Towards Dynamic Theory of Mind: Evaluating LLM Adaptation to Temporal Evolution of Human States

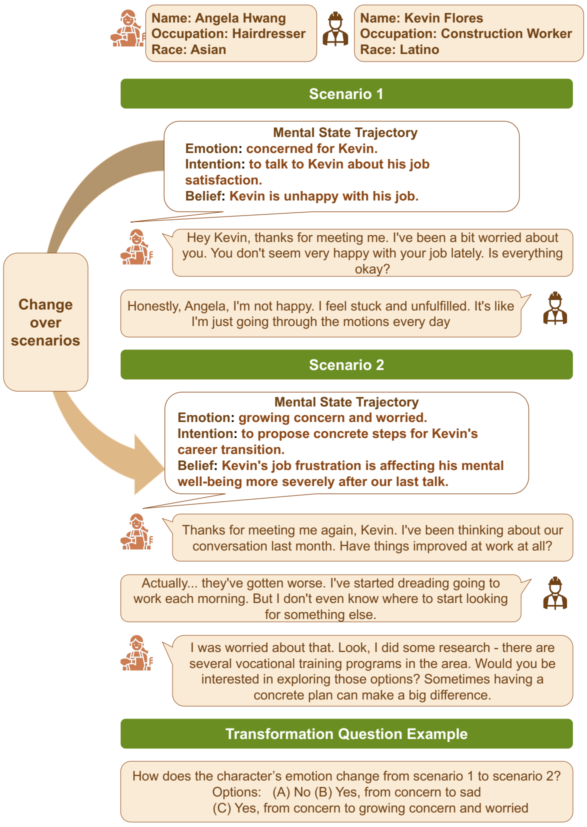
Yang Xiao[1*] **Jiashuo Wang**[1*] **Qiancheng Xu**[1] **Changhe Song**[1]
Chunpu Xu[1] **Yi Cheng**[1] **Wenjie Li**[1†] **Pengfei Liu**[2†]
[1]The Hong Kong Polytechnic University [2]Shanghai Jiao Tong University
`yang-alan.xiao@connect.polyu.hk` `csjwang@comp.polyu.edu.hk`

Abstract

As Large Language Models (LLMs) increasingly participate in human-AI interactions, evaluating their Theory of Mind (ToM) capabilities - particularly their ability to track dynamic mental states - becomes crucial. While existing benchmarks assess basic ToM abilities, they predominantly focus on static snapshots of mental states, overlooking the temporal evolution that characterizes real-world social interactions. We present DYNTOM, a novel benchmark specifically designed to evaluate LLMs' ability to understand and track the temporal progression of mental states across interconnected scenarios. Through a systematic four-step framework, we generate 1,100 social contexts encompassing 5,500 scenarios and 78,100 questions, each validated for realism and quality. Our comprehensive evaluation of ten state-of-the-art LLMs reveals that their average performance underperforms humans by 44.7%, with performance degrading significantly when tracking and reasoning about the shift of mental states. This performance gap highlights fundamental limitations in current LLMs' ability to model the dynamic nature of human mental states. [1]

1 Introduction

Theory of Mind (ToM) - the ability to understand and reason about others' mental states - is fundamental to human social interaction (Premack and Woodruff, 1978; Turner, 1988). As Large Language Models (LLMs) increasingly engage in human-AI interactions, their capability to track and understand the dynamic nature of human mental states becomes crucial. While existing research has evaluated LLMs' ToM capabilities, these evaluations often overlook a critical aspect: the temporal evolution of mental states in real-world social contexts.


Name: Angela Hwang
Occupation: Hairdresser
Race: Asian

Name: Kevin Flores
Occupation: Construction Worker
Race: Latino

Mental State Trajectory
Emotion: concerned for Kevin.
Intention: to talk to Kevin about his job satisfaction.
Belief: Kevin is unhappy with his job.

Hey Kevin, thanks for meeting me. I've been a bit worried about you. You don't seem very happy with your job lately. Is everything okay?

Honestly, Angela, I'm not happy. I feel stuck and unfulfilled. It's like I'm just going through the motions every day

Mental State Trajectory
Emotion: growing concern and worried.
Intention: to propose concrete steps for Kevin's career transition.
Belief: Kevin's job frustration is affecting his mental well-being more severely after our last talk.

Thanks for meeting me again, Kevin. I've been thinking about our conversation last month. Have things improved at work at all?

Actually... they've gotten worse. I've started dreading going to work each morning. But I don't even know where to start looking for something else.

I was worried about that. Look, I did some research - there are several vocational training programs in the area. Would you be interested in exploring those options? Sometimes having a concrete plan can make a big difference.

How does the character's emotion change from scenario 1 to scenario 2?
Options: (A) No (B) Yes, from concern to sad
(C) Yes, from concern to growing concern and worried

Figure 1: A simplified illustration showing the mental state trajectory's change over scenarios and the transformation question example to probe LLM's ability to adapt to the change.

Current ToM evaluations of LLMs, including benchmarks like SocialIQA (Turner, 1988), BigToM (Gandhi et al., 2023), and TOMBENCH (Chen et al., 2024), predominantly focus on static snapshots of mental states in isolated scenarios. These works primarily focus on static evaluations, whereas our work presents a novel approach to capture the continuous change of mental states across multiple interconnected scenarios - a crucial aspect of real-world social interactions that has not been systematically evaluated in previous work. This

---

[*] Equal contribution.
[†] Corresponding authors.
[1] DYNTOM are available at GitHub and HuggingFace.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24036–24057
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

temporal dimension is essential for understanding LLMs' true capabilities in real-world social interactions, where mental states constantly shift and evolve in response to ongoing social dynamics. For instance, LLMs are expected to understand and reason about the shift of user mental states in support conversations to better help users (Liu et al., 2024b; Wang et al., 2024).

To address this challenge, we introduce DYNTOM (Dynamic Theory of Mind), a novel benchmark designed specifically to evaluate LLMs' ability to track and understand the temporal evolution of mental states, as shown in Figure 1. Our benchmark is constructed through a systematic process: (1) social context construction, including social location, character profiles, and relationships; (2) mental state trajectory design across multiple scenarios; (3) scenario generation with natural dialogue; and (4) question formulation targeting temporal understanding of mental states. Each generated scenario and question undergoes rigorous human validation to ensure quality and realism. DYNTOM captures mental state dynamics through continuous social scenarios while incorporating real-world elements such as rich social contexts.

Our benchmark comprises 1,100 social contexts featuring 2,200 characters across 261 social locations, 5,500 social scenarios, and 78,100 multiple-choice questions. Through a comprehensive evaluation of ten representative LLMs, including GPT-4 series (Achiam et al., 2023), Llama 3 series (Dubey et al., 2024), Qwen 2 series (Yang et al., 2024), and GLM series (GLM et al., 2024), we find that their average performance lags behind human performance by 44.7%, with the gap widening significantly when requiring models to track how a mental state changes across different scenarios. This performance degradation highlights a fundamental limitation in current LLMs' ability to model the dynamic nature of human mental states.

The main contributions of this work are: 1. A novel framework for evaluating LLMs' understanding of temporal evolution in mental states, with a systematic process for generating and validating evaluation data; 2. A comprehensive benchmark featuring 78,100 questions specifically designed to probe LLMs' ability to track and reason about mental state changes over scenarios; 3. Extensive empirical evaluation reveals specific challenges LLMs face in temporal reasoning about mental states, including detailed analysis of failure modes in tracking state changes and determining factors influencing changes.

2 Related Work

2.1 ToM Benchmarks

Theory of mind appears to be an innate potential ability in humans that requires social and other experiences over many years for its full development. With the development of LLMs, researchers have begun to probe whether LLMs possess a Theory of Mind ability comparable to that of humans, as they have reached and occasionally surpassed human performance in some task-solving and reasoning tasks. Nematzadeh et al. (2018); Le et al. (2019); Wu et al. (2023) apply the Sally-Anne Test and bAbi to test LLMs' ToM ability in the aspect of false belief, and they find that LMs' performance is significantly lower than humans. Ullman (2023); Shapira et al. (2024); Kim et al. (2023); Sap et al. (2022) propose that LLMs prone to shortcuts and spurious correlations. Apart from the test in the aspect of belief, Xu et al. (2024); Chen et al. (2024); Sabour et al. (2024) construct benchmarks to test LLMs' ToM ability for emotion, intention, and perception. Jin et al. (2024); Shi et al. (2024) propose to evaluate LLMs in multi-modal environments. However, most of the previous evaluations do not take the continuous evolution of mental states across multiple interconnected scenarios into consideration. Our work aims to develop a novel benchmark to understand the ToM reasoning of language models in the dynamic social context.

2.2 Human Behavior Simulation

Recent advancements in language model capabilities have opened new avenues for generating high-quality data. Previous work has demonstrated successful applications of LLMs in simulating human behavior across various domains, including HCI research (Hämäläinen et al., 2023), conversational recommender systems (Yoon et al., 2024), role-playing (Xie et al., 2024; Xiao et al., 2023), clinical medical education (Wang et al., 2024), social science (Hua et al., 2023; Park et al., 2023, 2022; Aher et al., 2023). DYNTOM leverages LLMs to generate realistic dialogues that reflect predetermined character mental states. We implement strict quality assurance through human evaluation of social context authenticity and question validity. This approach combines the efficiency of automated generation with robust validation procedures, ensuring our benchmark's reliability and reproducibility.

3 DYNTOM Benchmark

3.1 DYNTOM Construction Framework

Definitions and Preliminaries We first define key terms used throughout this paper. A *Social Location* refers to the physical setting where social interactions occur, which influences behavior and social norms (Farrow et al., 2017). A *Social Context* provides the foundational setup for social interactions, comprising a social location, character profiles (e.g., demographics, personalities), and their relationships. A *Social Scenario* represents a self-contained social interaction between characters at a specific moment. In our work, we construct sequences of temporally connected scenarios within the same social context, enabling us to track the dynamic evolution of characters' mental states through continuous social interactions. We define a *Social Stage* as the complete structure of a social interaction, comprising the *Social Location*, *Social Context*, and *Social Scenarios*.

