?
6 Sara: Oh gosh, trying to train a cat is a whole different ball game. But I did manage to teach her a few commands and tricks. She was quite an intelligent little furball.
7 Gianna: Hey guys, I'm back, couldn't miss out on more pet stories. Speaking of teaching and training pets, it is amazing how that further strengthens the bond between us and our pets, right?
8 Sara: Absolutely, Gianna! The fact that they trust us enough to learn from us is really special.
9 Javier: I can't agree more. I believe that's one of the ways Bruno conveyed his love and trust towards me. It also gave me a sense of responsibility towards him.
10 Gianna: Just like Chirpy. Once she began to imitate me, we connected in a way I never imagined. She would repeat words that I was studying for exams and that somehow made studying less stressful.
11 Javier: Pets are indeed lifesavers in so many ways.
12 Sara: They bring so much joy and laughter too into our lives. I mean, imagine a little kitten stuck in a vase! I couldn't have asked for a better stress buster during my college days.
13 Gianna: Totally, they all are so amazing in their unique ways. It's so nice to have these memories to look back on.

Figure 6: Dialogue structure of FanToM.

requires the models to answer all five ToM question types which require the same type of ToM reasoning.

C Our Prompt

C.1 Prompt for Interactive Dialogue Scenario

Constructing Temporal Space:
```
The following is a dialogue. Your task
is to add timeline to the dialogue.

Here are one rules: Each utterance
spoken by a character corresponds to a
moment t, Use \n as a delimiter, and the
 timeline is t1,t2,... ,tN.

Dialogue:
{dialogue}

Only output the dialogue content with
the added timeline, do not provide
explanations.
```

Temporal Belief State Chain Construction:
```
The following is a dialogue with a
timeline between multiple characters.
Your task is to only output the dialogue
 content on the timeline that the
character {character} can aware of.

Here are two rules:
If a character leaves the conversation
to do something else and then back after
 a few rounds of dialogue, they are
unaware of the content of the
conversation that took place during
their absence, but they aware of the
content of the conversation besides
their absence.
If a character don't leaves the
conversation to do something else and
then back after a few rounds of dialogue
. They are aware of all the content of
dialogue with all timeline.

Dialogue:
{dialogue}

What dialogue content on the timeline
does {character} aware of? Only output the
 dialogue content according to the above
 rules, do not provide an explanation.
```

Time-Aware Belief Question Answer with Belief Compression (First-order ToM questions):
```
The following is the belief states chain
 of character {name}. This is the content
 known to {name}:[{perspective}]
You are {name}.
Based on the above information, answer
the following question:
{question}
When answering questions, based on own
belief, simply focus on the information
of things asked in the question and
ignore other distracting factors. You
must choose one of the above choices.
```

Time-Aware Belief Question Answer without Belief Compression (Higher-order ToM questions):
```
The following is the belief states chain
 of character {name}. This is the content
 known to {name}:[{perspective}]
You are {name}.
Based on the above information, answer
the following question:
{question}
You must choose one of the above choices
.
```

Time-Aware Answerablity Question[List] Answer:
```
The following is the belief states chain
 of each character. This is the content
known to each character.
Each character only knows the contents
within their own belief state chain and
is unaware of the contents within the
belief state chain of other characters.
{final_text}
Question:
```

{target}
Based on the belief state chain of the above-mentioned characters, only output all the characters who know the precise correct answer to this question, do not provide an explanation.

Time-Aware Answerablity Question[Binary] Answer:
The following is the belief states chain of character {character}. This is the content known to {character}.
{binary_context}
Question:
{target}
Based on the belief state chain of character {character}, does {character} know the precise correct answer to this question? Answer yes or no. Answer:.

Time-Aware Infoaccess Question[List] Answer:
The following is the belief states chain of each character. This is the content known to each character.
Each character only knows the contents within their own belief state chain and is unaware of the contents within the belief state chain of other characters.
{final_text}
Target:
{target_q}
{target_a}
Question:
Based on the belief state chain of the above-mentioned characters, only output all the characters who know the target information, do not provide an explanation.

Time-Aware Infoaccess Question[Binary] Answer:
The following is the belief states chain of character {character}. This is the content known to {character}.
{binary_context}
Target:
{target_q}
{target_a}
Question:
Based on the belief state chain of character {character}, does {character} know the target information? Answer yes or no. Answer:.

