TrajToM: Evaluating Theory of Mind over Mental State Trajectories in Social Dialogues

Chenglong Wu*, **Xi Wang***, **Jinzhi Liao**†, **Xiang Zhao**†
National Key Laboratory of Big Data and Decision,
National University of Defense Technology, China
{liaojinzhi12, xiangzhao}@nudt.edu.cn

Abstract

Theory of Mind (ToM) is crucial for understanding social dialogue, where characters' mental states are rarely stated explicitly but evolve through interaction and shape later behavior. Existing ToM benchmarks focus mainly on selected dialogue turns or state changes across separate scenarios, leaving within-dialogue mental state trajectories underexplored. We introduce TrajToM[1], a benchmark for evaluating whether LLMs can reason over evolving emotions, beliefs, intentions, and desires in continuous social dialogue. Built with a Sketch-first construction paradigm, TrajToM contains 700 dialogues and 18,165 multiple-choice questions across mental state reasoning, behavioral reasoning, and trajectory reasoning. Experiments reveal a substantial human–model gap, with GPT-5.2 reaching 67.4% overall accuracy compared with 88.2% for humans, and show persistent bottlenecks in second-order reasoning, behavior explanation, and change direction assessment. We further examine MARTA as an exploratory trajectory-aligned process-supervision framework, showing that explicit supervision over mental state trajectories can improve ToM reasoning.

1 Introduction

In social dialogue, understanding a character's behavior often requires more than identifying what they think or feel at a single moment (Yamashita et al., 2025). Characters reveal mental states gradually through contextual cues, revise them as new information appears, and express these changes in later behavior (Schneider, 2001). This dynamic process is central to Theory of Mind (ToM), the ability to attribute mental states to others and use them to explain observed behavior (Premack and Woodruff, 1978; Baron-Cohen et al., 1985). In dialogue, such mental states include beliefs, emotions, intentions, and desires that are rarely stated explicitly but shape how characters respond, revise their attitudes, and coordinate with others (Kosinski, 2024; Wang et al., 2024). As LLMs are increasingly deployed in dialogue-based AI systems, such as assistive robots (Puig et al., 2023), AI tutors (Andriella et al., 2025), and healthcare chatbots (Agarwal et al., 2025), their ability to track evolving mental states during interaction warrants careful evaluation.

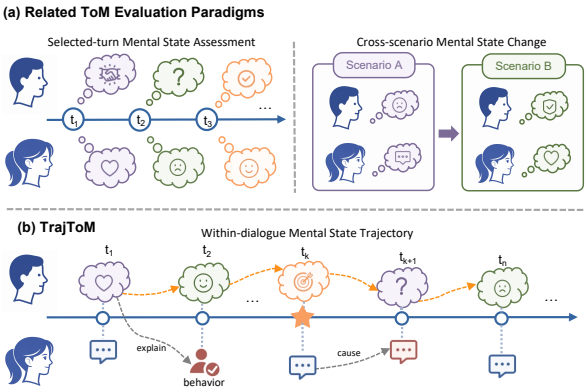


Figure 1: Comparison of ToM evaluation paradigms.

Existing ToM benchmarks have expanded LLM evaluation from controlled information-asymmetry tasks to richer narrative (Zhou et al., 2025), embodied (Fan et al., 2025), and dialogue-based settings (Liu et al., 2025; Kim et al., 2023). These benchmarks provide important tests of whether models can infer mental states from complex contexts. Within dialogue-based settings, phenomena such as negotiation and persuasion often depend on how characters' mental states evolve within the same interaction (Chan et al., 2024; Yu et al., 2025). As illustrated in Figure 1, however, dialogue-based evaluations commonly ask models to identify mental states at selected turns, while temporal evaluations may model state changes across separate

*Equal contribution.
†Corresponding authors.
[1]https://github.com/Louis1316/TrajToM-public.git

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37968–37990
October 24-29, 2026 ©2026 Association for Computational Linguistics

but connected scenarios (Xiao et al., 2025). Even when mental states are tracked across turns (Yu et al., 2025), turn-wise evaluation provides limited evidence that models can connect earlier cues, later behavior, and change triggers into a coherent mental state trajectory. Thus, within-dialogue trajectory-level reasoning remains underexplored, including whether LLMs can infer phase-specific states, explain behavior, and integrate distributed evidence to track what changes, when it changes, and why it changes across a continuous dialogue.

To address this gap, we introduce TrajToM, a benchmark that evaluates whether LLMs can reason over mental state trajectories within continuous social dialogue. TrajToM represents each dialogue as a structured trajectory over emotion, belief, intention, and desire. To capture how these states evolve, each dialogue is organized into four phases: opening, development, key moment, and closing. Based on this structure, TrajToM evaluates ToM reasoning at three levels: Mental State Reasoning (MSR), Behavioral Reasoning (BR), and Trajectory Reasoning (TR). Conceptually, MSR corresponds to explicit mental-state inference, BR to applied ToM, and TR extends these capabilities to the temporal evolution of mental states (Hutchins et al., 2016; Gu et al., 2026; Yadav et al., 2026). Together, these levels test phase-specific state inference, behavior explanation through latent mental states and information asymmetry, and trajectory-level change reasoning over key shifts, change directions, and change causes.

Because trajectory-level ToM involves latent and evolving mental states, ground-truth verifiability is critical. TrajToM adopts a Sketch-first paradigm that specifies phase-level states, key shifts, triggers, and causal links before dialogue realization, making questions and answers traceable to structured Sketch fields. Automatic filtering and human review validate construction quality, Sketch–dialogue consistency, and question answerability. The resulting benchmark contains 700 dialogues and 18,165 multiple-choice questions, covering a range of social dialogue scenarios and trajectory-based ToM reasoning skills.

Experiments on a range of LLMs reveal a substantial human–model gap, with GPT-5.2 reaching 67.4% overall accuracy compared with 88.2% for humans. Models especially struggle with second-order reasoning, behavior explanation, and change direction assessment, suggesting limitations in integrating distributed evidence for mental state evolution. We further examine MARTA as an exploratory trajectory-aligned process-supervision framework that leverages structured mental state trajectories to supervise intermediate reasoning, demonstrating the potential of trajectory-aligned supervision for ToM reasoning. Our contributions are summarized as follows:

- We introduce TrajToM, a benchmark for evaluating within-dialogue mental state trajectory reasoning in continuous social dialogue.
- We construct TrajToM with a Sketch-first paradigm that specifies mental state trajectories before dialogue realization, yielding 700 dialogues and 18,165 traceable questions across MSR, BR, and TR.
- We evaluate a range of LLMs on TrajToM, revealing systematic bottlenecks in second-order reasoning, behavior explanation, and trajectory change reasoning, and further study MARTA to show the potential of trajectory-aligned process supervision.

2 Related Work

2.1 ToM Benchmarks

Existing ToM benchmarks have expanded from Sally–Anne-style tasks (Le et al., 2019) to narrative (Zhou et al., 2025), multimodal (Jin et al., 2024; Shi et al., 2025), embodied (Fan et al., 2025), and dialogue-based settings. Dialogue-based benchmarks further evaluate belief reasoning (Kim et al., 2023), strategic and tactful communication (Chan et al., 2024; Yu et al., 2025; Liu et al., 2025), role-played interaction (Shinoda et al., 2025), and recommendation (Li et al., 2026). However, as summarized in Table 1, most dialogue-based benchmarks cover only part of the required abilities, and few explicitly evaluate trajectory change reasoning over mental state changes. Although recent benchmarks consider mental state dynamics across separate scenarios (Xiao et al., 2025) or track states across dialogue turns (Yu et al., 2025), they rarely require models to reason about what changes, when it changes, and why it changes within a continuous dialogue. TrajToM complements these benchmarks by jointly assessing phase-specific mental state inference, behavior-grounded explanation, and trajectory-level change reasoning.

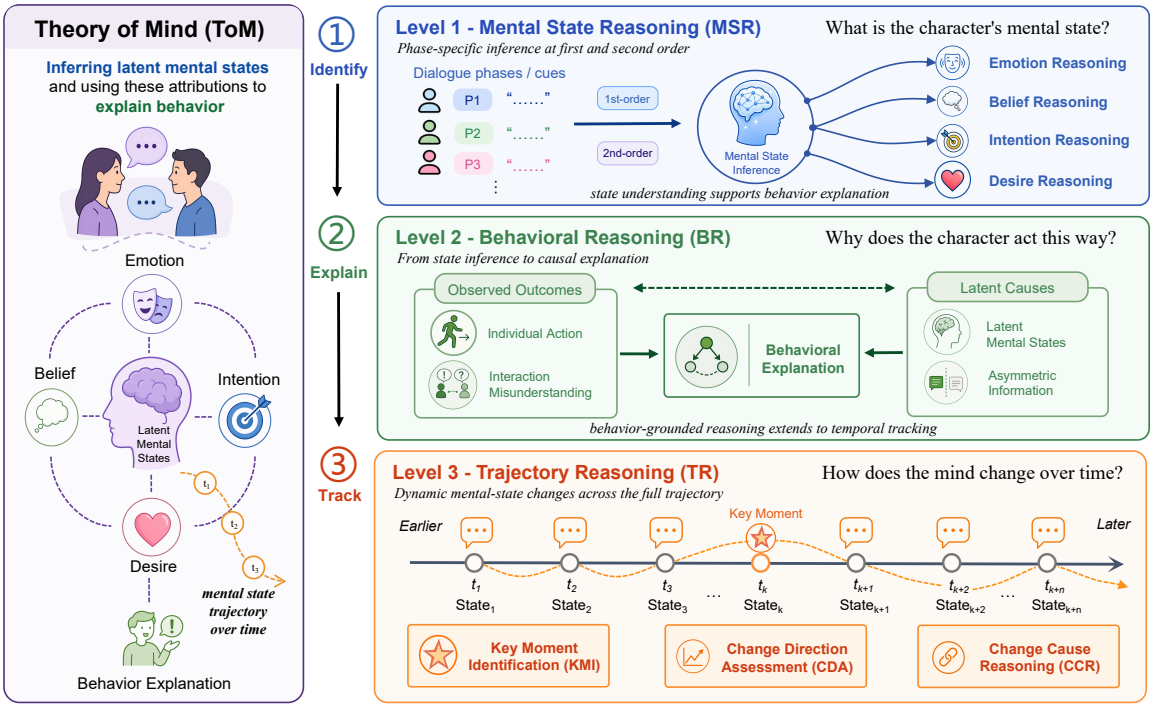


Figure 2: Overview of our TrajToM benchmark.

Benchmark	2nd	BR	TCR	#D	#Q
FANToM (Kim et al., 2023)	✓	✗	✗	256	10,000
NegotiationToM (Chan et al., 2024)	✓	✓	✗	395	13,800
ToMATO (Shinoda et al., 2025)	✓	✗	✗	753	5,400
TactfulToM (Liu et al., 2025)	✓	✗	✗	100	6,700
PersuasiveToM (Yu et al., 2025)	✓	✓	✗	525	18,878
RecToM (Li et al., 2026)	✗	✓	✗	336	20,524
TrajToM (ours)	✓	✓	✓	700	18,165

Table 1: Comparison with other dialogue-based ToM benchmarks.

Note. 2nd = second-order ToM; BR = behavioral reasoning; TCR = trajectory change reasoning over mental state changes; #D = number of dialogues; #Q = number of questions.

