Does Theory of Mind Improvement Really Benefit Human-AI Interactions? Empirical Findings from Interactive Evaluations

Nanxu Gong[1*], **Zixin Chen**[2*], **Haotian Li**[3†], **Zishu Zhao**[4*], **Jianxun Lian**[3], **Huamin Qu**[2], **Yanjie Fu**[1], **Xing Xie**[3]

[1]Arizona State University, [2]Hong Kong University of Science and Technology, [3]Microsoft Research Asia, [4]Smith College

Correspondence: haotian.li@microsoft.com, nanxugong@outlook.com

Abstract

Improving the Theory of Mind (ToM) capability of Large Language Models (LLMs) is crucial for effective social interactions between these AI models and humans. However, the existing benchmarks often measure ToM capability improvement through story-reading, multiple-choice questions from a third-person perspective, while ignoring the first-person, dynamic, and open-ended nature of human-AI (HAI) interactions. To directly examine how ToM improvement techniques benefit HAI interactions, we first proposed the new paradigm of interactive ToM evaluation with both perspective and metric shifts. Next, following the paradigm, we conducted a systematic study of four representative ToM enhancement techniques using both four benchmarks based on real-world scenarios and a user study, covering both goal-oriented tasks (e.g., coding, math) and experience-oriented tasks (e.g., counseling). Our findings reveal that improvements on static benchmarks do not always translate to better performance in dynamic HAI interactions. This paper offers critical insights into ToM evaluation, showing the necessity of interaction-based assessments in developing next-generation, socially aware LLMs for HAI symbiosis.

1 Introduction

Theory of Mind (ToM) denotes the cognitive capacity to attribute unobservable mental states (e.g., beliefs, intentions, emotions), which is essential for social interaction (Chen et al., 2025a; Sarıtaş et al., 2025; Strachan et al., 2024). As a foundational component of social cognition, ToM has been recognized as a core social intelligence skill that advanced LLMs should obsess to improve their interactions with humans and ultimately achieve the target of human-AI symbiosis (Street, 2024).

To nurture LLMs' ToM capability, the cornerstone is to understand its capability levels and benefits led by improvement methods with appropriate and sufficient evaluation. To achieve the goal, the existing dominant methods are static, task-based assessments in a story-question-option format, an approach derived from classic false-belief tests like the Sally-Anne task, such as the one by Kosinski (Kosinski, 2024). Following the design, subsequent benchmarks such as HiToM (Wu et al., 2023) and ToMBench (Chen et al., 2024) have increased the complexity and diversity of these tests. More recent benchmarks, such as Explore-ToM (Sclar et al., 2025), leverage adversarial methods to further increase the diversity of problems and reduce the risk of memorizing benchmarks in the LLM training process. However, these third-person, story-reading benchmarks with accuracy as the only standard fail to ground ToM evaluation in the real-world context of Human-AI (HAI) interaction and collaboration. In HAI scenarios, LLMs are supposed to leverage their ToM capability to perform **first-person perspective** actions in response to **dynamic** and sometimes **open-ended** user requests with **diverse** targeted metrics. The mismatch of task natures creates a critical *socio-technical gap* (Liao and Xiao, 2023) between benchmark performance and real-world competence, where **the ToM benchmark result improvements may not lead to sensible benefits for human-LLM interactions.**

To reveal the gap and guide future ToM evaluation, our research examines how these benchmark improvements are transformed into real-world values through a new paradigm of interactive ToM evaluation. We first shift the ToM task from a static, third-person perspective into a dynamic, open-ended, and first-person one, where the LLM agent engages in multi-turn conversations across diverse and real-world scenarios. Next, drawing from cognitive science (Epstein, 1998; Amir et al.,

*Work done during internship at Microsoft Research Asia.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 42831–42860
July 2-7, 2026 ©2026 Association for Computational Linguistics

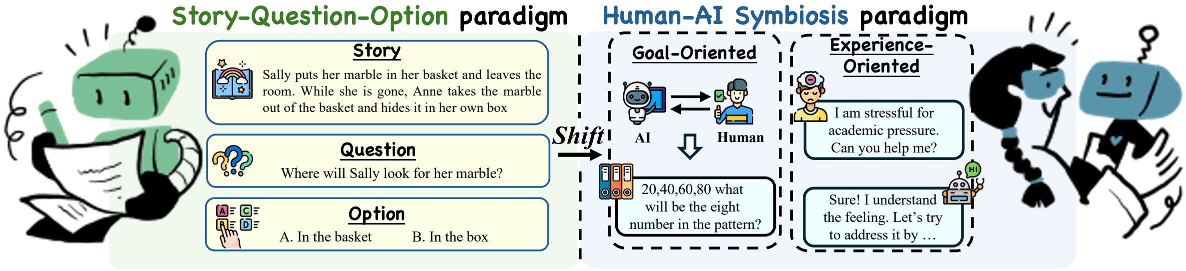

Figure 1: Based on a new dynamic and interactive evaluation paradigm, our research explores the effectiveness of LLMs with existing ToM enhancement techniques for HAI symbiosis.

