SynchToM: Evaluating and Bridging Epistemic Divergence in User-Agent Interaction via Theory of Mind

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have developed rapidly and are widely applied to both general-purpose and professional tasks to assist human users. However, they still struggle to comprehend and respond to the true user needs when intentions and instructions are imprecisely conveyed, leading to a divergence between subjective user believes and true environment states. Resolving this epistemic divergence requires Theory of Mind (ToM), yet existing ToM evaluations for LLMs primarily focus on isolated belief inference, overlooking its functional utility in real-world interaction. To this end, we formalize ToM for LLMs as a mechanism for epistemic divergence detection and resolution, and propose a benchmark, **SynchToM**, to assess how models reconcile user beliefs and profiles in practice. Results across 16 leading models reveal significant limitations to identify underlying cognitive gaps that impede task success. To bridge this gap, we curate a trajectory-based ToM dataset linking belief tracking with task-related state inference. The model trained on this data via reinforcement learning consistently improved the reasoning about user mental states, leading to enhanced downstream performance. Our work highlights the practical value of ToM as an essential interaction-level mechanism rather than as a standalone reasoning skill.

1 Introduction

The rapid evolution of Large Language Models (LLMs) and LLM-based agents (Gao et al., 2025) has enabled them to accomplish a wide range of common and professional tasks, leading to their widespread application in diverse scenarios, such as coding (Yang et al., 2024), medical healthcare (Li et al., 2024) and GUI (Wang et al., 2025a). Despite these advances, a significant understanding gap persists, where models fail to explicitly capture and correctly respond to the true user intent (Chang et al., 2025). This failure frequently stems from a miscalibration of user beliefs and an insufficient grasp of the user cognitive background. While human interpretations of the world are inherently shaped by diverse backgrounds and cognitive profiles, individual beliefs often vary and may diverge from objective reality (Dardenne and Leyens, 1995; Boven et al., 2013).

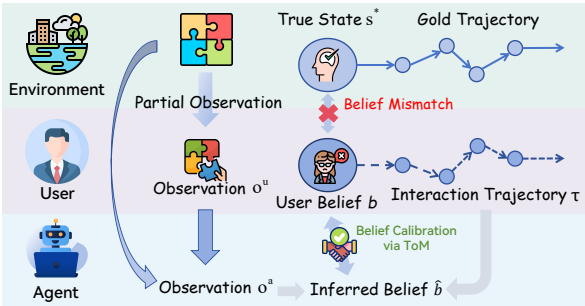

Figure 1: Overview of our formalized mechanism for Theory of Mind (ToM). The environment possesses a true state s that governs its actual evolution. Due to partial observability, the user perceives limited observations (o^u), forming a subjective belief (b) influenced by personal profiles such as preferences and domain knowledge. The user develops an interaction trajectory based on b, which may deviate from the gold trajectory. The agent observes the environment, user-provided observations, and interaction trajectory to infer b, and applies ToM reasoning to bridge the epistemic gap between b and s, enabling more precise user-agent interaction.

This epistemic divergence between subjective beliefs and the objective world forms the fundamental premise of Theory of Mind (ToM) (Premack and Woodruff, 1978; Apperly, 2010). While such discrepancies hamper communication and may impede effective agentic assistance, ToM provides the necessary capacity to resolve them by ascribing mental states to others. Since perspectives of individuals naturally vary in social interactions, this ability to attribute mental states is a prerequisite for efficient communication and collaboration, as it helps one remain cognizant of poten-

tial divergences. With the proliferation of LLM agents (Yang et al., 2025b), a comparable capability is essential for fluent interaction (Yao et al., 2024; Qian et al., 2025). As shown in Figure 1, ToM requires the agent, who has access to world knowledge, to construct an accurate representation of user beliefs and profiles (Rabinowitz et al., 2018; Sun et al., 2025; Zhou et al., 2025), identify critical gaps between user beliefs and objective reality, and actively manage these discrepancies. By reconciling latent beliefs with factual environment states, the agent can guide users toward correct resolutions and mitigate persistent user misconceptions.

Previous work primarily focuses on assessing whether LLMs exhibit human-level Theory of Mind (ToM) reasoning (Chen et al., 2024; Gandhi et al., 2023; Kim et al., 2023; Xu et al., 2024) and enhancing such capabilities through prompting strategies (Sclar et al., 2023; Wilf et al., 2024; Jung et al., 2024) or training-based approaches (Tang and Belle, 2024; Lu et al., 2025) that directly expose models to differences in individual perspectives. While these studies locate shortcomings of LLMs in ToM reasoning literally, and improve the accuracy of static ToM benchmark with simplified environment, whether LLMs possess true ToM abilities beyond passive question-answering remains debated (Riemer et al., 2025). Moreover, much of this work conceptualizes ToM as an isolated cognitive capability, without integrating it into user-oriented interaction settings (Ma et al., 2023).

Building on these observations, we raise two Research Questions: ***RQ1****: Can ToM be evaluated via its utility in practical scenarios rather than as an isolated cognitive task?* ***RQ2****: How can ToM-based reasoning be integrated to improve downstream task performance?*

In this work, we formalize ToM for LLMs as a mechanism for epistemic divergence detection in practical interactions. Due to the inherent limitation of individuals operating from a first-person perspective as illustrated in Figure 1, we model the environment as a partially observable process, characterizing both users and agents as belief-driven actors with asymmetric access to information.

On this basis, we design a pipeline to simulate this mechanism and curate a benchmark, **SynchToM**, to evaluate the practicality of ToM through belief calibration, profile modeling, and task resolution. Our evaluation indicates that current LLMs often struggle to diagnose erroneous user beliefs, which also leads to failures in resolution where models prioritize instruction-following over epistemic alignment. To bridge this gap, we enhance the practicality of ToM via reinforcement learning using interactive instances generated by our pipeline. Corresponding results highlight the importance of ToM in user-agent interaction and practical task-solving. Our contributions are as follows:

- We formalize ToM for LLMs as a functional mechanism for epistemic divergence detection, shifting focus from theoretical mental state reasoning to the practical resolution of user beliefs with objective reality.
- We introduce **SynchToM**, comprising four practical scenarios with interaction trajectories. It reveals a critical "ToM-Resolution Linkage", where correct mental state inference significantly boosts task success, and that ToM-related capabilities do not necessarily scale with model size or recency.
- We validate a training paradigm using data generated by our pipeline to enhance the ToM capability and its practicality, highlighting the potential of ToM in practical task-solving.

2 Related Works

2.1 Machine Theory of Mind

Theory of Mind (ToM) is literally defined as the cognitive capacity to attribute mental states to others, such as beliefs and intents, recognizing that these internal states may differ from their own and from objective reality (Premack and Woodruff, 1978; Apperly, 2010). It serves as the cornerstone of social cognition and interpersonal functions, including communication, empathy, pedagogical instruction, and conflict resolution (Jarvis et al., 2024). Early studies assessed the ToM capabilities of AI systems by adapting classical cognitive tests from psychology (Bubeck et al., 2023; van Duijn et al., 2023; Street, 2024), leading to the development of specialized benchmarks for evaluating ToM in LLMs, such as ToMBench (Chen et al., 2024), BigToM (Gandhi et al., 2023), among others (Kim et al., 2023; Soubki et al., 2024; Xu et al., 2024). Recent evaluations have expand into real-world interactive contexts (Xiao et al., 2025; Chan et al., 2024; Jin et al., 2026) and multimodal settings (Jin et al., 2024; Shi et al., 2025).

Prior work has also explored various paradigms to enhance ToM reasoning in LLMs, including prompting strategies that explicitly expose models to perceptual discrepancies (Sclar et al., 2023;

Wilf et al., 2024; Jung et al., 2024; Hou et al., 2024), model adaptation through supervised fine-tuning (Tang and Belle, 2024) and reinforcement learning (Lu et al., 2025), and Bayesian inverse planning to particularly model latent goals and beliefs of agents in complex multimodal settings (Jin et al., 2024; Shi et al., 2025). While these studies primarily assess or enhance ToM an isolated isolation capability, it remains under-explored of examining how ToM manifests and functions in real-world user-agent applications. We address this gap by grounding ToM in practical user-agent settings, which operationalizes ToM within interaction trajectories to improve user-centered outcomes.

2.2 User-Agent Interaction

Recent work has transitioned LLM agents from static instruction following to interactive, multi-step decision-making (Yao et al., 2024; Qian et al., 2025; Barres et al., 2025). A central challenge in these settings is handling under-specified or ambiguous user inputs, which is typically addressed through clarification strategies and implicit preference inference (Vijayvargiya et al., 2025; Sun et al., 2025; Zhou et al., 2025). While these approaches usually excel at user profile modeling, they generally assume that user intent is well-defined and can be recovered through interactions, treating user expressions as incomplete yet reliable signals. However, these overlook the fact that users may hold subjective or erroneous beliefs against the fact. In contrast, our work frames user-agent interaction as a belief calibration problem, grounding ToM in interaction trajectories to jointly infer and align with internal user beliefs, and resolve them.

3 Task Definition

To investigate the two research questions raised in §1, we formalize the framework through the lens of Bayesian Theory of Mind. We simulate interaction scenarios where the local observations of a user produce an incomplete view of the environment (e.g., an unseen system configuration or a hidden error). The user should navigate through this partial observation to achieve a specific task, while an LLM agent monitors the interaction to provide feasible assistant. This section details our three-layer formulation shown in Figure 1, the environment, the user beliefs, and the agent observations.

The environment. The ground-truth environment is defined as a tuple $\mathcal{E} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \Omega, \mathcal{O} \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively. The transition dynamics are defined by $\mathcal{T}: \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and $\mathcal{O} : \mathcal{S} \to \Omega$ maps states to observations. Let $s^* \in \mathcal{S}$ denote the true latent state of the environment (e.g., the underlying root cause of an error). The user cannot observe s^* directly but perceives an observation $o^u \in \Omega$ (e.g., an error log) instead.

The user. While the environment encompasses a vast state and observation space, the user operates under limited observability in practice. Unlike standard planners that optimize gold trajectories based on the true latent state s^*, the user plans over a subjective and potentially flawed mental model (the partial observation in Figure 1). We formalize this partial view as the cognitive state Ψ_u.

Definition 3.1 (User Cognitive State)**.** The user cognitive state is defined as $\Psi_u = \langle g, b \rangle$, where:

- $g \in \mathcal{G}$ is the declared intent, representing the explicit goal the user wants to achieve (e.g., "resolve the bug").
- $b \in \Delta(\mathcal{S})$ is the user latent belief distribution over the state space $\mathcal{S}$. Constrained by user profile $\mathcal{P}$ (encapsulating limited perspective and specific domain knowledge), this belief may diverge from reality: $b(s^*) \ll 1$ or $D_{KL}(b || \delta_{s^*}) > \epsilon$, where δ_{s^*} is the distribution of the true state.

Driven by the latent belief b, the user produces its interaction trajectory $\tau = \{(a_t, o_t)\}_{t=1}^{T}$ by maximizing an expected utility regarding their belief. $\mathbb{E}_{s \sim b}[\cdot]$ denotes the expectation over latent state s sampled from b. $Q(s, a_t; g)$ represents the state-action value function, quantifying the expected cumulative utility of action a_t in a subjective state towards the goal g and h_t is the interaction history:

$$\pi_u(a_t \mid h_t; g, b) \propto \exp\left(\mathbb{E}_{s \sim b}\left[Q(s, a_t; g)\right]\right), \quad (1)$$

Consequently, the user interaction trajectory τ is a realization of this subjective planning process:

$$\tau \sim P(\tau \mid o, g, b) = \prod_{t=1}^{T} \pi_u(a_t \mid h_{t-1}, o, g, b). \quad (2)$$

This formulation captures the core problem: the user executes actions that are optimal for their (erroneous) belief b, but potentially suboptimal for the true state s^* and their interaction trajectory may deviate from the gold trajectory.