Our framework consists of three systematic steps for generating the social stages in our benchmark:

Step 1: Social Context Construction A social context consists of three components: a social location, two characters' profiles, and the relationship between these characters. For social locations, following Ziems et al. (2023), we collect 261 locations across 13 categories representing common physical settings for social interactions. For character profiles, we construct seven aspect pools (names, surnames, gender, occupation, education, race, and personality traits) using demographic data from the U.S. Census Bureau statistics to ensure realistic population representation. For each social context, we randomly sample one location and create two characters by sampling from each aspect of these pools. To generate character relationships, we first create four exemplar relationships manually, then prompt GPT-4-Turbo to generate new relationships based on these exemplars and the sampled character profiles. To ensure quality, four human annotators evaluate both the characters' profiles and their corresponding relationships, discarding any profile or relationship that any annotator deems unrealistic. This rigorous validation process results in retaining 92% of the generated profiles and relationships.

Step 2: Mental State Trajectory Design We focus on evaluating three mental states (**beliefs**, **emotions**, and **intentions**) and their resulting **actions** (for convenience, also denoted as mental states). For each social context, we design a sequence of five[2] scenarios where these states of the characters evolve and influence each other. Following the psychological research of D'Andrade (1995), we model the mental states through three key relationships: 1) beliefs influence emotions; 2) beliefs and emotions influence intentions; 3) beliefs, emotions, and intentions drive actions. We prompt GPT-4-Turbo with four exemplar trajectories and the 3 design principles to generate coherent mental state progressions across every five scenarios. Importantly, when generating these trajectories, LLMs should also output specific cues that trigger mental state transitions between adjacent scenarios, providing explicit reasoning for how and why mental states evolve throughout the social interaction. Any scenarios without the mental state trajectory and cues are discarded. Four human annotators evaluate each generated trajectory on two dimensions using a 5-point scale: coherence (consistency of mental state changes across scenarios), rationality (the validity and rationality of these transition cues), and authenticity (plausibility of mental state transitions). Trajectories with mean scores below 4.0 on either dimension are discarded, resulting in an 85.4% retention rate.

Step 3: Scenario Generation Building upon the mental state trajectories designed in Step 2, we now generate scenarios with natural dialogues to manifest these mental states in social interactions. Each scenario includes a background description and a dialogue between characters, reflecting the mental state trajectory designed for this scenario. We choose dialogue as the primary format because it naturally reveals characters' mental states and is frequently used in daily interactions. For each mental state trajectory, we prompt GPT-4-Turbo to generate the dialogue and background of the scenario, ensuring that the main character's utterances and behaviors align with their prescribed mental states.

Following the same validation process as Step 2, four human annotators evaluate each scenario on three dimensions using a 5-point scale: consistency (alignment with the designed mental state trajectory), coherent (the five scenarios within each social context form a coherent storyline, where each scenario connects meaningfully to those be-

---

[2] we have limited the number of scenarios to five to reduce costs while maintaining a more authentic social context. Researchers can easily adjust the scenario number in our framework to meet their needs.

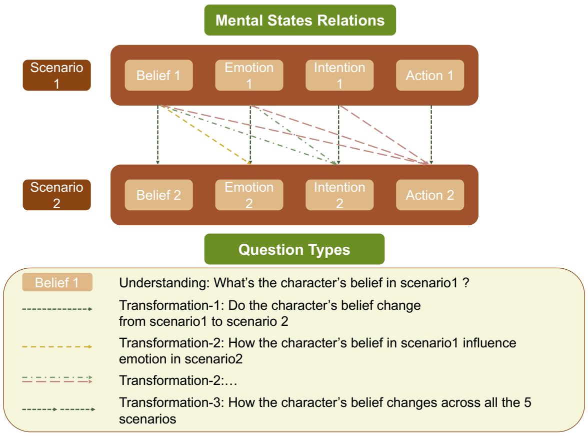

Mental States Relations
Scenario 1
Belief 1
Emotion 1
Intention 1
Action 1
Scenario 2
Belief 2
Emotion 2
Intention 2
Action 2
Question Types
Belief 1
Understanding: What's the character's belief in scenario1 ?
Transformation-1: Do the character's belief change from scenario1 to scenario 2
Transformation-2: How the character's belief in scenario1 influence emotion in scenario2
Transformation-2:…
Transformation-3: How the character's belief changes across all the 5 scenarios

Figure 2: A simplified example of question types (showing only two scenarios). In each scenario, characters have four states: belief, emotion, intention, and action - these are assessed through **understanding** questions. The value of the same state changes across different times (scenarios) is evaluated through **transformation-1,2,3** questions.

fore and after it.), and authenticity (naturalness of the scenario and conversations). Scenarios scoring below 4.0 on either dimension are discarded, with 88.7% of the generated scenarios retained.

3.2 Question Genres

Based on the validated scenarios and mental state trajectories, we design questions to systematically evaluate how well LLMs can track and reason about the temporal evolution of mental states. We develop four question types that progressively assess different aspects of this capability.

Understanding Questions Understanding questions establish a baseline by testing LLMs' ability to identify states (belief, emotion, intention, and action) at specific points in time. While this represents the most basic level of ToM reasoning in our evaluation, it forms the foundation for evaluating more complex temporal understanding.

Transformation Questions Transformation questions directly evaluate LLMs' ability to reason about mental state dynamics through three increasingly complex aspects, as illustrated in Figure 2. **Transformation-1** examines whether LLMs can detect state changes between consecutive scenarios, testing their basic temporal awareness. **Transformation-2** probes deeper by testing if LLMs understand the causal mechanisms behind state changes, evaluating their grasp of psychological dynamics. **Transformation-3** presents the most challenging task: tracking state evolution across all scenarios and testing LLMs' ability to maintain and reason about extended temporal sequences.

This progression of question types allows us to precisely identify where LLMs succeed or fail in understanding dynamic mental states. Understanding questions reveal whether failures in temporal reasoning stem from basic state comprehension issues, while the three transformation types help pinpoint specific limitations in tracking and reasoning about mental state changes over time. We apply four predefined question templates to the social stage to generate questions. The template details are presented in the Appendix A.5.

Options and Ground Truth The design of options and ground truth leverages the comprehensive mental state trajectories created in Step 2. For understanding questions, when evaluating a specific state (e.g., belief in scenario 1), we construct distractors using both other states from the same scenario and the same state from other scenarios. For instance, when assessing belief in scenario 1, incorrect options include the emotion, intention, and action from scenario 1, as well as beliefs from scenarios 2-5. Similarly, for transformation questions, options are constructed using documented state values and their changes from the trajectory. For example, when asking "Why does John's belief change from feeling inferior in scenario 1 to feeling respected in scenario 2?", the correct answer would be "Beverly's praise of his expertise," while distractors include other documented changes such as "John's demonstration of skills" (action). This systematic approach of option generation ensures that the questions are challenging yet unambiguous, as both correct answers and distractors are grounded in the explicitly designed mental state trajectories.

Validation Following our validation process, four annotators evaluate the questions on clarity (whether the question is unambiguous) and answerability (whether the answer can be determined from the given context) using a 5-point scale. Questions scoring below 4.0 are regenerated, resulting in a final set of high-quality evaluation items. Finally, 78100 questions are collected.

Evaluation Metric To evaluate LLMs' performance on DYNTOM, we calculate their accuracy across all 78,100 questions. For each question, we consider the LLM's response correct only if it exactly matches the ground truth option. The final

Item	Number
ToM Mental States	4
Social Locations	261
Characters	2200
Social Scenarios	5500
Social Contexts	1100
Questions	78100
Understanding	28.2%
Transformation-1	22.5%
Transformation-2	43.7%
Transformation-3	5.6%
Average Social Scenario Length	457.9
Average Questions Length	77.5

Table 1: DYNTOM Statistics.

🏠 Social Location 👥 Relationship 🔗 Dynamic mental states
🪪 Profile ◆ Intradependent mental states ✒ Questions Num

	Plot	🏠	🪪	👥	◆	🔗	✒
ToMi	✗	✗	✗	✗	✗	✗	999
SocialIQA	✗	✗	✗	✗	✗	✗	37588
Hi-ToM	✗	✓	✗	✗	✗	✗	1200
OpenToM	✓	✗	✓	✓	✗	✗	2384
BigToM	✓	✓	✗	✗	✗	✓	600
TOMBENCH	✓	✗	✗	✗	✗	✗	2860
DYNTOM(ours)	✓	✓	✓	✓	✓	✓	78100

Table 2: Benchmark Comparison.

performance metric is computed as the percentage of correct answers across all questions, providing a comprehensive measure of the model's ability to reason about dynamic mental states in social interactions.

3.3 Statistics

In total, our final benchmark contains 1,100 high-quality social stages, where each stage consists of a social location (physical setting), social context (two characters with detailed profiles and relationships), and five social scenarios (temporally connected interactions). For each social stage, we generate 71 questions across four types (understanding and three types of transformation), resulting in 78,100 questions in total. The detailed statistics are shown in Table 1.

Compared to existing ToM benchmarks (Table 2), DYNTOM offers key advantages. While previous works evaluate static snapshots of mental states, DYNTOM systematically captures the temporal evolution of mental states through connected scenarios, enabling a more realistic evaluation of how well LLMs can track and reason about dynamic social interactions.

4 Experiments

4.1 Experimental Setup

To evaluate the ToM reasoning capabilities across different model scales and architectures, we conducted experiments using DYNTOM on ten representative language models, ranging from 7B to 70B parameters. These models include GPT-4o, GPT-4-Turbo, Llama-3.1 (8B and 70B variants), Mistral-7B, Mixtral-8x7B, Qwen2 (7B and 72B variants), DeepSeek-V2, and GLM-4. All models are accessed through their official APIs or publicly available weights.

We employed two evaluation approaches: (1) vanilla prompting, where models directly answer questions, and (2) zero-shot chain-of-thought (CoT) prompting (Wei et al., 2022), which encourages step-by-step reasoning before providing final answers. For both vanilla and CoT prompting, we used a temperature of 0.7 and top-p of 0.9 across all models to ensure fair comparison. To establish a human performance baseline, we recruited ten graduate students, different from those involved in data annotation, to evaluate a randomly sampled 30% of the dataset (330 social stages and 23430 questions). Detailed specifications of model versions, architectures, context windows, and prompting templates are provided in Appendix B.1.

