AnaToM: A Dataset Generation Framework for Evaluating Theory of Mind Reasoning Toward the Anatomy of Difficulty through Structurally Controlled Story Generation

Jundai Suzuki **Ryoma Ishigaki** **Eisaku Maeda**
Tokyo Denki University
{24amj20@ms, 24amj02@ms, maeda.e@mail}.dendai.ac.jp

Abstract

Evaluating Theory of Mind (ToM) in Large Language Models (LLMs) is an important area of research for understanding the social intelligence of AI. Recent ToM benchmarks have made significant strides in enhancing the complexity, comprehensiveness, and practicality of evaluation. However, while the focus has been on constructing "more difficult" or "more comprehensive" tasks, there has been insufficient systematic analysis of the structural factors that inherently determine the difficulty of ToM reasoning—that is, "what" makes reasoning difficult. To address this challenge, we propose a new dataset generation framework for ToM evaluation named AnaToM. To realize an "Anatomy of Difficulty" in ToM reasoning, AnaToM strictly controls structural parameters such as the number of entities and the timeline in a story. This parameter control enables the isolation and identification of factors affecting the ToM of LLMs, allowing for a more precise examination of their reasoning mechanisms. The proposed framework provides a systematic methodology for diagnosing the limits of LLM reasoning abilities and offers new guidelines for future benchmark design.

1 Introduction

As AI agents become increasingly prevalent in society, their social intelligence, particularly their ability to smoothly interact and collaborate with humans, is of growing importance. At the core of this social intelligence is the ability to infer mental states such as intentions and beliefs from others' words and actions, known as "Theory of Mind" (ToM) (Premack and Woodruff, 1978). ToM has been extensively studied in cognitive science, and false belief tasks, in particular, have played a crucial role in understanding human ToM development and its neurological basis (Wimmer and Perner, 1983; Baron-Cohen et al., 1985).

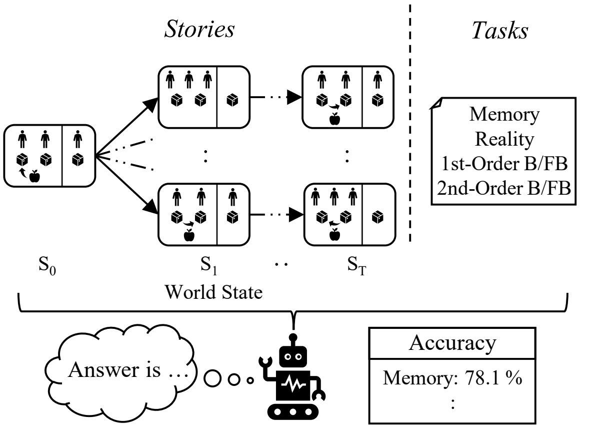

Figure 1: Overview of the proposed framework, AnaToM, which diagnoses ToM of LLMs using stories generated by parametrically controlling entity composition and timeline.

In recent years, this classical framework has been extended to evaluate the capabilities of Large Language Models (LLMs). There is an active debate as to whether LLMs possess ToM (Sap et al., 2022; Ullman, 2023; Ma et al., 2023; Sarıtaş et al., 2025), and evaluation methods have evolved rapidly. An early benchmark, ToMi (Le et al., 2019), highlighted the problem of pattern learning in template-based tasks, laying the foundation for subsequent research. Motivated by this issue, recent studies have pursued refinement through multifaceted approaches, including comprehensiveness (OpenToM (Xu et al., 2024), ToMBench (Chen et al., 2024)), higher-order reasoning (Hi-ToM (Wu et al., 2023)), conversational formats (ToMATO (Shinoda et al., 2025)), real-world data (CommonToM (Soubki et al., 2024)), and adversarial data generation by LLMs (ExploreToM (Sclar et al., 2025), BigToM (Gandhi et al., 2023)).

These prior studies have significantly contributed to the multifaceted evaluation of ToM in LLMs. However, their focus has primarily been on constructing "more difficult" or "more compre-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 244–257
December 20-24, 2025 ©2025 Association for Computational Linguistics

hensive" tasks, and a systematic analysis of the structural factors that inherently determine the difficulty of ToM reasoning—that is, what makes the reasoning difficult—is still insufficient. As a result, it has been difficult to clearly answer the question of which factors affect the belief-tracking ability of LLMs, and to what extent. This calls for an evaluation paradigm that can isolate the constituent elements of task difficulty and measure their individual effects.

To establish this new evaluation paradigm, our research proposes a new framework named AnaToM.[1] Its purpose is to deconstruct and analyze complex reasoning tasks into their fundamental structural factors, akin to an "Anatomy of Difficulty" in ToM reasoning. To achieve this goal, AnaToM enables the strict control of structural parameters considered to influence the cognitive load in LLM's ToM reasoning, such as entity composition and timelines (Figure 1). We deliberately employ template-based synthetic data generation to intentionally control for semantic variables, such as linguistic diversity, thereby isolating their confounding effects on reasoning difficulty. This approach is complementary to existing evaluations that use naturalistic datasets focused on ecological validity, and positions our framework as a "diagnostic tool" to identify which structural factors cause a model's failure. Through this approach, our research aims to establish a methodological foundation for precisely analyzing the limits of ToM in LLMs and the causes of their failures, thereby providing new guidelines for future benchmark design.

2 Related Works

2.1 The Current State of ToM Evaluation in LLMs

Early computational attempts to handle the "beliefs" of others involved formal inference using modal logic; however, its application was limited by issues such as logical omniscience and computational complexity (Isozaki and Katsuno, 1996). In contrast, the emergence of LLMs has established a new paradigm for evaluating ToM as a reading comprehension capability for natural language scenarios.

The evolution of benchmarks within this paradigm can be characterized by the ongoing challenge of mitigating "shortcut learning" in models. Early ToM benchmarks (Grant et al., 2017) and the ToM-bAbi benchmark (Nematzadeh et al., 2018), which introduced second-order beliefs, used templated stories based on the bAbi dataset (Weston et al., 2015). However, the simple structure of these benchmarks raised concerns that models could solve the tasks not through genuine ToM reasoning, but by learning superficial patterns like word co-occurrence. To address this issue, ToMi (Le et al., 2019) made it more difficult for models to rely on simple heuristics by randomizing story elements such as the timeline and the characters involved, thereby laying the foundation for subsequent research. Our framework, AnaToM, is inspired by and utilizes some structural components from the ToMi dataset, which is available under the MIT License.