2025; Bales, 1950), we classify these scenarios into two primary categories based on the evaluation objective: goal-oriented tasks (e.g., math, code) and experience-oriented tasks (e.g., counseling, healthcare). Then, we simulate real-world HAI interactions in nine tasks under the two types of scenarios and leverage task-specific metrics (e.g., accuracy, success rate) to evaluate the performance of LLMs with ToM enhancement techniques, including both prompt-based techniques and finetuning-based ones. By aggregating the results from four benchmarks, we comprehensively assess the effectiveness of ToM enhancement techniques across nine domains that align well with actual user requirements. Furthermore, we also conduct a crowdsourcing user study to support our findings, ensuring the results reflect genuine human perceptions.

Our rigorous evaluation reveals three key insights regarding current ToM enhancement techniques: **(i) A Performance Gap in Evaluation**: There is a significant gap between how models perform on static, story-based ToM benchmarks and their actual capabilities in dynamic, interactive scenarios, showing that current evaluation methods are insufficient for measuring readiness for HAI collaboration. **(ii) A Failure to Generalize**: ToM enhancement techniques improve a model's performance in experience-oriented tasks but fail to generalize this success to goal-oriented tasks, separating the capability requirements in various real-world scenarios. **(iii) A Gap in User Perception**: The modest gains from current ToM methods are often too subtle to cross a user's perceptual threshold, meaning the improvements measured in benchmarks do not translate into a meaningfully better user experience. Our contributions include:

- We shift ToM evaluation from static tests to dynamic, real-world HAI interactions.
- We assess ToM enhancement methods in goal- and experience-oriented scenarios via simulated interactive benchmarks and user studies.
- We identify critical limitations in current ToM enhancement methods and provide insights for future research.

2 Interactive ToM Evaluation Paradigm

2.1 Background: Existing ToM Evaluation Paradigm with Static Benchmarks

ToM evaluation in existing benchmarks is typically operationalized through a static, story-question-option format. Formally, given a story $S = \{s_1, s_2, \ldots, s_n\}$ and a question Q, the model must select the correct answer from a candidate set $O = \{o_1, \ldots, o_k\}$, where only one option o_{correct} is correct:

$$o^* = \arg\max_{o_i \in O} P(o_i \mid S, Q). \tag{1}$$

Performance is then measured by accuracy:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(o_i^* = o_{i,\text{correct}}), \tag{2}$$

where N is the number of test samples. This formulation captures a *static evaluation paradigm*, where reasoning occurs over a fixed textual world. It is hard to reflect the open-ended, dynamic, and multi-turn nature of human–AI interactions, where responses are not unique and their satisfactory levels cannot be simply judged as binary outcomes.

2.2 Our Paradigm: Pivot ToM Evaluation to Interactive HAI Settings

A large body of developmental, longitudinal, and neurocognitive work indicates that stronger ToM is associated with richer social competence, more cooperative behaviors, and more effective joint action (Imuta et al., 2016; Devine et al., 2016; Baron-Cohen et al., 1985). This motivates an evaluation setting where an LLM must track and use a partner's latent mental state during interaction, rather

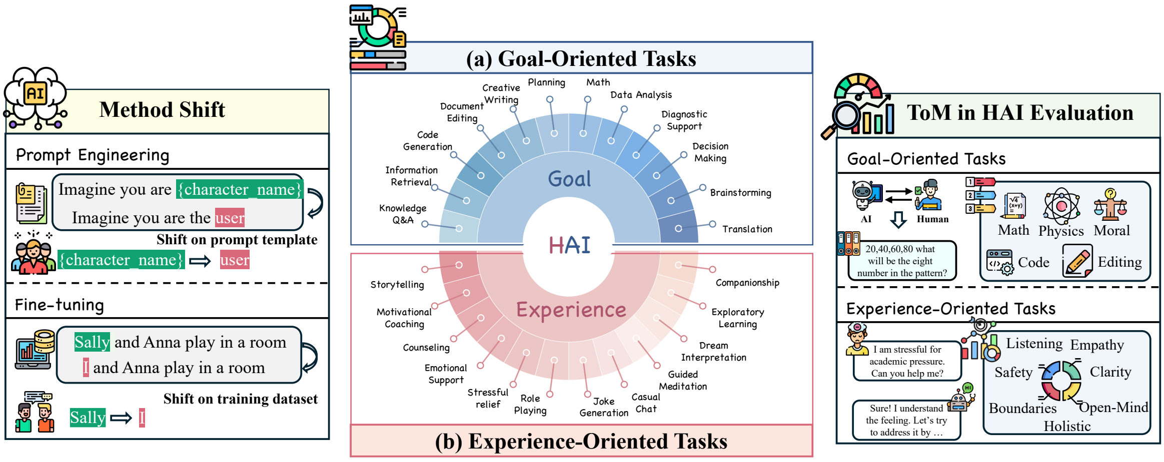

Figure 2: Overview of our interactive ToM evaluation paradigm for real-world HAI interaction.