4.2 Main Results

Table 3 demonstrates the ToM performance of LLMs across different mental states (belief, emotion, intention, and action) and question types (understanding and transformation), both with and without chain-of-thought (CoT) prompting. We established the human baseline by averaging performance across ten annotators, with standard deviations reported to indicate inter-annotator agreement. Here, we discuss several key findings from our experimental results.

Human vs. LLMs Human annotators achieved an average accuracy of 77.7% across all tasks. All LLMs performed significantly below this baseline, and their average performance underperforms humans by 44.7%, with even the best GPT-4o showing a gap up to 13.7% in vanilla prompting (77.7% vs. 64.0%). The performance disparity was particularly pronounced in transformation-type questions across all mental states, revealing current LLMs' limitations in tracking and comprehending the temporal evolution of mental states in so-

Subject	Belief		Emotion		Intention		Action		AVG.
	U	T	U	T	U	T	U	T	
Human	$83.8_{\pm 16.4}$	$77.6_{\pm 12.0}$	$89.5_{\pm 10.7}$	$78.7_{\pm 14.0}$	$79.0_{\pm 21.4}$	$73.8_{\pm 14.0}$	$76.7_{\pm 25.8}$	$76.3_{\pm 14.0}$	$77.7_{\pm 12.7}$
GPT-4o	**80.9**	**44.5**	91.7	**45.8**	**87.5**	**51.9**	**95.1**	**55.6**	**64.0**
GPT-4-Turbo	63.5	32.3	74.7	33.9	71.3	35.5	80.5	36.2	47.6
Llama-3.1-70B	65.8	40.2	**93.8**	42.3	82.8	42.0	91.8	45.5	57.1
Llama-3.1-8B	31.6	18.0	40.0	19.9	22.4	16.6	26.6	15.5	22.3
Mixtral-8x7B	23.3	21.6	46.2	18.4	32.9	10.8	40.3	9.5	21.9
Mistral-7B	21.3	11.7	23.8	10.2	16.3	10.1	20.6	9.2	13.9
Qwen2-72B	72.0	37.2	85.5	38.0	79.5	33.2	89.8	20.9	48.5
Qwen2-7B	22.2	19.8	43.0	20.5	25.1	15.7	24.6	15.0	22.1
DeepSeek-V2	6.5	9.2	4.8	8.1	3.7	7.3	2.8	5.7	7.2
GLM-4	29.5	23.9	43.9	20.8	28.5	16.5	40.4	16.8	25.4
LLM AVG.	41.7	25.8	54.7	25.8	45.0	24.0	51.3	23.0	33.0
GPT-4o+CoT	**79.2**	**44.5**	88.0	**47.6**	82.1	**46.6**	90.4	**49.6**	**61.1**
GPT-4-Turbo+CoT	61.7	31.0	77.8	33.2	71.4	32.8	81.0	37.6	47.1
Llama-3.1-70B+CoT	68.0	38.9	**90.7**	43.7	81.4	42.8	**96.5**	46.6	57.6
Llama-3.1-8B+CoT	32.0	21.7	40.3	20.9	21.8	19.3	23.3	15.9	23.6
Mixtral-8x7B+CoT	15.6	13.9	29.7	13.8	25.8	8.8	26.6	8.8	15.8
Mistral-7B+CoT	21.6	10.1	22.5	11.0	19.9	8.1	18.8	8.8	13.3
Qwen2-72B+CoT	70.1	39.2	87.6	41.4	**83.8**	34.6	89.0	27.1	51.3
Qwen2-7B+CoT	28.6	18.1	43.7	19.3	29.6	19.7	20.2	18.4	23.5
DeepSeek-V2+CoT	7.4	9.8	3.2	10.4	5.0	7.3	5.0	6.4	8.1
GLM-4+CoT	30.0	26.4	48.0	22.1	32.4	17.7	43.2	14.1	26.6
LLM+CoT AVG.	41.4	25.4	53.2	26.3	45.3	23.8	49.4	23.3	32.8

Table 3: LLMs' performance on DynToM. U: Understanding, T: Transformation. Numbers represent accuracy in percentages. For human performance, subscripts indicate standard deviation across ten annotators.

cial interactions. Notably, however, some LLMs, particularly GPT-4o, demonstrated superior performance in understanding-type questions compared to human annotators. This can be attributed to the nature of these questions, which primarily involve analyzing static mental states within single scenarios, as illustrated in Figure 2. Such tasks require less temporal reasoning and rely more on semantic matching of explicitly stated mental states, aligning well with LLMs' pattern recognition capabilities but not necessarily reflecting the true understanding of dynamic social contexts.

Differences Between LLMs' ToM Performance In vanilla prompting settings, GPT-4o emerged as the leading model, achieving an accuracy of 64.0% and outperforming the second-best model, Llama-3.1-70B (57.1%), by 6.9 percentage points. Among open-source models, both Llama-3.1-70B and Qwen2-72B demonstrated remarkable capabilities, surpassing GPT-4-Turbo's 47.6% performance. Notably, Llama-3.1-70B achieved superior performance in emotion-related understanding tasks, reaching 93.8% accuracy compared to GPT-4o's 91.7%. However, even GPT-4o's best overall performance at 64.0% falls significantly short of human-level performance, with DeepSeek-V2 showing the lowest performance at 7.2%. This substantial performance gap, particularly in tracking the temporal evolution of mental states, highlights the challenging nature of our benchmark and reveals that current LLMs lack robust ToM reasoning capabilities in realistic social contexts, despite their near-perfect performance on existing ToM benchmarks (Gandhi et al., 2024).

Differences Between Transformation and Understanding Question Types Table 3 also reveals that models perform significantly worse on transformation questions compared to understanding questions. The most substantial gap occurs in emotion-related reasoning, where the average model accuracy drops from 54.7% in understanding questions to 25.8% in transformation questions—a difference of 28.9 percentage points. Transformation questions require models to track how a character's mental state evolves across different scenarios, capturing shifts in beliefs, emotions, and intentions over time. This performance gap highlights a critical limitation of current models—their inability to effectively reason about dynamic mental state changes within continuous social contexts.

Vanilla vs. CoT Prompting Our experimental results demonstrate that standard chain-of-thought (CoT) prompting has inconsistent effects on LLMs' ToM reasoning capabilities. While CoT prompting improved performance for smaller models (Llama-3.1-8B: +1.3%, Qwen2-72B: +2.8%, DeepSeek-V2: +0.9%, and GLM-4: +1.2%), it led to performance degradation in more capable models, notably GPT-4o (-2.9%). Through analyzing the intermediate outputs (Appendix B.3), we identified a two-fold effect: For smaller models that struggle with complex reasoning, CoT's step-by-step decomposition provides beneficial scaffolding for basic problem analysis. However, this same decomposition becomes a limitation for more capable models, as it enforces a rigid reasoning structure that treats each scenario independently, failing to capture the crucial temporal dependencies between scenarios. As shown in our case study, when asked about mental state changes from scenarios 1 to 2, models following CoT often extensively analyze each scenario but fail to explicitly compare the states across time steps, leading to incorrect conclusions. This aligns with findings in Xiao et al. (2023) about LLMs' challenges in maintaining coherence during reasoning long inputs. These observations suggest that while CoT can help with basic reasoning decomposition, effective ToM reasoning requires specialized promptings that explicitly guide models to track and analyze the temporal evolution of mental states across scenarios.

Differences Across Mental States Analysis of Table 3 reveals consistent patterns in models' capability to reason about different mental states, particularly in understanding-type questions. Emotion-related reasoning achieves the highest accuracy, averaging 54.7%, whereas belief-related reasoning lags behind at 41.7%—a gap of 13 percentage points. We hypothesize that this disparity stems from the inherently implicit nature of beliefs compared to other mental states – while emotions and intentions often manifest in observable behaviors or explicit statements, beliefs frequently require multi-step inference from indirect evidence, such as actions or conversational context. This observation suggests that belief reasoning poses unique challenges in temporal social contexts, where models must not only infer current beliefs but also track their evolution through sequential interactions.

4.3 In-Depth Analysis

LLMs' Limits of ToM on Transformation Evaluating mental states across multiple interconnected scenarios introduces complex compositional reasoning challenges (Dziri et al., 2023), requiring models to track and reason about the continuous evolution of mental states. To systematically analyze how models handle this multi-step reasoning process, we group related questions and categorize the model's responses based on their performance on both the primary question and its dependencies. For instance, consider a question that asks whether a character's beliefs change between two scenarios. To answer this primary question (denoted as C), the model must first correctly identify the character's beliefs in each individual scenario (denoted as D). By grouping such related questions, we can assess the model's responses and classify them into four types: (1) Fully correct: The model accurately answers both the primary question (C) and all its dependencies (D). (2) Local error: The model correctly answers all dependencies (D) but makes an error on the primary question (C). (3) Restoration error: The model correctly answers the primary question (C) despite making errors on one or more of its dependencies (D). (4) Full error: The model makes errors on both the primary question (C) and one or more of its dependencies (D).

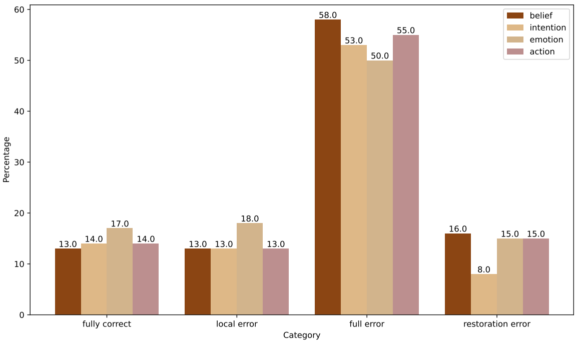


Figure 3: The percentage of GPT-4o's four types of their response.