2.2 ToM Enhancement

Existing ToM enhancement methods mainly follow three lines. Prompt-based methods guide models through perspective simulation (Wilf et al., 2024; Jung et al., 2024), query decomposition (Hou et al., 2024), or intermediate state tracking (Sarangi et al., 2025; Wu et al., 2026a), but are sensitive to prompt design and do not update the model's reasoning policy. Structured methods use Bayesian (Jin et al., 2024; Shi et al., 2025; Zhang et al., 2025b), symbolic (Wu et al., 2025), or logic-based (Tang and Belle, 2024) mechanisms to represent mental states, but require explicit state or transition assumptions. Recent RL-based methods use answer, format, or process rewards for ToM and social reasoning (Lu et al., 2025; Wu et al., 2026b), but these rewards are not explicitly aligned with latent mental state trajectories. MARTA complements these methods by exploring trajectory-aligned process supervision across dialogue phases.

3 TrajToM Benchmark

3.1 Benchmark Formulation

Task Definition. Following existing ToM benchmarks (Chan et al., 2024; Jin et al., 2024), TrajToM formulates each item as a single-answer multiple-choice question. Each instance is denoted as $(D, Q, \mathcal{A}, y)$, where D is a two-character social dialogue, Q is the question, $\mathcal{A}$ is the candidate set, and y is the gold answer. The model predicts $\hat{y} \in \mathcal{A}$ and is evaluated by accuracy. Questions target mental states, behaviors, and state changes, framing ToM evaluation as reasoning over **mental state trajectories** within a continuous dialogue.

Mental State Trajectory. A mental state trajectory represents a character's phase-level mental states and their transitions across a dialogue. Following prior work on mental and affective state reasoning (Chen et al., 2024; Tong et al., 2026), we characterize each trajectory along four dimensions: *emotion*, *belief*, *intention*, and *desire*, covering affective states (Ortony et al., 1988) and belief–desire–intention (BDI) attribution (Bratman, 1987). Each dialogue is segmented into four functional phases: *opening*, *development*, *key moment*, and

closing, which scaffold initial states, evidence accumulation, major shifts, and final states.

Question Taxonomy. As illustrated in Figure 2, TrajToM organizes ToM evaluation into three levels that move from phase-specific state inference to behavior explanation and trajectory change reasoning, with examples in Appendix A. This hierarchy avoids evaluating mental state change in isolation, since models must first infer phase-specific states, connect these states to observable behavior, and then integrate distributed evidence to reason about when, how, and why states change. **Level 1: Mental State Reasoning (MSR)** assesses phase-specific mental state inference at both first and second order, covering emotion, belief, intention, and desire. **Level 2: Behavioral Reasoning (BR)** tests whether LLMs can explain actions and misunderstandings by considering latent mental states and information asymmetry. **Level 3: Trajectory Reasoning (TR)** operationalizes trajectory change reasoning through *key moment identification* (KMI), *change direction assessment* (CDA), and *change cause reasoning* (CCR). KMI locates the turning point of a significant shift, CDA characterizes how a mental dimension evolves across phases, and CCR identifies the statement or event that triggers the shift.

3.2 Benchmark Construction

A key challenge in LLM-generated ToM benchmarks is ensuring that gold labels for latent mental states are supported by dialogue evidence. If scenarios, questions, and answers are generated jointly without explicit structural constraints, the resulting "correct" answers may reflect LLM hallucinations or self-consistent assumptions rather than inferable evidence (Long et al., 2024; Shapira et al., 2024). To improve the verifiability of gold answers, we adopt a **Sketch-first paradigm**, in which we first specify a structured mental state trajectory as a *Sketch*, then generate the dialogue under this trajectory, and finally derive questions whose answers are traceable to Sketch fields. We further validate profile plausibility, Sketch–dialogue consistency, and question answerability so that retained answers are supported by dialogue evidence. As shown in Figure 3, TrajToM is constructed through four stages: profile generation, Sketch construction, dialogue generation, and question generation, with additional details in Appendix B.

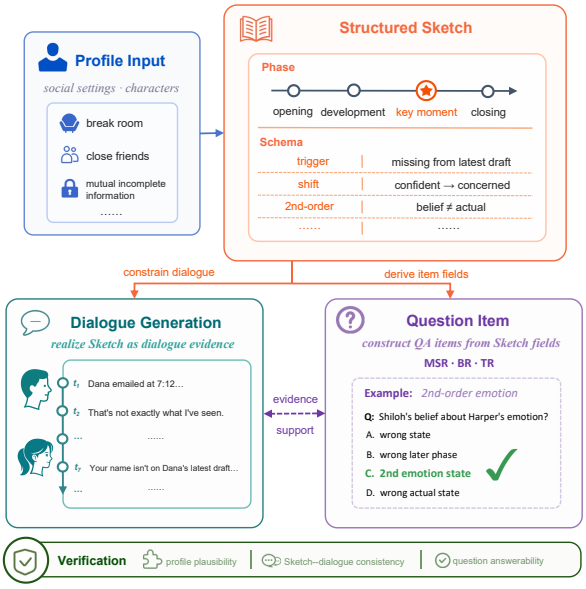


Figure 3: Overview of the Sketch-first construction pipeline.

Profile Generation. Each scenario begins with a profile specifying two characters and their social situation. Inspired by conditional sampling strategies (Zhang et al., 2025a), we construct profiles from ToM-relevant social context, character, and relationship attributes drawn from social scenario taxonomies and character profile conventions (Ziems et al., 2023; Zhou et al., 2024). These attributes define a large combinatorial space, and we sample a controlled subset while filtering incompatible pairings, such as first-meeting scenarios with pre-existing intimate relationships.

Sketch Construction. Given a profile, we use an LLM to produce a *Sketch*, a structured JSON trajectory that specifies phase-level mental states, key shifts, and causal links before dialogue generation. For each character c, phase ϕ, and mental state dimension d, the Sketch records a first-order state. It also specifies the key moment, its triggering event, and causal chains linking observable behaviors to latent mental states. For second-order reasoning, it stores $s^{(2)}_{c\to c',\phi,d} = (\textit{believed}, \textit{actual}, \textit{trigger})$, where *believed* is c's belief about c''s state on d at ϕ, *actual* is the Sketch-specified state of c', and *trigger* is the supporting cue later realized in dialogue. This trigger makes second-order answers traceable to explicit evidence. We filter invalid Sketches by structural completeness and consistency with the information-asymmetry setting, and five graduate researchers familiar with ToM reviewed trajectory plausibility and trigger clarity.

Dialogue Generation. The Sketch guides the generation of an approximately 18-turn dialogue between the two characters, realizing the specified phase-level mental states, causal chains, and key moment. To preserve ToM difficulty while maintaining answerability, characters reveal mental states through dialogue-grounded cues, such as behavior, implication, and conversational tension, while direct self-disclosure and inner monologue are disallowed. The key moment must also emerge naturally rather than being explicitly announced. Human reviewers assess Sketch–dialogue consistency and conversational naturalness, and failed dialogues are regenerated.

Question Generation. To ensure answer fidelity, gold answers are first determined from Sketch fields, while the LLM only verbalizes questions and options. Task-specific templates select the queried character, phase, mental-state dimension, and reasoning target. For MSR, first-order answers are extracted from phase-level states, while second-order answers are obtained from the Sketch tuple $s^{(2)}_{c\to c',\phi,d}$. For BR, answers are derived from Sketch-specified causal chains that explain stable behavior patterns or interaction-level misunderstandings. For TR, answers are derived from key moments, state shifts, and triggers, corresponding to key moment identification, change direction assessment, and change cause reasoning. Distractors are constructed from structurally related but incorrect alternatives, such as states from another phase, character, mental dimension, or an incorrect causal mechanism. Reviewers check item clarity, answerability, and distractor validity, and all invalid items are subsequently discarded.

3.3 Dataset Statistics

TrajToM comprises 700 dialogues and 18,165 multiple-choice questions, averaging 18.0 turns and 25.9 questions per dialogue. As shown in Figure 4, it covers three levels: MSR with 9,095 questions (50.1%), BR with 2,100 questions (11.6%), and TR with 6,970 questions (38.4%). It includes eight subtypes: four MSR dimensions, one BR category, and three TR subtypes. BR is further organized into three balanced behavior-explanation patterns. Among order-sensitive MSR and BR questions, 7,000 are first-order and 4,195 are second-order. Additional BR-pattern and phase/span statistics are provided in Appendix C.

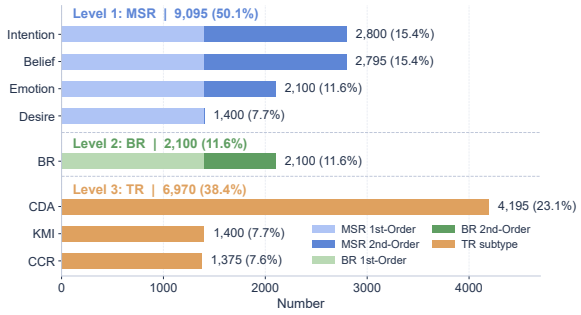


Figure 4: Distribution of TrajToM across reasoning levels, subtypes, and reasoning orders.

4 MARTA: Trajectory-Aligned Process Supervision

Beyond evaluation, the structured Sketches in TrajToM also provide a natural source of process-level supervision for improving ToM reasoning. We further study MARTA (Mental State Trajectory Alignment for ToM Reasoning), an exploratory training framework that complements TrajToM by using Sketch fields for trajectory-aligned training without relying on LLM judges or additional reward models. Outcome-only optimization may improve answer accuracy, but it does not ensure that the model reaches correct answers through trajectory-consistent reasoning. For each input $(D, Q, \mathcal{A})$, the model generates a reasoning chain τ and an answer $\hat{y}$. The associated Sketch S is used only for reward computation during training, and inference uses only $(D, Q, \mathcal{A})$.

4.1 Supervised Fine-Tuning for Structured Reasoning

We first perform supervised fine-tuning (SFT) to initialize the model with a structured ToM reasoning format. For each input $x = (D, Q, \mathcal{A})$, the target response r consists of a reasoning chain τ enclosed by `<think>` tags and a gold option $y^\star$ enclosed by `<answer>` tags. The reasoning chain follows three stages: `[Evidence]`, `[Inference]`, and `[Conclusion]`, which cite dialogue evidence, infer latent mental states or behavior explanations, and state the final conclusion. SFT targets are converted from answer explanations, with details provided in Appendix D.1. The model is trained with the standard negative log-likelihood objective:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,r)} \log \pi_\theta(r \mid x), \qquad (1)$$

where π_θ denotes the model policy.

4.2 Trajectory-Aligned GRPO

SFT initializes the reasoning format, but it does not explicitly optimize whether the reasoning conclusion is supported by the mental state trajectory underlying the answer. We therefore further refine the model with Group Relative Policy Optimization (GRPO) (Guo et al., 2025), using TrajToM Sketch fields as verifiable targets for process-level rewards. For each input, the policy samples N candidate responses, each containing a reasoning chain and a predicted answer, and updates the policy with the standard clipped GRPO objective.

The total reward is defined as

$$R_{\text{total}} = R_{\text{fmt}} \cdot (R_{\text{out}} + \lambda R_{\text{proc}}) \cdot R_{\text{len}}, \quad (2)$$

where R_{fmt} checks the required output format, R_{out} evaluates answer correctness, R_{proc} rewards conclusions that are aligned with Sketch-derived evidence and structured reasoning, and R_{len} penalizes overly long or repetitive responses. The coefficient λ is an outcome-calibrated trust weight computed from group-level answer correctness, reducing the influence of process rewards when outcome signals are unreliable.