2.2 Advancement and Diversification of ToM Benchmarks

Building on the challenges identified by ToMi, recent research has evolved through multifaceted approaches to enhance the validity and reliability of ToM evaluation. The first direction is the pursuit of comprehensiveness and practicality. As a preceding large-scale benchmark, SocialIQa (Sap et al., 2019) measures commonsense reasoning abilities about motivations and emotions in everyday social situations through crowdsourcing. More recently, OpenToM (Xu et al., 2024) and ToMBench (Chen et al., 2024), based on insights from psychology (Beaudoin et al., 2020), have enhanced evaluation comprehensiveness by including diverse mental states beyond belief. The second is an increase in complexity and realism. Hi-ToM (Wu et al., 2023) addresses higher-order belief reasoning, while ToMATO (Shinoda et al., 2025) introduced information asymmetry and character personalities within a conversational format. As another approach to pursuing scenario realism, some research addresses the issue of relying on synthetic data. Common-ToM (Soubki et al., 2024), the first attempt of its kind, evaluates ToM by tracking changes in common ground based on actual spoken dialogue data (Markowska et al., 2023), enabling the measurement of capabilities in more natural contexts. Furthermore, BigToM (Gandhi et al., 2023) and ExploreToM (Sclar et al., 2025) achieved more diverse and, at times, intentionally difficult adversarial evaluations by prompting LLMs to generate the stories themselves. Addi-

[1] The code and dataset for our framework are available under the MIT License at https://github.com/dsml-lab/AnaToM.

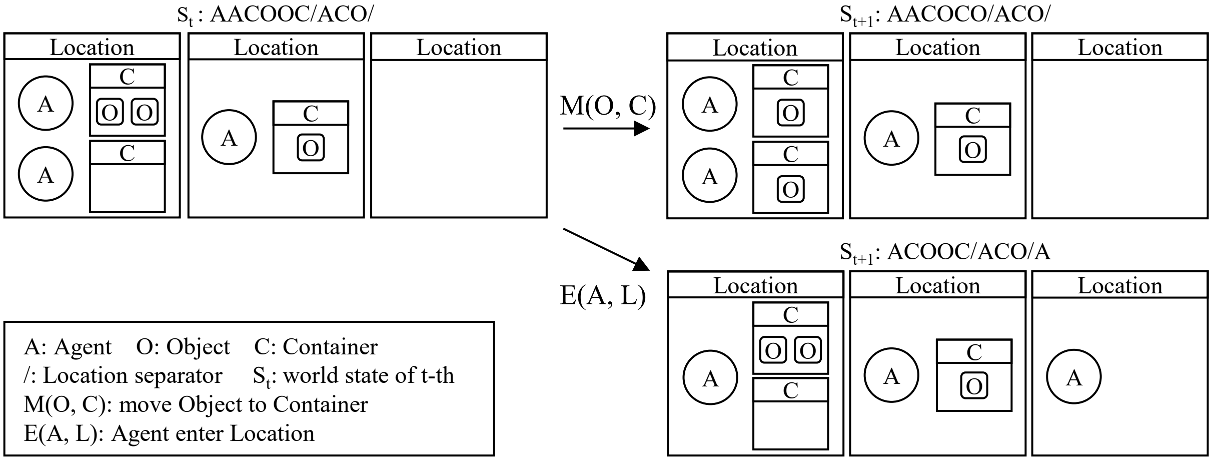
S_t : AACOOC/ACO/
Location
Location
Location
A
A
C
O O
C
A
C
O
M(O, C)
E(A, L)
S_t+1: AACOCO/ACO/
S_t+1: ACOOC/ACO/A
A: Agent O: Object C: Container
/: Location separator S_t: world state of t-th
M(O, C): move Object to Container
E(A, L): Agent enter Location

Figure 2: An example of state representation and action application in our framework. A state S_t is represented as the arrangement of entities (A: Agent, O: Object, C: Container) for each Location (separated by the delimiter /). It shows the transition from a state S_t to the next state S_{t+1} by applying an action $E(A, L)$ (Agent movement) or $M(O, C)$ (Object manipulation).

tionally, a method has been proposed to evaluate the ToM of LLMs through multi-agent cooperative tasks, focusing on dynamic belief states that are difficult to measure with static QA formats (Li et al., 2023). A third direction focuses on specific social reasoning abilities. Research such as FauxPas-EAI (Shapira et al., 2023) targets more specific and advanced social skills, such as the recognition of a "faux pas" (Baron-Cohen et al., 1999).

2.3 Relationship between Evaluation Methods and Model Capabilities

Regarding ToM in LLMs, findings vary depending on the evaluation method. While there are reports that recent large models show performance comparable to human children on classic false-belief tasks (Kosinski, 2024), it has also been pointed out that performance drops significantly with only slight alterations to the tasks (Ullman, 2023), leaving the question of whether models have acquired genuine ToM still under debate. Furthermore, it has been shown that prompting methods like Chain-of-Thought (Wei et al., 2023) significantly improve ToM performance (Moghaddam and Honey, 2023), suggesting that ToM in LLMs is highly context-dependent and rely heavily on the evaluation method.

Additionally, two-stage prompting frameworks like SimToM, which involve filtering the context from the character's perspective before answering the question, have also been proposed, reporting performance improvements over standard Chain-of-Thought (Wilf et al., 2024).

2.4 Positioning of This Research

While ToM evaluation for LLMs has progressed significantly toward greater comprehensiveness, complexity, and realism, the perspective of analyzing "structural factors" is not, in itself, unique to our work. For instance, LogicBench (Parmar et al., 2024) proposes a benchmark that systematically controls structural complexity to evaluate formal logical reasoning abilities in LLMs, spanning 25 distinct reasoning patterns across propositional, first-order, and non-monotonic logics.