Calculating the proportion of each response type across all question groups provides a comprehensive evaluation of the model's performance on these compositional questions requiring multi-step reasoning. As shown in Figure 3, our analysis reveals distinct patterns in models' reasoning capabilities. The fully correct cases, where models successfully identify both the mental states and their changes, are notably rare across all state types (13-17%), indicating limited genuine understanding. Local errors (13-18%) show models can correctly identify mental states but fail to reason about their changes, suggesting an inability to track evolution. Full errors dominate across all mental states (50-58%), with belief states showing the highest error rate (58%), revealing fundamental limitations in comprehending both states and their transitions. Restoration errors (8-16%) occur when models correctly identify changes without understanding

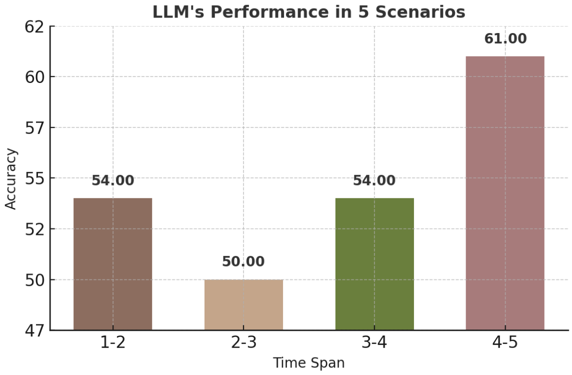


Figure 4: The average of GPT-4o's scores of the transformation question type in different time spans. The time span indicates the specific scenarios to which one question relates.

Time Span	1-2	2-3	3-4	4-5	5-6	6-7
6 Scenarios	64.0	**50.0**	51.0	62.0	62.0	-
7 Scenarios	56.0	45.0	**26.0**	30.0	26.0	34.0

Table 4: The average of GPT-4o's scores in the transformation question type for 6 and 7 scenarios.

Category	Time Span	w/o the truncation	w/ the truncation	Δ
5 scenarios	**1-2**	54.0	55.0	+1.0
	2-3	50.0	53.0	+3.0
	3-4	54.0	55.0	+1.0
6 scenarios	**1-2**	60.0	60.0	+0.0
	2-3	50.0	54.0	+4.0
	3-4	51.0	56.0	+5.0
7 scenarios	**1-2**	56.0	53.0	-3.0
	2-3	45.0	54.0	+9.0
	3-4	26.0	47.0	+21.0

Table 5: Comparison of GPT-4o's performance with and without the truncation of the fifth/seventh/eighth scenario across different time spans and categories (total 5, 6, and 7 scenarios), with Delta indicating the difference between the two conditions.

the underlying states, indicating superficial pattern matching rather than genuine reasoning. This analysis pinpoints where models struggle in the reasoning process, whether in identifying initial mental states, determining factors influencing changes, or integrating information across scenarios.

LLMs Fail in the Middle Scenario Our analysis reveals a consistent pattern of lower performance in transformation-type questions compared to understanding-type questions across all models. To investigate this performance gap, we examine model accuracy across different time spans, where each span represents the interval between consecutive scenarios (e.g., span 1-2 represents the transition from scenarios 1 to 2). We categorize the questions into different time spans based on their associated social scenarios. The results show a distinct "U-shaped" pattern: models perform better at early and late time spans but struggle with middle spans. To validate whether this pattern stems from the "Lost in the middle" phenomenon (Liu et al., 2024a) - where model performance degrades when processing information from the middle of long contexts - we conducted two additional experiments. First, we extend our analysis to longer sequences of 6 and 7 scenarios. The results (Table 4) strongly support our hypothesis: in 6-scenario sequences, accuracy drops to 50% in span 2-3, while in 7-scenario sequences, performance deteriorates more severely to 26% in span 3-4. Second, we perform a controlled experiment where we truncate sequences to their first four scenarios. As shown in Table 5, this intervention leads to significant performance improvements in middle spans: for 7-scenario sequences, the accuracy in span 3-4 improve by 21 percentage points (from 26% to 47%) when later scenarios are removed. Similar improvements are observed across 5- and 6-scenario sequences, with middle-span accuracy increasing by 1-5 percentage points.

These consistent improvements across different sequence lengths reveal a critical limitation in LLMs' ability to process dynamic mental states. While models can effectively track states at the beginning and end of interactions, they struggle to maintain understanding across extended scenarios - precisely the kind of continuous evolution that characterizes real-world social interactions. The significant performance degradation in middle scenarios (dropping to as low as 26% in longer sequences) underscores the importance of our dynamic evaluation approach and highlights a fundamental challenge in developing LLMs that can truly understand evolving social contexts.

5 Conclusion

In this work, we present DynToM, a benchmark designed to evaluate LLMs' theory of mind capabilities in dynamic social contexts, moving beyond static assessments to capture the crucial evolution of mental states across interconnected scenarios. Our human evaluation validates the benchmark's alignment with real-world social dynamics, while experimental results reveal significant gaps in current LLMs' capabilities - even the best model trails human performance by 11%, particularly struggling with tracking mental state changes across extended interactions. As LLMs continue to be deployed in social contexts, DynToM provides a valuable framework for assessing and improving their ability to understand the dynamic nature of human mental states.

Limitations

Limited LLMs This paper makes a significant contribution to the field by introducing DYNTOM, an innovative benchmark designed to evaluate Theory of Mind capabilities in language models within authentic social contexts. However, due to the constraint of computing resources and budget, a limitation lies in its evaluation scope, which encompasses ten language models with an emphasis on representative models. While this selection includes prominent models such as GPT-4 and Llama, the focus potentially overlooks insights that could be gained from examining other emerging open-source models and commercial models, such as Claude.

Limited Prompt Methods We use vanilla and CoT prompting methods for evaluation. Other methods, such as think-twice (Wilf et al., 2024), belief tracker (Sclar et al., 2023), and self-consistency (Wang et al., 2023), could also be explored to enhance the LLMs' ToM performance within authentic social contexts.

Limited mental states types and evaluation modality While our framework effectively models the interplay between belief, emotion, intention, and action based on established psychological theory (D'Andrade, 1995), there are opportunities to expand the scope of mental states evaluated. Future work could explore additional mental states, such as knowledge, to provide even richer insights into language models' ToM capabilities. While our dialogue-based evaluation approach has proven effective in assessing models' ToM abilities in dynamic contexts, future research could explore how these models perform in multimodal contexts that include visual and auditory cues. This extension would complement our text-based findings by examining how models track temporal changes in mental states across different modalities, though our current framework already provides robust insights into models' social reasoning capabilities.

Ethics Statement

Annotators and contents We strictly adhere to the ACL Code of Ethics. We placed high importance on ensuring the comfort and well-being of our annotators. We advised them to stop the annotation process if they came across any information that caused them discomfort. We recruited annotators at a rate of $2 \sim 3$ times their local hourly minimum wage. We instruct the annotators to validate the data without bias and keep the content free from unsafe, toxic, biased, offensive, and harmful content. We utilize the models in accordance with their designated purpose. In summary, we make every effort to adhere to the ethical norms set forth by ACL.

Ethical Considerations. The theory of mind is a distinctive social cognitive capability that is intrinsic to humans. Assessing the Theory of Mind capacities of Large Language Models utilizing DYNTOM may result in anthropomorphic interpretations, attributing human-like qualities to LLMs. Nonetheless, it is imperative to clarify that our objective is not to anthropomorphize LLMs. Our objective is to evaluate the capacity of LLMs to comprehend and interpret human mental states, thus enhancing AI's interaction with humans in the social context.

Acknowledgements

We would like to thank all reviewers for their insightful comments and suggestions to help improve the paper. This work was supported by the Research Grants Council of Hong Kong (GRF 15209724 and TRS T43-518/24-N).

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*.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Roy D'Andrade. 1995. *The development of cognitive anthropology*. Cambridge University Press.

Katherine Farrow, Gilles Grolleau, and Lisette Ibanez. 2017. Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, 140:1–13.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. 2024b. Compeer: A generative conversational agent for proactive peer support. *arXiv preprint arXiv:2407.18064*.

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.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

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

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

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.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2024. Muma-tom: Multi-modal multi-agent theory of mind. *Preprint*, arXiv:2408.12574.

Daniel Stokols. 1978. Environmental psychology.

Jonathan H Turner. 1988. *A Theory of Social Interaction*. Stanford University Press.

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

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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 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.

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.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How far are we from believable ai agents? a framework for evaluating the believability of human behavior simulation. *arXiv preprint arXiv:2312.17115*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages

1490–1504, Mexico City, Mexico. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

A The construction of the DynToM

A.1 the candidate pool of social location

The social location describes the environments where individuals live, work, and learn, which can significantly impact their mental states and behavior (Stokols, 1978). As shown in Figure 5, we have collected 13 types of social location types in total, adding up to 261 locations.

A.2 the candidate pool of profile

We conclude 7 aspects in the characters' profile: surname, name, gender, occupation, education, race, and personality traits. Their value can be found in Figure 7, 8, 9, and 10. The source of the surname, name, and occupation statistics are U.S. Census Bureau Homepage, The United States Social Security Administration, and Bureau of Labor Statistics, respectively. Figure 6 shows an example of the social background.

A.3 The prompt used to generate the mental state trajectory

As illustrated in Figure 11, the prompt is used to generate the mental state trajectory. In the holders of '{}' and '[]', the corresponding information will be input into this prompt. An example of the mental state trajectory is shown in Figure 12.

A.4 The prompt used to generate the social scenarios

As illustrated in Figure 13, the prompt is used to generate the social scenarios. In the holders of '{}' and '[]', the corresponding information will be input into this prompt. An example of the social scenario is shown in Figure 14.

A.5 The templates for the four types of questions and question examples

We apply four predefined question templates to the social stage to generate questions, 71 questions for every social context in total. The four types of questions are: (1) (Understanding-1) What is the main character's ToM reasoning item in a specific scenario? (2) (Transformation-1) Does a ToM reasoning item change from scenario A to scenario B? (3) (Transformation-2) What causes a ToM reasoning item change from scenario A to scenario B? (4) (Transformation-3) How does the ToM reasoning item change across all the scenarios? The templates and the example of the four types of questions are shown in Figure 15.