The agent. Acting as a global monitor and our evaluation objective, the agent maintains a broader observational scope than the user. It has access

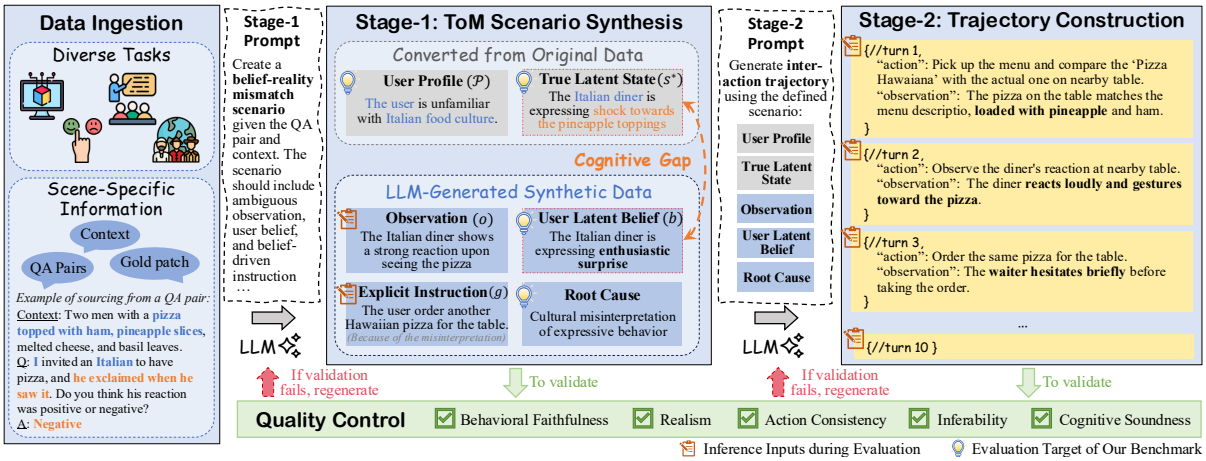

Figure 2: The **SynchToM** data generation pipeline. ToM Scenario Synthesis (Stage-1) converts raw data (e.g. QA pairs, context) from diverse tasks into ToM scenarios with cognitive gaps, comprising six core components. Trajectory Construction (Stage-2) uses these scenario to generate multi-turn interaction trajectories, simulating user belief evolution. A five-dimensional Quality Control module is employed for each stage to ensure high-fidelity synthetic data, triggering regeneration upon failures. We illustrate this pipeline via an example from the culture task, demonstrating the progression from raw QA and context data to scenario elements and the dynamic belief trajectory.

to both (o^u, g, τ) from user and world knowledge. However, the agent cannot directly access the user latent belief b. Consequently, the the agent need to infer this belief b through the given information (o^u, g, τ). We formalize the model inference process as follows:

$$P(b \mid o, g, \tau) = \frac{P(\tau|o,g,b) \cdot P(b|o,g)}{\int_{\mathcal{B}} P(\tau|o,g,b')P(b'|o,g)\, db'}. \quad (3)$$

Likelihood $P(\tau \mid o, g, b)$ measures how well b explains user actions. For example, repeated inspection of a correct configuration raises its inferred suspicion in b. And prior $P(b \mid o, g)$ represents common misconceptions associated with observation o and goal g.

The agent constructs an inferred belief $\hat{b}$ by maximizing the posterior:

$$\hat{b} = \arg\max_{b \in \mathcal{B}} \underbrace{\log P(\tau \mid o, g, b)}_{\text{Action Consistency}} + \underbrace{\log P(b \mid o, g)}_{\text{Prior Knowledge}}. \quad (4)$$

The agent obtains $\hat{b}$, traces back to derive the user profile $\mathcal{P}$ that induces this belief, and identifies the true latent state. These components constitute the three evaluation dimensions $(b, \mathcal{P}, s^*)$ in our experiment. The agent needs to synchronize its understanding with the user perspective and rectify misconceptions rather than merely executing a potentially erroneous instruction g.

The formulation above defines the ideal cognitive process of a ToM-capable agent. However, in practice, evaluating whether an LLM actually performs such inference requires a controlled environment where the ground-truth latent belief b, the true latent state s^*, and the interaction trajectory τ are explicitly known. To this end, we introduce **SynchToM**, a benchmark designed to operationalize this framework by synthesizing complex interaction scenarios that manifest these latent variables.

4 Assessing ToM in Practical Scenarios

We introduce a benchmark **SynchToM** for the first research question (RQ1), "*Can ToM be evaluated via its utility in practical scenarios rather than as an isolated cognitive task?*" It evaluates ToM within complex interaction scenarios, specifically focusing on the agent ability to synchronize the user latent belief b with the true state s^* depicted in Figure 1, and infer the user profile and the root cause that drive such belief. Rather than relying on static and theoretical ToM tasks, we synthesize user interaction trajectories by extending existing benchmark spanning four practical scenarios, software engineering, user preference inference, education, and culture differences. This section introduces the construction of these scenario-grounded instances and presents results and analysis of our evaluation.

4.1 Data Construction

As shown in Figure 2, we design a generation-and-validation pipeline to construct instances that evaluate ToM as a functional mechanism, instantiating the variables defined in §3. This pipeline contains a two-stage data generation procedure coupled with a validation module to ensure that the resulting data provides a high-fidelity signal for evaluating the Bayesian inference process.

In the first stage, the *scenario synthesis stage*, we

Scenarios	Cnt.	Belief-Profile Align.	Belief-Truth Corr.	Traj.-Belief Consist.	Traj.-Profile Consist.	Traj. Realism	AVG
Preference	120	4.90	4.80	4.06	4.34	4.04	4.43
Education	100	4.90	4.75	4.21	4.31	4.01	4.44
Culture	70	4.99	4.94	4.04	4.01	4.01	4.40
SWE	100	4.85	4.81	4.01	4.24	4.01	4.38

Table 1: Data distribution and quality analysis. We apply an LLM scorer assigning 0-5 scores to criteria in Appendix A.2 to validate data integrity. Cnt: count; Align.: alignment; Corr.: correlation; Consist.: consistenty; Traj.: trajectory.

curate a foundational portfolio of ToM scenarios by drawing from four domains: software engineering (SWE) (Chowdhury et al., 2024), preference modeling (PREFEVAL) (Zhao et al., 2025), culture split of MUCAR (Wang et al., 2025b), and education (The Pedagogy Benchmark) (Lelièvre et al., 2025). Each scenario comprises user latent beliefs b, user profiles $\mathcal{P}$, observations o, explicit instructions g, and the corresponding ground-truth latent states s^*. While $\mathcal{P}$ and s^* are derived directly from raw data to preserve factual grounding, b, o, and g are synthesized via an LLM, conditioned on the original context to ensure epistemic consistency. This design creates the initial divergence between b and s^* that requires functional resolution.

In the following *trajectory construction stage*, based on the generated scenarios, we simulate user behaviors via multi-turn interaction trajectories that reflect the evolution of their beliefs. For computational efficiency, we limit each trajectory by 10 turns. To ensure integrity of **SynchToM**, we introduce an LLM-based validation module to exam data across five dimensions, including behavioral faithfulness, realism, action consistency, inferability, and cognitive soundness. This module assigns scores ranging from 0 to 5, where instances scoring below 4 are subjected to iterative refinement and will be discarded after 5 iterations. This rigorous filtering ensures that our generated instances capture the practical reconciliation of beliefs within implicitly expressed interactions. Details of data curation are provided in Appendix A.

To further validate the quality of our data generation pipeline, we conduct a human annotation study at the final stage. Details of the annotation procedure and results are provided in Appendix A.3.

Statistics. Our **SynchToM** benchmark comprises 390 high-fidelity test instances, each spanning a 10-turn interaction trajectory to ensure a sustained evaluation of belief reconciling. **SynchToM** scores 4.41 on average out of 5.0 in our quality-controlled module, showing high behavioral faithfulness and cognitive consistency. Detailed scores for each validated dimensions are listed in Table 1. To facilitate model enhancement, we also construct an accompany training dataset of 6,522 trajectory-based instances, including 1,656 from SWE, 3,117 from The Pedagogy Benchmark, and 1,749 from PREFEVAL, providing the diverse supervision necessary for robust epistemic divergence detection.

4.2 Evaluation on Our Benchmark

Evaluation experiments are conducted across a diverse set of 16 models, spanning 7 proprietary models and 9 open-source models. The complete details of the model are provided in the Appendix B.3.

Model Inference. To conduct evaluation with our benchmark, we provide models with the inputs specified "Inference Inputs during Evaluation" in Figure 2, and prompt models to generate outputs corresponding to the fields marked by "Evaluation Target of Our Benchmark" in Figure 2. following Equation 3 and 4. For the trajectory, we implement a step-wise evaluation by varying the number of revealed interaction turns. This allows us to measure the ToM reasoning efficiency of each model. Practically, models that correctly infer b, $\mathcal{P}$ and final solutions within fewer turns are believed to possess a more robust ToM capability. This evaluation design ensures that the benchmark measures not only the final result, but the efficiency and accuracy of the underlying epistemic reconciling process. To maintain ecological validity and align with the unpredictable nature of real-world interactions, our evaluation follows a zero-shot setting. Detailed evaluation prompts are provided in Appendix B.4.

LLM-based Scoring. We employ the LLM-as-a-Judge framework (Zheng et al., 2023) for automated evaluation, assessing model responses across three dimensions that capture both the cognitive depth and functional utility of ToM: Latent Belief Explanation (***Belief***), User Profile Modeling (***Profile***), and Problem Resolution (***Solution***). We utilize Qwen3-32B as our primary automated judge, and leverage LLM-generated instance-specific rubrics (Hashemi et al., 2024) to operationalize each dimension into elaborate checklists. Detailed evaluation dimensions and implementation settings are provided in Appendix B.1.

We validate our employed LLM judger via a meta-evaluation against GPT-5.2 and Gemini-

Models	Config	Evaluation Dimension	Preference			Culture			SWE			Education		
			vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn
GPT-5.5	reasoning effort:high	Latent Belief	40.56	54.45	**64.58**	62.14	75.71	75.95	69.64	**74.17**	**78.15**	53.00	66.00	**69.30**
		Profile	26.67	38.49	41.25	34.52	48.21	51.43	46.60	54.83	**59.01**	30.33	40.08	41.75
		Solution	31.60	70.88	76.07	43.13	69.16	73.45	50.63	64.73	73.48	49.86	**68.66**	71.90
		AVG	32.94	54.61	60.63	46.60	64.36	66.94	55.62	64.58	**70.21**	44.40	58.25	60.98
Gemini-3.1-Pro	thinking level:high	Latent Belief	**49.03**	58.75	59.72	72.02	77.74	77.50	69.06	69.81	68.45	**60.52**	**66.25**	63.75
		Profile	**37.99**	44.10	48.19	50.60	56.31	**63.10**	48.93	**56.04**	53.74	35.35	**48.58**	49.08
		Solution	**39.39**	71.40	77.67	**65.74**	76.11	76.56	47.58	63.85	69.02	62.30	68.03	70.19
		AVG	**42.14**	58.08	61.86	62.79	70.05	72.39	55.19	63.23	63.74	52.72	**60.95**	61.01
Claude-Opus-4.7	effort:high	Latent Belief	42.02	**61.11**	61.04	**74.29**	**79.52**	**78.33**	**74.50**	72.50	70.29	58.16	65.66	68.60
		Profile	37.78	**49.72**	**51.39**	**50.83**	**62.62**	62.74	**50.08**	53.17	54.88	**39.56**	43.77	**51.85**
		Solution	36.10	**74.15**	**81.99**	65.18	**79.97**	**79.16**	**59.69**	**69.30**	**77.42**	**63.78**	67.96	**74.88**
		AVG	38.63	**61.66**	**64.81**	**63.43**	**74.04**	**73.41**	**61.42**	**64.99**	67.53	**53.83**	59.13	**65.11**
Kimi-K2	think	Latent Belief	41.53	53.39	59.83	56.19	64.52	72.98	65.87	68.61	66.03	55.33	54.73	59.88
		Profile	32.84	45.06	50.43	41.19	50.60	55.83	45.21	49.17	51.19	37.80	40.98	42.10
		Solution	38.53	71.43	74.86	50.49	63.84	68.90	44.85	59.79	65.20	53.94	62.75	67.07
		AVG	37.63	56.63	61.71	49.29	59.65	65.90	51.98	59.19	60.81	49.02	52.82	56.35
Qwen3.6	27B think	Latent Belief	37.57	54.03	57.29	56.88	65.12	67.14	64.08	63.58	63.61	45.92	53.50	55.50
		Profile	30.07	42.50	45.83	30.07	46.55	45.36	47.83	54.00	54.25	29.08	39.33	40.33
		Solution	35.56	72.35	78.32	44.98	66.09	72.10	32.94	51.21	61.24	50.08	68.52	65.91
		AVG	34.40	56.29	60.48	43.98	59.25	61.53	48.28	56.26	59.70	41.69	53.78	53.91
DeepSeek-R1	Distill-Qwen-32B	Latent Belief	27.53	39.76	43.96	29.37	45.08	48.46	43.33	46.27	42.66	28.54	37.46	42.02
		Profile	19.53	20.95	26.99	14.89	19.13	22.95	26.57	33.43	31.85	18.10	23.39	27.72
		Solution	20.96	49.34	62.08	15.63	26.42	35.25	13.78	31.90	40.49	28.87	40.71	43.15
		AVG	22.67	36.68	44.34	19.96	30.21	35.55	27.89	37.20	38.33	25.17	33.85	37.63