The process reward is computed as

$$R_{\text{proc}} = w_{\text{conc}} R_{\text{conc}} + w_{\text{struct}} R_{\text{struct}}, \quad (3)$$

where R_{conc} compares the conclusion span with the corresponding Sketch-derived target, and R_{struct} checks the required `[Evidence]`, `[Inference]`, and `[Conclusion]` stages. Additional details of the GRPO objective and reward computation are provided in Appendix D.2.

5 Experiments

5.1 Experimental Setup

Models and Baselines. We evaluate twelve representative LLMs covering both API-accessed and open-weight models: Claude-Haiku-4.5, Claude-Sonnet-4.5, DeepSeek-V3.2, Gemini-2.5-Flash (Gemini Team, 2025), GPT-5.2, Llama-3.1-8B, Llama-3.3-70B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Mixtral-8x7B, Qwen3.5-4B, Qwen3.5-9B, and Qwen3.5-35B. For controlled comparison on the same base model, we compare Qwen3.5-9B with zero-shot CoT prompting (Wei et al., 2022), SFT, and MARTA. MARTA is trained with $N = 8$ rollouts per prompt on 4 NVIDIA A800 80GB GPUs.

Evaluation. We use multiple-choice accuracy as the primary metric. To avoid leakage across questions derived from the same dialogue, we split TrajToM at the dialogue level. The test set contains 350 dialogues and 9,067 questions, while the remaining 350 dialogues are used for training and development. For human performance, independent graduate participants not involved in dataset construction answer a stratified sample of the test set across reasoning levels and subtypes under the same input format as models. Details of model versions, prompts, hyperparameters, data splits, and human evaluation are provided in Appendix E.

5.2 Main Results

Table 2 reports subtype-level results on TrajToM, while Figure 5 summarizes performance across the three reasoning levels.

TrajToM Exposes a Persistent Human–Model Gap. TrajToM remains challenging for current LLMs. The human baseline achieves 88.2 overall accuracy, whereas the strongest general-purpose model, GPT-5.2, reaches 67.4. Even MARTA, trained with trajectory-aligned supervision, obtains 72.5 and still leaves a clear gap to humans. This suggests that strong language understanding alone is insufficient for robust ToM reasoning in continuous social dialogues, where models must infer and update latent mental states across dialogue phases rather than answer from isolated textual cues.

Scaling Improves ToM Performance but Does Not Eliminate Core Difficulty. Among general-purpose models, GPT-5.2 achieves the best overall performance, followed by Claude-Sonnet-4.5 and Llama-3.3-70B. Open-weight models show clear scale-related trends, with Llama-3.3-70B substantially outperforming Llama-3.1-8B and Qwen3.5-35B improving over Qwen3.5-9B and Qwen3.5-4B, suggesting that larger and stronger models generally perform better on TrajToM. However, all remain far below the human baseline, indicating that model scaling alone does not fully address trajectory-based ToM reasoning.

Subtype Results Reveal Uneven ToM Difficulty. Performance varies substantially across subtypes. Although TR shows relatively high level-wise accuracy in Figure 5, Table 2 shows that this mainly comes from KMI and CCR, which are often anchored in salient turning points or triggering events. CDA is harder because it requires comparing men-

Category	Model	MSR				BR	TR			Overall
		Emotion	Belief	Intention	Desire	BR	KMI	CDA	CCR	Avg.
Human	Human	86.4	79.2	84.8	96.8	88.3	97.1	82.7	90.1	88.2
API Models	Claude-Haiku-4.5	57.0	48.1	58.8	92.4	64.8	86.0	58.8	78.1	63.8
	Claude-Sonnet-4.5	58.3	50.7	59.0	93.7	63.3	91.1	63.9	81.3	66.1
	Gemini-2.5-Flash	48.5	38.6	51.1	92.0	57.6	83.3	54.7	69.4	57.5
	DeepSeek-V3.2	52.8	45.4	57.7	92.7	55.1	85.1	55.8	77.9	60.9
	GPT-5.2	60.7	**53.7**	**64.9**	94.1	65.4	86.6	62.6	80.6	67.4
Open Weight	Llama-3.1-8B	37.6	35.1	36.4	93.4	44.3	72.1	42.1	53.4	47.0
	Llama-3.3-70B	57.2	50.5	59.9	93.4	58.0	87.9	60.4	84.0	64.6
	Mistral-7B	30.1	25.5	26.7	59.7	27.8	51.9	26.8	33.6	32.1
	Mixtral-8x7B	38.1	31.2	33.4	71.3	36.9	45.6	37.8	29.6	38.6
	Qwen3.5-4B	43.4	34.1	42.4	89.9	43.3	69.1	46.2	59.4	49.2
	Qwen3.5-9B	43.7	36.1	41.5	92.1	42.5	78.6	51.0	71.7	52.3
	Qwen3.5-35B	47.9	40.4	52.9	93.3	55.9	84.7	54.0	77.9	58.5
Qwen3.5-9B Variants	+ CoT	46.6 (+2.9)	43.3 (+7.2)	45.0 (+3.5)	91.1 (-1.0)	41.2 (-1.3)	82.9 (+4.3)	51.6 (+0.6)	75.6 (+3.9)	54.8 (+2.5)
	+ SFT	53.2 (+9.5)	45.3 (+9.2)	54.3 (+12.8)	94.7 (+2.6)	61.3 (+18.8)	91.6 (+13.0)	66.4 (+15.4)	81.6 (+9.9)	64.4 (+12.1)
	+ MARTA	**70.0** (+26.3)	52.0 (+15.9)	63.4 (+21.9)	**95.6** (+3.5)	**68.3** (+25.8)	**96.0** (+17.4)	**75.8** (+24.8)	**84.9** (+13.2)	**72.5** (+20.2)

Table 2: Main results on TrajToM. Overall is micro-averaged accuracy over all questions. Values in parentheses denote absolute gains over Qwen3.5-9B. The best and second-best model results, excluding humans, are marked in **bold** and underline.

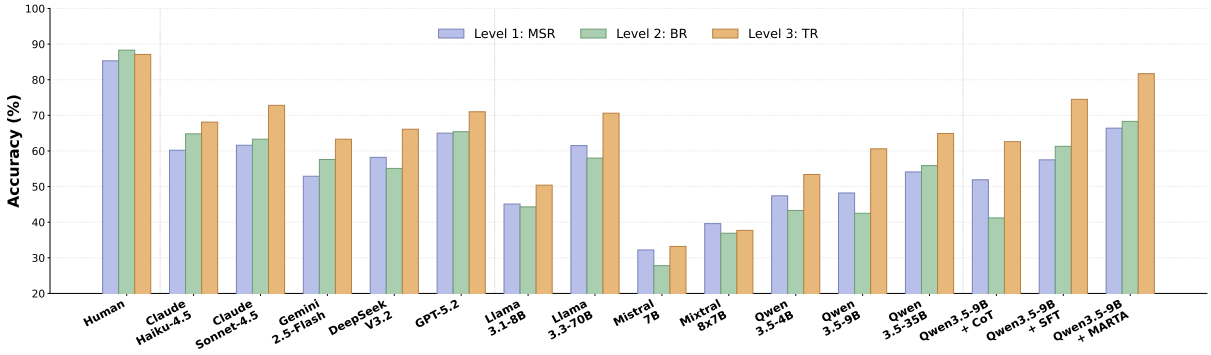


Figure 5: Level-wise performance on TrajToM. Scores are question-count-weighted averages over subtypes within each level.

tal states across phases and identifying the direction of change. For example, GPT-5.2 drops from 86.6 on KMI and 80.6 on CCR to 62.6 on CDA, while Qwen3.5-9B drops from 78.6 and 71.7 to 51.0. Thus, high TR-level scores mainly reflect success in locating salient events or local causal cues, rather than full trajectory-level tracking.

MSR shows a similar imbalance. Desire reasoning is easier, with GPT-5.2 and Qwen3.5-9B reaching 94.1 and 92.1, likely because stable motivations recur across the dialogue. In contrast, belief and intention reasoning are more context-sensitive: GPT-5.2 drops to 53.7 and 64.9, while Qwen3.5-9B drops to 36.1 and 41.5. BR is also difficult, with GPT-5.2 reaching 65.4 and Qwen3.5-9B 42.5, suggesting weakness in linking latent mental states to observable behavior under information asymmetry. Together, these results show that success on stable motivations, salient events, or local triggers does not imply robust ToM reasoning.

MARTA Provides Consistent Gains over Prompting and SFT. For controlled comparison on Qwen3.5-9B, CoT prompting only modestly improves overall accuracy from 52.3 to 54.8, suggesting that generic step-by-step reasoning is insufficient when the key challenge is tracking mental state evolution across dialogue phases. SFT reaches 64.4, showing that a structured reasoning format helps organize ToM inference. MARTA further improves accuracy to 72.5, outperforming the base model by 20.2 points and SFT by 8.1 points. The gains are especially clear on BR and level-wise TR, where MARTA improves from 42.5 to 68.3 and from 60.6 to 81.7, respectively. These results support the motivation of MARTA: Sketch-derived process supervision improves answer selection and encourages reasoning conclusions that are more consistent with the latent mental state trajectory

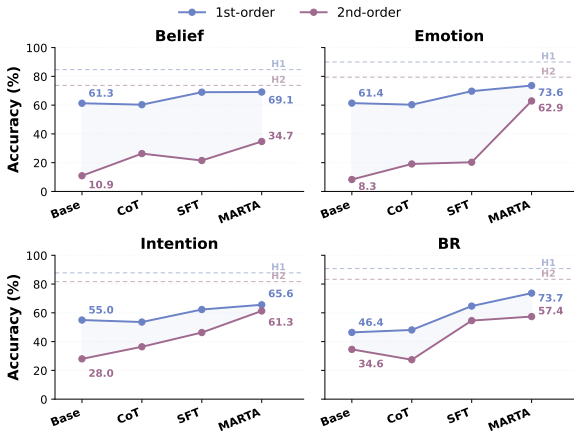


Figure 6: Order-sensitive performance across MSR and BR subtypes. Solid lines show Qwen3.5-9B variants, dashed lines indicate human references, and shaded regions denote first-to-second order gaps. H1/H2 denote human first-/second-order performance.

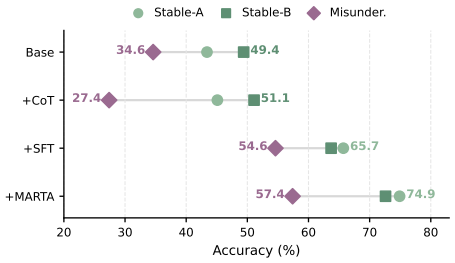


Figure 7: Behavioral reasoning breakdown across three BR patterns. Stable-A/B denote first-order stable behavior explanations, while Misunder. denotes second-order interaction-level misunderstanding.

Setting	Reward	Overall	2nd
SFT	–	64.4	35.1
Answer-only	R_{out}	$66.4_{+2.0}$	$42.6_{+7.5}$
w/o λ	$R_{\text{out}} + R_{\text{proc}}$	$71.3_{+6.9}$	$51.5_{+16.4}$
MARTA	$R_{\text{out}} + \lambda R_{\text{proc}}$	$\mathbf{72.5}_{+8.1}$	$\mathbf{52.1}_{+17.0}$

Table 3: Ablation results of MARTA. Subscripts denote absolute gains over SFT.

underlying the dialogue.