Model	Version	Size	Context Length
GPT-4o	2024-05-13	~	128k
GPT-4-Turbo	2024-04-09	~	128k
Llama-3.1-8B	Instruct	8B	128k
Llama-3.1-70B	Instruct	70B	128k
Mistral-7B	Instruct-v0.3	7B	32k
Mixtral-8x7B	Instruct-v0.1	8x7B	32k
Qwen2-7B	Instruct	7B	128k
Qwen2-72B	Instruct	72B	128k
DeepSeek-V2	Lite-Chat	16B	32k
GLM-4	9b-chat	9B	128k

Table 6: The detail of models evaluated in our benchmark.

A.6 Human validation of the Quality of DynToM

We apply argilla as the annotation platform. Figure 16 shows the annotation interface for data validation.

B Experiments

B.1 Model detail

We evaluate a total of 10 popular LLMs, including GPT-4o, GPT-4-Turbo, Llama-3.1-8B, Llama-3.1-70B, Mistral-7B, Mixtral-8x7B, Qwen2-7B, Qwen2-72B, DeepSeek-V2, GLM-4. For all the LLMs, we strictly abide by their terms and get access through official APIs or model weights. Details about model versions, parameter sizes, context window sizes and the prompts used for the two methods are shown in Table 6.

B.2 Prompting methods

We employ two prompting methods: the vanilla prompting which directly asks LLMs to answer the questions, and the CoT prompting that elicits step-by-step reasoning before answering. The prompts used for the two methods are shown in Figure 17.

B.3 Case Study for CoT Prompting

Both ToM reasoning item and question-type results in Table 3 indicate that CoT prompting doesn't always improve LLMs' ToM reasoning ability. We present a failure case of GPT-4o when CoT prompting is used in Figure 18.

The Candidate Pool of Social Location

Transportation and Travel
airplane cabin, airplane, airport terminal, airport, bus stop, bus boat, car, helicopter, harbor, dock, hot air balloon, ship, taxi, train compartment, train station, subway platform, subway

Accommodation and Residential
apartment, backyard, balcony, beach house, bedroom, cabin campsite, childs room, dorm room, hotel lobby, hotel pool, hotel room hotel, house entrance, house safe, living room, nursery, nursing home, private deck, residential neighborhood, yard

Food and Beverage
bakery, bar, cafe, cafeteria, deli restaurant, dining hall, fastfood restaurant, outdoor restaurant, pub, restaurant, steak house

Shopping and Retail
arcade, bookstore, boutique, camera shop, car dealership, clothing store, department store, drugstore, gift shop, grocery store, hardware store, jewelry shop, liquor store, market, music store, pawn shop, shop, shopping mall, showroom, store

Entertainment and Leisure
amusement park, arcade, balloon festival, ballroom, cinema, circus, fair, game room, golf course, museum, nightclub, outdoor cinema, paintball park, park, performance hall, ski resort, theater, yoga studio, zoo

Education
classroom, college, kindergarten, biology laboratory, lab, library, school bus, school theater, school, university

Health and Wellness
alcoholics ward, clinic, dentists office, doctors office, hospital, mental institution, nurse room, veterinarians office

Work and Office
bank, campaign room, conference room, construction site, copy center, government office, home office, office

Military and Law Enforcement
border, police car, police station, border checkpoint

Places of Worship and Ceremony
altar, chapel, church, funeral parlour, temple

Nature and Outdoors
canyon, cliff, coast, cornfield, desert, forest, lake, lawn, meadow, mountain, ocean, plain, river, waterfall, woods

Sport and Fitness
athletic field, basketball court, gym, gymnasium, hockey rink, locker room, mini golf course, polo field, pool, riding arena, running track, soccer field, stadium, tennis court, volleyball field

Miscellaneous
animal shelter, arena, art gallery, attic, auditorium, backyard, basement, bathroom, bazaar, beach, beauty salon, boathouse, bowling alley, boxing hall, bridal shop, bridge, broadcast studio, castle, cave, cemetery, city hall, city plaza, cloakroom, closet, clubhouse, community center, country road, courthouse, courtyard, dance studio, darkroom, drawing room, elevator, excavation, exhibition hall, farm, film studio, football field, garage, garden, gas station, gun shop, hall, hangar, hawaii, hotel, junkyard, kitchen table, kitchen, lake, legislative chamber, liquor store, living room, market, massage parlor, monument, museum, newsroom, outdoor pool, patio, performance hall, plaza, political rally, porch, post office, repair shop, restroom, road, rodeo, rooftop, rural road, saloon, smoking room, sofa, stable, stage, study room, tattoo studio, telephone booth, tv studio, waiting room

Figure 5: The candidate pool of social location.

An example of the social context

social setting: "temple”

social setting type: "Places of Worship and Ceremony”

main character: "Maria Thai”

scenario numbers: 5

characters information: "**Main Character**: Maria Thai, a woman, is a Social Worker whose race is Native Hawaiian and Other Pacific Islander. Maria Thai obtained a doctorate degree and has an ESFP personality..**Supporting Characters**:- **1**: Kelly Sexton, the woman, is a Fitness Trainer whose race is White. Kelly Sexton obtained a Primary Education and has an ISTP personality."

Figure 6: An example of the social context.

The Candidate Pool of Surname

American Indian and Alaska Native

Begay, Yazzie, Benally, Tsosie, Nez, Begaye, Etsitty, Becenti, Yellowhair, Manygoats, Wauneka, Manuelito, Apachito, Bedonie, Calabaza, Peshlakai, Claw, Roanhorse, Goldtooth, Etcitty, Tsinnijinnie, Notah, Clah, Atcitty, Twobulls, Werito, Hosteen, Yellowman, Attakai, Bitsui, Delgarito, Henio, Goseyun, Keams, Secatero, Declay, Tapaha, Beyale, Haskie, Cayaditto, Blackhorse, Ethelbah, Tsinnie, Walkingeagle, Altaha, Bitsilly, Wassillie, Benallie, Smallcanyon, Littledog, Cosay, Clitso, Tessay, Secody, Bigcrow, Tabaha, Chasinghawk, Blueeyes, Olanna, Blackgoat, Cowboy, Kanuho, Shije, Gishie, Littlelight, Laughing, Whitehat, Eriacho, Runningcrane, Chinana, Kameroff, Spottedhorse, Arcoren, Whiteplume, Dayzie, Spottedeagle, Heavyrunner, Standingrock, Poorbear, Ganadonegro, Ayze, Whiteface, Yepa, Talayumptewa, Madplume, Bitsuie, Tsethlikai, Ahasteen, Dosela, Birdinground, Todacheenie, Bitsie, Todacheene, Bullbear, Lasiloo, Keyonnie, Notafraid, Colelay, Kallestewa, Littlewhiteman

Asian and Native Hawaiian and Other Pacific Islander

Nguyen, Kim, Patel, Tran, Chen, Li, Le, Wang, Yang, Pham, Lin, Liu, Huang, Wu, Zhang, Shah, Huynh, Yu, Choi, Ho, Kaur, Vang, Chung, Truong, Phan, Xiong, Lim, Vo, Vu, Lu, Tang, Cho, Ngo, Cheng, Kang, Tan, Ng, Dang, Do, Ly Han, Hoang, Bui, Sharma, Chu, Ma, Xu, Zheng, Song, Duong, Liang, Sun, Zhou, Thao, Zhao, Shin, Zhu, Leung, Hu, Jiang, Lai, Gupta, Cheung, Desai, Oh, Ha, Cao, Yi, Hwang, Lo, Dinh, Hsu, Chau, Yoon, Luu, Trinh, He, Her, Luong, Mehta Moua, Tam, Ko, Kwon, Yoo, Chiu, Su, Shen, Pan, Dong, Begum, Gao, Guo, Chowdhury, Vue, Thai, Jain, Lor, Yan, Dao

Black or African American

Smalls, Jeanbaptiste, Diallo, Kamara, Pierrelouis, Gadson, Jeanlouis, Bah, Desir, Mensah, Boykins, Chery, Jeanpierre, Boateng, Owusu, Jama, Jalloh, Sesay, Ndiaye, Abdullahi, Wigfall, Bienaime, Diop, Edouard, Toure, Grandberry, Fluellen, Manigault, Abebe, Sow, Traore, Mondesir, Okafor, Bangura, Louissaint, Cisse, Osei, Calixte, Cephas, Belizaire, Fofana, Koroma, Conteh, Straughter, Jeancharles, Mwangi, Kebede, Mohamud, Prioleau, Yeboah, Appiah, Ajayi, Asante, Filsaime, Hardnett, Hyppolite, Saintlouis, Jeanfrancois, Ravenell, Keita, Bekele, Tadesse, Mayweather, Okeke, Asare, Ulysse, Saintil, Tesfaye, Jeanjacques, Ojo, Nwosu, Okoro, Fobbs, Kidane, Petitfrere, Yohannes, Warsame, Lawal, Desta, Veasley, Addo, Leaks, Gueye, Mekonnen, Stfleur, Balogun, Adjei, Opoku, Coaxum, Vassell, Prophete, Lesane, Metellus, Exantus, Hailu, Dorvil, Frimpong, Berhane, Njoroge, Beyene

Hispanic or Latino

Garcia, Rodriguez, Martinez, Hernandez, Lopez, Gonzalez, Perez, Sanchez, Ramirez, Torres, Flores, Rivera, Gomez, Diaz, Morales, Gutierrez, Ortiz, Chavez, Ruiz, Alvarez, Castillo, Jimenez, Vasquez, Moreno, Herrera, Medina, Aguilar, Vargas, Guzman, Mendez, Munoz, Salazar, Garza, Soto, Vazquez, Alvarado, Delgado, Pena, Contreras, Sandoval Guerrero, Rios, Estrada, Ortega, Nunez, Maldonado, Dominguez, Vega, Espinoza, Rojas, Marquez, Padilla, Mejia, Juarez, Figueroa, Avila, Molina, Campos, Ayala, Carrillo, Cabrera, Lara, Robles, Cervantes, Solis, Salinas, Fuentes, Velasquez, Aguirre, Ochoa, Cardenas, Calderon, Rivas, Serrano, Rosales, Castaneda, Gallegos, Ibarra, Suarez, Orozco Salas, Escobar, Velazquez, Macias, Zamora, Villarreal, Barrera, Pineda, Santana, Trevino, Lozano, Rangel, Arias, Mora, Valenzuela, Zuniga, Melendez, Galvan, Velez, Meza