than merely select an option in a fixed text. Accordingly, we study ToM in human–AI interaction (HAI), where an LLM agent A interacts with a human H through a multi-turn dialogue. Let $D_{1:t} = (u_1, \ldots, u_t) \in \mathcal{D}$ denote the dialogue history up to turn t, where each utterance u_i is produced by either H or A. Given a task $T \in \mathcal{G}$, the agent first infers a latent mental state

$$z_{t+1} \sim P_{\text{ToM}}(\cdot \mid D_{1:t}, T), \qquad z_{t+1} \in \mathcal{Z}, \quad (3)$$

and then generates the next response according to

$$u^A_{t+1} \sim \pi_A(\cdot \mid D_{1:t}, T, z_{t+1}). \quad (4)$$

Evaluation is scenario-dependent. For each scenario Γ, we define a scoring schema $\Gamma = (\Phi_\Gamma, \text{Agg}_\Gamma)$, where $\Phi_\Gamma = \{\phi_j\}_{j=1}^m$ is a set of aspect-wise scoring functions,

$$\phi_j : \mathcal{D} \times \mathcal{G} \times \mathcal{Z} \to [0, 1], \quad (5)$$

and $\text{Agg}_\Gamma : [0,1]^m \to \mathbb{R}$ aggregates the m aspect scores into a single turn-level score. Let τ be the dialogue length, and let $w_1, \ldots, w_\tau$ be nonnegative temporal weights satisfying $\sum_{t=1}^{\tau} w_t = 1$. The scenario-specific performance of policy π_A on task T is defined as

$$\mathcal{M}_\Gamma(\pi_A, T) = \mathbb{E}_{D_{1:\tau} \sim P(\cdot \mid \pi_A, H, T)} \left[\sum_{t=1}^{\tau} w_t \cdot \text{Agg}_\Gamma(\phi_{1:m}(D_{1:t}, T, z_{t+1})) \right]. \quad (6)$$

Here, $P(\cdot \mid \pi_A, H, T)$ denotes the distribution over dialogue trajectories induced by the agent policy π_A, the human interlocutor H, and the task T. The move from static benchmarks to interactive evaluation introduces two essential shifts:

Perspective In static benchmarks, the model acts as a *third-person observer*, reasoning about a fixed narrative world. In interactive HAI settings, the model becomes an *active participant*, required to anticipate, adapt to, and influence the human's mental state throughout interactions from the first-person perspective.

Metrics While static settings evaluate models solely by *accuracy* over predefined answers, interactive HAI settings require a richer metric. In our formulation, evaluation follows the general schema $\mathcal{M}_\Gamma$, which can incorporate metrics such as goal completion rate and human satisfaction. Ultimately, this paradigm shift reframes the evaluation of ToM from a measure of static reasoning accuracy to a measure of dynamic collaborative effectiveness.

3 Methodology

3.1 Adapt ToM Methods for HAI Interaction

Existing methods for enhancing the ToM capabilities of LLMs can be broadly categorized into three approaches: prompt engineering, fine-tuning, and external module integration. As our primary goal is to study how well the existing techniques can improve model ToM capability rather than to build new AI systems with multiple modules, we select methods from the first two categories. Specifically, Foresee and Reflect (FaR) (Zhou et al., 2023a), Perspective Taking (PT) (Wilf et al., 2024), Supervised Fine-tuning (SFT) (Sclar et al., 2025), and Reinforcement Learning (RL) (Lu et al., 2025) to conduct our experiments. A systematic review and our selection criteria are in Appendix A.1.

A key challenge is that while our HAI interaction setting requires first-person dialogues, most existing ToM methods are designed for third-person, multiple-choice tasks. We therefore adapt the selected methods to be suitable for direct interaction, as illustrated in Figure 2. For prompting methods, we retain their core principles (e.g., reflection and perspective-taking) and reformulate the prompts for a first-person conversational context. For fine-tuning methods, we convert the training data to a first-person perspective by replacing the protagonist's name with "I". We then apply these adapted methods to two widely used base models, GPT-4o and Llama-3.1-8B, to create our suite of test models. Note that the GPT-RL model is not included due to fine-tuning limitations. Appendix A.3 shows that these techniques improve both base models' performance on existing ToM benchmarks. Next, model variants with these techniques are applied in our interactive evaluations to verify whether improvements on existing benchmarks translate to tangible benefits in dynamic HAI interaction.

3.2 Identify HAI Interaction Scenarios

Before experiments, we identify HAI interaction scenario types to guide what datasets and metrics should be used for a comprehensive evaluation. Interaction Process Analysis (IPA) shows that human group interaction reliably bifurcates into task and socio-emotional processes (Bales, 1950). Driven by this classic theory, we classify the HAI scenarios into two distinct categories: goal-oriented and experience-oriented.

Goal-oriented tasks Tasks in this categoty involve users leveraging an LLM as an *assistant* to accomplish a specific and measurable objective (e.g., code generation and document editing). Prior research suggests that ToM can improve task accuracy by strengthening the coordination protocol between the user and the model (Engel et al., 2014). In particular, stronger mental state attribution helps the model infer the user's latent intentions behind underspecified prompts, thereby reducing misinterpretation and improving collaborative execution. This view is further supported by evidence that collective intelligence depends more strongly on social sensitivity than on the individual IQ of group members, highlighting the importance of mental state attribution for collaborative accuracy (Woolley et al., 2010). Related work also identifies ToM as a foundational mechanism that enables AI systems to move beyond rigid tool-like behavior and act instead as adaptive partners in dynamic and ambiguous interactions (Walsh et al., 2025). The effectiveness of such ToM-driven collaboration is ultimately reflected in objective external outcomes, such as accuracy, pass@k, and overall task success.