5.3 In-Depth Analysis

Order-sensitive Evaluation Reveals a Clear Second-order Bottleneck. To examine higher-order ToM reasoning, we analyze order-sensitive questions in MSR and BR, covering belief, emotion, intention, and behavioral reasoning. As shown in Figure 6, Qwen3.5-9B exhibits a clear first-to-second order gap across all four subtypes, especially in belief and emotion reasoning. This is because second-order questions require inferring one character's understanding of another character's mental state, rather than directly inferring a single character's state from dialogue evidence. CoT prompting provides limited and unstable improvements, suggesting that generic step-by-step reasoning does not reliably resolve nested mental state inference. SFT improves several results, but the second-order gap remains evident. By contrast, MARTA consistently improves second-order performance across all subtypes, increasing aggregated second-order accuracy from 20.1 to 52.1 and reducing the first-to-second order gap from 35.9 to 18.4 points. Nevertheless, MARTA still remains below the human second-order reference, showing that second-order ToM remains a key bottleneck even with trajectory-aligned process supervision.

Interaction-level Misunderstanding Remains Hard for BR. We further examine BR performance across its three behavior-explanation patterns. As shown in Figure 7, Qwen3.5-9B performs worse on Misunderstanding than on Stable-A and Stable-B, indicating that behavioral reasoning becomes harder when it depends on reciprocal misinterpretation under asymmetric information. CoT slightly improves the two stable behavior patterns but further hurts Misunderstanding. SFT and MARTA improve all three patterns, but Misunderstanding remains the lowest-scoring pattern, showing that interaction-level misunderstanding is particularly challenging.

Process Supervision Contributes beyond Answer-only Optimization. To assess the role of reward components, we conduct an ablation study on MARTA (Table 3). Compared with SFT, MARTA substantially improves both overall accuracy and second-order reasoning, showing that its gains are not merely due to learning a structured reasoning format. The answer-only variant improves over SFT but remains below full MARTA, indicating that outcome rewards alone are insufficient for aligning the reasoning process. Removing the outcome-calibrated trust weight λ also slightly degrades performance, suggesting that process rewards benefit from calibration by answer correctness. These results show that Sketch-derived process supervision provides benefits beyond answer-only optimization, especially for second-order ToM reasoning.

6 Conclusion

We present TrajToM, a dialogue-based benchmark for evaluating LLMs' ToM through evolving mental state trajectories. TrajToM covers mental state reasoning, behavioral reasoning, and trajectory reasoning over beliefs, emotions, intentions, and desires in continuous interactions. Experiments reveal a substantial human–model gap, with persistent bottlenecks in second-order reasoning, behavior explanation, and change direction assessment. We further show that MARTA improves over CoT, SFT, and answer-only optimization through trajectory-aligned process supervision. Overall, TrajToM provides a fine-grained framework for studying evolving mental states in social dialogue.

Limitations

Synthetic Data Artifacts and Generator Bias. TrajToM uses a Sketch-first paradigm to make ground-truth mental states traceable. However, all LLM-based generation steps use a single generator model, which may introduce stylistic artifacts or model-family bias. Although Sketch constraints, automatic checks, and human review improve answer traceability, they may not remove all generator-specific artifacts. Future work can use multiple generator models or human-written dialogues to further diversify the benchmark.

Scope of Social Interaction. TrajToM focuses on continuous two-character social dialogues. This controlled setting supports trajectory-level evaluation, but it does not cover group dynamics, multilingual interactions, multimodal cues, or embodied social environments.

Multiple-choice Evaluation. TrajToM uses multiple-choice questions for scalable and objective evaluation. This format evaluates recognition of the correct option and may not fully capture open-ended explanation quality, interactive decision-making, or free-form social reasoning.

Training Scope of MARTA. MARTA is evaluated on one open-weight base model and relies on structured trajectory evidence for process supervision. Although Sketches are not provided as model input, the method still requires trajectory annotations during training. Its effectiveness on larger models, other model families, and datasets without trajectory annotations remains to be further explored.

Ethical Considerations

Anthropomorphism. ToM is a human social-cognitive ability, and evaluating LLMs with ToM benchmarks may risk anthropomorphic interpretations. We do not claim that LLMs possess human-like mental states or genuine ToM. TrajToM provides a functional evaluation of whether models can infer characters' beliefs, emotions, intentions, and desires from dialogue evidence.

Data and Human Review. TrajToM contains synthetic social dialogues involving interpersonal goals, information asymmetry, misunderstandings, and emotional changes. The dialogues do not contain personal data from real individuals. We check the data for personally identifying information and remove or revise examples that may contain real names, contact information, addresses, or user names. We apply automatic filtering and human review to reduce unsafe, offensive, or harmful content and to improve data quality. Reviewers were informed of the research purpose, could stop reviewing uncomfortable content, and were compensated according to the estimated workload.

Responsible Use. TrajToM and MARTA are intended for research on ToM evaluation and reasoning improvement. Although better ToM reasoning may benefit dialogue agents, education, and assistive systems, it may also create dual-use risks, including misuse for manipulation, profiling, unwanted persuasion, or inference of users' private mental states without consent. We therefore encourage responsible use and discourage applying these resources to infer or influence users' private mental states without consent, especially in high-stakes or user-facing settings.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Nos. U25B2047, 62272469, 72301284).

References

Ishita Agarwal, V. Sakthivel, and P. Prakash. 2025. Toward inclusive healthcare: An LLM-based multimodal chatbot for preliminary diagnosis. *IEEE Access*, 13:136420–136432.

Antonio Andriella, Giovanni Falcone, and Silvia Rossi. 2025. Enhancing robot assistive behaviour by mentalising user intent and beliefs with reinforcement

learning. *International Journal of Social Robotics*, 17(10):2293–2309.

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

Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, Mass.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.

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.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. 2025. SoMiToM: Evaluating multi-perspective theory of mind in embodied social interactions. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

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 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2026. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. In *International Conference on Learning Representations*, volume 2026, pages 50252–50297.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. TimeToM: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11532–11547, Bangkok, Thailand. Association for Computational Linguistics.

Tiffany L. Hutchins, Patricia A. Prelock, Hope Morris, Joy Benner, Timothy LaVigne, and Betsy Hoza. 2016. Explicit vs. applied theory of mind competence: A comparison of typically developing males, males with ASD, and males with ADHD. *Research in Autism Spectrum Disorders*, 21:94–108.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

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.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809, Miami, Florida, USA. 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.

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.

Mengfan Li, Xuanhua Shi, and Yang Deng. 2026. RecToM: A benchmark for evaluating machine theory

of mind in LLM-based conversational recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31636–31644. AAAI Press.

Yiwei Liu, Emma Jane Pretty, Jiahao Huang, and Saku Sugawara. 2025. TactfulToM: Do LLMs have the theory of mind ability to understand white lies? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25043–25061, Suzhou, China. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. 2025. ToM-RL: Reinforcement learning unlocks theory of mind in small LLMs. *arXiv preprint arXiv:2504.01698*.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.

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

Xavier Puig, Tianmin Shu, Joshua B. Tenenbaum, and Antonio Torralba. 2023. NOPA: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7628–7634.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10228–10241, Abu Dhabi, UAE. Association for Computational Linguistics.

Ralf Schneider. 2001. Toward a cognitive theory of literary character: The dynamics of mental-model construction. *Style*, 35(4):607–639.

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. 2025. MuMA-ToM: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1510–1519.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1520–1528.

Weizhi Tang and Vaishak Belle. 2024. ToM-LM: Delegating theory of mind reasoning to external symbolic executors in large language models. In *Neural-Symbolic Learning and Reasoning*, pages 245–257, Cham. Springer Nature Switzerland.

Haibo Tong, Zeyang Yue, Feifei Zhao, Erliang Lin, Lu Jia, Ruolin Chen, Yinqian Sun, Qian Zhang, and Yi Zeng. 2026. CogToM: A comprehensive theory of mind benchmark inspired by human cognition for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31383–31425, San Diego, California, United States. Association for Computational Linguistics.

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*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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.

Chenglong Wu, Jinzhi Liao, and Xiang Zhao. 2026a. DMS-ToM: Enhancing LLMs' theory of mind via dramaturgy-driven mental-state process. In *Database Systems for Advanced Applications: 31st International Conference, DASFAA 2026, Jeju, South Korea, April 27–30, 2026, Proceedings, Part III*, pages 357–373, Berlin, Heidelberg. Springer-Verlag.

Jincenzi Wu, Yuxuan Lei, Jianxun Lian, Yitian Huang, Lexin Zhou, Haotian Li, Xing Xie, and Helen Meng. 2026b. Social-R1: Towards human-like social reasoning in LLMs. *arXiv preprint arXiv:2603.09249*.

Yuheng Wu, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. 2025. DEL-ToM: Inference-time scaling for theory-of-mind reasoning via dynamic epistemic logic. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11383–11397, Suzhou, China. Association for Computational Linguistics.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2025. Towards dynamic theory of mind: Evaluating LLM adaptation to temporal evolution of human states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24036–24057, Vienna, Austria. Association for Computational Linguistics.

Neemesh Yadav, Palakorn Achananuparp, Jing Jiang, and Ee-Peng Lim. 2026. DialToM: A theory of mind benchmark for forecasting state-driven dialogue trajectories. *Preprint*, arXiv:2604.20443.

Masahiro Yamashita, Rieko Kubo, and Shinji Nishimoto. 2025. Conversational content is organized across multiple timescales in the brain. *Nature Human Behaviour*, 9:2066–2078.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. 2025. PersuasiveToM: A benchmark for evaluating machine theory of mind in persuasive dialogues. *arXiv preprint arXiv:2502.21017*.

Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. 2025a. MemSim: A bayesian simulator for evaluating memory of LLM-based personal assistants. In *Advances in Neural Information Processing Systems*, volume 38, pages 90475–90511. Curran Associates, Inc.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. 2025b. AutoToM: Scaling model-based mental inference via automated agent modeling. In *Advances in Neural Information Processing Systems*, volume 38, pages 129659–129699. Curran Associates, Inc.

Chulun Zhou, Qiujing Wang, Mo Yu, Xiaoqian Yue, Rui Lu, Jiangnan Li, Yifan Zhou, Shunchi Zhang, Jie Zhou, and Wai Lam. 2025. The essence of contextual understanding in theory of mind: A study on question answering with story characters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22612–22631, Vienna, Austria. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pages 40975–41019.

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 Task Taxonomy and Examples

This section provides additional details and representative examples for the TrajToM task taxonomy. For compactness, we show example questions and gold answers rather than full dialogues or full answer explanations. All examples are originally multiple-choice questions with one gold answer.

Mental State Reasoning. MSR evaluates whether models can infer characters' mental states from dialogue evidence. First-order MSR asks about a character's own emotion, belief, intention, or desire, while second-order MSR asks about one character's understanding of another character's mental state. Table 4 provides representative examples.

Behavioral Reasoning. BR evaluates whether models can explain observable actions and misunderstandings using latent mental states and information asymmetry. It includes stable behavior explanations for individual characters and interaction-level misunderstanding explanations. Table 5 provides representative examples.

Trajectory Reasoning. TR evaluates whether models can reason about mental-state changes across dialogue phases. KMI asks when a significant shift occurs, CDA asks how a mental state changes, and CCR asks why the change occurs. Table 6 provides representative examples.