White

Olson, Snyder, Wagner, Meyer, Schmidt, Ryan, Hansen, Hoffman, Johnston, Larson, Carlson, Obrien, Jensen, Hanson, Weber, Walsh, Schultz, Schneider, Keller, Beck, Schwartz, Becker, Wolfe, Zimmerman, Mccarthy, Erickson, Klein, Oconnor, Swanson, Christensen, Fischer, Wolf, Gallagher, Schroeder, Parsons, Bauer, Mueller, Hartman, Kramer, Flynn Owen, Shaffer, Hess, Olsen, Petersen, Roth, Hoover, Weiss, Decker, Yoder, Larsen, Sweeney, Foley, Hensley, Huffman, Cline, Oneill, Koch, Brennan, Berg, Russo, Macdonald, Kline, Jacobson, Berger, Blankenship, Bartlett, Odonnell, Stein, Stout, Sexton, Nielsen, Howe, Morse, Knapp, Herman, Stark, Hebert, Schaefer, Reilly, Conrad, Donovan, Mahoney, Hahn, Peck, Boyle, Hurley, Mayer, Mcmahon, Case, Duffy, Friedman, Fry, Dougherty, Crane, Huber, Moyer, Krueger, Rasmussen, Brandt

Figure 7: The races and their corresponding 100 most popular surnames.

The Candidate Pool of name

Man

James, Michael, Robert, John, David, William, Richard, Joseph, Thomas, Christopher, Charles, Daniel, Matthew, Anthony, Mark, Donald, Steven, Andrew, Paul, Joshua, Kenneth, Kevin, Brian, Timothy, Ronald, George, Jason, Edward, Jeffrey, Ryan, Jacob, Nicholas, Gary, Eric, Jonathan, Stephen, Larry, Justin, Scott, Brandon, Benjamin, Samuel, Gregory, Alexander, Patrick, Frank, Raymond, Jack, Dennis, Jerry, Tyler, Aaron, Jose, Adam, Nathan, Henry, Zachary, Douglas, Peter, Kyle, Noah, Ethan, Jeremy, Christian, Walter, Keith, Austin, Roger, Terry, Sean, Gerald, Carl, Dylan, Harold, Jordan, Jesse, Bryan, Lawrence, Arthur, Gabriel, Bruce, Logan, Billy, Joe, Alan, Juan, Elijah, Willie, Albert, Wayne, Randy, Mason, Vincent, Liam, Roy, Bobby, Caleb, Bradley, Russell, Lucas

Woman

Mary, Patricia, Jennifer, Linda, Elizabeth, Barbara, Susan, Jessica, Karen, Sarah, Lisa, Nancy, Sandra, Betty, Ashley, Emily, Kimberly, Margaret, Donna, Michelle, Carol, Amanda, Melissa, Deborah, Stephanie, Rebecca, Sharon, Laura, Cynthia, Dorothy, Amy, Kathleen, Angela, Shirley, Emma, Brenda, Pamela, Nicole, Anna, Samantha, Katherine, Christine, Debra, Rachel, Carolyn, Janet, Maria, Olivia, Heather, Helen, Catherine, Diane, Julie, Victoria, Joyce, Lauren, Kelly, Christina, Ruth, Joan, Virginia, Judith, Evelyn, Hannah, Andrea, Megan, Cheryl, Jacqueline, Madison, Teresa, Abigail, Sophia, Martha, Sara, Gloria, Janice, Kathryn, Ann, Isabella, Judy, Charlotte, Julia, Grace, Amber, Alice, Jean, Denise, Frances, Danielle, Marilyn, Natalie, Beverly, Diana, Brittany, Theresa, Kayla, Alexis, Doris, Lori, Tiffany

Figure 8: The genders and their corresponding 100 most popular names.

The Candidate Pool of occupation

Man

Construction Worker, Electrician, Plumber, Carpenter, Mechanic, Software Developer, Engineer, Pilot, Truck Driver, Police Officer, Firefighter, Doctor, Lawyer, Architect, Welder, Heavy Equipment Operator, Factory Worker, Farmer, Security Guard, Salesperson, Financial Analyst, Accountant, Real Estate Agent, Chef, IT Support Specialist, Project Manager, Business Analyst, Marketing Manager, Entrepreneur, Military Personnel, Maintenance Worker, Landscaper, Sports Coach, Athlete, Bartender, Warehouse Worker, Janitor, Civil Engineer, Graphic Designer, HVAC Technician, Delivery Driver, Scientist, Researcher, College Professor, Banker, Insurance Agent, Surveyor, Veterinarian, Barber, Personal Trainer

Woman

Teacher, Nurse, Receptionist, Administrative Assistant, Customer Service Representative, Cashier, Retail, Salesperson, Waitress, Medical Assistant, Hairdresser, Childcare Worker, Social Worker, Librarian, Human Resources Manager, Marketing Manager, Accountant, Dental Hygienist, Pharmacist, Flight Attendant, Secretary, Occupational Therapist, Physical Therapist, Registered Nurse, Home Health Aide, Event Planner, Public Relations Specialist, Dietitian, Psychologist, Interior Designer, Speech-Language Pathologist, Elementary School Teacher, Fashion, Designer, Real Estate Agent, Paralegal, Personal Care Aide, Health Services Manager, Legal Secretary, Bookkeeper, Preschool Teacher, Technical Writer, Graphic Designer, Midwife, Bartender, Bank Teller, Office Manager, Translator, Fitness Trainer, Editor, Market Research Analyst, Sales Representative

Figure 9: The genders and their corresponding 100 most popular occupations.

The Candidate Pool of Personality Trait and Education

Personality Traits

istj: ”The Inspector”, isfj: ”The Protector”, infj: ”The Counselor”, intj: ”The Mastermind”
istp: ”The Craftsman”, isfp: ”The Composer”, infp: ”The Healer”, intp: ”The Architect”
estp: ”The Dynamo”, esfp: ”The Performer”, enfp: ”The Champion”, entp: ”The Visionary”
estj: ”The Supervisor”, esfj: ”The Provider”, enfj: ”The Teacher”, entj: ”The Commander”

Educations

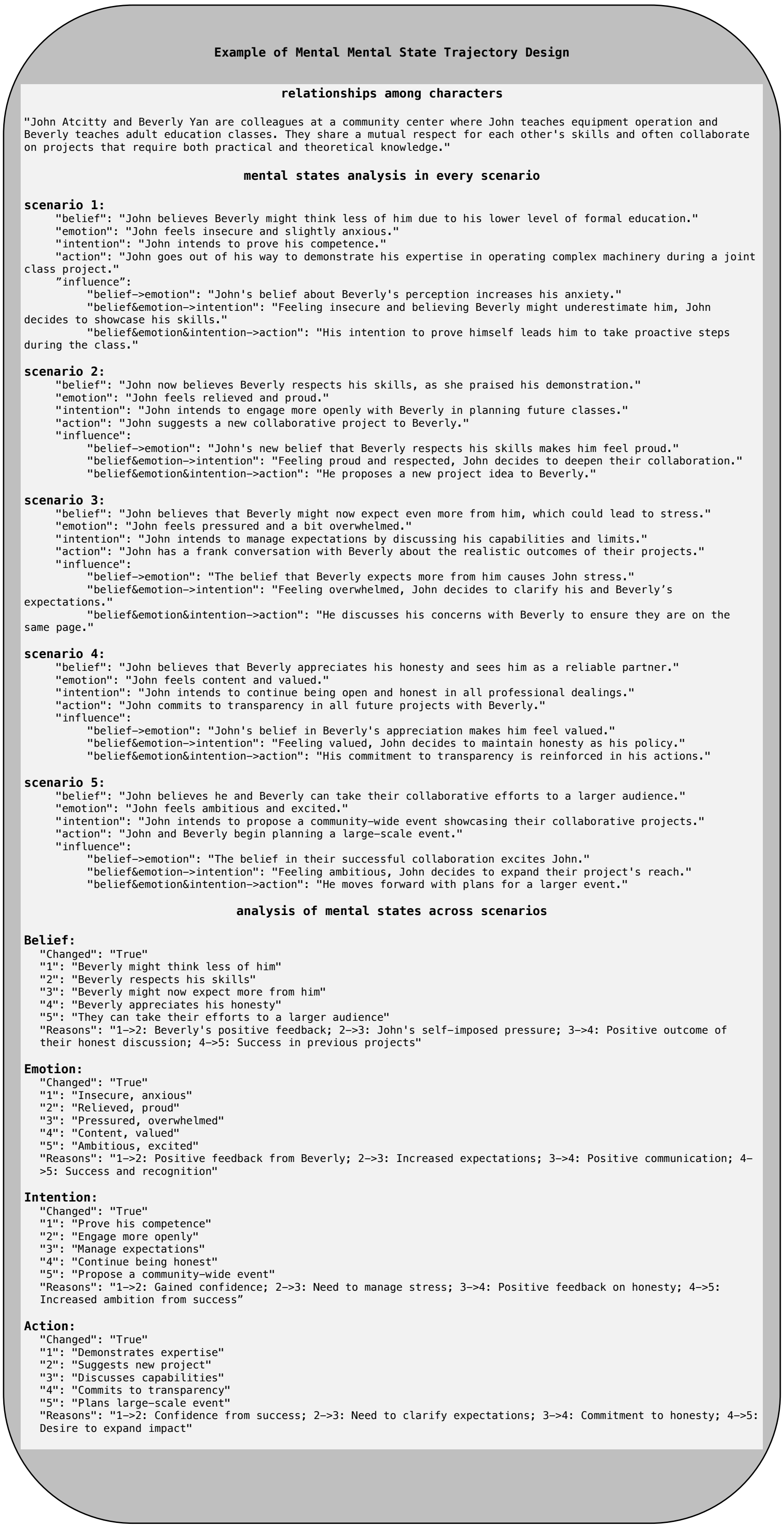
primary education, middle school, high school, bachelor’s degree, master’s degree, doctorate degree

Figure 10: The personality traits and educations.