However, the logical reasoning evaluated by LogicBench operates in an extensional context, dealing only with the objective truth values of propositions. In sharp contrast, the ToM reasoning addressed by our research requires an intentional context that models an agent's subjective mental state. For example, in a false-belief task, a statement about "Agent A's belief" can be true independently of the objective truth of that belief's content. Whereas the difficulty evaluated in LogicBench arises from the combinatorial complexity of logical rules, the difficulty AnaToM analyzes as an "Anatomy of Difficulty" arises from the cognitive load specific to ToM, such as managing multiple perspectives and tracking their dynamic changes.

This perspective—analyzing the "combinatorial complexity" specific to ToM, which differs fundamentally from the evaluation axis of formal logic like LogicBench—has been largely overlooked in the mainstream of ToM research. As a result, when a model fails, it has become difficult to disentangle whether the cause is a lack of higher-order social commonsense, such as understanding a Faux Pas, or a limitation in the fundamental combinatorial ability to track numerous entities and timelines.

3 AnaToM

In this work, we propose AnaToM, a novel benchmark generation framework that enables an "Anatomy of Difficulty" for ToM reasoning in LLMs. AnaToM aims to more precisely diagnose the reasoning capabilities of LLMs by explicitly treating the structural complexity of ToM tasks as parameters.

3.1 Design Principles

AnaToM is based on the following three design principles to overcome the challenges in diagnosability faced by prior work.

The first design principle is parametric control. This defines the structural elements of a story that govern its difficulty as independently manipulable continuous or discrete parameters, rather than viewing task difficulty as a "difficult/easy" binary. This makes it possible to identify how a model's performance changes in response to specific parameter values, such as thresholds where performance drops sharply.

The second principle is formal definition and deterministic generation. To avoid the inherent ambiguity, reproducibility issues, and potential "preference bias" associated with LLM-based scenario generation, our framework does not involve LLMs in the generation of the story's logical structure. Instead, the world "state," agent "actions," and their interactions are formally defined (Section 3.2.2). All stories are generated programmatically and deterministically based on these definitions, which ensures logical consistency, reproducibility, and a foundational absence of any LLM-induced generation bias.

The third principle is diagnosability, where each generated task is clearly linked to the combination of structural parameters that produced it. This allows for the identification of which structural factors caused a model's failure, based on its performance evaluation results.

3.2 Formal Definition of the Framework

The formal definitions of the world state and actions, which form the core of AnaToM, are presented below. This formalization ensures that all scenarios are constructed based on consistent rules.

3.2.1 World State Representation

In a ToM scenario, the world is defined by a set of entities and their attributes. Entities consist of four types: Agent ($AGT = A_1, \ldots, A_{N_A}$), Object ($OBJ = O_1, \ldots, O_{N_O}$), Container ($CON = C_1, \ldots, C_{N_C}$), and Location ($LOC = L_1, \ldots, L_{N_L}$), where the number of elements in each set is denoted by N_A, N_O, N_C, N_L. The relationships between these entities are governed by attributes. Agents and Containers have a Location as an attribute, represented as $loc(A)$ ($\in LOC$), $loc(C)$ ($\in LOC$). Objects have a Container as an attribute, represented as $con(O)$ ($\in CON$).

The world state S_t at a given time t is defined as the set of attribute values for all entities at that moment. A state is a "snapshot" that completely describes the physical arrangement of the world. An entire story is represented as a trajectory that begins from an initial state S_0 and transitions through the state space via a series of actions (S_0, S_1, ..., S_T) (Figure 2).

3.2.2 Definition of Actions

An action is defined as a deterministic function that transitions a state S_t to the next state S_{t+1}. In this framework, we define two types of actions: Enter/Exit and Move.

The Enter/Exit action $E(A, L)$ moves agent A from its current location $loc(A)$ to a different location L (where $L \neq loc(A)$), updating the value of the attribute $loc(A)$ to L. This action plays a crucial role in generating false belief scenarios by changing the perceptual state of agent A.

The Move action $M(O, C)$ involves an agent A moving an object O from its current container $con(O)$ to a different container C (where $C \neq con(O)$). As a precondition for physical plausibility, this action requires that the acting agent A, the source container $con(O)$ holding the target object, and the destination container C all exist in the same location ($loc(A) = loc(con(O)) = loc(C)$). This action updates the value of the attribute $con(O)$ to C, causing a central change in the world state that is the subject of the agents' beliefs.

3.2.3 State Transitions as a Markov Chain

The state space and the set of actions defined in Section 3.2.2 form the basis for modeling the dynamics of the entire system as a Markov chain. Specifically, each state S corresponds to a node in the Markov chain. By applying an action ACT

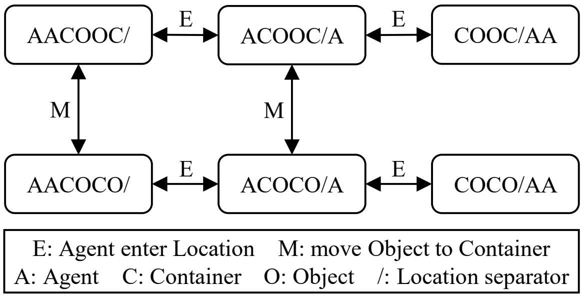

Figure 3: A Markov chain representation of the state space for the case where $N_A, N_O, N_C, N_L = 2$. Each state deterministically transitions via actions E (Enter/Exit) and M (Move).

(a general placeholder for any action from the set {E(Enter/Exit), M(Move)} defined in Section 3.2.2) from a state S the system deterministically transitions to the next state S_{t+1}.

This transition can be regarded as a directed edge connecting nodes within the state space (Figure 3). In this framework, story generation is defined as the process of selecting a specific path in this state space, starting from an initial state S_0. The timeline composition parameters determine the length and shape of this path. This formalization enables the analysis of a story's structural properties, such as the reachability of states where a false belief occurs, or the minimum number of timeline events required to perform a specific inference.

3.3 Structural Parameter Space

AnaToM defines the difficulty of a story as a parameter space consisting of the following three aspects. This is the most important contribution of our framework, as it enables the "deconstruction" and reconstruction of difficulty.