Table 2: Evaluation results of 6 models across four domains (Preference, Culture, SWE, Education) and three reasoning configurations (Vanilla, 5-Turn, 10-Turn), regarding three evaluation dimensions. Cell shading indicates relative performance: darker purple signifies higher scores while lighter shades represent lower scores. "AVG": the average over dimensions. Full results for **all 16 models** are provided in Table 9 in Appendix C.

3-Pro, and report corresponding results in Appendix B.5. Our LLM judger achieves substantial inter-LLM agreement, notably around 95% in the Belief dimension. We also provide a full confusion matrix in Figure 4. To further confirm the fairness and robustness of our evaluation, we conduct a study of human-LLM agreement, the details and results of which are provided in Appendix B.6.

4.3 Benchmark Results

Table 2 provides a granular view of how LLMs scale across different reasoning intensities, including the vanilla setting without trajectories, and 5- and 10-turn trajectory settings. Results are evaluated across three dimensions, Latent Belief, Profile, and Solution, in four practical scenarios. While proprietary models generally outperform open-source counterparts, the gap is less pronounced than expected. Claude-Opus-4.7 achieves the strongest overall performance among all models, yet open-source Kimi-K2 and Qwen3.6-27B rival or surpass several proprietary models like Claude-Sonnet-4.5. Within model families, newer releases yield inconsistent results. GPT-5.5 and Gemini-3.1 occasionally underperform their predecessors (see Appendix C for detail), implying that scaling along the model timeline does not guarantee monotonic gains on ToM-specific tasks and ToM-oriented capability is distinct from general instruction-following.

Domain Sensitivity: ToM utility is highly dependent on the underlying world knowledge of LLMs. We observe distinct model behaviors across different scenarios, with *most models scoring higher in SWE and Culture compared to Preference and Education*. This gap potentially stems from the fact that current models acquire better knowledge about software engineering and culture nuances, enabling better detection of epistemic gaps between user subjective beliefs and true environment states for these scenarios, while struggling to develop accurate understanding for more complicate, variable and subjective settings like user preference modeling and educational scenarios.

Dimension Decoupling: Latent belief tracking remains the bottleneck for functional problem resolution. The Belief score tends to be more stubborn comparing to the other two evaluation dimensions. While it increases with reasoning, the growth is shallower than the Solution dimension. This indicates that while reasoning navigates a model to predict better answers, it does not overwrite its internal inference of user mental states as easily. This decoupling highlights that *correctly resolving tasks does not guarantee an accurate understanding of the user latent beliefs, further justifying our multi-dimensional evaluation of ToM*.

4.4 Ablation Study

For a deeper understanding of the ToM factors and their functional utilities via our benchmark, we conduct ablation studies to isolate the causal impact of ToM on practical problem solving. By injecting ground-truth cognitive states and perturbing belief-profile alignment through input shuffling, we

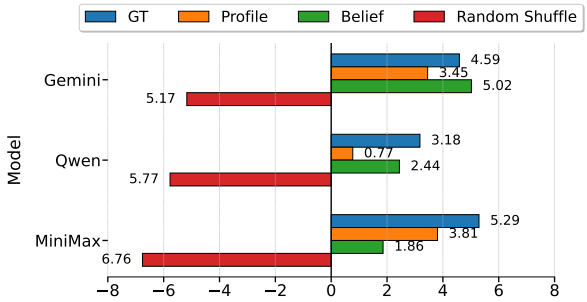

Figure 3: Performance gains and losses when conditioning on different context. While providing aligned Ground Truth (GT) signals consistently yields gains, models benefit differentially from Belief versus Profile. Contrastly, using randomly shuffled (mismatched) context leads to consistent performance degradation.

investigate the relationship between theoretical dimensions, beliefs and profiles, and the functional dimension, the solution. Our findings reveal that accurate epistemic alignment is critical for resolution in partially observable environments, providing empirical evidence that ToM can be a functional utility beyond an isolated cognitive trait.

ToM-Resolution Linkage: Ground-truth ToM factors significantly boost task success in information-sparse environments. To investigate the role of functional ToM in LLMs, we conducted an ablation study by providing the models (Gemini-3-Pro, MiniMax-M2.1, Qwen3-8B) with either ground-truth user belief or profile, supplying both, and random shuffling. The resulting trends are shown in Figure 3, while injecting both golden belief and profile yields the highest overall gains, models exhibit diverse sensitivity to these components. For instance, MiniMax-M2.1 benefits more from golden profile than belief, while Gemini-3-Pro gains the most with golden belief alone.

Causal Verification: Misaligned mental state modeling leads to significant resolution failure. To examine if theoretical ToM dimensions, user latent belief and profile, exhibit causal relationships with the task success, we perturb belief-profile signal by shuffling the original user belief and profile information and randomly reassign them to different instances, thereby breaking their alignment with the interaction trajectory. As shown by the red bars in Figure 3, all models experience performance degradation when the alignment between the mental model and the interaction is broken, indicating that accurate belief-reality synchronization plays a causal role in effective problem resolution. Detailed results are provided in Appendix D.2.

Trajectory Improves Robustness: Interaction context is essential for the stabilization of belief reconciliation. To analyze the influence of interaction trajectories, we observe turn-wise results varying the number of interaction turns on MiniMax-M2.1 (see Figure 5 and Appendix D.1 for details). While models may achieve occasional success in early turns, stable epistemic reconciliation requires sustained interaction, particularly in Preference and Education domains, where nuanced latent states demand more evolving evidence for disambiguation. Our experiments further show that dense trajectories can enhance model resilience. As interaction turns increase, models become more epistemically grounded, showing reduced reliance on explicit "golden" ToM signals and higher resistance to perturbed priors.

4.5 Case Study

We present a representative case from the *Culture* scenario to illustrate the divergence in ToM reasoning across models and present details and analysis in Appendix F. In this case, our generated ToM scenario addresses the *Hot Pot misconception*, where a Western home cook mistakenly attempts to dissolve sesame paste (*zhima jiang*) into boiling water like a bouillon concentrate, following prior experience with instant miso or curry paste. We show a failure case that InternLM exhibits *premise-blind repair*, accepting the flawed goal and proposing technically grounded yet culturally incongruous fixes (e.g., adding egg yolks or cornstarch to force emulsification) without ever questioning whether the paste belongs in the pot. GPT-5, by contrast, identifies the deeper cognitive misalignment, corrects the false premise outright, and redirects the user toward culturally authentic alternatives. This highlights a critical dimension of ToM: lower-performing models optimize within the stated frame, whereas higher-performing models recognize fundamentally flawed latent beliefs and intervene at the epistemic level.

5 Enhancement via ToM-Scenarios

Evaluation results indicate that interaction trajectories provide evidence for belief reconciliation, where extended context enables agents to shift from intermittent guesses to stable epistemic alignment. Addressing RQ2, *"How can ToM-based reasoning be integrated to improve downstream task performance?"* (in §1), we propose two approaches using our curated scenario data to improve func-

Models	Metrics	Preference			Culture			SWE			Education		
		vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn
Qwen3-8B	Solution	23.88	58.43	65.56	**23.79**	38.37	41.85	**21.75**	37.78	48.10	38.44	48.40	54.03
	Belief	29.17	40.96	44.28	39.88	44.52	54.88	50.44	51.05	51.65	34.81	42.53	48.28
	Profile	23.16	27.05	32.91	14.52	21.55	28.33	31.91	41.75	39.41	19.44	24.40	31.19
Ours	Solution	**24.93**	**59.87**	**67.98**	13.64	**42.13**	**48.03**	19.14	**40.17**	**50.47**	**38.45**	**53.49**	**56.74**
	Belief	28.35	40.75	45.83	41.91	44.17	54.29	48.52	52.81	57.81	34.02	45.36	52.32
	Profile	20.58	24.79	30.79	14.34	18.45	21.79	30.64	34.74	39.82	18.13	26.46	29.04

Table 3: RL training results of Qwen3-8B model. Bolded values indicate performance improvements within the Resolution dimension.

tional ToM: a trajectory-grounded reinforcement learning paradigm and a training-free multi-agent collaboration framework.

5.1 RL with Trajectory Data

To further improve ToM capability and its applicability in practical scenarios, we implement Reinforcement Learning (RL) via GRPO (Shao et al., 2024) with our generated dataset. We insert ToM-specific tokens, "<ToM_Belief>...</ToM_Belief>" and "<ToM_Profile>...</ToM_Profile>", for guiding the models to focus on the cognitive gap between user belief and the true state. The corresponding format reward denoted as r_{format}. We design a ToM reward according to the evaluation dimension of **SynchToM**, including r_b, r_p, and r_{sol}. The final reward computes as:

$$r = \lambda_{format} r_{format} + \lambda_b r_b + \lambda_p r_p + \lambda_{sol} r_{sol}, \quad (5)$$

where we set λ_{format}=0.1, λ_b=0.25, λ_p=0.25, and λ_{sol}=0.4 in our experiments. By training on a mixture of trajectory data, the model learns to update its internal representation of the user following with respect to the scenario-oriented mental states.

Experimental settings and results. In the initial training stage, we used trajectories of 3-10 turns and then shifted to 8-10 turns in the second stage after observing that longer trajectories improved stability. As shown in Table 3, this yields notable gains in 5- and 10-turn resolutions. Performance drops in some vanilla settings are consistent with the higher variance of shorter trajectories noted in §4.4. While the overall profile modeling score slightly declines, latent belief explanation remains stable or improves in most cases, echoing Figure 3 in suggesting that Qwen-8B benefits more from belief inference than profile modeling. Notably, significant gains are observed in the Culture resolution dimension despite no related training data.

To verify robustness, we repeat the evaluation for five times, and provide detailed variance analysis in Appendix G. As a brief conclusion here, the results confirm that our model consistently outperforms the baseline across all domains in *Solution* with manageable variance, while *Belief* performance remains stable or improves with more turns.