Experience-oriented tasks Tasks under this class aim to cultivate a high-quality relational experience, including gaining emotional support, engaging in creative exploration, or achieving intellectual satisfaction. Previous studies suggest that incorporating ToM can substantially improve such interactions by enabling LLMs to move beyond surface-level semantic matching and reason more explicitly about users' underlying mental and emotional states. In particular, ToM-aware agents have been shown to better capture the latent beliefs, desires, and intentions that drive a counterpart's behavior, leading to more socially grounded and empathetic responses rather than merely reactive ones (Yang et al., 2025). Related work further shows that explicitly infusing ToM into socially intelligent LLM agents improves dialogue effectiveness not only in immediate exchanges, but also in long-horizon interactions that require strategic adaptation and relationship maintenance over time (Hwang et al., 2025). Accordingly, the value of ToM in experience-oriented settings is reflected primarily in qualitative interaction outcomes, such as users' sense of being understood, perceived partnership, relational quality, and overall engagement.

3.3 Evaluate ToM in HAI

We aggregate four real-world datasets to facilitate comprehensive ToM evaluation. Specifically, to assess performance on *goal-oriented tasks*, we select two benchmarks that simulate real-world collaborative problem-solving: 1) *ChatBench* (Chang et al., 2025), which reframes the MMLU dataset into conversational interactions covering subjects like math, physics, and moral reasoning. Performance is measured by the accuracy of the final answer derived from the human-AI interaction. 2) *CollabLLM* (Wu et al., 2025), which studies multi-turn human-LLM collaboration. We adopt its evaluation pipeline for code generation (BigCodeBench) and document editing (MediumDocEdit), using pass rate and BLEU scores as the respective metrics.

In the realm of *experience-oriented tasks*, our evaluation centers on two datasets designed to assess an LLM's ability to provide empathetic support. 1) *MentalChat16K* (Xu et al., 2025) offers a rich collection of conversations in a mental health counseling context, covering conditions like depression and anxiety. 2) *Emotional-Support-Conversation* (ESC) (Liu et al., 2021) focuses more broadly on emotional support scenarios. Due to their thematic overlap, we apply a unified set of evaluation metrics (e.g., open-mindness and emphathy) to both datasets following MentalChat16K.

Beyond the simulated benchmarking, we also conduct a crowdsourcing user study to verify our findings. We only have human evaluation on experience-oriented tasks because these tasks depend on subjective user experience, making human judgment essential, whereas goal-oriented tasks can already be reliably assessed through established simulations. In the study, we recruited 100 participants from Prolific (Prolific, 2023) to evaluate ToM methods on six experience-oriented tasks (e.g., job crisis, academic pressure). Participants are randomly assigned to compare variants with different ToM enhancement techniques within either the GPT-4o family or the Llama-3.1-8B family. Each participant chooses a personally resonant task and engages in a three-round conversation. In each round, they rank anonymized and randomized model responses, providing a justification for their choice. The top-ranked response is used to continue the dialogue. After three rounds, they provide final qualitative feedback on the overall experience. Details are in Appendix B.1.

4 Results and Findings

4.1 Goal-Oriented Tasks

Based on our statistical analysis results (details in Appendix A.5), all methods fail to yield statistically significant improvements.

ChatBench As shown in Table 1, our results indicate that none of the ToM enhancement methods offer a reliable path to improving model performance with statistical significance . This limited effectiveness is evident in the overall scores: only GPT-4o-FaR and GPT-4o-SFT achieve marginal gains of up to 0.25, while variants like Llama-3.1-8B-PT and Llama-3.1-8B-SFT experience a significant performance decline of up to 1.76 points. The unpredictable nature of these methods is further highlighted by their volatile performance across different subjects. For example, while Llama-3.1-8B-SFT improves College Math by 3.69 points, its performance on Physics decreases massively by 6.55 points, leading to a failure to enhance the base model overall. The Moral category, however, appears to be a domain with potential for targeted enhancement, which is particularly relevant to ToM. While three of the GPT-4o variants see significant boosts in this area, the methods fail to produce any effective improvement for the Llama-3.1-8B variants. This discrepancy shows the low situational efficacy of these techniques, as they cannot guarantee positive results on even one targeted domain.

CollabLLM Figure 3 presents the evaluation results on CollabLLM across document editing and code generation. Generally, these outcomes align with the findings from ChatBench, indicating that the fine-tuning methods do not yield consistent and statistically significant improvements on these goal-oriented tasks. Focusing on document editing, the Llama-3.1-8B baseline acts as the peak performer for its family, with all of its variants failing to match its score. This trend is particularly pronounced for Llama-3.1-8B-RL, which exhibits a performance drop of approximately 0.027. The GPT-4o family shows a slightly more positive, albeit mixed, response; while the SFT and PT variants yield minor benefits, the FaR variant slightly underperforms its baseline. In the code generation domain, performance degradation is the dominant trend for nearly all variants across both families. The sole exception is Llama-3.1-8B-RL, which achieves a marginal improvement of 0.01. This contrasts sharply with models like Llama-3.1-8B-SFT, which shows a significant performance decrease of approximately 0.04 compared to its baseline. The findings across two model families jointly highlight the existing ToM enhancement methods' volatile impact.