First is entity composition, which determines the static complexity of the story. The parameters are the sizes of each entity set: N_A, N_O, N_C, N_L These directly govern the total amount of information that a model must track and manage.

Second is combinatorial complexity. The number of entities determines the number of possible initial state combinations, which increases non-linearly. This number of combinations can be modeled as a "balls and bins problem," where n indistinguishable balls are placed into m indistinguishable bins. For example, the number of combinations for placing N_C containers into N_L locations can be formulated as a function $f(N_L, N_C)$. As a concrete example, in the base configuration used in this study, where $N_A, N_O, N_C, N_L = 3$, the total number of initial entity placement patterns is 97. This parameter is a potential difficulty factor in how LLMs comprehend the initial state, and our framework makes it possible to analyze the impact of this combinatorial complexity on a model's reasoning.

Third is timeline composition, which determines the dynamic complexity of the story. This includes two parameters. One is the total number of actions T that occur in the story; this value is related to the memory load required for a model to continuously update beliefs. The other is the sequence of actions $ACTs = (ACT_1, ACT_2, \ldots, ACT_T)$. Even with the same set of actions, a different sequence can result in a completely different final belief state for an agent. In particular, the relative order of object movements and agent movements is a crucial parameter that determines whether a false belief is formed.

3.4 Generation Process

AnaToM generates ToM evaluation datasets through a five-step process based on the defined parameters. The process begins by setting specific conditions from the parameter space defined in Section 3.3, corresponding to the hypothesis being tested (e.g., to measure the ability to track agent A, only N_A is varied from 2 to 5 while other parameters are held constant). Next, based on the set entity composition, a specific initial state S_0 is sampled from the combinatorial space. Subsequently, based on the set timeline, the sequence of actions defined in Section 3.2.2 is applied, and the world state transitions deterministically from S_0 to S_T. Afterward, this formal sequence of state transitions $(S_0, S_1, \ldots, S_T)$ is converted into a natural language story using predefined templates. Finally, based on the final state S_T and the perceptual information of each agent, tasks for tracking facts in the story (Memory, Reality) as well as belief and false belief tasks (1st/2nd-order belief/false belief) are automatically generated. For example, a false belief task concerning a certain agent A is generated only if that agent perceived the initial move of an object O but did not perceive its final move.

AnaToM makes it possible to systematically and quantitatively evaluate the impact of structural factors, such as an increase in the number of agents or changes in the timeline, on the ToM reasoning of LLMs.

Table 1: ToM evaluation results from AnaToM. Values are accuracy (%). Model abbreviations: 8B (Llama-3-8B-Instruct), 70B (Llama-3-70B-Instruct), 4o-m (GPT-4o-mini), 4.1-m (GPT-4.1-mini). GPT scores (4o-m, 4.1-m) are the average ± standard deviation of 3 runs. Llama scores were identical across all 3 runs for 8B and all 2 runs for 70B.

	Llama-3		**GPT**	
task type	**8B**	**70B**	**4o-m**	**4.1-m**
Memory	70.8	92.9	99.7 ±0.0	99.9 ±0.0
Reality	97.1	88.3	99.3 ±0.1	99.7 ±0.1
1st-order true belief	45.1	52.0	57.4 ±0.1	82.9 ±0.2
1st-order false belief	40.4	27.5	29.7 ±0.2	74.1 ±0.4
2nd-order true belief	27.2	35.6	42.6 ±0.3	69.4 ±0.2
2nd-order false belief	35.9	55.4	3.7 ±0.1	54.6 ±0.2
overall accuracy	**52.8**	**58.6**	**55.4** ±0.1	**80.1** ±0.1

4 Experiments

4.1 Dataset Construction Process

This experiment aims to precisely diagnose the limits of ToM in LLMs, for which we constructed an evaluation dataset by applying AnaToM. In this construction, we did not merely generate stories, but applied specific procedures and filtering criteria to extract instances particularly useful for probing the limits of LLM reasoning capabilities.

First, as structural parameters, we defined seven settings for the entity composition parameters (N_A, N_O, N_C), varying each from 3 to 5 (e.g., $N_A = 4, N_O = 3, N_C = 3$), while the number of locations N_L was fixed at 3. For the timeline, the total number of actions T was fixed at 4, and we limited the generation to five specific action sequence patterns that can produce false beliefs (e.g., Move → Exit/Enter → Move → Exit/Enter). Based on these parameters, initial states were sampled from the combinatorial space described in Section 3.3. The initial belief of each agent was then initialized based on a perception-based rule, assuming they fully grasp the correct location of any object present in the same location as themselves.

Next, during the story generation process, we introduced a simple lookahead heuristic for action selection to avoid "dead-end" states where subsequent actions would become physically impossible. This ensured that only stories that could be completed to the end of the sequence were generated. From the pool of candidate stories for each setting obtained through this process, we finally randomly sampled 1,000 instances to be used as the evaluation dataset for this experiment.

4.2 Evaluated Models

In this experiment, we evaluated several representative, widely used LLMs. The selection was based on diversity in model architecture, parameter size, and developer. As a representative of open-source models, we selected Llama-3-8B-Instruct (Grattafiori et al., 2024), an 8-billion parameter, instruction-tuned model from the Llama-3 family developed by Meta. To analyze the impact of model scale on ToM reasoning, we also included its larger family member, Llama-3-70B-Instruct (Grattafiori et al., 2024). Additionally, to evaluate the performance of a model family considered to have state-of-the-art reasoning capabilities, we selected the high-performance commercial models GPT-4o-mini and GPT-4.1-mini, which are based on the GPT-4 architecture developed by OpenAI, allowing for a comparison between versions. All evaluations in this study were conducted via the official APIs for each model, using settings that facilitate deterministic outputs (e.g., temperature = 0.0).

4.3 Evaluation Metrics

The evaluation metric is accuracy. Accuracy is calculated for each task type (Memory, Reality, 1st/2nd-order belief/false belief) and for each parameter setting. This allows for a detailed analysis of which structural factors affect which types of reasoning, and to what extent.