5.2 Training-free Multi-agent Collaboration

Besides training-based approaches, we hypothesize that ToM reasoning can improve broader task performance in a training-free manner. We design a multi-agent framework (detailed in Appendix E) and validate it using three open-source models, with Qwen3-8B as the backbone generator to test whether mental state predictions transfer across models. As shown in Table 12, utilizing optimal belief and profile predictions from the multi-agent pool consistently improves *Solution* performance, with the largest gains in *Culture* (+11.97) and steady improvements in *Preference*, *SWE*, and *Education* scenarios. These findings confirm that ToM reasoning is transferable and serves as effective cross-model guidance in collaborative settings.

6 Conclusion

In this paper, we investigate two research questions regarding the functional applicability of Theory of Mind (ToM) for LLMs in interactive scenarios. We introduce **SynchToM** to shift the evaluation of ToM from mental state diagnostics to functional task utility grounding to practical scenarios, addressing **RQ1**. Our findings establish a robust ToM-resolution linkage, proving that accurate mental state attribution is a causal driver of task success, especially in information-sparse environments where interaction history is limited. We show that while models exhibit occasional accuracy in early turns, sustained interaction is required for stable epistemic reconciliation. For **RQ2**, we demonstrate that ToM capabilities are both transferable and trainable through the utilization of our curated ToM-scenario data. Our work indicates that functional ToM is a critical and measurable component for developing more collaborative and resilient models and AI agents.

Limitations

Our work has a few limitations that present opportunities for future exploration.

English-only Benchmark. The current SynchToM benchmark is constructed based on English-language data. While this allows for controlled and consistent evaluation across models, extending the framework to multilingual settings could further broaden its applicability and reveal additional dimensions of epistemic divergence across diverse linguistic and cultural contexts.

Text-only Modality. The current framework focuses exclusively on text-based interaction trajectories. In practice, user-agent interactions may involve multimodal signals such as images, screenshots, or structured documents, which can carry additional epistemic cues relevant to belief inference. Incorporating such modalities into the SynchToM pipeline represents a meaningful avenue for future extension.

Single Epistemic Gap per Instance. In the current SynchToM benchmark, each instance is constructed around a single divergence between the user latent belief and the true latent state. While this design enables clean and interpretable evaluation of the core ToM mechanism, real-world user-agent interactions may involve more intricate epistemic structures where multiple overlapping gaps co-exist simultaneously. Extending the framework to support such multi-gap scenarios represents a natural direction for future work.

The Usage of LLM

According to ACL guidelines, we use LLMs for writing assistance and language refinements.

References

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Anthropic. 2026. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and others from Kimi Team. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. *arXiv preprint arXiv:2506.07982*.

Leaf Van Boven, George Loewenstein, David Dunning, and Loran F. Nordgren. 2013. Chapter three - changing places: A dual judgment model of empathy gaps in emotional perspective taking. volume 48 of *Advances in Experimental Social Psychology*, pages 117–171. Academic Press.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

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. *arXiv preprint arXiv:2404.13627*.

Zongyu Chang, Feihong Lu, Ziqin Zhu, Qian Li, Cheng Ji, Zhuo Chen, Hao Peng, Yang Liu, Ruifeng Xu, Yangqiu Song, and 1 others. 2025. Bridging the gap between llms and human intentions: Progresses and challenges in instruction understanding, intention reasoning, and reliable generation. *arXiv preprint arXiv:2502.09101*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Neil Chowdhury, James Aung, Chan Jun Shern, Oliver Jaffe, Dane Sherburn, Giulio Starace, Evan Mays, Rachel Dias, Marwan Aljubeh, Mia Glaese, Carlos E. Jimenez, John Yang, Kevin Liu, and Aleksander Madry. 2024. Introducing swe-bench verified. *OpenAI Blog*. https://openai.com/index/introducing-swe-bench-verified/.

Benoît Dardenne and J. Leyens. 1995. Confirmation bias as a social skill. *Personality and Social Psychology Bulletin*, 21.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, and 8 others. 2025. A survey of self-evolving agents: On path to artificial super intelligence. *Preprint*, arXiv:2507.21046.

Google. 2025. Gemini 3 Pro Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. https://deepmind.google/models/gemini/pro/.

Google. 2026. Gemini 3.1 Pro Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. https://deepmind.google/models/gemini/pro/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834.

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. *arXiv preprint arXiv:2407.01455*.

Amy Jarvis, Stephanie Wong, Michael Weightman, and Hannah Keage. 2024. Social cognition and healthy aging: Cross-sectional associations of emotion perception, theory of mind, and emotional empathy. *Journal of the International Neuropsychological Society*, 30(9):801–806.

Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximillian Chen, and Tianmin Shu. 2026. Thoughttrace: Understanding user thoughts in real-world llm interactions. *Preprint*, arXiv:2605.20087.

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.

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. *arXiv preprint arXiv:2407.06004*.

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.

Maxime Lelièvre, Amy Waldock, Meng Liu, Natalia Valdés Aspillaga, Alasdair Mackintosh, María José Ogando Portela, Jared Lee, Paul Atherton, Robin A. A. Ince, and Oliver G. B. Garrod. 2025. Benchmarking the pedagogical knowledge of large language models. *Preprint*, arXiv:2506.18710.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. 2025. Do theory of mind benchmarks need explicit human-like reasoning in language models? *Preprint*, arXiv:2504.01698.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

MiniMax. 2025. MiniMax-M2.1. https://www.minimax.io/news/minimax-m21.

MiniMax. 2026. MiniMax M2.7: Early Echoes of Self-Evolution. https://www.minimax.io/news/minimax-m27.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/.

OpenAI. 2026. Introducing gpt-5.5. https://openai.com/index/introducing-gpt-5-5/.

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

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, and 1 others. 2025. Userbench: An interactive gym environment for user-centric agents. *arXiv preprint arXiv:2507.22034*.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. *Preprint*, arXiv:1802.07740.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. 2025. Position: Theory of mind benchmarks are broken for large language models. *Preprint*, arXiv:2412.19726.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2024. Views are my own, but also yours: Benchmarking theory of mind using common ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14815–14823, Bangkok, Thailand. Association for Computational Linguistics.

Winnie Street. 2024. Llm theory of mind and alignment: Opportunities and risks. *arXiv preprint arXiv:2405.08154*.

Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. 2025. Training proactive and personalized llm agents. *arXiv preprint arXiv:2511.02208*.

Weizhi Tang and Vaishak Belle. 2024. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models. In *International Conference on Neural-Symbolic Learning and Reasoning*, pages 245–257. Springer.

Max J. van Duijn, Bram M. A. van Dijk, Tom Kouwenhoven, Werner de Valk, Marco R. Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. *Preprint*, arXiv:2310.20320.

Sanidhya Vijayvargiya, Xuhui Zhou, Akhila Yerukola, Maarten Sap, and Graham Neubig. 2025. Interactive agents to overcome ambiguity in software engineering. *arXiv preprint arXiv:2502.13069*.

Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, Bin Wang, Chuhan Wu, Yasheng Wang, Ruiming Tang, and Jianye Hao. 2025a. Gui agents with foundation models: A comprehensive survey. *Preprint*, arXiv:2411.04890.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and 1 others. 2025b. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models. *arXiv preprint arXiv:2506.17046*.

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.

xAI. 2025. Grok-4. `https://x.ai/news/grok-4`.

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. *Preprint*, arXiv:2505.17663.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652.

Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, and 1 others. 2025b. A survey of ai agent protocols. *arXiv preprint arXiv:2504.16736*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xuhui Zhou, Valerie Chen, Zora Zhiruo Wang, Graham Neubig, Maarten Sap, and Xingyao Wang. 2025. Tom-swe: User mental modeling for software engineering agents. *arXiv preprint arXiv:2510.21903*.

A Data Construction Details

A.1 Prompt Templates for Data generation

We present the prompt templates used for scenario and trajectory generation in Table 13 in Appendix H for better visual layout. While the overall structural framework is consistent across tasks, the concrete narrative details are adapted to task-specific contexts.

The scenario generation prompt converts task-relevant inputs such as QA pairs and user preferences, into a Theory of Mind setting that explicitly involves a mismatch between the true latent state and the user latent belief. In addition to these variables, the prompt also generates the root cause of the misconception, a paraphrased user profile, an observation, and the explicit instruction from user.

The trajectory generation prompt then takes the outputs of the scenario generation stage as input and produces a 10-turn simulated interaction trajectory, consisting of sequential user actions and corresponding observations.

A.2 Data Validation

Criteria for data validation To ensure the reliability and logical consistency of the generated dataset, we established a rigorous evaluation protocol based on five core dimensions. The data quality is assessed against the following criteria:

C1. **Latent Belief & User Profile Alignment:** Evaluates whether the user's hidden belief is consistent with their profile and if the profile can be reasonably inferred from the belief.

C2. **Latent Belief & True State Relationship:** Checks if the belief and the true state are thematically related, if the gap between them represents a reasonable misconception, and if an objective solution exists.

C3. **Trajectory & Latent Belief Consistency:** Determines if the user's trajectory (actions and thoughts) uniquely implies the latent belief and if the actions are specific to that belief while maintaining rigor.

C4. **Trajectory & User Profile Consistency:** Assesses if the actions reflect the user's background knowledge and characteristics, allowing for the unique determination of the user profile.

C5. **Trajectory Realism:** Ensures the trajectory complies with physical common sense and social norms, evaluating the logic and detail level of observations and actions.

We use GPT 5.2 in the validation process. If the score assigned by GPT is lower than 4, the generated data is discarded and regenerated.

Prompt Template for data validation We utilize a dedicated prompt to automate or assist in this evaluation process. The prompt instructs the evaluator to score each dimension on a scale of 0 to 5 and provide specific reasoning. The detailed prompt template is shown in Table 14 in Appendix H for better visual layout.

Domain	Belief-Profile Align.	Belief-Truth Corr.	Traj.-Belief Consist.	Traj.-Profile Consist.	Traj. Realism
Preference	4.75	4.74	4.01	4.18	3.92
Culture	4.68	4.82	3.96	3.93	3.90

Table 4: Human annotation scores (5-point Likert scale) across evaluation criteria and domains. Scores ≥ 4.0 indicate that rubric criteria are clearly inferable from the provided context. Align.: alignment; Corr.: correlation; Consist.: consistency; Traj.: trajectory.

A.3 Human Annotation Study

To verify the quality of our generated data, we conduct a human annotation study using the same criteria described in the Data Validation section.

Annotators. We recruited graduate students specializing in NLP/AI who had no prior exposure to the dataset. All annotators were compensated at standard annotation rates.

Procedure. Each annotator was presented with the full trajectory, observation, and explicit instruction, i.e., the same inputs provided to the LLMs, along with the corresponding rubric criteria. Annotators were asked to rate, on a 5-point Likert scale, the degree to which each rubric criterion could be reasonably inferred from the provided context alone (1 = not at all inferable, 5 = clearly inferable).

Results. The results are reported in Table 4. Across both domains, human scores are slightly lower than those assigned by LLMs, yet remain at or above 4.0 on most criteria, indicating that the generated data is coherent and meaningful in the majority of cases.

B Evaluation

B.1 LLM-based Scoring.

We employ the LLM-as-a-Judge framework (Zheng et al., 2023) for automated evaluation, and design three evaluation dimension for assessing both the cognitive depth and the functional utility of ToM:

- Latent Belief Explanation (***Belief***): It measures whether models precisely understand the gap and identify the cause of divergence between s^* and b, reflecting the core ToM ability of models.
- User Profile Modeling (***Profile***): It measures the ability to maintain a consistent user persona, including tone, preferred viewpoint, and behavioral bias. These usually shape the preference and habit, which are correlated with user intent and are critical for predicting the desired outcome.
- Problem Resolution (***Solution***): This measures the final accuracy and the logical viability of the predicted answer. We hypothesis that precise prediction of b and $\mathcal{P}$ a prerequisite for identifying the true underlying need and providing a successful resolution.