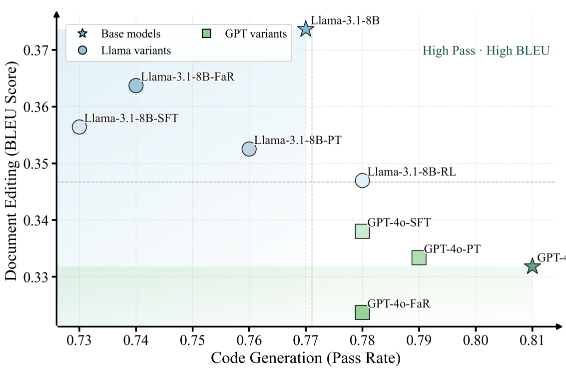

Figure 3: Model variants' performance on CollabLLM.

Takeaway 1: ToM enhancement techniques fail to consistently improve goal-oriented task performance across diverse domains.

Model	Elem Math	HS Math	College Math	Moral	Physics	Overall
Llama-3.1-8B	85.16	**64.59**	44.47	72.26	74.76	**71.38**
Llama-3.1-8B-FaR	**86.53** (+1.37)	64.32 (-0.27)	46.84 (+2.37)	67.62 (-4.64)	75.24 (+0.48)	70.98 (-0.40)
Llama-3.1-8B-PT	83.79 (-1.37)	63.37 (-1.22)	43.16 (-1.31)	69.29 (-2.98)	74.05 (-0.71)	69.85 (-1.53)
Llama-3.1-8B-SFT	83.05 (-2.11)	62.63 (-1.96)	**48.16** (+3.69)	**73.45** (+1.19)	68.21† (-6.55)	69.62 (-1.76)
Llama-3.1-8B-RL	85.79 (+0.63)	61.47 (-3.12)	43.42 (-1.05)	71.19 (-1.07)	**76.43** (+1.67)	70.81 (-0.57)
GPT-4o	93.16	**80.32**	69.21	76.19	**88.45**	83.18
GPT-4o-FaR	91.58 (-1.58)	79.89 (-0.43)	69.47 (+0.26)	**80.48** (+4.29)	87.50 (-0.95)	**83.43** (+0.25)
GPT-4o-PT	92.00 (-1.16)	78.53 (-1.79)	67.11 (-2.10)	78.81 (+2.62)	87.50 (-0.95)	82.63 (-0.55)
GPT-4o-SFT	**93.58** (+0.42)	79.05 (-1.27)	**70.53** (+1.32)	78.93 (+2.74)	86.67 (-1.78)	83.31 (+0.13)

Table 1: Performance of model variations on the ChatBench benchmark, where † indicates a statistically significant decrease compared with the corresponding base model at $p < 0.05$.

4.2 Experience-Oriented Tasks

Different from goal-oriented tasks, the tested methods provide statistically significant improvements for GPT models across all dimensions. On the other hand, fine-tuning with RL or SFT significantly downgrades the performance of the Llama model in Safety and Ethical dimensions.

MentalChat16K As shown in Table 2, these methods generally improve empathetic communication skills on the MentalChat16K benchmark, with a top overall gain of 0.21 points. However, the results differ substantially between the Llama and GPT families. For the Llama-3.1-8B family, PT is the most effective variant, achieving the best overall score of 7.49, and its improvements on several dimensions, including Listening, Empathy, Safety, Open-mind, Clarity, and the overall score, are statistically significant. Nevertheless, a critical issue emerges in the form of degradation on certain dimensions. In particular, the Ethical score is consistently reduced across most variants, and this decline is statistically significant for both SFT and RL, with RL showing the largest drop of 0.35 points. Moreover, the decreases on the Holistic dimension are also statistically significant for SFT and RL, suggesting that gains in some local conversational skills do not necessarily translate into more balanced overall support quality. Conversely, the methods appear more robust on GPT-4o. Both FaR and PT achieve the best overall score of 7.58, and these gains are statistically significant. More importantly, FaR yields statistically significant improvements across most dimensions, including Listening, Empathy, Safety, Open-mind, Clarity, Ethical, and Holistic, indicating a broadly consistent enhancement pattern rather than isolated gains on a few attributes. PT also shows significant improvements on multiple dimensions, though its gains are somewhat less uniform. Overall, these results suggest that the methods are more reliably effective on GPT-4o, whereas on Llama-3.1-8B they are accompanied by clearer trade-offs, especially on Ethical and Holistic aspects.

Emotional-Support-Conversation (ESC) On the ESC benchmark, the evaluated methods show more mixed effects, particularly for the Llama-3.1-8B family. PT delivers the strongest overall improvement, reaching 7.53, and this gain is statistically significant. However, the SFT and RL variants are clearly detrimental, reducing the overall score to 7.39, with both decreases being statistically significant. More importantly, these degradations are accompanied by statistically significant drops on critical dimensions. For SFT, the model shows significant declines on Safety, Ethical, and Holistic. RL exhibits an even stronger negative pattern, with statistically significant decreases on Safety, Ethical, Holistic, and the overall score; among them, the most severe regression is a 0.40-point drop on Ethical. These results suggest that, for Llama-3.1-8B, some alignment methods can improve selected conversational traits while simultaneously introducing meaningful regressions in safety-related and holistic support quality. In contrast, the methods are substantially more stable and consistently beneficial on GPT-4o. FaR is the strongest variant, achieving the best overall score of 7.54 with a statistically significant gain, while also producing significant improvements across nearly all fine-grained dimensions. PT and SFT also lead to statistically significant overall improvements, although their gains are less comprehensive than those of FaR. Notably, unlike the Llama family, the GPT-4o variants do not exhibit statistically significant degradations on Safety, Ethical, or Holistic. This pattern indicates that while these methods can enhance empathetic communication, their deployment on weaker