Table 2: Accuracy (%) for each entity composition, grouped by model family. Abbreviations: 8B (Llama-3-8B-Instruct), 70B (Llama-3-70B-Instruct), 4o-m (GPT-4o-mini), 4.1-m (GPT-4.1-mini). Scores for Llama models (8B, 70B) were identical across all 3 runs (8B) and 2 runs (70B), respectively. Scores for GPT models (4o-m, 4.1-m) are the average of 3 runs, and the Coefficient of Variation is at most 1.3%.

	1st-order true belief				1st-order false belief				2nd-order true belief				2nd-order false belief			
	Llama-3		GPT		Llama-3		GPT		Llama-3		GPT		Llama-3		GPT	
setting	**8B**	**70B**	**4o-m**	**4.1-m**	**8B**	**70B**	**4o-m**	**4.1-m**	**8B**	**70B**	**4o-m**	**4.1-m**	**8B**	**70B**	**4o-m**	**4.1-m**
A3_O3_C3 (base)	47.3	59.5	64.3	85.4	42.5	29.8	33.2	77.0	27.3	38.7	43.6	72.0	41.0	56.1	3.3	57.6
agents +																
A4_O3_C3	40.1	45.9	52.2	78.5	37.6	24.9	26.0	70.3	21.2	28.1	41.8	68.1	35.3	54.3	4.7	52.3
A5_O3_C3	37.2	41.9	46.8	78.2	34.6	22.6	21.2	63.8	20.8	24.1	36.4	63.0	29.0	46.6	4.3	51.3
objects +																
A3_O4_C3	44.2	53.0	55.7	85.6	41.3	29.1	33.3	78.0	26.6	38.3	45.6	73.6	38.3	57.7	2.6	56.7
A3_O5_C3	45.8	47.9	53.4	86.0	42.0	26.6	26.6	76.8	28.5	37.1	45.6	75.5	38.6	55.5	2.4	52.3
containers +																
A3_O3_C4	47.7	54.5	63.1	83.5	41.2	29.8	32.7	76.4	31.2	39.8	42.3	66.1	34.7	58.6	3.8	57.0
A3_O3_C5	53.7	61.6	66.2	83.0	43.7	29.8	34.9	76.8	34.9	42.9	40.4	67.8	34.6	59.1	4.5	54.9

Table 3: Accuracy (%) for each event sequence, grouped by model family. Abbreviations: 8B (Llama-3-8B-Instruct), 70B (Llama-3-70B-Instruct), 4o-m (GPT-4o-mini), 4.1-m (GPT-4.1-mini). Scores for Llama models (8B, 70B) were identical across all 3 runs (8B) and 2 runs (70B), respectively. Scores for GPT models (4o-m, 4.1-m) are the average of 3 runs, and the Coefficient of Variation is at most 1.3%.

	1st-order true belief				1st-order false belief				2nd-order true belief				2nd-order false belief			
	Llama-3		GPT		Llama-3		GPT		Llama-3		GPT		Llama-3		GPT	
timeline	**8B**	**70B**	**4o-m**	**4.1-m**	**8B**	**70B**	**4o-m**	**4.1-m**	**8B**	**70B**	**4o-m**	**4.1-m**	**8B**	**70B**	**4o-m**	**4.1-m**
E -> E -> M -> M	48.8	50.4	58.7	78.7	21.8	4.7	21.5	82.0	20.3	22.8	38.9	70.3	22.3	39.3	0.5	55.9
E -> M -> E -> M	41.0	51.4	53.6	84.8	49.8	41.8	43.1	73.9	34.3	44.1	43.9	67.5	39.8	63.6	9.0	50.8
E -> M -> M -> E	50.4	57.6	59.2	87.5	26.7	10.1	15.2	77.7	21.2	31.8	31.1	79.6	22.0	37.2	0.9	50.4
M -> E -> E -> M	44.8	50.1	59.9	78.8	49.3	39.2	37.8	77.5	29.8	38.2	51.6	64.0	46.8	72.6	3.8	68.1
M -> E -> M -> E	41.0	50.8	55.3	84.4	53.4	40.6	30.6	60.5	30.2	40.5	47.1	65.9	47.7	63.6	4.0	48.1

5 Results

In this section, we report the results of experiments conducted using the dataset constructed in Section 4 and analyze the impact of structural factors on the ToM reasoning of LLMs.

5.1 Overall Performance Overview

First, we provide an overview of the overall performance of the models evaluated in this experiment. Table 1 summarizes the average accuracy for each task type across the entire dataset. Several important trends can be observed from Table 1. First, GPT-4o-mini and GPT-4.1-mini achieved high accuracy rates exceeding 99% on the fact-tracking tasks, Memory and Reality. However, the performance of these same models dropped significantly on first- and second-order belief and false belief tasks. Notably, GPT-4o-mini showed extremely low performance on second-order false belief tasks at 3.7%, confirming a significant gap between its fact-tracking ability and its ToM reasoning capabilities. Llama-3-8B-Instruct performed lower than the GPT-based models even on the fact-tracking tasks and tended to struggle further with the belief tasks. Meanwhile, Llama-3-70B-Instruct, with its scaled-up parameter size, showed improved performance over the 8B model in terms of overall accuracy and on higher-order belief reasoning (Table1). However, its scores on the Reality and 1st-order false belief tasks decreased, suggesting that simply scaling up the model does not lead to uniform improvements across all ToM-related capabilities.

5.2 Analysis of the Impact of Structural Factors

Next, we analyze in detail the impact of the structural factors that determine the difficulty of ToM reasoning. This section focuses on the effects of entity composition and timeline composition on performance in the particularly challenging belief and false belief tasks.

5.2.1 Impact of Entity Composition

Table 2 shows the accuracy rates on first- and second-order belief and false belief tasks when the number of each entity (Agent, Object, Container) was individually increased from the baseline setting $(N_A, N_O, N_C = 3)$.