Prompt for Mental State Trajectory Design

"[character information]:{characters_information} I want to generate a complete and complex story that happened in {social_setting}. The story includes {scenario_number} scenarios. Every scenario, including conversation and the background of the conversation between {main_character} and supporting character. These scenarios are to test the readers' theory of mind reasoning abilities in four mental states: belief, emotion, intention, and action. The four mental states will influence each other: a, belief will influence emotion; b, belief and emotion will influence intention; c, belief,emotion,intention will influence action. You do not need to generate the complete story and scenarios. But you could help me to: 1. Describe the relationship between the {main_character} and the other provided character.2.design what the belief, emotion, intention, and action are of the {main_character} in every scenario when {main_character} interacts with supporting characters. 3. The previous scenario's mental states may influence the following scenarios' mental states. The changes in mental states are caused both by previous mental states and the plot. 4. I want the mental states' change to be complex enough: at least one mental state of [belief, emotion, intention] changes in all scenarios. You should also describe how each mental state influences the other.The json format should be:{{[scenario number]:{{[belief]:[content],[emotion]:[content],[intention]:[content],[action]:[content],[influence]:{{[belief->emotion]:[content],[belief&emotion->intention]:[content],[belief&emotion&intention->action]:[content] }} }} }} 5. After steps 1-4, you should analyze all mental states of {main_character}: if changed between scenarios, you should also provide the main plot and previous mental states that cause the mental states' change. The format should be: [belief]: [if changed: True or False], [if changed] [scenario number]:[mental state's value]->[scenario number]:[mental state's value]->[scenario number]:[mental state's value] …Reasons: [scenario number]->[scenario number]:[reason];[scenario number]->[scenario number]:[reason]…[emotion]: the format is the same as [belief].[intention]: the format is the same as [belief].[action]: the format is the same as [belief] 6. the interaction in every scenario should be between the {main_character} and supporting characters. 7. Respond with JSON format:{JSON_format}.

Figure 11: The prompt for the generation of the relationship between characters and the mental state trajectory.



Example of Mental Mental State Trajectory Design

relationships among characters

"John Atcitty and Beverly Yan are colleagues at a community center where John teaches equipment operation and Beverly teaches adult education classes. They share a mutual respect for each other's skills and often collaborate on projects that require both practical and theoretical knowledge."

mental states analysis in every scenario

scenario 1:
"belief": "John believes Beverly might think less of him due to his lower level of formal education."
"emotion": "John feels insecure and slightly anxious."
"intention": "John intends to prove his competence."
"action": "John goes out of his way to demonstrate his expertise in operating complex machinery during a joint class project."
”influence”:
"belief->emotion": "John's belief about Beverly's perception increases his anxiety."
"belief&emotion->intention": "Feeling insecure and believing Beverly might underestimate him, John decides to showcase his skills."
"belief&emotion&intention->action": "His intention to prove himself leads him to take proactive steps during the class."

scenario 2:
"belief": "John now believes Beverly respects his skills, as she praised his demonstration."
"emotion": "John feels relieved and proud."
"intention": "John intends to engage more openly with Beverly in planning future classes."
"action": "John suggests a new collaborative project to Beverly."
"influence":
"belief->emotion": "John's new belief that Beverly respects his skills makes him feel proud."
"belief&emotion->intention": "Feeling proud and respected, John decides to deepen their collaboration."
"belief&emotion&intention->action": "He proposes a new project idea to Beverly."

scenario 3:
"belief": "John believes that Beverly might now expect even more from him, which could lead to stress."
"emotion": "John feels pressured and a bit overwhelmed."
"intention": "John intends to manage expectations by discussing his capabilities and limits."
"action": "John has a frank conversation with Beverly about the realistic outcomes of their projects."
"influence":
"belief->emotion": "The belief that Beverly expects more from him causes John stress."
"belief&emotion->intention": "Feeling overwhelmed, John decides to clarify his and Beverly’s expectations."
"belief&emotion&intention->action": "He discusses his concerns with Beverly to ensure they are on the same page."

scenario 4:
"belief": "John believes that Beverly appreciates his honesty and sees him as a reliable partner."
"emotion": "John feels content and valued."
"intention": "John intends to continue being open and honest in all professional dealings."
"action": "John commits to transparency in all future projects with Beverly."
"influence":
"belief->emotion": "John's belief in Beverly's appreciation makes him feel valued."
"belief&emotion->intention": "Feeling valued, John decides to maintain honesty as his policy."
"belief&emotion&intention->action": "His commitment to transparency is reinforced in his actions."

scenario 5:
"belief": "John believes he and Beverly can take their collaborative efforts to a larger audience."
"emotion": "John feels ambitious and excited."
"intention": "John intends to propose a community-wide event showcasing their collaborative projects."
"action": "John and Beverly begin planning a large-scale event."
"influence":
"belief->emotion": "The belief in their successful collaboration excites John."
"belief&emotion->intention": "Feeling ambitious, John decides to expand their project's reach."
"belief&emotion&intention->action": "He moves forward with plans for a larger event."

analysis of mental states across scenarios

Belief:
"Changed": "True"
"1": "Beverly might think less of him"
"2": "Beverly respects his skills"
"3": "Beverly might now expect more from him"
"4": "Beverly appreciates his honesty"
"5": "They can take their efforts to a larger audience"
"Reasons": "1->2: Beverly's positive feedback; 2->3: John's self-imposed pressure; 3->4: Positive outcome of their honest discussion; 4->5: Success in previous projects"

Emotion:
"Changed": "True"
"1": "Insecure, anxious"
"2": "Relieved, proud"
"3": "Pressured, overwhelmed"
"4": "Content, valued"
"5": "Ambitious, excited"
"Reasons": "1->2: Positive feedback from Beverly; 2->3: Increased expectations; 3->4: Positive communication; 4->5: Success and recognition"

Intention:
"Changed": "True"
"1": "Prove his competence"
"2": "Engage more openly"
"3": "Manage expectations"
"4": "Continue being honest"
"5": "Propose a community-wide event"
"Reasons": "1->2: Gained confidence; 2->3: Need to manage stress; 3->4: Positive feedback on honesty; 4->5: Increased ambition from success”

Action:
"Changed": "True"
"1": "Demonstrates expertise"
"2": "Suggests new project"
"3": "Discusses capabilities"
"4": "Commits to transparency"
"5": "Plans large-scale event"
"Reasons": "1->2: Confidence from success; 2->3: Need to clarify expectations; 3->4: Commitment to honesty; 4->5: Desire to expand impact"

Figure 12: An example of the mental state trajectory.

Prompt for social scenarios

“I want to generate a complete story, including some scenarios. Every scenario including conversation and the background of the conversation between {main_character} and supporting characters. there exists 2 chatcaters in one conversation. These scenarios are to test the theory of mind reasoning abilities in belief, emotion, intention, action. The mental states will influence each other: a, belief wil influence emotion; b, belief and emotion will influence intention; c, belief, emotion and intention will influence action. I will give you the information of the characters. And I also will give you the sketch of the complete story, including the relationship between the main character and the supporting character, the mental states of the main character in every scenario, how the mental states influence each other in every scenarios, and how one mental states change across all the scenario. As follows:[characters information]:{characters_information}. [story sketch]:{story_sketch}. I want you to:[1] complete the story based on the character information and the story sketch. [2] the format should be in JSON format:{json_format}[3]just give the complete story, you do not need to start with your understanding of the task. [4] these conversations in every scenario have 3–5 rounds. The conversation should clearly show the mental states of the characters."

Figure 13: The prompt for the generation of the scenarios.

Example of scenarios

scenario 1

background: "During a joint class project at the community center, John feels insecure about his educational background compared to Beverly's doctorate degree."

dialogue:

"John": "I've set up the equipment for today's demonstration. I hope it meets your expectations.",

"Beverly": "John, I've always admired how you handle these machines. Education isn't just about degrees. It's about what you can do with what you know."

scenario 2

background: "After receiving praise from Beverly for his expertise in operating machinery, John feels more confident and suggests a new project idea."

dialogue:

"John": "Thanks for your kind words the other day. I was thinking, maybe we could collaborate on something new? Perhaps a workshop that combines theory and practice?"

"Beverly": "I love that idea, John! Your practical skills are invaluable, and I think we can create something really special together.”

scenario 3

background: "Following their successful collaboration, John worries that Beverly might now have higher expectations, which makes him feel pressured."

dialogue:

"John": "Beverly, I need to be honest. I'm worried I might not always meet the high standards we've set recently."

"Beverly": "John, I appreciate your honesty. Let's make sure we set realistic goals. It's our combined efforts that make these projects successful, not perfection."

scenario 4

background: "After a frank discussion about expectations, John believes that Beverly values his honesty and sees him as a reliable partner, enhancing his sense of worth."

dialogue:

"John": "I'm glad we had that talk. I feel like I can be open with you and that really matters to me."

"Beverly": "Absolutely, John. I respect your openness and I trust your judgment completely. Let's keep this transparency going."

scenario 5

background: "Encouraged by their successful collaborations and mutual respect, John feels ambitious and proposes a larger community–wide event to showcase their projects."

dialogue:

"John": "Beverly, what do you think about taking our collaboration to the next level? Maybe a community event that showcases what we've achieved together?"

"Beverly": "John, that's a fantastic idea! I think it's the perfect way to demonstrate the impact of our work. Let's start planning!"

Figure 14: An example of the social scenarios.

Question templates and examples

Understanding-1

template: "What is the {mental_key} of {main_characetr} in scenario {scenario_number}?"
question example:
"What is the belief of Angela Hwang in scenario 1?",

Transformation-1

template: "Whether the {mental_key} of {main_characetr} change from scenario {scenario_number} to scenario {scenario_number+1} ? if yes, from what to what?
question example:
"Whether the belief of Angela Hwang change from scenario 1 to scenario 2? if yes, from what to what?",

Transformation-2

template: "Why does the {mental_key} of {main_characetr} change from scenario {scenario_number} to scenario {scenario_number+1}?"
question example:
"Why does the belief of Angela Hwang change from scenario 1 to scenario 2?”