B Benchmark Construction Details

Unless otherwise specified, all LLM-based data generation steps in benchmark construction use GPT-5.4. Human reviewers were recruited from graduate students with NLP, AI, or related research backgrounds. They were informed that the data consisted of synthetic social dialogues created for research purposes. They were asked to follow the review criteria described in the corresponding construction stages, including profile plausi-

Subtype	Order	What It Tests	Example Question	Gold Answer
Emotion	1st	Affective state at a dialogue phase.	Right after Harper interprets Dana's early-morning email, what emotional state is Harper most likely in?	Alert and mildly excited.
Belief	1st	Belief about what is true at a dialogue phase.	Before Shiloh shows the printed draft, what does Harper believe the meeting will likely involve?	A process-change announcement indicated by Dana's email.
Intention	1st	Intended action or communicative goal at a dialogue phase.	When Harper cites Dana's early-morning email and says nobody should walk in "flat-footed," what does Harper intend to accomplish?	To use the email as credible evidence.
Desire	1st	Stable motivation or goal across the dialogue.	Across the conversation, what underlying desire best explains Harper's behavior?	To keep the team informed and aligned before the meeting.
Belief	2nd	One character's belief about another character's mental state.	Before Shiloh says Harper is missing from Dana's draft, what does Shiloh think Harper is feeling?	Shiloh thinks Harper feels smugly certain.

Table 4: Examples of MSR subtasks.

Pattern	Order	What It Tests	Example Question	Gold Answer
Stable-A	1st	Explains character A's stable behavior from latent mental and information states.	Why does Harper keep citing Dana's 7:12 email and speaking with certainty about what the meeting will involve?	Harper treats Dana's early email as authoritative and time-sensitive evidence.
Stable-B	1st	Explains character B's stable behavior from latent mental and information states.	Why does Shiloh respond with cautious comparisons and gentle redirects instead of revealing the draft issue immediately?	Shiloh wants to avoid conflict while privately knowing about the revised draft.
Misunderstanding	2nd	Explains interaction-level misunderstanding under asymmetric information.	Why do Harper and Shiloh talk past each other when Shiloh says the folder does not match the email and Harper doubles down on Dana's timestamp?	They hold different hidden facts and misread each other.

Table 5: Examples of BR behavior-explanation patterns.

bility, Sketch coherence, dialogue fidelity, question answerability, and distractor validity. They were also asked to flag unsafe or offensive content and possible personally identifying information. Reviewers were informed that some examples might involve interpersonal conflict, misunderstanding, or emotional changes, and that they could stop reviewing any content that made them uncomfortable. Reviewers were compensated at a local-currency equivalent of approximately USD 10 per hour, which was determined based on pilot estimates of review time and local standards for comparable graduate student research-assistant work.

B.1 Profile Schema

Each TrajToM instance starts from a profile that specifies the social scenario, the two characters, their relationship, and the information-asymmetry structure. The profile is used as the conditioning input for Sketch construction and dialogue generation. It ensures that each dialogue is grounded in a coherent social situation before mental-state trajectories are specified. The profile attributes are sampled conditionally: after a social scenario type is selected, compatible locations, relationship types, character goals, and information-asymmetry patterns are sampled under predefined compatibility constraints. Table 7 summarizes the main profile fields with representative values; the actual sampling pool contains a broader set of compatible values for each field.

Profile Filtering. We apply rule-based filters to remove logically incompatible profiles before Sketch construction. A profile is discarded if its relationship type conflicts with the scenario setting, if the sampled goals cannot support a plausible social tension, or if the information-asymmetry structure cannot support later belief divergence or misunderstanding. For example, first-meeting scenarios are not paired with pre-existing intimate relationships, and misunderstanding-oriented profiles must include private or unevenly distributed information that can plausibly affect how both characters interpret the interaction. Human reviewers further check profile plausibility and discard or revise profiles with incoherent goals, unnatural relationships, or underspecified information asymmetry.

Profile-to-Sketch Prompt. The Profile-to-Sketch prompt contains two parts. The system prompt defines the global constraints for constructing a coherent psychological trajectory, while the user prompt provides the validated profile, required Sketch components, and output schema. A simplified prompt template is shown below.

System Prompt.
You are a story architect for Theory of Mind research. Given a character profile, generate a structured psycho-

Subtype	What It Tests	Example Question	Gold Answer
KMI	Locates the turning point of a mental-state shift.	At what point does Harper stop pressing Shiloh to accept the email's reading and turn to checking whether Dana's draft changed Harper's role?	When Shiloh reveals that Harper's name is missing from Dana's latest draft.
CDA	Assesses the direction of mental-state change.	How does Harper's understanding of the meeting materials shift over the course of the conversation?	From an email-based process-change expectation to concern about Harper's missing draft role.
CCR	Identifies the trigger of a mental-state change.	Which statement or event causes Harper to shift from pushing the 7:12 email to checking Shiloh's page before confronting Dana?	Shiloh states that Harper's name is missing from Dana's latest draft.

Table 6: Examples of TR subtasks.

Profile Attribute	Description	Representative Values
Scenario-level attributes		
Social scenario type	Broad type of social interaction.	workplace professional, family relationship, friendship dynamics, social misunderstanding, competitive conflict
Setting location	Physical or virtual location of the conversation.	office break room, coffee shop, campus library, conference room, online video call
Information asymmetry	Distribution of private or incomplete information between the two characters.	A/B has hidden information, mutual incomplete information, false belief by A/B, mutual misunderstanding
Character-level attributes		
Character names	Gender-neutral names used to instantiate the two characters.	Harper, Shiloh, Jordan, Riley, Morgan, Casey
Personality seed	Lightweight personality-control attribute for diversifying character styles.	INTJ, ENFP, ISFJ, ESTP, and other MBTI-style labels
Social goal	Communicative or social goal of each character.	influence belief, resolve misunderstanding, defend reputation, disclose information, assert autonomy
Relationship-level attributes		
Relationship type	Social relation between the two characters.	colleagues, close friends, roommates, siblings, romantic partners, acquaintances
Relationship closeness	Degree of emotional closeness between the characters.	very close, close, moderately close, distant, strained
Power dynamic	Relative authority or influence between the characters.	equal, slightly unequal, mentor–mentee, authority–subordinate

Table 7: Profile schema used for TrajToM construction. Representative values are shown for compactness.

logical Sketch that will guide dialogue generation and serve as ground-truth annotation for ToM evaluation. The Sketch should describe the two characters' emotion, belief, intention, and desire across four phases: opening, development, key moment, and closing. It should also specify information asymmetry, causal relationships, and second-order mental states.
Requirements.

- Output only valid JSON matching the provided schema.
- Ensure that mental states evolve coherently across phases.
- Make the key event sufficient to support the intended state changes.
- Represent second-order states as plausible misreadings rather than perfect access to the other character's actual state.

User Prompt.
Profile. `{profile_json}`
Sketch Requirements.

1. `information_asymmetry`
2. `phase_states` for opening, development, key moment, and closing
3. a concrete key-moment trigger and the resulting state shifts
4. a three-step `causal_chain`
5. `second_order_states` with the perceived state, actual state, and observable trigger

Output Schema. `{schema_json}`

B.2 Sketch Schema

Given a profile, TrajToM constructs a Sketch before dialogue generation. The Sketch is a structured JSON trajectory that specifies information asymmetry, phase-level mental states, key-moment shifts, causal chains, and second-order mental-state beliefs. It serves as the intermediate representation that connects the social profile, generated di-

alogue, and question answers. The schema enforces required fields and disallows unspecified fields, which helps reduce underspecified or inconsistent trajectories. Table 8 summarizes the main Sketch fields.

For second-order reasoning, entries in `second_order_states` instantiate the tuple $s^{(2)}_{c\to c',\phi,d} = (\textit{believed}, \textit{actual}, \textit{trigger})$. Here, *believed* denotes what character c thinks about character c''s mental state, *actual* denotes the Sketch-specified state of c', and *trigger* denotes the observable cue that supports this belief or misread. For trajectory reasoning, the `key_moment` phase records both the triggering event and the resulting shifts in emotion, belief, and intention. The desire field is treated as stable by default and is filled in the shift structure only when the character's deeper motivation genuinely changes.

Sketch Filtering and Review. Invalid Sketches are discarded or regenerated if they miss required fields, contain underspecified mental states, lack a clear key-moment trigger, or conflict with the sampled information-asymmetry setting. For phase-level states, we check whether each character has complete emotion, belief, intention, and desire descriptions in the required phases. For key moments, we check whether the trigger can plausibly cause the specified state shifts. For second-order states, we check whether the believed state differs from the actual state when a misread or false belief is intended, and whether the trigger can later be realized as explicit dialogue evidence. Human reviewers further examine trajectory plausibility, causal-chain coherence, and trigger clarity before dialogue generation.

B.3 Dialogue Generation Details

Each dialogue is generated from a paired profile and Sketch. The profile provides character attributes, relationship information, social goals, and the situation context, while the Sketch provides the information-asymmetry structure, phase-level mental states, causal chains, key moment, and second-order beliefs. The generator is instructed to realize the Sketch through natural conversational cues, such as word choice, hesitation, disagreement, implication, and shifts in conversational focus. To preserve ToM difficulty, characters are not allowed to directly state their own mental states, beliefs, or intentions, and inner monologue or explicit narration is disallowed. The key moment is required to emerge naturally from the interaction rather than being announced as a turning point.

Dialogue Generation Prompt. The dialogue generation prompt contains two parts. The system prompt defines global writing constraints, while the user prompt provides the profile, a Sketch-derived psychological blueprint, and output requirements. This design makes the Sketch a constraint blueprint rather than a dialogue script: the generator must realize the specified mental-state trajectory through natural conversational evidence. A simplified prompt template is shown below.

System Prompt.
You are a dialogue writer for Theory-of-Mind research. Write a realistic, psychologically grounded conversation between two characters. Write only the dialogue, with no narration, stage directions, or inner monologue. Characters must not directly state their own mental states, beliefs, or intentions. Psychological depth should emerge from what characters say and how they respond. Follow the psychological blueprint faithfully, especially the key-moment trigger. Output only a valid JSON array.

User Prompt.
Characters and Scenario.
`[Person A]`: name, personality seed, social goal. `[Person B]`: name, personality seed, social goal. `[Relationship]`: relationship type, closeness, power dynamic. `[Setting]`: scenario type, location, situation context.

Psychological Blueprint from Sketch.
`[Information Asymmetry]`: private information held by A and/or B, plus any shared false assumption.
`[Phase-level Mental States]`: each character's emotion, belief, intention, and desire across opening, development, key moment, and closing.
`[Key Moment]`: key-moment type, triggering statement/event, and the resulting state shifts.
`[Causal Chains]`: causal explanations for A's stable behavior, B's stable behavior, and the interaction-level misread.
`[Second-order Beliefs]`: what each character thinks about the other's mental states, including the observable trigger for each misread.

Dialogue Requirements.
Write a natural dialogue between A and B. The dialogue must realize the information asymmetry, phase-level states, causal chains, second-order misreads, and key moment specified above. Express mental states through dialogue-grounded cues such as word choice, hesitation, disagreement, implication, and shifts in conversational focus. Do not directly reveal inner states. The key moment should emerge naturally rather than being explicitly announced. After the key moment, reflect the state shifts specified in the Sketch.

Output Format.
Return a JSON array of turn-level objects: `[{"turn": 0, "speaker": "...", "text": "..." }, ...]`.

Automatic Validation and Human Review. After generation, each dialogue is automatically checked for basic structural validity, including turn count, speaker alternation, and non-empty utterances. Dialogues that fail these checks are regenerated or discarded. Human reviewers further assess whether the dialogue faithfully realizes the Sketch, preserves the intended information asymmetry, expresses mental states through dialogue-grounded cues rather than direct self-disclosure, and introduces the key moment naturally. Dialogues are regenerated if they omit required Sketch content, reveal mental states too explicitly, or introduce unsupported state changes.