The most notable point from these results is that the increase in the number of agents (N_A) degrades performance more consistently than any other factor. For example, in GPT-4.1-mini, when the number of agents increased from 3 to 5, the accuracy on second-order false belief tasks dropped from 57.6% to 51.3%. This trend was similarly observed in the scaled-up Llama-3-70B (dropping from 56.1% to 46.6%), suggesting that an increased number of agents constitutes a common cognitive load for many models. However, an exception was noted for GPT-4o-mini on the 2nd-order false belief task; with its extremely low baseline accuracy of 3.3%, a clear degradation trend was not observed. On the other hand, the performance degradation from an increase in the number of objects (N_O) or containers (N_C) was limited in comparison.

5.2.2 Impact of Timeline Composition

Table 3 shows the accuracy on false belief tasks for each of the five different action sequence (timeline) patterns. The timeline sequence was also confirmed to be a factor that significantly affects task difficulty. In particular, in sequences such as E -> M -> M -> E and E -> E -> M -> M, the accuracy of GPT-4o-mini on second-order false belief tasks plummeted to below 1%, confirming that specific timelines have a catastrophic impact on the model's performance. Llama-3-70B also exhibited its lowest performance on these same sequences (37.2% and 39.3% respectively), revealing that vulnerability to specific timeline structures persists even with increased model size.

6 Discussion

In this section, we discuss new insights into the ToM reasoning of LLMs based on the experimental results obtained in Section 5. We also describe the implications of our research and future prospects.

6.1 Principal Findings of This Research

The experiments in this study demonstrated that AnaToM is effective for precisely diagnosing the ToM reasoning capabilities of LLMs, yielding three primary findings.

First, a clear disparity was observed across all evaluated models between their performance on fact-tracking tasks (Memory, Reality) and their performance on belief-reasoning tasks that require inferring others' mental states (especially false beliefs). This suggests that even for high-performing models, fact-tracking ability does not directly translate to advanced ToM reasoning.

Second, among the structural factors that determine the difficulty of ToM reasoning, the number of agents, N_A, was found to have the most dominant impact. Compared to an increase in the number of objects (N_O) or containers (N_C), an increase in the number of agents consistently caused the most significant performance degradation, regardless of the model type or the order of the task.

Third, it was found that the impact of the timeline on performance differs qualitatively depending on the model's capability level. Less capable models showed a significant performance drop in sequences where state changes occurred consecutively while an agent was absent. On the other hand, the most high-performing model, while handling such sequences, faced relative difficulty with different, more complex sequences in which the types of actions frequently alternated.

6.2 Identifying Cognitive Bottlenecks in LLM's ToM

These findings indicate the existence of multiple cognitive bottlenecks in the ToM reasoning of LLMs.

The first bottleneck is the limitation of multi-agent tracking capabilities. The finding that an increase in the number of agents is a primary factor in performance degradation indicates that a fundamental constraint in the ToM of LLMs lies in the ability to simultaneously track and manage multiple perspectives. Objects and containers, which constitute the physical state of the world, are passive elements belonging to a single "reality" that the model must track. In contrast, agents are active entities, each capable of holding their own unique beliefs, and as their number increases, the model must manage and update multiple different mental states in parallel. We posit that this increase in cognitive load exposes a fundamental limitation of current LLM architectures.

In addition, the different responses among models to the timeline composition reveal a second bottleneck: a hierarchy in the ability to maintain beliefs in a dynamic context. The sequences that Llama-3-8B, Llama-3-70B, and GPT-4o-mini struggled with were those in which Move actions occurred consecutively, such as E -> M -> M -> E and E -> E -> M -> M. Whereas an Exit/Enter action updates the location information of a single agent, a Move action is a more complex operation that changes the state of an object and simultaneously requires updating the belief states of all agents present in that location. Therefore, this result suggests that models have a fundamental challenge in their ability to accurately integrate information when complex actions that have a compound effect on belief states occur consecutively.

On the other hand, the highest-performing model, GPT-4.1-mini, showed a different pattern of difficulty. While GPT-4.1-mini processed sequences that other models struggled with, such as E -> E -> M -> M, with a high accuracy of 81.1%, its performance dropped the most on the M -> E -> M -> E sequence where Move and Exit/Enter events alternate, with an accuracy of 60.4%. The reasoning required by this sequence is more complex than mere information integration. In this sequence, the perceptual state of a certain agent (let's call it B) frequently changes from "present (perceivable)" -> "absent (imperceptible)" -> "present again (perceivable again)." To handle this dynamic change, the model needs to flexibly switch the rules of inference on which information to base its belief updates. For example, it must treat "directly perceived information" as the source of belief while the agent is present, and "the last seen memory" as the source while the agent is absent. The fact that GPT-4.1-mini faced relative difficulty with this sequence suggests that while the model has, to some extent, overcome the bottleneck of simple state change integration, it faces a new challenge in a more advanced capability: the switching and management of the reasoning process according to the situation. In other words, this indicates that as a model's capabilities improve, the nature of the difficulties it faces shifts to a higher level of complexity.

7 Conclusion

In this work, we addressed the problem that the structural factors defining the difficulty of ToM evaluation in LLMs have not been sufficiently analyzed. To tackle this issue, we proposed AnaToM, a novel benchmark generation framework that enables an "Anatomy of Difficulty" for ToM reasoning. AnaToM strictly controls structural parameters such as the number of entities and the timeline in a story.

Experiments using this framework revealed that a major bottleneck in the ToM reasoning of LLMs lies in the limits of multi-agent tracking capabilities, which stems from the number of agents to be tracked. Furthermore, it was shown that the nature of timelines perceived as difficult changes qualitatively depending on the model's capability level.

The approach proposed in our research promotes the elucidation of LLM reasoning mechanisms and presents a new path toward a deeper understanding of their capabilities and limitations. The findings of this research, which analyzes the difficulty of ToM by deconstructing it into its constituent components, are expected to contribute to the development of more advanced and robust AI with social intelligence.

Ethical Consideration

This research, which analyzes ToM in LLMs, involves several ethical considerations.

First is the risk of misunderstanding model capabilities and of anthropomorphism. This study analyzes specific text-processing patterns in LLMs and does not suggest that the models possess mental states equivalent to those of humans.

Second is the potential for misuse of the research. The findings from this study carry a risk of being applied to the development of adversarial attacks that intentionally manipulate models, or to applications that deceive users.