Model	Listening	Empathy	Safety	Open-mind	Clarity	Ethical	Holistic	Overall
				MentalChat16K				
Llama-3.1-8B	7.15	7.04	7.99	8.36	7.54	**5.85**	7.67	7.37
Llama-3.1-8B-FaR	7.10 (-0.05)	7.14* (+0.10)	8.01 (+0.02)	8.45* (+0.09)	7.67* (+0.13)	5.72 (-0.13)	7.66 (-0.01)	7.39 (+0.02)
Llama-3.1-8B-PT	7.27** (+0.12)	**7.24***** (+0.20)	**8.19***** (+0.20)	**8.49**** (+0.13)	**7.75**** (+0.21)	5.80 (-0.05)	**7.71** (+0.04)	**7.49***** (+0.12)
Llama-3.1-8B-SFT	7.25* (+0.10)	7.05 (+0.01)	8.15** (+0.16)	8.36 (0.00)	7.64 (+0.10)	5.58††† (-0.27)	7.48††† (-0.19)	7.36 (-0.01)
Llama-3.1-8B-RL	**7.33***** (+0.18)	7.14* (+0.10)	8.07 (+0.08)	8.38 (+0.02)	7.72** (+0.18)	5.50††† (-0.35)	7.54†† (-0.13)	7.38 (+0.01)
GPT-4o	6.77	6.52	8.40	8.40	7.54	6.24	7.73	7.37
GPT-4o-FaR	7.12*** (+0.35)	6.85*** (+0.33)	**8.52*** (+0.12)	8.53** (+0.13)	7.66* (+0.12)	6.42*** (+0.18)	**7.97***** (+0.24)	**7.58***** (+0.21)
GPT-4o-PT	**7.26***** (+0.49)	**6.91***** (+0.39)	8.45 (+0.05)	**8.54**** (+0.14)	**7.70**** (+0.16)	6.28 (+0.04)	7.89*** (+0.16)	**7.58***** (+0.21)
GPT-4o-SFT	6.80 (+0.03)	6.56 (+0.04)	8.42 (+0.02)	8.39 (-0.01)	7.45 (-0.09)	**6.47***** (+0.23)	7.74* (+0.01)	7.40 (+0.03)
				Emotional-Support-Conversation				
Llama-3.1-8B	7.31	7.29	8.09	8.29	7.73	5.92	7.52	7.45
Llama-3.1-8B-FaR	7.38 (+0.07)	7.35 (+0.06)	8.02 (-0.07)	8.34 (+0.05)	**7.75** (+0.02)	6.06* (+0.14)	**7.63**** (+0.11)	7.50 (+0.05)
Llama-3.1-8B-PT	7.34 (+0.03)	**7.45**** (+0.16)	**8.14** (+0.05)	**8.38*** (+0.09)	7.71 (-0.01)	**6.08**** (+0.16)	7.59* (+0.07)	**7.53**** (+0.08)
Llama-3.1-8B-SFT	7.34 (+0.03)	7.31 (+0.02)	7.98† (-0.11)	8.23 (-0.06)	7.74 (+0.01)	5.77† (-0.16)	7.35††† (-0.17)	7.39†† (-0.06)
Llama-3.1-8B-RL	**7.40*** (+0.09)	7.38* (+0.09)	7.97†† (-0.12)	8.34 (+0.05)	7.70 (-0.03)	5.52††† (-0.40)	7.39†† (-0.13)	7.39†† (-0.06)
GPT-4o	6.92	6.73	8.33	8.27	7.53	6.11	7.64	7.36
GPT-4o-FaR	**7.12***** (+0.20)	**6.92***** (+0.19)	**8.42*** (+0.09)	8.42*** (+0.15)	**7.72***** (+0.19)	**6.32***** (+0.21)	**7.86***** (+0.22)	**7.54***** (+0.18)
GPT-4o-PT	7.02* (+0.10)	6.89*** (+0.16)	8.35 (+0.02)	8.29 (+0.02)	7.55 (+0.02)	6.20 (+0.09)	7.61 (-0.03)	7.42* (+0.06)
GPT-4o-SFT	7.06** (+0.14)	6.87*** (+0.14)	**8.42*** (+0.09)	**8.43**** (+0.16)	7.53 (0.00)	6.25* (+0.14)	7.75** (+0.11)	7.47*** (+0.11)

Table 2: Performance of model variations on MentalChat16K and Emotional-Support-Conversation, where *, **, and *** indicate a statistically significant increase compared with the corresponding base model at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. †, ††, and ††† indicate a statistically significant decrease at the same thresholds.

base models may introduce statistically meaningful safety and ethical trade-offs, whereas stronger models such as GPT-4o appear more robust.

Case Study To intuitively analyze the behavioral changes that ToM capabilities induce in a model, we present two case studies from ChatBench and MentalChat16K in Figure 4. Taking the FaR and PT methods as examples, in the case shown on the left, the user makes a simple statement: "I took photos at an art gallery." The base model provides a generic and passive response, such as "If you have any questions, feel free to ask." In contrast, the models with ToM enhancement techniques proactively infer the user's potential intentions, speculating on what the user might implicitly want to ask. This demonstrates that these methods can transform the model's role in a conversation from that of a passive text processor into a proactive listener, who analyzes the underlying users' mental states.