Sketch Component	Description	Used For
Instance identifier	Links each Sketch to its profile, dialogue, and generated questions.	Instance tracking
Information asymmetry	Records private or unevenly distributed information, including facts known only to A, facts known only to B, and shared false assumptions.	Belief divergence; BR misunderstanding; consistency checks
Phase-level states	Stores phase-level states for opening, development, key moment, and closing. Standard phases contain each character's emotion, belief, intention, and desire.	First-order MSR; CDA state comparison
Key moment	Specifies the turning-point trigger, key-moment type, and state shifts for both characters.	KMI; CDA shifts; CCR triggers
Behavioral causal chains	Provides three causal explanations: A's stable behavior, B's stable behavior, and the interaction-level misread caused by information asymmetry or value difference.	BR Stable-A/B; BR Misunderstanding
Second-order states	Records what each character thinks about the other's mental states. Each non-null entry contains the believed state, actual state, and observable trigger.	Second-order MSR; false-belief evaluation

Table 8: Main components in the Sketch schema.

B.4 Question Generation Details

Question generation is designed to preserve answer fidelity. For each item, the gold answer is determined from the Sketch before the question and options are phrased. The LLM is used only to verbalize the question and answer options in natural language, while task-specific templates select the queried character, mental-state dimension, dialogue phase or phase span, and reasoning target. Table 9 summarizes the answer sources and distractor strategies for each question type.

Question Planning. Before verbalization, each question is first instantiated from a task-specific template that determines the queried character, reasoning target, and dialogue anchor. This planning step specifies what the question should ask before the LLM rewrites it into a natural multiple-choice format.

For MSR, first-order templates ask about a target character's emotion, belief, or intention at a concrete moment in the dialogue. Instead of exposing abstract phase labels such as *opening*, *development*, or *closing*, the question stem anchors the target state to an observable exchange, such as when a character first raises an issue, challenges another character, or reacts to new information. Desire questions are handled differently: because desire represents a more stable motivation, the stem asks about the character's underlying desire or goal across the whole conversation rather than at a single moment. Second-order MSR questions are instantiated from fixed perspective-taking templates, asking what one character thinks another character is feeling, believing, or trying to do before a key-moment event.

For BR, question planning starts from observable behavior rather than mental-state labels. Stable-A and Stable-B templates ask why one character repeatedly shows a consistent behavior pattern across the dialogue. Misunderstanding templates ask why both characters talk past each other at a specific exchange. This interaction-level misunderstanding is treated as one reciprocal phenomenon, rather than being split into two separate questions about each character's individual misread.

For TR, templates focus on observable change. KMI questions describe the contrast between a character's behavior before and after a shift and ask when the shift occurs. CDA questions ask how a character's understanding of a concrete content domain changes across the dialogue or between selected phases, without naming the underlying mental-state dimension in the stem. CCR questions describe an observable change in stance or behavior and ask why that change occurs. For BR and TR, question stems avoid explicit mental-state terms such as *belief*, *desire*, *intention*, and *emotion*, so that models must infer the relevant mental-state structure from dialogue evidence rather than from lexical cues in the question.

Question Generation Prompt. We use task-specific prompts for different question types. For compactness, we show a simplified template that abstracts the common structure across MSR, BR, and TR prompts.

Question Type	Gold Answer Source	Distractor Strategy
MSR, first-order	Emotion, belief, and intention from sampled phase-level states; desire as a cross-phase stable motivation.	Same-dimension states from another phase or character; plausible but globally inconsistent goals for desire.
MSR, second-order	Believed state from second-order mental state beliefs; actual state and trigger for explanation and validation.	Other second-order beliefs or true phase-level states with incorrect perspective or timing.
BR Stable-A/B	Causal chains explaining character A's or B's stable behavior.	Dimension mismatch, time mismatch, or perspective mismatch.
BR Misunderstanding	Causal chain describing the bidirectional misread between two characters.	One-sided attribution, swapped misread directions, or shifted misread intensity.
TR-KMI	Key-moment trigger; state shifts used to verify the turning point.	Too-early, too-late, or high-pressure but non-turning moments.
TR-CDA	Comparison of selected phase-level mental states.	Wrong dimension, wrong character, or incorrect endpoint phase.
TR-CCR	Key-moment trigger, key-moment type, and resulting state shift.	Stable-goal attribution, premature-state attribution, or correct trigger with wrong causal mechanism.

Table 9: Answer sources and distractor strategies for TrajToM question types.

Input.
`[Scenario Context]`: social setting, characters, relationship, and situation context.
`[Dialogue]`: generated dialogue with turn-level utterances.
`[Question Plan]`: question type, target character, reasoning order, mental-state dimension or trajectory target, and dialogue anchor.
`[Gold Answer]`: pre-determined answer extracted from the Sketch. The correct answer must be preserved when phrased as an option.
`[Distractor Seeds]`: structurally related but incorrect alternatives derived from Sketch fields. The model may only rephrase these seeds for fluency and parallel style.

Task.
Write one four-option multiple-choice question by verbalizing the provided question plan, gold answer, and distractor seeds. The question stem should be anchored to observable dialogue evidence rather than to hidden Sketch labels. Do not directly mention phase names such as *opening*, *development*, or *closing* in the stem. For BR and TR questions, avoid explicit mental-state words such as *belief*, *desire*, *intention*, and *emotion* in the stem. The correct option must preserve the pre-determined gold answer. Distractors should be plausible on a first reading but invalid under the Sketch. Do not introduce new mental states, events, or causal facts that are not supported by the dialogue or Sketch. All options should be parallel in style and length.

Task-specific Constraints.
MSR: anchor first-order questions to a concrete dialogue moment; for desire, ask about the underlying goal across the whole conversation; for second-order questions, ask what one character thinks another character is feeling, believing, or trying to do before a key-moment event.
BR: anchor the stem to observable behavior patterns; for misunderstanding questions, treat the reciprocal misread as a single interaction-level phenomenon rather than two separate subquestions.
TR: for KMI, ask when an observable shift occurs; for CDA, ask how a concrete content-domain understanding changes; for CCR, ask why an observable stance or behavior changes.

Output Format.
Return only a valid JSON object: `{"question": "...", "option_A": "...", "option_B": "...", "option_C": "...", "option_D": "...", "correct_answer": "...", "explanation": "..."}`.

Answer Derivation. Gold answers are derived from different Sketch components depending on the question type. For first-order MSR, emotion, belief, and intention are anchored to sampled dialogue phases, while desire is treated as a cross-phase stable motivation. For second-order MSR, the answer corresponds to one character's belief about another character's mental state rather than the actual state. For BR, answers come from causal chains that explain either one character's stable behavior or reciprocal misunderstanding. For TR, answers are derived from key-moment triggers, phase-level state comparisons, and the causal mechanism associated with state changes.

Distractor Construction. Distractors are structurally related to the gold answer but invalid under the Sketch. For MSR, distractors are drawn from incorrect phases, characters, dimensions, or perspectives. For BR, distractors target common attribution errors, such as using the wrong mental-state dimension, relying on an outdated state, or explaining one character's behavior with another character's private state. For TR, distractors target temporal and causal confusions, including selecting a non-turning pressure peak, using the consequence as the turning point, reversing the change direction, or pairing the correct trigger with an incorrect causal mechanism.

Question Filtering and Review. Generated questions are automatically checked to ensure that the gold answer appears among the options, that the options refer to the intended target character, and that the explanation does not contradict the gold answer. Human reviewers further check item clarity, dialogue-grounded answerability, and distractor validity. Items are discarded if the answer depends on unstated assumptions, if multiple options are reasonably defensible, or if distractors are either trivially wrong or indistinguishable from the gold answer.

Pattern	Order	#Q
Stable-A	1st	700
Stable-B	1st	700
Misunderstanding	2nd	700
Total	–	2,100

Table 10: Statistics of BR behavior-explanation patterns.

Question Type	Phase / Span	#Q
MSR	Opening	1,388
	Development	1,382
	Closing	1,430
TR	Key moment	2,775
	Opening–Development	1,395
	Opening–Closing	2,800

Table 11: Statistics of question groups with explicit phase or span annotations.

C Dataset Statistics

This section provides additional statistics that are not expanded in the main text, including the internal structure of BR questions and the task-specific phase/span annotations used in TrajToM.

Behavioral Reasoning Patterns. BR contains three behavior-explanation patterns: Stable-A, Stable-B, and Misunderstanding. Here we report their distribution statistics. The three patterns are evenly balanced, with 700 questions each, as shown in Table 10.

Phase and Span Annotations. Phase/span annotations in TrajToM are task-specific rather than uniformly defined across all question types. For first-order MSR questions, emotion, belief, and intention are anchored to one of three dialogue phases: opening, development, and closing. Desire questions are treated as cross-phase MSR questions because they reflect relatively stable motivations across the dialogue rather than states anchored to a single phase. Second-order MSR questions are perspective-dependent because they ask about one character's understanding of another character's mental state. BR questions are dialogue-level because they explain behavior patterns or misunderstandings over the interaction. For TR questions, annotations correspond to key moments or phase spans, such as opening–development and opening–closing. Table 11 summarizes the question groups with explicit phase or span annotations.

D MARTA Method Details

D.1 SFT Target Construction

We construct SFT targets by converting the answer explanations in TrajToM into a structured reasoning format. The goal is to initialize the model with an explicit reasoning style while keeping the gold answer fixed. The conversion follows three steps.

Step 1: Explanation-to-reasoning conversion. For each training item, we provide an LLM with the question, candidate options, gold answer, and original explanation. The LLM rewrites the explanation into a structured response with a `<think>` block containing `[Evidence]`, `[Inference]`, and `[Conclusion]`, followed by an `<answer>` tag. The `[Evidence]` stage cites dialogue evidence using message identifiers such as `[mid X]`. The `[Inference]` stage explains how the evidence supports a latent mental state, behavior-level explanation, or trajectory-level change. The `[Conclusion]` stage states the reasoning conclusion that supports the gold option. The LLM does not infer the answer during this conversion, since the gold option is fixed in the prompt.

Step 2: Target validation. We validate the generated targets for format compliance and consistency with the fixed gold answer. Targets that fail validation are excluded from the final SFT set.

Step 3: Instruction-format conversion. Validated targets are converted into instruction-style SFT examples. Each example contains an instruction, an input field with the dialogue, question, and options, and an output field with the structured target response. Evidence is anchored through dialogue message identifiers to make the reasoning traceable to the input dialogue. Table 12 summarizes the conversion prompt.

D.2 GRPO Objective and Reward Computation

For each input $x = (D, Q, \mathcal{A})$, the policy samples a group of N candidate responses $\{o_i\}_{i=1}^{N}$. Each response contains a reasoning chain and a predicted answer. Given the scalar reward R_i for response o_i, we compute the group-normalized advantage as

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \tag{4}$$

where μ_R and σ_R are the mean and standard deviation of rewards within the sampled group, and ϵ is a small constant for numerical stability.

Component	Prompt Template
Instruction	Convert the following answer explanation into a structured reasoning format in English.
Input fields	Question: {question}; Options: {options}; Gold Answer: {gold_answer}; Explanation: {explanation}.
Output format	The output contains a <think> block with three stages, [Evidence], [Inference], and [Conclusion], followed by <answer>{gold_answer}</answer>.
Constraints	Cite dialogue evidence with [mid X] references, do not change the gold answer, keep the reasoning concise, and output only the formatted response in English.

Table 12: Prompt template used to convert answer explanations into structured SFT targets.