All data used in this study are synthetic data and does not infringe on personal privacy.

Limitations

This study has several limitations, which suggest directions for future research.

First, this research focuses on "belief" reasoning within ToM and limits actions to deterministic physical movements. This was a deliberate design choice, serving as a first step in our novel structural analysis approach to ToM evaluation, intended to ensure maximum control and reproducibility. As a result, this framework does not address the broader range of mental states integral to real-world social situations, such as intentions, desires, and emotions, nor does it capture more ambiguous, probabilistic interactions. Extending the framework to these aspects is a crucial future direction that can build upon the methodological foundation established here.

Second, the use of templates in the natural language realization of stories restricts linguistic diversity. This represents a key design trade-off necessary to achieve our research goal of an “Anatomy of Difficulty”. By intentionally excluding semantic variables like linguistic diversity, we guarantee that performance differences can be purely attributed to the structural factors being analyzed, such as the number of agents or the timeline. Consequently, this framework is not intended to measure ecological validity, and model robustness to more natural and diverse linguistic expressions is not evaluated in this study.

Finally, the models evaluated in this experiment were limited to four types, and the exploration of the parameter space was not exhaustive. Further experiments with a wider range of models are desirable to confirm whether the observed trends can be generalized to a broader class of LLMs.

References

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

Simon Baron-Cohen, Michelle O’Riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29(5):407–418.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. 2020. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Kanishk Gandhi, J.-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA. Curran Associates Inc.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How Can Memory-Augmented Neural Networks Pass a False-Belief Task? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39, pages 427–432.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. The Llama 3 Herd of Models. *arXiv preprint* arXiv:2407.21783. Version 3.

Hideki Isozaki and Hirofumi Katsuno. 1996. A semantic characterization of an algorithm for estimating others’ beliefs from observation. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96, Portland, Oregon, USA, August 4-8, 1996, Volume 1*, pages 543–549. AAAI Press / The MIT Press.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the Evaluation of Theory of Mind through Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding Common Ground: Annotating and Predicting Common Ground in Spoken Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. *arXiv preprint* arXiv:2304.11490. Version 3.

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

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

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

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Karahan Sarıtaş, Kıvanç Tezören, and Yavuz Durmazkeser. 2025. A Systematic Review on the Evaluation of Large Language Models in Theory of Mind Tasks. *arXiv preprint* arXiv:2502.08796. Version 1.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2025. Explore Theory of Mind: program-guided adversarial data generation for theory of mind reasoning. In *International Conference on Representation Learning*, volume 2025, pages 67635–67660.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How Well Do Large Language Models Perform on Faux Pas Tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. ToMATO: Verbalizing the Mental States of Role-Playing LLMs for Benchmarking Theory of Mind. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 1520–1528. AAAI Press.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2024. Views Are My Own, but Also Yours: Benchmarking Theory of Mind Using Common Ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14815–14823, Bangkok, Thailand. Association for Computational Linguistics.

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint* arXiv:2302.08399. Version 5.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint* arXiv:2201.11903. Version 6.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv preprint* arXiv:1502.05698. Version 10.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

A Example Appendix

A.1 Full List of Initial Placement Patterns

The following is a complete list of the 97 combinatorially possible initial placement patterns for the baseline entity composition ($N_A, N_O, N_C, N_L = 3$) used in this study. Each string represents the placement of Agents A, Containers C, and Objects O into three locations (delimited by /).

```
AAA// (1/3)
AAACOOOCC//, AAA/COOOCC/
AAACOOOC/C/, AAAC/COOOC/, AAA/COOOC/C
AAACOOO/CC/, AAACC/COOO/, AAA/COOO/CC
AAACOOO/C/C, AAAC/COOO/C
AAACOOCOC//, AAA/COOCOC/
AAACOOCO/C/, AAAC/COOCO/, AAA/COOCO/C
AAACOOC/CO/, AAACO/COOC/, AAA/COOC/CO
AAACOO/COC/, AAACOC/COO/, AAA/COO/COC
AAACOO/CO/C, AAACO/COO/C, AAAC/COO/CO
AAACOCOCO//, AAA/COCOCO/
AAACOCO/CO/, AAACO/COCO/, AAA/COCO/CO
AAACO/CO/CO
```

```
AA/A/ (2/3)
AACOOOCC/A/, AA/ACOOOCC/, AA/A/COOOCC
AACOOOC/AC/, AACOOOC/A/C, AAC/ACOOOC/,
AAC/A/COOOC
```

```
AA/ACOOOC/C, AA/AC/COOOC
AACOOO/ACC/, AACOOO/A/CC, AACC/ACOOO/,
AACC/A/COOO
AA/ACOOO/CC, AA/ACC/COOO
AACOOO/AC/C, AAC/ACOOO/C, AAC/AC/COOO
AACOOCOC/A/, AA/ACOOCOC/, AA/A/COOCOC
AACOOCO/AC/, AACOOCO/A/C, AAC/ACOOCO/,
AAC/A/COOCO
AA/ACOOCO/C, AA/AC/COOCO
AACOOC/ACO/, AACOOC/A/CO, AACO/ACOOC/,
AACO/A/COOC
AA/ACOOC/CO, AA/ACO/COOC
AACOO/ACOC/, AACOO/A/COC, AACOC/ACOO/,
AACOC/A/COO
AA/ACOO/COC, AA/ACOC/COO
AACOO/ACO/C, AACOO/AC/CO, AACO/ACOO/C,
AACO/AC/COO
AAC/ACOO/CO, AAC/ACO/COO
AACOCOCO/A/, AA/ACOCOCO/, AA/A/COCOCO
AACOCO/ACO/, AACOCO/A/CO, AACO/ACOCO/,
AACO/A/COCO
AA/ACOCO/CO, AA/ACO/COCO
AACO/ACO/CO
```

```
A/A/A (3/3)
ACOOOCC/A/A
ACOOOC/AC/A
ACOOO/ACC/A
ACOOO/AC/AC
ACOOCOC/A/A
ACOOCO/AC/A
ACOOC/ACO/A
ACOO/ACOC/A, ACOO/ACO/AC
ACOCOCO/A/A
ACOCO/ACO/A
ACO/ACO/ACO
```

A.2 Dataset Distribution by Experimental Setting

This section clarifies the distribution of task types for each experimental setting. The dataset consists of 7,000 unique stories (instances) in total. Each story is associated with exactly one question for each of the 6 task types (Memory, Reality, 1st-order true/false, 2nd-order true/false), totaling 6 questions per story.