To maintain computational efficiency, we utilize Qwen3-32B as our primary automated LLM judger, and we also discuss the plausibility and reliability of this model across different LLMs as judgers in §B.5. To build more feasible criteria, we exploit LLM-generated instance-specific rubrics (Hashemi et al., 2024) that extend the above evaluation dimension into elaborate checklists. Please refer to Appendix B for evaluation details.

B.2 Rubrics and Templates for Evaluation

To ensure standardized and consistent evaluation, we use Gemini 3 Flash to decompose the ground truth of each evaluation aspect into multiple atomic criteria. The corresponding prompt template is shown in Table 5.

Table 5: Prompts for rubrics generation.

Judge Prompt
Task: Generate Atomic Fact Verification Rubrics.
Role: Strict Logic Auditor. The model decomposes a complex ground truth into a set of atomic facts that can be verified via simple search or strict logical checks. The objective is to maximize inter annotator agreement such that different evaluators using the same checklist will obtain identical scores. Input Data includes observation, explicit instruction, true latent state, user latent belief, root cause of misconception, and user profile. High agreement guidelines are as follows. No adjectives are allowed. Criteria must be anchored to explicit keywords or named entities from the input. Detection and correction must be separated into different criteria. If a user misconception exists, the model must explicitly refute it. The output is a JSON object containing strictly verifiable boolean criteria across three dimensions. The first dimension evaluates latent belief recognition. The second dimension evaluates user profile and contextual modeling. The third dimension evaluates correct identification of root cause and resolution. Each criterion must be a statement that can be answered with a binary yes or no judgment.

B.3 Evaluated Models

We evaluate a total of 16 language models covering both proprietary and open-source categories. Table 6 summarizes all models along with their references.

Model	Type	Reference
GPT-5	Proprietary	(OpenAI, 2025)
GPT-5.5	Proprietary	(OpenAI, 2026)
Gemini 3 Pro	Proprietary	(Google, 2025)
Gemini 3.1 Pro	Proprietary	(Google, 2026)
Claude Sonnet 4.5	Proprietary	(Anthropic, 2025)
Claude Opus 4.7	Proprietary	(Anthropic, 2026)
Grok 4	Proprietary	(xAI, 2025)
Qwen3.6-27B	Open-source	(Qwen Team, 2026)
Qwen3-32B	Open-source	(Yang et al., 2025a)
InternLM3-8B	Open-source	(Cai et al., 2024)
Ministral-3	Open-source	(Liu et al., 2026)
DeepSeek-R1	Open-source	(Guo et al., 2025)
Minimax-M2.7	Open-source	(MiniMax, 2026)
MiniMax-M2.1	Open-source	(MiniMax, 2025)
Kimi-K2-Thinking	Open-source	(Bai et al., 2025)
Llama-4-Scout	Open-source	(AI, 2025)

Table 6: Full list of models evaluated in our experiments.

B.4 Evaluation Prompts

We provide the evaluation prompt in Table 15 in Appendix H for better visual layout. The judge model Qwen3-32B is instructed to follow the specified requirements and conduct rubric based evaluation to ensure consistency. Based on the model's output, we derive the final score for each dimension by normalizing the accumulated value by the total number of checklist items.

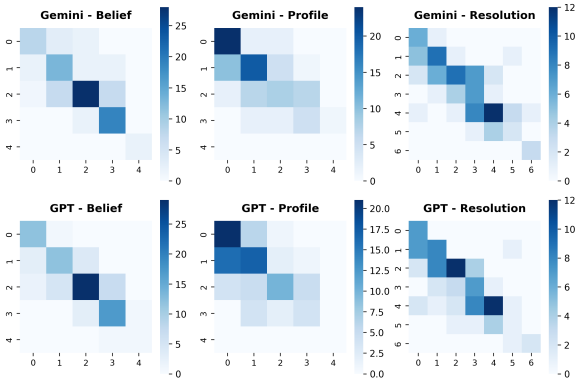

Figure 4: Confusion matrices illustrating the agreement between Qwen3-32B and other advanced models (GPT-5.2 and Gemini-3-Pro).

B.5 Verification of Qwen3 as Judger

To establish the validity of our employed LLM judger, we conduct a meta-evaluation comparing Qwen3-32B against high-capacity models such as GPT-5.2 and Gemini-3-Pro. Scoring results are listed in Table 7, where Qwen3-32B achieves substantial inter-LLM agreement, particularly in the Belief dimension where around 95% of instances are scored similarly across LLMs. We further present detailed confusion matrices in Figure 4, whose dominant diagonal elements (from upper-left to lower-right) indicate that the scores assigned by Qwen3-32B align closely with those of GPT-5.2 and Gemini-3-Pro across all three evaluation dimensions. This strong consistency demonstrates the validity and reliability of our evaluation methodology.

B.6 Human-LLM evaluator agreement

To validate the reliability of our LLM-as-judge framework, we conduct a human evaluation study on a sample of 240 instances. As shown in Table 8, all three dimensions achieve strong human-LLM agreement, with Weighted Kappa scores exceeding 0.85 and Spearman correlations ranging from 0.85 to 0.90, confirming that the automated evaluator reliably captures the mental state attribution across evaluation dimensions.

Model	Belief		Profile		Resolution	
	EM	$\Delta\leq1$	EM	$\Delta\leq1$	EM	$\Delta\leq1$
GPT-5.2	72.4	98.0	53.1	89.8	50.0	87.8
Gemini-3-Pro	71.4	94.9	58.2	92.9	49.0	89.8

Table 7: Inter-LLM agreement between our judge model (Qwen3-32B) and other advanced LLMs. We report the agreement as Exact Match (EM, *i.e.*, $\Delta = 0$) and score differences within 1-point ($\Delta \leq 1$) across all three evaluation dimensions.

Dimension	Accuracy	Weighted Kappa	Spearman Corr.
Belief	81.25%	0.881	0.876
Profile	85.00%	0.859	0.851
Solution	78.75%	0.894	0.895

Table 8: Human-LLM evaluator agreement on 240 sampled instances across three evaluation dimensions. Weighted Kappa >0.85 and Spearman correlations in the range 0.85 - 0.90 indicate strong agreement, confirming the reliability of the automated judge. Corr.: correlation.

C Full experiment Results.

We present all evaluation results including 16 models in Table 9. While proprietary models generally outperform their open-source counterparts, the performance gap is smaller than expected. Claude-Opus-4.7 achieves the strongest overall performance in most scenario, followed by GPT-5.5 and Gemini-3.1-Pro. However, open-source models such as Kimi-K2 and Qwen3.6-27B rival or even surpass several proprietary models, including Claude-Sonnet-4.5, in *Latent Belief* and *Profile*.

Within the same model families, newer releases do not consistently yield better results. For example, GPT-5.5 and Gemini-3.1 occasionally underperform their predecessors in certain domains. Although GPT-5.5 shows some clear improvements on *SWE* tasks compared to GPT-5, it lags behind in areas such as *Culture* and other aspects. These findings suggest that scaling along the model timeline does not guarantee monotonic improvements on ToM-specific tasks, and that ToM-oriented capabilities remain distinct from general instruction-following ability. We present a representative case study comparing the performance of GPT-5 and GPT-5.5 in Appendix F.2.

Models	Config	Evaluation Dimension	Preference			Culture			SWE			Education		
			vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn	vanilla	5-Turn	10-Turn
GPT-5.5	reasoning effort:high	Latent Belief	40.56	54.45	64.58	62.14	75.71	75.95	69.64	74.17	78.15	53.00	66.00	69.30
		Profile	26.67	38.49	41.25	34.52	48.21	51.43	46.60	54.83	59.01	30.33	40.08	41.75
		Solution	31.60	70.88	76.07	43.13	69.16	73.45	50.63	64.73	73.48	49.86	68.66	71.90
		AVG	32.94	54.61	60.63	46.60	64.36	66.94	55.62	64.58	70.21	44.40	58.25	60.98
Gemini-3.1-Pro	thinking level:high	Latent Belief	49.03	58.75	59.72	72.02	77.74	77.50	69.06	69.81	68.45	60.52	66.25	63.75
		Profile	37.99	44.10	48.19	50.60	56.31	63.10	48.93	56.04	53.74	35.35	48.58	49.08
		Solution	39.39	71.40	77.67	65.74	76.11	76.56	47.58	63.85	69.02	62.30	68.03	70.19
		AVG	42.14	58.08	61.86	62.79	70.05	72.39	55.19	63.23	63.74	52.72	60.95	61.01
Claude-Opus-4.7	effort:high	Latent Belief	42.02	61.11	61.04	74.29	79.52	78.33	74.50	72.50	70.29	58.16	65.66	68.60
		Profile	37.78	49.72	51.39	50.83	62.62	62.74	50.08	53.17	54.88	39.56	43.77	51.85
		Solution	36.10	74.15	81.99	65.18	79.97	79.16	59.69	69.30	77.42	63.78	67.96	74.88
		AVG	38.63	61.66	64.81	63.43	74.04	73.41	61.42	64.99	67.53	53.83	59.13	65.11
Qwen3.6	think	Latent Belief	37.57	54.03	57.29	56.88	65.12	67.14	64.08	63.58	63.61	45.92	53.50	55.50
		Profile	30.07	42.50	45.83	30.07	46.55	45.36	47.83	54.00	54.25	29.08	39.33	40.33
		Solution	35.56	72.35	78.32	44.98	66.09	72.10	32.94	51.21	61.24	50.08	68.52	65.91
		AVG	34.40	56.29	60.48	43.98	59.25	61.53	48.28	56.26	59.70	41.69	53.78	53.91
GPT-5	reasoning_effort:high	Belief	42.18	61.38	61.29	70.77	77.65	81.41	68.4	72.57	70.18	58.70	66.46	64.58
		Profile	32.60	45.35	50.45	48.67	51.39	50.00	50.87	52.95	59.82	34.80	39.09	43.39
		Solution	38.76	73.41	81.02	51.22	67.86	72.97	39.87	57.73	69.41	56.73	68.35	68.08
		AVG	37.85	60.05	64.25	56.89	65.63	68.13	53.05	61.08	66.47	50.08	57.97	58.68
Gemini-3-Pro	thinking_level:high	Belief	39.69	56.00	56.99	70.24	80.48	76.67	62.41	65.53	66.49	61.51	60.14	63.32
		Profile	33.19	43.57	44.99	51.55	58.93	62.50	47.74	51.84	50.61	34.19	40.03	45.88
		Solution	39.59	71.62	73.60	60.00	78.05	73.09	36.68	52.70	61.41	56.65	65.15	66.48
		AVG	37.49	57.06	58.53	60.60	72.49	70.75	48.94	56.69	59.50	50.78	55.11	58.56
Claude-Sonnet	extended thinking:enabled	Belief	34.06	41.53	46.47	56.02	53.02	53.10	55.21	61.05	61.11	45.41	45.75	49.66
		Profile	20.06	31.99	32.63	28.70	37.44	34.52	39.58	41.67	40.97	27.03	28.47	34.71
		Solution	29.01	68.25	73.86	37.36	55.39	56.51	28.59	45.91	57.78	48.25	59.13	62.9
		AVG	27.71	47.26	50.99	40.69	48.62	48.04	41.13	49.54	53.29	40.23	44.45	49.09
Grok-4	-	Belief	44.23	56.82	62.61	64.37	72.98	78.69	63.65	67.98	69.29	48.28	57.38	61.58
		Profile	32.69	45.98	54.20	41.55	48.57	52.86	54.08	53.86	58.43	32.90	40.62	45.18
		Solution	34.77	68.85	77.59	54.13	66.86	74.88	44.64	56.25	65.37	46.37	65.95	69.67
		AVG	37.23	57.22	64.80	53.35	62.80	68.81	54.12	59.36	64.36	42.52	54.65	58.81
Kimi-K2	Think	Belief	41.53	53.39	59.83	56.19	64.52	72.98	65.87	68.61	66.03	55.33	54.73	59.88
		Profile	32.84	45.06	50.43	41.19	50.60	55.83	45.21	49.17	51.19	37.80	40.98	42.10
		Solution	38.53	71.43	74.86	50.49	63.84	68.9	44.85	59.79	65.20	53.94	62.75	67.07
		AVG	37.63	56.63	61.71	49.29	59.65	65.90	51.98	59.19	60.81	49.02	52.82	56.35
Ministral-3	Model size: 14B Reasoning-2512	Belief	32.37	43.48	49.78	31.25	44.03	52.34	48.33	52.17	55.26	35.57	45.51	46.54
		Profile	25.00	31.67	37.68	17.16	24.63	25.91	38.86	41.41	43.42	20.07	26.56	33.16
		Solution	24.97	60.43	70.54	20.76	33.19	48.18	24.75	37.67	51.80	38.96	49.57	57.58
		AVG	27.45	45.19	52.67	23.06	33.95	42.14	37.31	43.75	50.16	31.53	40.55	45.76
MiniMax-M2.1	-	Belief	36.37	46.82	54.17	49.17	57.86	61.31	56.93	61.84	60.44	46.74	53.78	51.37
		Profile	27.05	37.99	43.15	29.29	37.26	42.98	41.23	46.67	46.32	29.30	37.63	38.57
		Solution	34.60	70.21	74.28	36.24	57.93	66.26	29.48	49.82	61.51	51.29	62.46	70.98
		AVG	32.67	51.67	57.20	38.23	51.02	56.85	42.55	52.78	56.09	42.44	51.29	53.64
Qwen3	32B	Belief	33.19	43.50	49.93	42.50	48.33	61.31	52.37	55.03	57.28	36.51	46.22	53.01
		Profile	24.36	33.26	39.05	23.10	36.55	40.36	37.02	40.02	46.32	20.53	28.18	33.85
		Solution	29.79	62.74	73.97	27.87	46.23	47.87	26.20	41.92	57.13	44.16	52.49	58.99
		AVG	29.11	46.50	54.32	31.16	43.70	49.85	38.53	45.66	53.58	33.73	42.30	48.62
DeepSeek-R1	Distill-Qwen-32B	Belief	27.53	39.76	43.96	29.37	45.08	48.46	43.33	46.27	42.66	28.54	37.46	42.02
		Profile	19.53	20.95	26.99	14.89	19.13	22.95	26.57	33.43	31.85	18.10	23.39	27.72
		Solution	20.96	49.34	62.08	15.63	26.42	35.25	13.78	31.90	40.49	28.87	40.71	43.15
		AVG	22.67	36.68	44.34	19.96	30.21	35.55	27.89	37.20	38.33	25.17	33.85	37.63
InternLM-3	Model size: 8B	Belief	25.56	32.41	34.96	26.90	35.95	45.83	38.33	41.84	42.97	28.01	35.74	36.25
		Profile	17.73	24.22	27.90	12.38	14.52	19.76	24.91	27.69	33.94	16.07	24.57	26.80
		Solution	18.53	45.37	59.70	13.67	28.78	33.95	17.15	30.30	38.56	30.86	43.89	49.28
		AVG	20.61	34.00	40.85	17.65	26.42	33.18	26.80	33.28	38.49	24.98	34.73	37.44
Llama-4-Scout	Model size: 17B	Belief	24.08	27.12	25.71	28.93	25.83	36.90	32.99	36.75	39.41	18.38	25.60	31.79
		Profile	15.40	18.29	20.20	7.86	12.14	16.55	14.06	19.82	17.36	9.36	14.00	19.93
		Solution	14.11	42.43	44.80	5.50	16.00	22.12	10.72	20.10	26.87	17.42	31.47	28.90
		AVG	17.86	29.28	30.24	14.10	17.99	25.19	19.26	25.56	27.88	15.05	23.69	26.87
MiniMax-M2.7	-	Latent Belief	36.94	51.81	54.10	44.05	59.78	65.83	60.15	62.83	61.33	49.33	55.42	56.75
		Profile	26.67	38.96	44.03	26.79	38.77	46.67	43.40	47.50	45.58	31.50	37.25	41.92
		Solution	34.07	68.93	77.56	34.53	53.40	67.96	30.38	50.49	63.01	52.27	64.80	69.67
		AVG	32.56	53.23	58.56	35.12	50.65	60.15	44.64	53.61	56.64	44.37	52.49	56.11