Time-Aware Belief Solver:
The following is the belief states chain of character {name}. This is the content known to {name}:[{perspective}]
You are {name}.
Based on the above information, answer the following question:
{question}
Answer:{answer}
Feedback: The event corresponding to the period of belief communication between characters {character1}, {character2} and {character3}: {common_belief} Based on this information, the answer we get to the question:{question} is [{answer2}]
Considering this feedback, answer the question: {question} again. Keep your answer concise, one sentence is enough. You must choose one of the above choices .

C.2 Prompt for Reading Comprehension Scenario

Constructing Temporal Space:
The following is a story. Your task is to add timeline to the story.

Here are one rules: Each sentence corresponds to a moment t, Use \n as a delimiter, and the timeline is t1,t2,... ,tN.

Story:
{story}

Only output the story with the added timeline, do not provide explanations.

Temporal Belief State Chain Construction:
The following is a sequence of events with a timeline about some characters, that takes place in multiple locations. Your job is to output only the events on the timeline that character {character} can aware of.

Here are a few commonsense rules:
1. If a character is in a certain room/location, they will be aware of all other events happening in that room. This includes other characters entering or leaving the location, the locations of objects within it, and whether someone has moved an object to another location.
2. If a character leaves a location and is no longer there, they will no longer be aware of any events occurring at that location. However, they can re-enter the location.
3. A character is aware of all the events that they do.

Story:
{story}

What events on the timeline does {character} aware of? Only output the events according to the above rules, do not provide an explanation.

Time-Aware Belief Question Answer with Belief Compression (First-order ToM questions):
Belief Compression: The following is information from the perspective of the character, {character}.

Perspective:
{perspective}

Output the remaining perspective information after removing the events of characters enter or leave/exit the room /location, do not provide an explanation .

```
Time-Aware Belief Question Answer:
{perspective2}
You are {name}.
Based on the above information, answer
the following question:
{question}
Keep your answer concise, one sentence
is enough. You must choose one of the
above choices.
```

Time-Aware Belief Question Answer without Belief Compression (Higher-order ToM questions):

```
{perspective}
You are {name}.
Based on the above information, answer
the following question:
{question}
Keep your answer concise, one sentence
is enough. You must choose one of the
above choices.
```

Time-Aware Belief Solver:

```
Perspective1: {perspective}
You are {name}.
Based on the above information, answer
the following question:
{question}
Answer1:{answer}
Feedback Perspective2: The event
corresponding to the period of belief
communication between characters {
questionSubject} and {questionObject}: {
common_belief} Based on this information,
the answer we get to the question:{
question} is Answer2: {answer2}
Consider Perspective1, Feedback
Perspective2 and their answers, answer
the question: {question} again. Keep your
 answer concise, one sentence is enough.
 You must choose onea of the above
choices.
```

D Experiments

D.1 The Effect of Constructing Temporal Space

ToMI \ Model	L-7b	w/t	L-13b	w/t
Total	44.50	58.80 ↑14.30	51.00	60.90↑9.90
True-Belief	50.75	**73.00 ↑22.25**	50.25	60.00↑9.75
False-Belief	28.25	30.00 ↑1.75	39.25	52.00↑12.75
Mem-Real	64.50	**88.00 ↑23.50**	76.00	83.50 ↑7.50
First-Order	39.00	52.75 ↑13.75	54.75	58.50↑3.75
Second-Order	40.00	50.25 ↑10.25	34.75	52.00↑17.25

Table 4: Performance comparison of the Llama2 series models in 0-shot and 0-shot with timeline settings under ToMI benchmark.

D.2 Full Results

In Table 5 and 6, we present the full results of TIMETOM on the ToMI, BigToM, and FanToM benchmarks.