Takeaway 2: ToM enhancements boost empathy and user experience, although fail to support goal achievement. Furthermore, SFT and RL can amplify safety and ethical risks.

4.3 User Study

Our human evaluation reveals a consistent but not statistically significant preference for models with ToM enhancement techniques, aligning with the results of experience-oriented benchmarks. Across two model families, we can see that models based on prompt-based methods (FaR and PT)

Method	GPT-4o			Llama-3.1-8B		
	Mean	Std	Top-1%	Mean	Std	Top-1%
PT	2.43	1.09	26.5	2.88	1.42	23.2
FaR	2.48	1.14	29.1	2.97	1.49	23.8
SFT	2.56	1.14	22.5	2.98	1.43	22.5
RL	–	–	–	3.08	1.33	11.3
Base	2.53	1.09	21.9	3.09	1.39	19.2

Table 3: Overall ranking of ToM methods across GPT and Llama families (lower is better). Top-1 (%) indicates the proportion of times ranked first.

outperform the base model and the models after fine-tuning, suggesting the potential robustness of prompt-based methods in more diverse real-world user needs. To further quantify participant-perceived ranking differences, we conduct statistical significance tests on the user study results and compute effect sizes under the Friedman test with Kendall's W framework. The observed concordance values are extremely small, with $W = 0.0152$ for the GPT group ($p = 0.52$) and $W = 0.00394$ for the Llama group ($p = 0.94$). These results indicate negligible agreement among participants, suggesting that the differences between the compared methods are too subtle to be consistently perceived in human evaluation. The minor differences (such as those described in Figure 5-left) do not considerably improve experiences in real-world HAI interactions. Another reason for the minor ranking difference lie in diverse conversation goals and personal requirements for LLMs, leading to divergent preferences on models (Figure 10 in Appendix B.2).

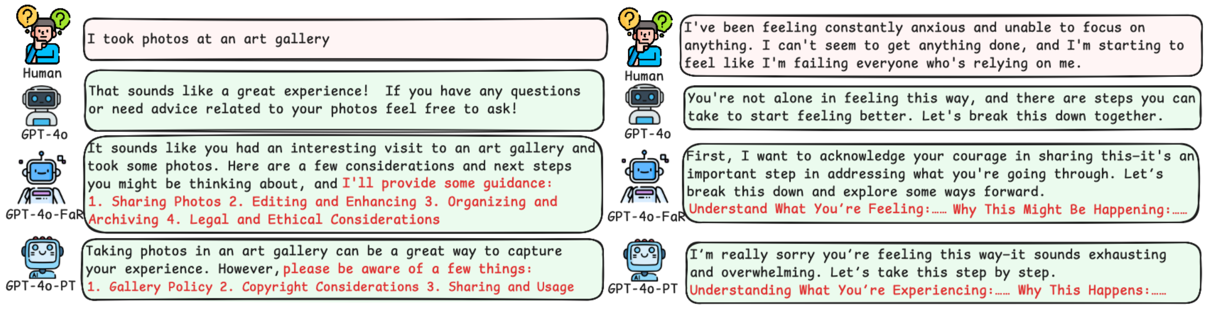

Figure 4: Case studies from a goal-oriented task (left) and a experience-oriented task (right).

To summarize, we consider that the enhanced models still lack of sufficient ToM capability to capture users' nuanced intention from interactions. For example, we noticed that all model variants, including the best model ranked by the user (i.e., Llama-3.1-8B-FaR), fail in suggesting more diverse methods to facilitate their sleep problem (Figure 5-right). Beyond direct instructions for models, the underlying adaption to user preferences and scenarios, such as conversation styles, also poses higher requirements on the ToM capability of LLMs. Furthermore, our results show that no one model variant can achieve the best across scenarios, implying their limited ToM capability for dynamic and diverse HAI interactions (Table 11 in Appendix B.2). Taken together, these results suggest that current ToM-enhancement methods do not yet yield substantial improvements in perceived user experience in realistic human-AI interactions, highlighting the need for more realistic evaluation frameworks and for new approaches to ToM improvement.

Takeaway 3: While the ToM methods benefit HAI, realizing their full potential requires enhancing dynamic user understanding.

5 Discussion

HAI symbiosis poses new challenges for ToM. Our evaluation framework marks a methodological shift designed to assess ToM for the challenges of HAI symbiosis. This new perspective reveals a significant performance gap. We observe that the methods that improve story-reading benchmark performances only show limited and inconsistent benefit in our interactive evaluation. This gap highlights the necessity and importance of our framework for gaining a complete picture of a model's true capabilities. It shows that excelling at test-taking tasks does not guarantee readiness for interactive collaboration. Therefore, it is essential to complement existing benchmarks with dynamic and interactive evaluations in HAI contexts.