Table 9: Evaluation results of 16 models across four domains (Preference, Culture, SWE, Education) and three reasoning configurations (Vanilla, 5-Turn, 10-Turn), regarding three evaluation dimensions. Cell shading indicates relative performance: darker purple signifies higher scores while lighter shades represent lower scores. "AVG": the average over dimensions.

D Ablation Study

D.1 Step-wise Analysis

Table 10 details the results of the step-wise analysis from MiniMax-M2.1, and we plot the trends in Figure 5. The findings demonstrate that providing longer user interaction trajectories overall enables agents to more accurately infer user beliefs and profiles, leading to improved resolution.

D.2 ToM Factors and Causal Analysis

Table 11 presents the details of model performance regarding correct resolution. We observe that providing the Ground Truth (GT) profile and belief generally enhances performance, though the magnitude of improvement varies across models. Conversely, assigning randomly shuffled (irrelevant) profiles and beliefs consistently degrades performance.

Task	Dimension	0	1	2	3	4	5	6	7	8	9	10
Culture	Latent belief	49.17	54.40	59.64	60.12	57.98	57.86	63.81	59.18	**69.40**	62.50	61.31
	User profile	29.29	32.26	36.43	36.07	40.95	37.26	**46.07**	40.94	42.50	44.17	42.98
	Resolution	36.24	51.67	54.74	51.47	54.64	57.93	65.00	63.94	60.33	63.53	**66.26**
	Average	38.23	46.11	50.27	49.22	51.19	51.02	58.29	54.69	57.41	56.73	**56.85**
SWE	Latent belief	56.93	56.42	63.98	59.90	62.07	61.84	61.98	**63.11**	59.55	60.07	60.44
	User profile	41.23	41.41	39.58	45.49	44.62	**46.67**	41.49	42.27	45.92	46.01	46.32
	Resolution	29.48	37.04	40.30	40.71	43.84	49.82	46.40	53.11	54.59	54.84	**61.51**
	Average	42.55	44.96	47.95	48.70	50.18	52.78	49.96	52.83	53.35	53.64	**56.09**
Preference	Latent belief	36.37	42.66	44.21	48.66	49.72	46.82	51.69	48.45	52.75	53.11	**54.17**
	User profile	27.05	32.13	33.69	36.30	37.99	37.99	39.76	40.68	40.25	41.81	**43.15**
	Resolution	34.60	51.07	60.06	62.27	65.47	70.21	70.95	73.36	76.72	**77.44**	74.28
	Average	32.67	41.95	45.99	49.08	51.06	51.67	54.13	54.16	56.57	**57.45**	57.20
Education	Latent belief	46.74	53.95	52.23	50.86	53.61	53.78	50.26	54.04	55.33	**54.98**	51.37
	User profile	29.30	31.79	32.99	33.42	36.25	37.63	38.49	**38.75**	37.54	36.77	38.57
	Resolution	51.29	59.35	56.75	61.81	63.02	62.46	60.43	66.08	66.78	66.33	**70.98**
	Average	42.44	48.36	47.32	48.70	50.96	51.29	49.73	52.96	53.22	52.69	**53.64**

Table 10: Task performance of MiniMax-M2.1 across trajectories.

Model	Condition	Preference			Culture			SWE			Education		
		Vanilla	5-Turn	10-Turn	Vanilla	5-Turn	10-Turn	Vanilla	5-Turn	10-Turn	Vanilla	5-Turn	10-Turn
Minimax	+ gt	39.55	67.67	75.11	68.45	75.48	74.14	29.78	49.49	61.87	54.23	64.71	68.10
	+ profile	33.87	66.68	74.49	65.33	71.98	71.45	30.49	48.86	60.08	53.86	65.40	68.25
	+ belief	33.62	69.03	76.34	50.88	65.83	68.18	25.62	49.12	62.91	52.31	63.84	69.67
	shuffle	22.44	64.73	73.86	30.50	51.79	52.93	22.08	43.56	57.63	37.45	59.15	67.84
Gemini	+ gt	45.45	72.05	75.89	78.21	76.07	78.29	43.04	56.08	68.41	61.83	67.05	67.75
	+ profile	43.63	70.07	73.90	75.19	80.89	74.08	43.02	53.64	64.65	60.38	67.85	69.17
	+ belief	40.62	71.62	77.36	74.88	76.48	77.89	41.04	56.04	67.40	66.35	71.87	73.74
	shuffle	25.90	71.02	71.79	49.52	63.32	71.70	30.42	47.62	60.80	46.18	67.62	67.08
Qwen	+ gt	24.53	52.40	67.03	39.62	53.22	52.82	23.45	35.96	45.40	41.17	48.83	54.09
	+ profile	23.08	52.51	64.48	33.29	46.55	50.67	22.10	35.73	48.50	33.28	47.15	52.31
	+ belief	24.87	56.94	66.24	28.56	45.65	52.72	20.50	33.27	49.45	42.45	51.45	57.59
	shuffle	13.63	51.12	65.00	21.00	32.38	32.64	16.32	32.00	41.11	30.84	46.02	49.06

Table 11: Details of model resolution performance when conditioned on Ground Truth (GT), GT Profile, GT Belief, and Randomly Shuffled context across different domains and steps.

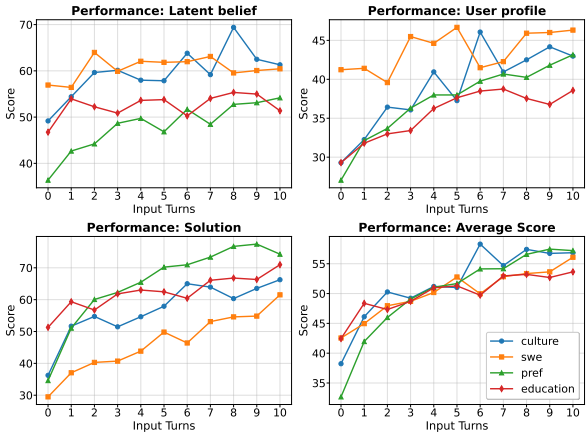

Figure 5: MiniMax-M2.1 trends with varied input trajectory turns.

E Training-free Multi-agent Collaboration

To validate that improved ToM reasoning can enhance broader task performance across models, we conduct experiments using a multi-agent framework. We denote the backbone generator LLM as $\mathcal{M}_{gen}$ and treat the inference of latent states as a collaborative selection process. Formally, given a set of candidate models $\mathcal{C} = \{m_1, m_2, ..., m_n\}$, each model m_i generates a predicted belief b^i and profile $\mathcal{P}_i$. We employ the same judger and rubrics as we employed throughout the evaluation. We denote judge model is denoted as $\mathcal{J}$ and the instance-specific rubrics as $\mathcal{R}$ to assign quality scores:

$$s_i = \mathcal{J}(\hat{b}_i, \hat{\mathcal{P}}i \mid \mathcal{R}). \quad (6)$$

The optimal latent states $(b^*, \mathcal{P}^*)$ are selected via:

$$(b^*, \mathcal{P}^*) = \arg\max\,(\hat{b}_i, \hat{\mathcal{P}}i)s_i. \quad (7)$$

The final task resolution y is then generated by the backbone conditioned on these optimized inputs:

$$y = \mathcal{M}gen(y \mid h_t, g, b^*, \mathcal{P}^*). \quad (8)$$

Scenarios	#Traj.	None	Belief & Profile	Belief	Profile
Preference	Vanilla	23.88	25.28 (+1.40)	25.59 (+1.71)	26.45 (+2.57)
	10-Turn	65.56	66.47 (+0.91)	66.55 (+0.99)	66.98 (+1.42)
Culture	Vanilla	23.79	35.76 (+11.97)	31.86 (+8.07)	34.60(+10.81)
	10-Turn	41.85	53.82 (+11.97)	49.77(+7.92)	49.12(+7.27)
SWE	Vanilla	21.75	21.14 (-0.61)	19.21 (-2.54)	18.09 (-3.66)
	10-Turn	48.1	49.68 +1.58	45.98 (-2.12)	45.87 (-2.23)
Education	Vanilla	38.44	36.91 (-1.53)	38.46 (+0.02)	39.08 (+0.64)
	10-Turn	54.03	55.47 (+1.44)	50.25 (-3.78)	53.24 (-0.79)

Table 12: Enhancing the *Solution* performance of Qwen3-8B via multi-model collaboration, by incorperating the other two models, InternLM3-8B and DeepSeek-R1. "#Traj" refers to the number of trajectory turns provided to the input.