We optimize the policy with the standard clipped GRPO objective. For each sampled response, we define

$$\ell_i(\theta) = \min\big(\rho_i A_i, \bar{\rho}_i A_i\big), \tag{5}$$

where $\rho_i = \pi_\theta(o_i|x)/\pi_{\text{old}}(o_i|x)$ and $\bar{\rho}_i = \text{clip}(\rho_i, 1-\epsilon_c, 1+\epsilon_c)$. The clipped loss is

$$\mathcal{L}_{\text{clip}}(\theta) = -\mathbb{E}_{x,\{o_i\}} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta). \tag{6}$$

The final loss adds KL regularization to the reference policy:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathcal{L}_{\text{clip}}(\theta) + \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{7}$$

Total Reward. For each sampled response, the total reward is

$$R_{\text{total}} = R_{\text{fmt}} \cdot (R_{\text{out}} + \lambda R_{\text{proc}}) \cdot R_{\text{len}}. \tag{8}$$

Here, R_{fmt} acts as a format gate, R_{len} controls length and repetition, and R_{out} and R_{proc} provide outcome-level and process-level supervision.

Format and Outcome Rewards. The format reward is binary: $R_{\text{fmt}}(o_i) = 1$ if the response follows the required <think> and <answer> structure, and 0 otherwise. The outcome reward is also binary: $R_{\text{out}}(\hat{y}, y^\star) = 1$ if the predicted option $\hat{y}$ matches the gold option $y^\star$, and 0 otherwise.

Outcome-Calibrated Trust Weight. The coefficient λ controls the contribution of process-level supervision. For a group of N sampled responses, we compute it from group-level answer correctness:

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} R_{\text{out}}^{(i)}. \tag{9}$$

Thus, process rewards are trusted more when the sampled group contains more correct answers, and their influence is reduced when outcome signals are unreliable.

Process Reward. The process reward combines conclusion alignment and reasoning-structure checking:

$$R_{\text{proc}} = w_{\text{conc}} R_{\text{conc}} + w_{\text{struct}} R_{\text{struct}}, \tag{10}$$

where $w_{\text{conc}} + w_{\text{struct}} = 1$. We set $w_{\text{conc}} = 0.7$ and $w_{\text{struct}} = 0.3$ to prioritize semantic conclusion alignment over format-level structure checking.

The conclusion reward R_{conc} measures whether the final conclusion is semantically aligned with the corresponding target derived from the Sketch. The Sketch-derived target is selected according to the question type: phase-level states for first-order MSR, second-order belief tuples for second-order MSR, causal chains for BR, and key moments, state shifts, or triggers for TR. The conclusion span is extracted from the [Conclusion] stage. We compute the cosine similarity s between the generated conclusion span and the Sketch-derived target using the `all-MiniLM-L6-v2` sentence embedding model (Wang et al., 2021), and define

$$R_{\text{conc}} = \min\left(\frac{\max(s, 0)}{0.65}, 1\right). \tag{11}$$

Here, 0.65 is the cosine-similarity threshold used for assigning full conclusion reward.

The structure reward checks whether the reasoning chain contains the required stages $\mathcal{T} = \{\texttt{[Evidence]}, \texttt{[Inference]}, \texttt{[Conclusion]}\}$:

$$R_{\text{struct}} = \frac{1}{3} \sum_{t \in \mathcal{T}} \mathbb{I}[t \in o_i]. \tag{12}$$

Length and Repetition Penalty. The length and repetition factor is defined as

$$R_{\text{len}} = R_{\text{rep}}(\rho) \cdot R_{\text{win}}(L), \tag{13}$$

where ρ is the 4-gram repetition rate and L is the response length. The repetition penalty is

$$R_{\text{rep}}(\rho) = \begin{cases} 1, & \rho \le 0.1, \\ \exp(-8(\rho - 0.1)), & \rho > 0.1. \end{cases} \tag{14}$$

Category	Model	Version	Context
API Models	Claude-Haiku-4.5	2025-10-16	200k
	Claude-Sonnet-4.5	2025-09-30	1M
	DeepSeek-V3.2	2025-12-01	131k
	Gemini-2.5-Flash	2025-06-17	1M
	GPT-5.2	2025-12-11	400k
Open Weight	Llama-3.1-8B	Instruct	128k
	Llama-3.3-70B	Instruct	128k
	Mistral-7B	Instruct-v0.1	4k
	Mixtral-8x7B	Instruct-v0.1	32k
	Qwen3.5-4B	Qwen3.5-4B	262k
	Qwen3.5-9B	Qwen3.5-9B	262k
	Qwen3.5-35B	Qwen3.5-35B-A3B	262k

Table 13: Details of the models evaluated in our benchmark.

Setting	Prompt Template
Direct	Read the following dialogue, then answer the multiple-choice question. The input contains `DIALOGUE`, `QUESTION`, and `OPTIONS`. Please respond with only the letter of the chosen answer, without any explanation.
Zero-shot CoT	Read the following dialogue, then answer the multiple-choice question. The input contains `DIALOGUE`, `QUESTION`, and `OPTIONS`. Let's think step by step. First analyze the dialogue and reason about the answer, then conclude with the final choice in the format `Answer: X`, where `X` is one of A, B, C, or D.

Table 14: Evaluation prompt templates used in our experiments.

The length-window reward is

$$R_{\text{win}}(L) = \sigma\left(\frac{L-200}{50}\right) \cdot \sigma\left(\frac{2500-L}{50}\right), \tag{15}$$

where $\sigma(\cdot)$ is the sigmoid function.

E Experimental Details

E.1 Model Versions and Access

We evaluate twelve representative LLMs, including both API-accessed and open-weight models. For compactness, Table 13 reports the model versions and context lengths used in our experiments.

E.2 Evaluation Prompts

All models are evaluated with the same input format, consisting of the dialogue, question, and candidate options. No model is given the Sketch, gold answer, or answer explanation during evaluation. For the standard setting, models are instructed to directly output one option letter. For the zero-shot CoT baseline, models are instructed to reason step by step and conclude with a final answer. The prompt templates are shown in Table 14.

Split	# Dialogues	# Questions
Train	315	8,189
Dev	35	909
Test	350	9,067
Total	700	18,165

Table 15: Dialogue-level data splits of TrajToM.

E.3 Data Splits

We split TrajToM at the dialogue level rather than the question level. All questions derived from the same dialogue are assigned to the same split, preventing leakage across training, development, and test sets. The training and development sets contain 350 dialogues in total, while the test set contains another 350 non-overlapping dialogues for final evaluation. Table 15 summarizes the dialogue-level split statistics.

For MARTA training, the SFT stage uses all 8,189 questions from the training split. Because GRPO requires multiple rollouts per prompt, we use a 4,000-question subset sampled from the training split for policy optimization. The development set is used to monitor training dynamics. The test set is held out and is not used for SFT target construction, GRPO training, or hyperparameter selection.

E.4 Training Hyperparameters

MARTA is trained in two stages: supervised fine-tuning (SFT) followed by trajectory-aligned GRPO. Table 16 summarizes the key hyperparameters used in both stages. The GRPO stage uses 3 NVIDIA A800 80GB GPUs for training and one additional A800 GPU for the vLLM rollout server. The 909-question development set is used for monitoring during GRPO training.

Hyperparameters were selected based on preliminary runs monitored on the development set and then kept fixed for the final test evaluation. We did not perform an exhaustive grid search; instead, we selected stable settings for SFT and GRPO based on development-set behavior and training stability. The held-out test set was not used for SFT target construction, GRPO training, or hyperparameter selection. Unless otherwise specified, all reported model results are obtained under fixed prompting, decoding, and evaluation settings. For API-accessed and open-weight baselines, we report results on the held-out test set under this fixed protocol. MARTA results are based on one complete

Hyperparameter	SFT Stage	GRPO Stage
Base model	Qwen3.5-9B	SFT-tuned model
Framework	LlamaFactory	TRL GRPOTrainer + vLLM
Training data	8,189 questions	4,000 questions
Learning rate	1.0×10^{-5}	1.0×10^{-6}
Epochs	3	1
Batch size	16	24 prompts/update
Precision	bfloat16	bfloat16
LoRA	rank 16	SFT adapter
Cutoff / completion length	2048	1024
Rollout samples N	–	8
KL coefficient β	–	0.005

Table 16: Key hyperparameters for the SFT and GRPO training stages.

SFT+GRPO training run.

Implementation Details. We implement MARTA training with PyTorch, Transformers, TRL, and vLLM. vLLM is used as the rollout server during GRPO. Evaluation is conducted with a custom accuracy-based script under fixed prompting, decoding, and answer-extraction settings. We provide package versions, configuration files, and evaluation scripts in the supplementary code. We do not modify the core implementations of these libraries.

Computational Budget. MARTA training is conducted on 4 NVIDIA A800 80GB GPUs. The SFT stage takes approximately 4 wall-clock hours on 4 GPUs, and each GRPO run takes approximately 12 wall-clock hours, with 3 GPUs used for policy optimization and 1 GPU used for the vLLM rollout server. Thus, one complete SFT+GRPO run requires approximately 16 wall-clock hours and 64 GPU hours. This estimate excludes preliminary debugging, failed runs, and API-accessed model evaluation.

E.5 Human Evaluation Protocol

To establish a human performance baseline, we sample 1,000 questions from the test set using stratified sampling across reasoning levels and question subtypes. The sampled questions were answered by graduate participants who were not involved in dataset construction. Participants were recruited from graduate students with NLP, AI, or related research backgrounds. Each question was answered independently by three participants, resulting in 3,000 human responses in total.

Participants were informed that the study evaluates understanding of synthetic social dialogues created for research purposes. They were told that some examples might involve interpersonal conflict, misunderstanding, or emotional changes, and that they could stop answering any content that made them uncomfortable. Participants received the same input format as models, including the dialogue, question, and candidate options. They were instructed to choose the single best answer based only on the provided dialogue. They did not receive Sketch annotations, gold answers, answer explanations, or model predictions. Participants were compensated at a local-currency equivalent of approximately USD 10 per hour, which was determined based on pilot estimates of task completion time and local standards for comparable graduate student research-assistant work.

Human performance is computed at the question level by majority vote over the three responses for each question. If no majority answer is obtained, the item is counted as incorrect. This protocol provides a human reference while keeping the evaluation format comparable to model evaluation.

E.6 Additional Results

Dialogue-Level Performance and Variation. Because multiple questions are derived from each dialogue, we additionally evaluate performance at the dialogue level. For each dialogue, we compute accuracy over all associated questions and then average these accuracies with equal weight across dialogues. Table 17 reports the mean, standard deviation, and distribution statistics of per-dialogue accuracy.

Across models, the standard deviation ranges from 8.4 to 10.2 percentage points, indicating meaningful but moderate cross-dialogue variation. The main model comparisons remain consistent with the question-level results. In particular, Qwen3.5-9B + MARTA achieves a mean dialogue-level accuracy of 72.4, compared with 64.4 for SFT and 52.2 for the base model, while maintaining a comparable standard deviation. These results suggest that the main performance patterns are not driven primarily by a small subset of dialogues.