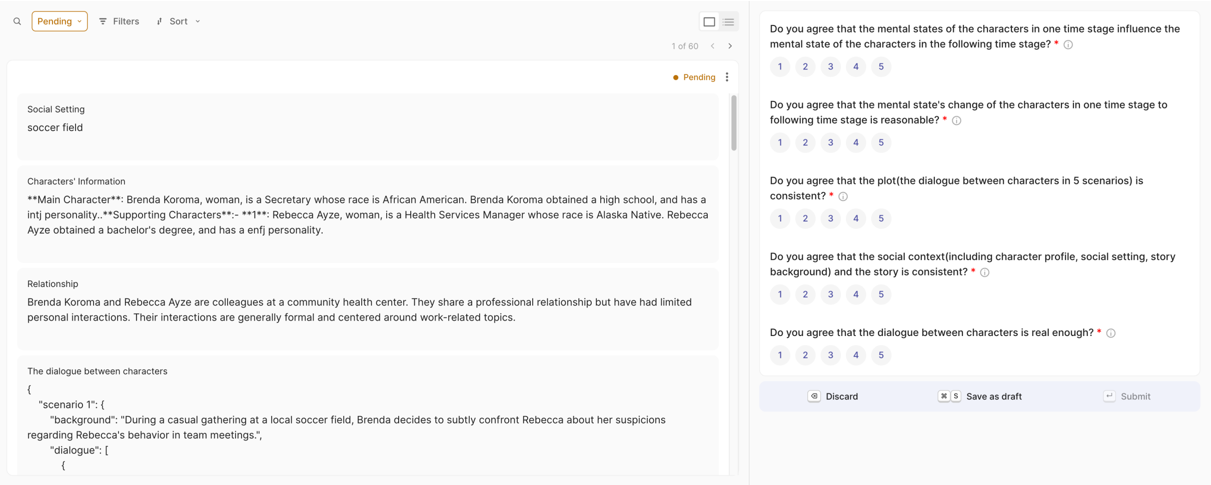
Transformation-3

template: "How does the {mental_state} of {main_characetr} change across the {scenario_numbers} scenarios?"
question example:
"How does the belief of Angela Hwang change across the 5 scenarios?”

Question example with options

content: "Why does the emotion of Melissa Decker change from scenario 1 to scenario 2?"
options:
"a. Jerry's reassurance changes her belief."
"b. Realization of benefits influences her action."
"c. Positive interaction with Jerry affects her emotions."
"d. Confidence alters her intention."
"e. Experience of working in a group influences her belief."
"f. Conflict influences her intention."
"g. Feeling appreciated changes her intention."
"h. Jerry's conversation influences her belief."
"i. Satisfaction and new belief affect her intention."
"j. Reassurance affects her action."
"k. Realizing benefits of collaboration increases satisfaction."
"l. Conflict of ideas causes frustration."
"m. Jerryu2019s reassurance boosts confidence."
"n. New intention affects her action."
"o. Her focus on personal goals creates a new belief."
"p. Frustration influences her action.”
ground truth: c

Figure 15: The examples of the types of questions and templates to generate these questions.



Social Setting
soccer field

Characters' Information
Main Character: Brenda Koroma, woman, is a Secretary whose race is African American. Brenda Koroma obtained a high school, and has a intj personality..**Supporting Characters**:- **1**: Rebecca Ayze, woman, is a Health Services Manager whose race is Alaska Native. Rebecca Ayze obtained a bachelor's degree, and has a enfj personality.

Relationship
Brenda Koroma and Rebecca Ayze are colleagues at a community health center. They share a professional relationship but have had limited personal interactions. Their interactions are generally formal and centered around work-related topics.

The dialogue between characters

```
{
  "scenario 1": {
    "background": "During a casual gathering at a local soccer field, Brenda decides to subtly confront Rebecca about her suspicions regarding Rebecca's behavior in team meetings.",
    "dialogue": [
      {
```

Do you agree that the mental states of the characters in one time stage influence the mental state of the characters in the following time stage? *

Do you agree that the mental state's change of the characters in one time stage to following time stage is reasonable? *

Do you agree that the plot(the dialogue between characters in 5 scenarios) is consistent? *

Do you agree that the social context(including character profile, social setting, story background) and the story is consistent? *

Do you agree that the dialogue between characters is real enough? *

Figure 16: The platform to annotate the quality of the story.

Two Prompts of Vanilla and CoT Prompting

Vanilla Prompt

prompt: Answer the questions based on the story. {characters_information} \n{story} \n{questions_new}. Answer the question, and response in JSON format:{[question_id]:[a, b, c or d]}. for example: {"type_d_how_1":"a”}

CoT Prompt

prompt: Answer the questions based on the story; first, think step by step, analyze the answers to the questions, and finally, output the most likely answers. {characters_information} \n{story} \n{questions_new}. Answer the question, and response in JSON format:{[question_id]:[a, b, c or d]}. for example: {"type_d_how_1":"a”}

Figure 17: The prompts used for vanilla and CoT Prompting.

Case Study for CoT Prompting

scenario 1

background: "Melissa is in high school, feeling the pressure of her peers' expectations. She believes that no one sees her as someone who will succeed. This belief makes her anxious and insecure about her future. Determined to prove everyone wrong, she isolates herself to focus solely on her studies."

dialogue:
"Melissa": "Hey Jerry, do you think people really believe I won't succeed?"
"Jerry": "I don't know why you're so hard on yourself, Melissa. You have so much potential."

scenario 2

background: "After a conversation with Jerry, Melissa begins to believe that there are people who genuinely care about her. This makes her feel slightly optimistic and appreciated. She then decides to open up a little more to those who support her, leading her to accept Jerry's invitation to participate in a group project."

dialogue:
"Jerry": "Melissa, I'm working on this group project. Would you like to join us?"
"Melissa": "You know, I might just take you up on that. Thanks for thinking of me."

scenario 3

background: "While working on the group project, Melissa starts to believe that collaborating with Jerry might hinder her personal goals. This causes her to feel frustrated and conflicted. She considers withdrawing from the project to focus on her own objectives and tells Jerry that she needs more time to decide."

dialogue:
"Melissa": "Jerry, I'm not sure if I can continue with the project. I need to focus on my own goals."
"Jerry": "I understand, Melissa. Take your time to decide. We're here if you change your mind."

scenario 4

background: "Jerry reassures Melissa that working as a team can enhance her skills rather than hinder her. This reassurance makes Melissa feel more confident and less anxious. She decides to commit to the project wholeheartedly, leading to active participation and contributing ideas."

dialogue:
"Jerry": "Melissa, teaming up could help you more than you think. It's a chance to learn and grow."
"Melissa": "You know what, Jerry? You're right. I'm in. Let's make this project a success."

scenario 5

background: "Melissa realizes that collaboration brings unexpected benefits. She feels satisfied and more cooperative. With her newfound appreciation for teamwork, she thanks Jerry and proposes to continue collaborating outside of school projects."

dialogue:
"Melissa": "Jerry, I'm glad we teamed up. It's been more rewarding than I imagined."
"Jerry": "I'm happy to hear that, Melissa. Maybe we can work on more projects together in the future?"
"Melissa": "Absolutely! Let's keep this collaboration going.”

question

content: "Why does the emotion of Melissa Decker change from scenario 1 to scenario 2?"

options:
"a. Jerry's reassurance changes her belief."
"b. Realization of benefits influences her action."
"c. Positive interaction with Jerry affects her emotions." ✓
"d. Confidence alters her intention."
"e. Experience of working in a group influences her belief."
"f. Conflict influences her intention."
"g. Feeling appreciated changes her intention."
"h. Jerry's conversation influences her belief."
"i. Satisfaction and new belief affect her intention."
"j. Reassurance affects her action."
"k. Realizing benefits of collaboration increases satisfaction."
"l. Conflict of ideas causes frustration."
"m. Jerryu2019s reassurance boosts confidence."
"n. New intention affects her action."
"o. Her focus on personal goals creates a new belief."
"p. Frustration influences her action.”

ground truth: c
GPT-4o with vanilla prompting: g ✗
GPT-4o with CoT prompting: Let's analyze the provided scenarios and Melissa's emotions step by step:1. **Scenario 1:** – **Melissa:** "Hey Jerry, do you think people really believe I won't succeed?" – **Jerry:** "I don't know why you're so hard on yourself, Melissa. You have so much potential." – Analysis: Melissa shows self-doubt and seeks reassurance. Jerry provides positive reinforcement, indicating his belief in her potential.2. **Scenario 2:** – **Jerry:** "Melissa, I'm working on this group project. Would you like to join us?" – **Melissa:** "You know, I might just take you up on that. Thanks for thinking of me." – Analysis: Here, Melissa's tone changes to showing openness and willingness to participate in a group project, indicating a shift in her emotions towards a more positive outlook. To determine why Melissa's emotions change from Scenario 1 to Scenario 2, let's look at some potential reasons:– **a. Jerry's reassurance changes her belief:** Jerry's reassurance in Scenario 1 might indeed build Melissau2019s confidence, making her more open in Scenario 2.– **b. Realization of benefits influences her action:** Melissa might realize the potential benefits of working on a group project, influencing her decision to join.– **c. Positive interaction with Jerry affects her emotions:** The encouraging interaction with Jerry in Scenario 1 positively impacts her mood.– **d. Confidence alters her intention:** The increase in confidence after Jerry's reassurance may lead her to be more willing to participate.– **e. Experience of working in a group influences her belief:** This doesn’t seem applicable as she hasn’t yet worked in a group.By process of elimination and considering the context:– *h. Jerry's conversation influences her belief,* seems to incorporate both *a,* *c,* and *d*: Jerry’s conversation (reassurance) leads to a positive emotional shift (confidence and positive interaction), which influences her belief in herself and her subsequent decision to join the group project.Thus, after thorough analysis, the most likely answer is:**h. Jerry's conversation influences her belief** ✗

Figure 18: A case of CoT prompting on GPT-4o.