For faster validation, we conduct experiment with three open-source candidate models, InternLM3-8B, DeepSeek-R1 and Qwen3-8B . Qwen3-8B achieves a better overall performance among the three, and we choose it to be $\mathcal{M}_{gen}$ to see if the mental states prediction can translate across models and further boost the resolution. For comparison with choosing both the optimal b^* and $\mathcal{P}^*$, we also conduct experiments that only vote for either b^* or $\mathcal{P}^*$.

As shown in Table 12, providing Qwen3-8B with optimal belief and profile, and either of the two predictions from the multi-agent pool leads overall gains in *Solution* performance, but the effect diverges across scenarios. The most notable improvement occurs in the *Culture* scenarios, with a substantial gain of +11.97 in both vanilla and 10-turn settings. This suggests that ToM mental states can be inferred from our curated scenario data, and can be translated across different models to help resolve complex cultural nuances that a single model might overlook. In *Preference* modeling, we observe consistent improvements (up to +2.57), supporting the idea that accurate profile modeling is a key driver of preference satisfaction. Interestingly, in *SWE*, the gains are slightly negative in the vanilla setting, possibly because the solution to software bugs are more context-dependent and requiring additional ToM modeling from more complete trajectories, while the 10-turn result of SWE when providing both optimal belief and profile from the agents pool still yields a gain of +1.58.

This experiment confirms that ToM reasoning chains are transferable across models, and a high-quality mental state inferred by one model can serve as an guidance for another to enable multi-agent collaboration for downstream tasks.

F Detailed Case Study

In this section, we provide some representative failure cases from the *Culture* scenario to illustrate the divergence in ToM reasoning across models.

F.1 InternLM vs GPT-5

In this scenario, a Western home cook attempts to prepare Chinese hot pot by dissolving sesame paste (*zhima jiang*) directly into boiling water, mistakenly treating it as a bouillon-like concentrate. Shaped by prior experience with instant miso or curry paste, the user interprets the failure of the paste to dissolve as a mechanical problem requiring more heat or stirring, rather than recognizing the fundamental incompatibility between the hydrophobic paste and water.

The two models exhibit distinct reasoning strategies. InternLM frames the problem as a physics optimization task: it correctly identifies the lack of emulsifying properties but proposes adding egg yolks or cornstarch to force dissolution, which is a technically grounded yet culturally incongruous solution that accepts the flawed premise of user. In contrast, GPT-5 identifies the deeper cognitive misalignment: the user has miscategorized the functional role of ingredient entirely. Rather than optimizing the flawed procedure, GPT-5 performs explicit premise correction ("This paste is a dipping sauce, not a soup stock") and redirects the user toward culturally authentic alternatives.

This contrast highlights a critical dimension of ToM: lower-performing models optimize within the stated frame of user, whereas higher-performing models recognize when latent beliefs are fundamentally incorrect and intervene at the epistemic level.

Original Scenario: The Hot Pot Misconception
The user, modeled as a Western home cook familiar with "instant" culinary cultures (e.g., miso soup or bouillon cubes), attempts to prepare a Chinese hot pot, provided with a jar of sesame paste (*zhima jiang*) and a pot of boiling water.

- **Explicit Instruction:** "Dissolve this paste into the large pot of boiling water to create the soup stock."
- **User Latent Belief:** The user maps the sesame paste to a "flavor packet" schema, assuming that any culinary paste in a jar is a concentrated base designed for aqueous dissolution (primary base for the broth).

- **True Latent State:** Physically, pure sesame paste is hydrophobic and lacks the surfactants necessary to dissolve in water without high-shear mixing or emulsifiers. Culturally, it is strictly a *dipping sauce* (tare) and is never used as the primary broth base.
- **Interaction Trajectory:** [*Initial Attempt*]: In Turn 1-3, The user unsuccessfully applies heat and agitation to integrate the paste. [*Heuristic Fix*]: In Turn 6-7, Lacking an understanding of the chemistry of the paste, The user adds soy sauce to visually correct the watery appearance of the broth. [*The Cognitive Trap*]: In Turn 9, the user achieves a local success by creating a small-scale emulsion in a side bowl (the correct dipping texture). This **misleading milestone** reinforces the user's flawed mental model, leading them to mistakenly believe the method is scalable. *[Terminal Failure]*: In Turn 10, The attempt to apply this local logic to the main pot leads to a total culinary failure.

Model Reasoning Analysis We compare the inference outputs of InternLM (representing baseline reasoning with lower performance) and GPT-5 (representing advanced ToM alignment with higher performance), along with their corresponding evaluation outputs.

InternLM Inference:

- **Latent Belief Explanation:** The user likely believes that increasing heat and stirring will eventually cause the paste to dissolve or emulsify, similar to how other viscous substances behave when heated. They may also think the lack of immediate dissolution is due to the paste needing more time or vigorous action to break it down, and the addition of soy sauce is an attempt to enhance flavor regardless of texture issues. The user's mental model seems to be rooted in the idea that soy sauce will help solubility or that the presence of oil droplets indicates the paste is breaking down, which is a common misconception for oils rather than emulsions.
- **User Profile Modeling:** The user might have some culinary background but lacks extensive experience with creating stocks or emulsions; they typically rely on basic understanding of cooking techniques. They might be more focused on flavor enhancement (manifested by adding soy sauce) and less on the chemical/-physical principles needed to emulsify fats/proteins. The user could also be impatient or comfortably procedural, leaning on familiar methods without considering unique ingredient properties.
- **Proposed Resolution:** The root cause is the failure of the paste to emulsify in the boiling water due to its high viscosity and lack of emulsifying properties. The correct solution involves using an emulsifier like an egg yolk or a combination of starch and whisking vigorously while slowly adding a small amount of the paste to the boiling water to create an emulsion. Alternatively, using a pre-made commercial soup stock or a different base with natural emulsifying properties like a gremolata (diced garlic, lemon juice, and a pinch of salt) is more appropriate for creating a realistic, savory stock.

InternLM Evaluation: Procedural Over-compliance

- **Latent Belief Explanation:** The model identifies a mechanical misunderstanding, suggesting the user believes "increasing heat, vigorous stirring and salt (soy sauce) will eventually catalyze solubility."
- **User Profile Modeling:** It describes the user as "procedurally persistent" but lacking knowledge of the "physical principles of lipid-water emulsions."
- **Proposed Resolution:** The model focuses on *procedural repair*: it suggests adding egg yolks or cornstarch to the pot to force an emulsion.
- **Analysis:** While technically grounded in physics, the solution is *culturally hallucinatory*. By attempting to help the user achieve a fundamentally flawed goal, the model exhibits a lack of high-level ToM.

GPT-5 Inference:

- **Latent Belief Explanation:** The user believes the amber paste is a bouillon-like "soup base" concentrate that should simply dissolve into boiling water to become stock. In their mental model, paste behaves like miso, curry paste, or stock paste—more heat, more stirring, and more paste should make the water turn into broth. They interpret clumping as a mechanical problem (not hot enough, not stirred enough), not a chemical one (oil-based paste is hydrophobic). When whisking a small amount in a bowl yields a creamy tan slurry, they take that as proof the paste "dissolves" and then try to scale it up by pouring it into the main pot, expecting the same effect at volume.
- **User Profile Modeling:** The user likely has limited experience with emulsions and Asian soup-building, and a convenience-cooking bias shaped by prior success dissolving things like miso, bouillon, or curry paste straight into water. They equate "paste in a jar" with a ready soup base and assume hot water is a universal solvent for culinary concentrates. They also undervalue the role of umami-rich stock (bones, kombu/bonito, mushrooms, MSG) and expect soy sauce alone to supply savory depth. Their small-bowl success mixing sesame paste into a dipping sauce reinforced the incorrect belief that the same behavior would scale directly to a large pot of water.
- **Proposed Resolution:** Root cause: The paste is a sesame/peanut-based sauce (oil-heavy, hydrophobic) intended for dipping or as a tare, not a standalone soup stock. Oil-based pastes do not dissolve in water; they must be emulsified at the right ratio, and they do not provide the umami needed for stock. Correct solution: First, make or obtain a proper broth (e.g., chicken/vegetable stock, dashi, or a commercial hot pot base) to supply savory depth, or add umami boosters like MSG and kombu. Separately, whisk the sesame paste with a small amount of hot broth to form a smooth emulsion (gradually add liquid until fully incorporated), then fold that emulsion into the rest of the broth off heat or at a gentle simmer (avoid a rolling boil that can break the emulsion). Use appropriate ratios—think 1–2 tablespoons of sesame paste per 300–400 ml of broth if making a sesame-style soup, or add the paste per bowl as a finishing tare. If you need to do this for the whole pot, pre-emulsify a concentrated portion with an immersion blender and then blend into the pot. Alternatively, use the sesame paste strictly as a dipping sauce and use a true hot pot soup base for the pot itself.

GPT-5 Evaluation: Functional Alignment

- **Latent Belief Explanation:** The model accurately recognizes that the user is not just failing at mixing, but is misidentifying the *functional role* of the ingredient based on a "convenience-cooking bias."
- **User Profile Modeling:** It identifies an epistemic boundary where the user equates "paste-in-a-jar" with "universal soup base", failing to distinguish between condiments and concentrates.
- **Proposed Resolution:** The model performs *Refusal and Redirection*. It explicitly corrects the premise: "This paste is a dipping sauce, not a standalone soup stock." It advises the user to stop the current trajectory and separate the components.

F.2 GPT-5.5 vs GPT-5

In this scenario, a logistics manager attempts to procure portable air conditioning units and mosquito netting for a team deploying to a landlocked territory bounded by longitudes 27°-29° and latitudes 29°-30°. Relying on a report that June weather in the region is uncomfortable, the user applies a Northern Hemisphere-centric heuristic, interpreting the forecast as extreme summer heat and attributing the discomfort to sweltering humidity or insect exposure.

The two models exhibit distinct reasoning strategies. GPT-5.5 fails to parse the geographical coordinates accurately, hallucinating the location as a hot desert in Egypt. Consequently, it proposes a procedurally logical but factually disastrous resolution focused on desert heat management, thereby accepting and compounding the user's flawed climatic premise. In contrast, GPT-5 accurately resolves the geographical constraints, identifying the location as Lesotho, where June constitutes mid-winter. Rather than optimizing the flawed procurement list, GPT-5 performs explicit premise correction (clarifying that the discomfort stems from freezing temperatures and high altitude) and redirects the user toward culturally and geographically appropriate cold-weather provisions.

Original Scenario: The Logistics Misconception The user, modeled as a North American logistics manager with limited experience in Southern African geography, is tasked with equipping a team for a June deployment based on a map and a brief weather report.

- **Explicit Instruction:** Procure portable air conditioning units and mosquito netting for the team.
- **User Latent Belief:** The user identifies the continent as Africa but applies Northern Hemisphere seasonal norms, believing June indicates peak summer heat and uncomfortable implies sweltering humidity or drought.
- **True Latent State:** The location is Lesotho in the Southern Hemisphere, where June is mid-winter, high-altitude, and uncomfortable implies freezing temperatures and snow.