Dialogue-level Bootstrap Uncertainty. To quantify uncertainty in the reported accuracies, we conduct a dialogue-level cluster bootstrap for all 15 models and method variants. We use 10,000 bootstrap replicates, sampling dialogue clusters with replacement while retaining all questions associated with each sampled dialogue, thereby preserving within-dialogue dependence. Table 18 reports the resulting 95% confidence intervals. The in-

Model	Mean	SD	Min	P25	Median	P75	Max
Claude-Haiku-4.5	63.8	10.2	26.9	57.7	65.4	69.2	88.5
Claude-Sonnet-4.5	66.1	8.8	42.3	61.5	65.4	73.1	88.5
Gemini-2.5-Flash	57.5	9.6	23.1	50.0	57.7	65.4	80.8
DeepSeek-V3.2	60.8	9.9	26.9	53.8	61.5	69.2	84.6
GPT-5.2	67.4	9.9	34.6	61.5	69.2	73.1	92.3
Llama-3.1-8B	47.0	9.0	19.2	42.3	46.2	53.8	73.1
Llama-3.3-70B	64.6	9.7	23.1	57.7	65.4	73.1	84.6
Mistral-7B	32.1	9.0	7.7	26.9	30.8	38.5	57.7
Mixtral-8x7B	38.6	9.5	11.5	30.8	38.5	46.2	65.4
Qwen3.5-4B	49.2	8.8	19.2	42.3	50.0	53.8	73.1
Qwen3.5-9B	52.2	8.4	26.9	46.2	53.8	57.7	73.1
Qwen3.5-35B	58.5	9.3	26.9	50.0	57.7	65.4	76.9
Qwen3.5-9B + CoT	54.7	9.1	30.8	50.0	53.8	61.5	80.8
Qwen3.5-9B + SFT	64.4	9.1	36.0	57.7	65.4	69.2	84.6
Qwen3.5-9B + MARTA	72.4	8.6	46.2	65.4	73.1	77.9	92.3

Table 17: Dialogue-level performance and cross-dialogue variation on TrajToM. Accuracy is computed over all questions within each dialogue and then summarized across dialogues with equal dialogue weights. All values are percentages.

Model	Accuracy	95% CI
GPT-5.2	67.4	[66.3, 68.4]
Claude-Sonnet-4.5	66.1	[65.2, 67.0]
Llama-3.3-70B	64.6	[63.6, 65.6]
Claude-Haiku-4.5	63.8	[62.7, 64.9]
DeepSeek-V3.2	60.8	[59.8, 61.9]
Qwen3.5-35B	58.5	[57.5, 59.5]
Gemini-2.5-Flash	57.5	[56.5, 58.5]
Qwen3.5-9B	52.3	[51.4, 53.1]
Qwen3.5-4B	49.2	[48.3, 50.2]
Llama-3.1-8B	47.0	[46.1, 48.0]
Mixtral-8x7B	38.6	[37.6, 39.6]
Mistral-7B	32.1	[31.1, 33.0]
Qwen3.5-9B + CoT	54.8	[53.8, 55.7]
Qwen3.5-9B + SFT	64.4	[63.5, 65.4]
Qwen3.5-9B + MARTA	72.4	[71.6, 73.3]

Table 18: Overall accuracy and 95% confidence intervals from dialogue-level cluster bootstrap with 10,000 replicates.

tervals have half-widths of approximately 0.9–1.1 percentage points, indicating that the overall results are stable while small differences between similarly performing models should be interpreted cautiously. In contrast, the confidence interval for Qwen3.5-9B + MARTA ([71.6, 73.3]) is clearly separated from that of Qwen3.5-9B + SFT ([63.5, 65.4]), supporting the robustness of MARTA's improvement over SFT.

Phase- and Span-sensitive Results. We further report results on question groups with explicit phase or span annotations. Since phase/span annotations are task-specific, phase-wise analysis is conducted for first-order MSR questions, while span-wise analysis is conducted for TR questions. Table 19 reports these results for all evaluated models and Qwen3.5-9B variants.

For first-order MSR, models generally achieve the highest accuracy on closing-phase questions. For example, Qwen3.5-9B improves from 44.4 on opening and 55.1 on development to 77.5 on closing, and a similar pattern appears across most model families. This suggests that later dialogue phases provide more accumulated behavioral and linguistic evidence for inferring mental states. By contrast, opening-phase questions are harder because models must infer early beliefs, emotions, or intentions from limited context, before the dialogue trajectory becomes fully observable.

For TR, the most difficult group is opening-to-development. Qwen3.5-9B reaches 75.2 on Key Moment and 62.9 on Open–Close, but only 27.1 on Open–Dev. This contrast indicates that models are relatively better at locating salient turning points or comparing clearly separated endpoint states, but struggle to track early and local mental-state shifts. CoT does not resolve this difficulty and even decreases Open–Dev performance from 27.1 to 21.7, suggesting that generic step-by-step prompting may not reliably capture fine-grained trajectory transitions.

MARTA improves all phase/span groups over Qwen3.5-9B. The largest gain appears on Open–Dev, where accuracy increases from 27.1 to 68.1. This improvement suggests that trajectory-aligned supervision is especially helpful when models need to identify subtle early-stage changes rather than rely on salient final-state cues. Nevertheless, the remaining gap between Open–Dev and other TR groups shows that local trajectory tracking remains a challenging aspect of ToM reasoning.

Order-disaggregated Results. Because the aggregate MSR and BR scores combine first- and second-order questions, whereas TR contains no second-order component, we additionally report order-specific performance for all models. For MSR, first- and second-order scores are computed over Emotion, Belief, and Intention, which contain both reasoning orders; Desire is reported separately because it has no second-order counterpart. Table 20 reports the results.

Across all models, MSR exhibits a substantial first-to-second-order drop of 18.4–42.0 percentage points, showing that second-order reasoning strongly affects the aggregate MSR score. The order effect in BR is smaller and less consistent, with several models showing similar first- and second-order performance. TR should therefore not be interpreted as inherently easier than MSR or BR

Category	Model	MSR Phase			TR Span		
		Opening	Development	Closing	Key Moment	Open–Dev	Open–Close
API Models	Claude-Haiku-4.5	53.3	61.6	89.0	82.1	30.0	73.1
	Claude-Sonnet-4.5	62.2	61.1	89.0	86.3	42.6	74.5
	Gemini-2.5-Flash	51.8	56.1	83.6	76.4	36.9	63.6
	DeepSeek-V3.2	50.5	59.1	87.0	81.5	28.9	69.2
	GPT-5.2	56.9	**63.8**	89.1	83.6	38.3	74.8
Open Weight	Llama-3.1-8B	38.4	40.5	79.0	62.8	18.3	54.1
	Llama-3.3-70B	59.3	58.7	**91.4**	85.9	24.3	78.5
	Mistral-7B	30.9	37.1	51.0	42.7	18.7	30.9
	Mixtral-8x7B	33.4	45.8	71.0	37.6	22.0	45.7
	Qwen3.5-4B	46.3	48.7	75.8	64.3	28.0	55.3
	Qwen3.5-9B	44.4	55.1	77.5	75.2	27.1	62.9
	Qwen3.5-35B	50.4	55.6	84.1	81.3	31.9	65.1
Qwen3.5-9B Variants	+ CoT	44.2 (-0.2)	53.3 (-1.8)	75.8 (-1.7)	79.3 (+4.1)	21.7 (-5.4)	66.4 (+3.5)
	+ SFT	56.9 (+12.5)	60.3 (+5.2)	83.0 (+5.5)	86.7 (+11.5)	46.8 (+19.7)	76.2 (+13.3)
	+ MARTA	**63.3** (+18.9)	62.3 (+7.2)	82.0 (+4.5)	**90.5** (+15.3)	**68.1** (+41.0)	**79.7** (+16.8)

Table 19: Phase- and span-sensitive results on TrajToM. MSR columns report first-order phase-anchored questions, and TR columns report key-moment or phase-span questions. Parentheses denote gains over Qwen3.5-9B.

Model	MSR				BR			TR
	o1	o2	Gap	Desire	o1	o2	Gap	Acc.
Claude-Haiku-4.5	68.1	37.9	30.2	92.4	64.6	**65.1**	-0.6	68.1
Claude-Sonnet-4.5	**71.0**	37.4	33.6	93.7	63.7	62.6	1.1	72.8
Gemini-2.5-Flash	64.0	24.1	39.9	92.0	57.6	57.7	-0.1	63.3
DeepSeek-V3.2	65.7	35.3	30.4	92.7	56.7	52.0	4.7	66.1
GPT-5.2	70.0	47.3	22.8	94.1	68.4	59.4	9.0	71.0
Llama-3.1-8B	52.8	16.4	36.3	93.4	46.7	39.4	7.3	50.4
Llama-3.3-70B	69.9	38.8	31.1	93.4	57.4	59.1	-1.7	70.6
Mistral-7B	39.8	12.1	27.6	59.7	28.4	26.6	1.9	33.2
Mixtral-8x7B	50.2	14.2	36.1	71.3	38.7	33.1	5.6	37.7
Qwen3.5-4B	57.0	18.8	38.2	89.9	51.1	27.7	23.4	53.4
Qwen3.5-9B	59.2	17.2	42.0	92.1	46.4	34.6	11.9	60.6
Qwen3.5-35B	63.5	27.2	36.3	93.3	58.7	50.3	8.4	64.9
Qwen3.5-9B + CoT	58.0	28.9	29.1	91.1	48.1	27.4	20.7	62.6
Qwen3.5-9B + SFT	67.0	31.2	35.8	94.7	64.7	54.6	10.1	74.5
Qwen3.5-9B + MARTA	69.4	**51.0**	18.4	**95.6**	**73.7**	57.4	16.3	**81.7**

Table 20: Order-disaggregated performance on TrajToM. MSR-o1 and MSR-o2 are computed over Emotion, Belief, and Intention, which contain both first- and second-order questions; Desire is reported separately because it has no second-order counterpart. Gap denotes the first-order minus second-order accuracy difference and is computed before rounding. TR contains no second-order questions. **Bold** and underlined values denote the best and second-best performance in each accuracy column, respectively.

based on aggregate scores alone, since it contains no second-order questions. These results indicate that reasoning order and task type capture distinct sources of difficulty.

E.7 Case Study

Figure 8 presents a representative second-order emotion reasoning example. Qwen3.5-9B+CoT generates a long chain but relies on urgency-related cues and predicts "impatient," while SFT follows the required format but shifts toward Riley's actual task state. MARTA instead grounds its reasoning in Shiloh's perspective and correctly infers that Shiloh attributes annoyance over a trivial slowdown to Riley. This case shows that MARTA improves not only reasoning format, but also evidence-grounded perspective alignment.

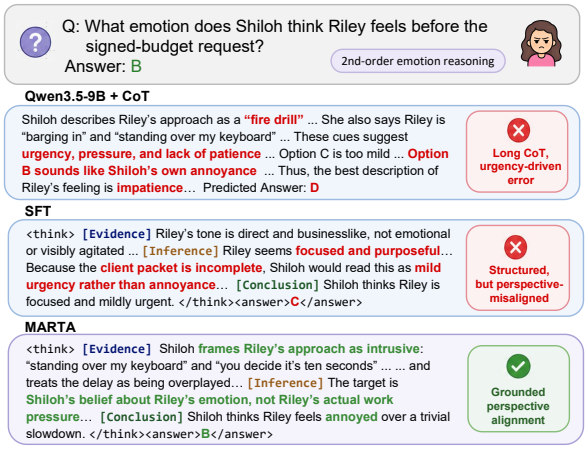


Figure 8: Case study of second-order emotion reasoning.

E.8 License and Terms of Use.

Upon release, the TrajToM dataset will be distributed under the CC BY-NC 4.0 license, and the accompanying evaluation scripts will be distributed under the MIT License. TrajToM is constructed from synthetic social dialogues and does not contain personal data from real individuals. The dataset and scripts are intended for research and evaluation purposes. They should not be used for psychological profiling, manipulation, surveillance, unwanted persuasion, or high-stakes decision-making. They should also not be used to infer or influence users' private mental states without consent.