Model	ToMI						BigTOM	
	Total	True-Belief	False-Belief	Mem-Real	First-Order	Second-Order	True-Belief	False-Belief
0-Shot								
Llama2-7b-chat	44.50	50.75	28.25	64.50	39.00	40.00	51.50	53.50
Llama2-13b-chat	51.00	50.25	39.25	76.00	54.75	34.75	64.00	46.50
GPT-3.5-turbo	68.60	54.25	67.25	100.00	68.75	52.75	87.50	69.50
GPT-4	66.50	90.75	25.50	100.00	50.75	65.50	96.00	99.00
0-Shot-CoT								
Llama2-7b-chat	43.70	58.75	24.00	53.00	45.00	37.75	61.50	39.50
Llama2-13b-chat	45.00	63.50	16.50	65.00	43.00	37.00	62.00	52.50
GPT-3.5-turbo	64.10	77.50	34.00	97.50	58.50	53.00	90.00	71.50
GPT-4	74.40	61.75	74.25	100.00	73.75	62.25	96.50	99.00
SIMTOM								
Llama2-7b-chat	48.10	46.50	40.00	67.50	47.25	39.25	37.50	75.00
Llama2-13b-chat	61.10	72.00	35.50	90.50	53.75	53.75	53.00	62.50
GPT-3.5-turbo	72.80	51.00	81.00	100.00	74.75	57.25	90.00	78.00
GPT-4	87.80	81.75	87.75	100.00	93.75	75.75	94.00	98.00
TIMETOM								
Llama2-7b-chat	64.30 **(+19.80,+16.20)**	67.00 **(+16.25, +20.50)**	47.25 **(+19.00, +7.25)**	93.00 **(+28.50, +25.50)**	56.50 **(+17.50, +9.25)**	57.75 **(+17.75, +18.50)**	53.00 **(+1.50, +15.50)**	84.50 **(+31.00, +9.50)**
Llama2-13b-chat	67.20 **(+16.20, +6.10)**	73.50 **(+23.25, +1.50)**	44.75 **(+5.50, +9.25)**	99.50 **(+23.50, +9.00)**	61.25 **(+6.50, +7.50)**	57.00 **(+22.25, +3.25)**	66.00 **(+2.00, +13.00)**	89.50 **(+43.00, +27.00)**
GPT-3.5-turbo	80.80 **(+12.20, +8.00)**	70.00 **(+15.75, +19.00)**	82.00 **(+14.75, +1.00)**	100.00 **(+0.00, +0.00)**	80.50 **(+11.75, +5.75)**	71.50 **(+18.75, +14.25)**	91.50 **(+4.00, +1.50)**	96.00 **(+26.50, +18.00)**
GPT-4	96.00 **(+29.50, +8.20)**	90.75 **(+0.00, +9.00)**	98.75 **(+73.25, +11.00)**	100.00 **(+0.00, +0.00)**	95.50 **(+44.75, +1.75)**	94.50 **(+29.00, +18.75)**	95.00 **(-1.00, +1.00)**	99.00 **(+0.00, +1.00)**

Table 5: The full results of TIMETOM on the ToMI, BigToM benchmarks. Mem-Real can be viewed as zeroth-order ToM question.

Model	ALL* Question Types	Belief Questions				Answerability Questions			Infoaccess Questions		
		Overall	First-order	Third-acyc	Third-cyc	All	List	Binary	All	List	Binary
0-Shot											
Llama2-70b-chat	0.0	6.5	8.7	0.0	5.7	4.3	30.4	60.4	8.7	21.7	75.4
GPT-4	8.7	76.2	73.0	77.1	85.7	23.5	44.3	73.8	23.5	28.7	90.3
0-Shot-CoT											
Llama2-70b-chat	3.5	69.7	64.3	77.1	80.0	11.3	45.2	66.8	13.9	47.0	72.8
GPT-4	10.4	75.1	73.0	74.3	82.9	25.2	48.7	75.6	34.8	47.8	89.4
TIMETOM											
Llama2-70b-chat	6.1 **(+6.1, +2.6)** **(× ∞, × 1.7)**	79.0 **(+72.5, +9.3)**	75.7 **(+67.0, +11.4)**	80.0 **(+80.0, +2.9)**	88.6 **(+82.9, +8.6)**	17.4 **(+13.1, +6.1)** **(× 4.0, × 1.5)**	51.3 **(+20.9, +6.1)**	69.0 **(+8.6, +2.2)**	15.7 **(+7.0, +1.8)** **(× 1.8, × 1.1)**	60.0 **(+38.3, +13.0)**	68.2 **(-7.2, -4.6)**
GPT-4	41.7 **(+33.0, +31.3)** **(× 4.8, × 4.0)**	93.0 **(+16.8, +17.9)**	93.1 **(+20.1, +20.1)**	94.3 **(+17.2, +20.0)**	91.5 **(+5.8, +8.6)**	51.3 **(+27.8, +26.1)** **(× 2.2, × 2.0)**	62.6 **(+18.3, +13.9)**	90.7 **(+16.9, +15.1)**	52.2 **(+28.7, +17.3)** **(× 2.2, × 1.5)**	63.5 **(+34.8, +15.7)**	92.0 **(+1.7, +2.6)**

Table 6: The full results of TIMETOM on the FanToM benchmark.