For the entity composition analysis (Table 4), these 7,000 stories are evenly divided into the 7 settings (e.g., A3_O3_C3, A4_O3_C3), resulting in 1,000 stories per setting. Consequently, the total number of questions for each setting is exactly 6,000 (1,000 stories × 6 task types), demonstrating that the task distribution is perfectly balanced across these conditions.

For the timeline analysis (Table 5), the same 7,000 stories are classified into 5 timeline settings. In this case, the stories are not evenly distributed (e.g., 1,357 stories for E->E->M->M). Therefore, the total number of questions for each timeline setting varies (e.g., 1,357 stories × 6 task types = 8,142 questions).

As both tables show, the distribution of task types is perfectly balanced within every single experimental setting (e.g., all 1,000 questions for A4_O3_C3 are split evenly across the 6 task types). This ensures that no bias is introduced by the task distribution.

Furthermore, a detailed breakdown of the 1,000 stories within the A3_O3_C3 (base) setting, categorized by their specific child patterns, is provided in Table 6.

Table 4: Distribution of question counts for each entity composition setting. This table demonstrates that the distribution of task types is perfectly balanced across all settings analyzed in Table 2 (1,000 instances per task type for each setting).

setting	Memory	Reality	1st-order true	1st-order false	2nd-order true	2nd-order false	total questions
A3_O3_C3 (base)	1000	1000	1000	1000	1000	1000	6000
A4_O3_C3 (agents +)	1000	1000	1000	1000	1000	1000	6000
A5_O3_C3 (agents +)	1000	1000	1000	1000	1000	1000	6000
A3_O4_C3 (objects +)	1000	1000	1000	1000	1000	1000	6000
A3_O5_C3 (objects +)	1000	1000	1000	1000	1000	1000	6000
A3_O3_C4 (containers +)	1000	1000	1000	1000	1000	1000	6000
A3_O3_C5 (containers +)	1000	1000	1000	1000	1000	1000	6000

Table 5: Distribution of question counts for each timeline setting. This demonstrates that task types are perfectly balanced within each timeline, and the total number of questions is reasonably balanced across timeline conditions analyzed in Table 3.

timeline setting	Memory	Reality	1st-order true	1st-order false	2nd-order true	2nd-order false	total questions
E -> E -> M -> M	1357	1357	1357	1357	1357	1357	8142
E -> M -> E -> M	1378	1378	1378	1378	1378	1378	8268
E -> M -> M -> E	1395	1395	1395	1395	1395	1395	8370
M -> E -> E -> M	1405	1405	1405	1405	1405	1405	8430
M -> E -> M -> E	1465	1465	1465	1465	1465	1465	8790

A.3 Experimental Details

A.3.1 Models and Parameter Settings

In this experiment, we evaluated Meta's Llama-3-8B-Instruct and Llama-3-70B-Instruct as representative open-source models, and OpenAI's GPT-4o-mini and GPT-4.1-mini as commercial API models.

To ensure reproducibility and enhance the determinism of the outputs, a common set of configurations was applied to all models. For inference with the Llama-3 models (8B and 70B), the `transformers` library's pipeline was used with the parameters `do_sample=False` and `temperature=0.0`. The 70B model was loaded with `torch_dtype=torch.bfloat16`. For the OpenAI models, the official API was utilized with `temperature=0.0` and `top_p=1.0`. The maximum number of new tokens to generate was limited to 50 for all models. The evaluation for Llama-3-8B, GPT-4o-mini, and GPT-4.1-mini (3 runs each, approx. 10 hours per run) and Llama-3-70B (2 runs, approx. 600 hours total) required a total computational budget of about 690 GPU hours.

A.3.2 Prompt Format

A common zero-shot prompt format was used for the evaluation. The prompt consists of a system prompt that assigns the role of a reading comprehension expert, and a user prompt that includes the story and the question.

System prompt:
You are an expert in reading comprehension. Answer the following question based ONLY on the text provided in the story. Provide only the answer, without any introductory phrases or explanations.

User prompt:
Please read the following story and answer the subsequent question.

— STORY —
{story_text}
— END OF STORY —

Question: {question}

For Llama-3 models (8B and 70B), the above content was converted into the model-specific chat format using the `tokenizer.apply_chat_template` method before being input.

A.4 Use of AI Assistants

In preparing this manuscript, AI assistants were used for coding support, Japanese-English translation, and text editing.

Table 6: Detailed breakdown of the 1,000 stories (total 6,000 questions) in the A3_O3_C3 (base) setting, grouped by parent (agent placement) and child (full entity) patterns.

parent pattern	child pattern (A/O/C)	instances (stories)	total questions
A/A/A	A/A/ACCCOOO	10	60
	A/AC/ACCOOO	7	42
	A/ACCO/ACOO	5	30
	A/ACCOO/ACO	7	42
	(subtotal for A/A/A)	(29)	(174)
AA/A	A/AACCCOOO/	300	1800
	AA/ACCCOOO/	22	132
	AAC/ACCOOO/	4	24
	AACCO/ACOO/	29	174
	AACCOO/ACO/	103	618
	AACCOOO/AC/	71	426
	AA/ACCO/COO	4	24
	AA/ACCOO/CO	5	30
	AA/ACCOOO/C	6	36
	AACO/ACCOO/	8	48
	AACOO/ACCO/	3	18
	A/AACCO/COO	37	222
	A/AACCOO/CO	120	720
	A/AACCOOO/C	90	540
	(subtotal for AA/A)	(802)	(4812)
AAA/	AAACCO/COO/	22	132
	AAACCOO/CO/	72	432
	AAACCOOO/C/	75	450
	(subtotal for AAA/)	(169)	(1014)
Total		**1000**	**6000**