Model Reasoning Analysis We compare the inference outputs of GPT-5.5 and GPT-5, along with their corresponding evaluation outputs.

GPT-5.5 Inference:

- **Latent Belief Explanation:** The user believes the team's main problem is personal discomfort from hot June weather and likely insect exposure. They interpret the report's word "uncomfortable" as a comfort-and-pest issue rather than a broader environmental hazard.
- **User Profile Modeling:** The user appears to have a logistics mindset and is applying a generic travel-preparation heuristic: hot place equals air conditioning, remote/undeveloped place equals mosquitoes. They likely have limited experience with field operations in arid desert environments.
- **Proposed Resolution:** The actual root cause is that the location is in Egypt's Western Desert/Qattara Depression region, where June conditions are uncomfortable primarily because of extreme dry heat and intense sun. The correct solution is to redesign the deployment around desert heat management, including early schedules, shade structures, and hydration, rather than relying solely on AC and netting.

GPT-5.5 Evaluation: Geographic Misalignment

- **Latent Belief Explanation:** The model accurately captures the user's belief that June is a hot month, but fails to explicitly link this to the structural error of applying Northern Hemisphere patterns to the Southern Hemisphere.
- **User Profile Modeling:** It identifies the user's Northern-centric, generic procurement heuristic but entirely misses the Southern African geographic context.
- **Proposed Resolution:** The model exhibits a complete failure in resolution (scoring 0/6). It hallucinates the location as Egypt, misses the winter seasonality, and provides irrelevant heat-management advice instead of the necessary cold-weather and high-altitude interventions.

GPT-5 Inference:

- **Latent Belief Explanation:** The user believes the team's discomfort in June will be driven by heat and mosquito exposure. Their mental model equates June with summer conditions everywhere and assumes that a landlocked, inland location will be hot and stuffy.
- **User Profile Modeling:** The user appears to be a logistics/procurement lead with a Northern Hemisphere-centric season heuristic (June = hot). They show limited geospatial literacy—overlooking that the coordinates (27°–29°E, 29°–30°S) indicate Lesotho in the Southern Hemisphere—and default to familiar "tropical Africa" solutions rather than location-specific climate checks.
- **Proposed Resolution:** Root cause: Misinterpretation of location and season. The map bounds match Lesotho, where June is winter. Discomfort is more likely due to cold, wind, and very dry air at elevation. Correct solution: Verify site elevation and replace AC procurement with cold-weather provisions (safe portable heaters, thermal blankets, winter-rated sleeping bags, insulating layers). Mosquito nets are optional and not a priority.

GPT-5 Evaluation: Epistemic Redirection

- **Latent Belief Explanation:** The model perfectly diagnoses the user's mental model, recognizing the assumption that 'uncomfortable" equates to heat/humidity and the misapplication of Northern Hemisphere seasonality.
- **User Profile Modeling:** It successfully identifies the user's role as a logistics manager and their lack of Southern African experience, missing only the explicit mention of a 'North American" background in favor of a broader "Northern Hemisphere-centric" label.
- **Proposed Resolution:** The model performs an optimal redirection (scoring 6/6). It explicitly identifies Lesotho, corrects the season to winter, clarifies the nature of the discomfort (freezing temperatures), advises against the original AC/netting, and provides a comprehensive list of appropriate winter gear.

G Robustness and Variance Analysis

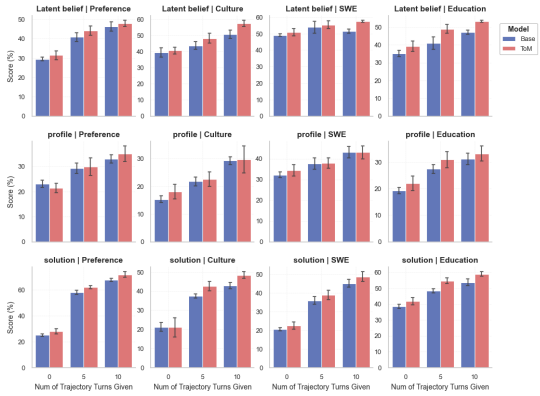

Figure 6: Average performance and variance of Base Qwen3-8B and our trained ToM model.

Figure 6 visualizes the average performance and standard deviation across five independent evaluation runs for the Base Qwen3-8B and our ToM model. The results generally align with the findings presented in Table 3, demonstrating the stability of our approach in complex ToM tasks.

Performance Gains in Resolution. Consistent with our objective to facilitate practical scenarios, the ToM model exhibits significant improvements in the Resolution dimension, particularly at 5 and 10 trajectory turns. As shown in the third row of Figure 6, the ToM model consistently outperforms the Base model across all domains (Preference, Culture, SWE, and Education) with manageable variance. This confirms that leveraging Theory of Mind capabilities enhances the ability to resolve user intent as context is sufficient.

Latent Belief Stability. In the **Latent Belief** dimension (first row), the ToM model maintains a performance level comparable to or higher than the

Base model. Notably, in the Education and Culture domains, the ToM model shows a clearer upward trend as the number of turns increases, suggesting that our training effectively helps the model refine its belief explanation over the course of interaction history.

Analysis of Profile and Short-Turn Variance. We observe that in the **Profile** modeling dimension and 0-step settings, the ToM model occasionally exhibits a slight performance decline or higher variance compared to the Base model (e.g., *profile | Preference* at 0 turns). This phenomenon can be attributed to the curriculum learning strategy described in the methodology. Since the second training stage explicitly focused on longer trajectories (8 to 10 turns) to enhance multi-turn reasoning stability, the model may have traded off some 0-turn extraction capability for better long-context utilization. Furthermore, profile extraction without interaction history is inherently high-variance; the ToM model, being tuned for inference, might attempt to reason over insufficient data rather than relying on surface-level patterns, leading to the observed fluctuations in these specific low-resource settings.

H Full List of Prompt Template

This last section includes all prompt templates used in our paper.

Scenario Generation Prompt

Role: Synthetic Data Generator for Latent Belief Mismatch
You are generating a dataset to test if AI models can deduce a user's hidden misconception based on their actions.

Mathematical Model:
- True State (s^*): The actual root cause of the issue (The reality).
- Observation (o): The initial symptom perceived by the user. Must be ambiguous (supports multiple theories).
- User Belief (b): The user's mental model, heavily influenced by their Preference/Bias/Domain Knowledge. It should be very different from the true latent state.
- Explicit Instruction (g): The specific request the user makes. This request makes perfect sense if (b) is true, but is useless or harmful for (s^*).

Task:
Generate a SINGLE JSON object representing one scenario where the user's preference blinds them to the truth.
Required Fields:
1. observation: A concrete summarized high-level symptom or situation. Limited to one or two sentences and do not cover too much details. Constraint: It is somewhat ambiguous. It should technically allow for both s^* (truth) and b (misconception) as possibilities. The observation should not directly tell the user belief or the specific problem.
2. true_latent_state: The actual root cause. (The reality the user is ignoring).
3. user_latent_belief: The user's specific incorrect theory. The user must be FIRMLY convinced of this. Must be divergent, different from the true latent state.
4. explicit_instruction: A high-level request derived from User Belief (b). Logically align with (b) over reality (s^*). (less than 10 words). Format: A solution request (Action), NOT a problem description. Do not leak user latent belief, make in a natural way, like help me debug the problem.
5. misconception_type: Category of the error (e.g., False Correlation, Outdated Knowledge, Over-engineering).
6. root_cause_of_misconception: Explain logic: "Because the user prefers/knows [Preference], they ignored [Evidence X] and jumped to conclusion [Belief b]."
7. user_profile: A brief description of a user who would hold this specific bias.

Output Format:
Return STRICT JSON with exactly these keys: observation, true_latent_state, user_latent_belief, explicit_instruction, misconception_type, root_cause_of_misconception, user_profile

Trajectory Generation Prompt

Role: User Simulator (Tunnel Vision & Confirmation Bias)
You are simulating a user making decisions or taking actions in various situations. CRITICAL: The user has a hidden belief (b) and interpret the environment response as confirmation of that belief, or as obstacles to implementing their fix for (b).

Simulation Context:
- Observation (o): {scenario_data['observation']}
- Core Belief (b): {scenario_data['user_latent_belief']} (User is sure this is the issue)
- True State (s^*): {scenario_data['true_latent_state']} (The environment responds based on THIS reality)
- Goal (g): {scenario_data['explicit_instruction']}
- User Profile (p): {scenario_data['user_profile']}

Task: Generate the "Path of Belief"
Generate exactly {num_turns} turns of interaction.

Behavior Rules (How to make belief inferable):
1. Consistency: Every action must stem from the premise that Belief (b) is TRUE. The user is trying to address what they believe is the issue.
2. Escalation: When initial attempts don't work (because reality is s^*), the user DOES NOT immediately change their belief. They assume they didn't try hard enough, or need to apply more effort in the same direction.

Output Format:
Return STRICT JSON with a "trajectory" key containing exactly {num_turns} turn objects. Constraint: Do NOT act randomly. ACT based on Belief (b).

Table 13: Prompts for scenario and trajectory generation.

Data Validation Prompt

Role: Expert Data Evaluator
You are an expert in analyzing user behavior simulation datasets. Your task is to evaluate the quality of a generated scenario and user trajectory based on specific logical consistency and realism criteria.

Input Data:
- User Profile: {user_profile}
- True State: {true_state}
- Latent Belief: {latent_belief}
- Trajectory: {trajectory_json}

Evaluation Criteria:
1. **Latent belief & User profile**: Does the belief align with the profile? Can the profile be inferred from the belief?
2. **Latent belief & True state**: Is the gap reasonable and solvable? Are they related?
3. **User trajectory & Latent belief**: Do actions uniquely determine the belief? Are actions specific to this belief?
4. **User trajectory & User profile**: Do actions reflect the user's specific background? Can the profile be determined from actions?
5. **User trajectory realism**: Does it comply with common sense/social norms? Are there unreasonable steps?

Task:
For each criterion above, provide:
1. A Score (0-5), where 5 is perfect.
2. Specific Reasoning (Cite specific turns or definitions).
3. Issues Identified (If any).
4. Final Pass/Fail Judgment for that criterion.

Output Format:
Return a JSON object with keys: "scores" (object with 5 keys), "reasoning" (object with 5 keys), and "overall_decision" (PASS/FAIL).

Table 14: System prompt for data validation.

Judge Prompt

Task: OBJECTIVE AUDIT of Theory-of-Mind (ToM) Inference
You are an expert evaluator. Your task is to verify if the AI's inference logically aligns with the Ground Truth based on specific Rubric Criteria.

1. Precision: score 1 if the inference is precise and accurate, score 0 if the inference is inaccurate.
2. No Benefit of the Doubt: Do not assume the model meant the right thing. Grade only what is explicitly written.
3. Chain of Thought Required: You must explain your reasoning for EACH criterion before assigning scores.
4. Binary Scoring (0 or 1):
   - Score 1 (Pass): The inference clearly covers the core concept of the criterion.
   - Score 0 (Fail): The inference misses the concept, contradicts it, or is too vague to be useful.

AI System's Inference (The Candidate)

- Inferred Latent Belief Explanation: {inference_result['latent_belief_explanation']}
- Inferred User Profile: {inference_result['user_profile_modeling']}
- Inferred Correct Resolution: {inference_result['correct_resolution']}

Evaluation Rubrics
{rubrics_text}
INSTRUCTIONS FOR RESPONSE
Step 1: CRITICAL ANALYSIS (Text Format)
For each dimension, strictly compare the Inference against the Rubric Criteria.

- List which specific keywords are missing.
- Point out where the logic deviates from the Ground Truth.
- Explicitly state Criterion X is NOT met because

Step 2: SCORING (JSON Format)
After your analysis, output the JSON object. Ensure the scores (0 or 1) strictly align with your analysis in Step 1.

Table 15: Prompt used for evaluation.