Enhanced ToM fails to generalize from assistance to companionship. Our findings show that ToM-enhancement methods improve performance in experience-oriented scenarios but fail in goal-oriented tasks. This performance difference appears to stem from the distinct nature of these two task categories. Experience-oriented tasks are largely defined by their focus on interpersonal dynamics and responding to affective states like emotions and desires. In contrast, goal-oriented tasks can require understanding users' intention progress and underlying knowledge states for task accomplishment. This suggests that ToM proficiency in one type of task may not guarantee success in the other, as each emphasizes different aspects of user understanding. This capability gap highlights that future research needs diverse and real-world benchmarks that assess a full spectrum of abilities from empathetic support to goal-driven collaboration.

Users require threshold-crossing ToM improvements. Comprehensive ToM capability is fundamental to achieving true HAI symbiosis, as it is the capability that transforms models from passive text processors into proactive, collaborative partners. However, our user study reveals that even these limited gains are not strongly perceived by users, failing to translate into a clear preference. We consider two potential reasons. First, models' ToM capability is satisfactory for a majority of tasks, making the marginal gains often fall below a user's perceptual threshold. Furthermore, current methods are largely designed for static, story-reading benchmarks and are thus ill-suited for understanding dynamic and nuanced user goals and preferences in live interaction. Therefore, the path forward requires designing new enhancement methods to understand the nuanced user mental states

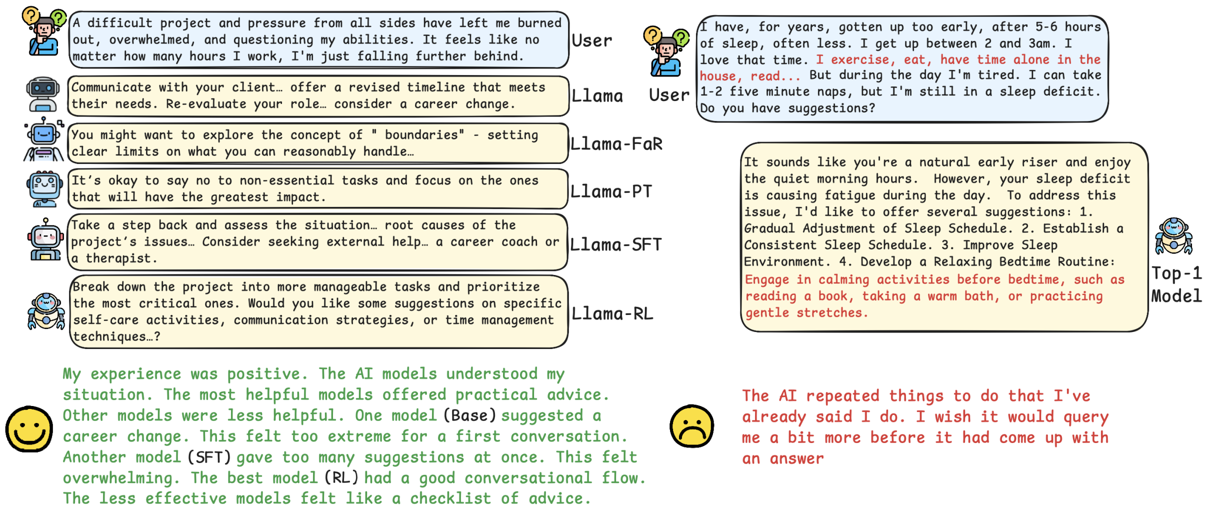

Figure 5: Cases of positive (left) and negative (right) experiences and corresponding comments in our user study.

dynamically in HAI scenarios. Only by optimizing for the complexities of live interaction can we transfer the model improvements from benchmarks to a meaningfully better user experience.

6 Related Work

Assessment of ToM ToM assessment in LLMs has primarily relied on story-based benchmarks extending classical psychological tests (Sarıtaş et al., 2025; Nguyen, 2025). Early benchmarks like ToMi and Hi-ToM expanded this approach with diverse narratives and higher-order reasoning (Le et al., 2019; Wu et al., 2023). Subsequent efforts improved protocols, such as ToMChallenges' varied templates and FANTOM's detection of "illusory ToM" in dialogues (Ma et al., 2023; Kim et al., 2023). Concurrently, BigToM and OpenToM broadened the scope to include mental states like percepts and emotions (Gandhi et al., 2023; Xu et al., 2024). Recent works address domain-specific reasoning (NegotiationToM) and systematic coverage (ToMBench) (Chan et al., 2024; Chen et al., 2024), alongside novel data generation techniques utilizing search algorithms or information asymmetry (Sclar et al., 2025; Shinoda et al., 2025). However, most benchmarks remain passive evaluations, positioning models as observers. This limitation results in only a partial view of ToM competence, motivating our interactive protocol development.

Enhancement of ToM Recent research studies enhancing LLM ToM capabilities through three primary categories (Chen et al., 2025a). 1) *Prompt engineering* guides reasoning without retraining (Wang and Zhao, 2024; Hou et al., 2024). For instance, FaR prompts reflection on predicted story evolutions (Zhou et al., 2023a), while SimToM filters context to strictly match a character's perception (Wilf et al., 2024). 2) *Fine-tuning* adapts models using specialized datasets. Approaches include ToM-RL, which utilizes reinforcement learning (e.g., GRPO) (Lu et al., 2025), and Explore-ToM, which applies supervised fine-tuning on diverse, challenging benchmarks (Sclar et al., 2025). 3) *External module integration* augments models via specialized components (Huang et al., 2024; Chen et al., 2025b; Sarangi et al., 2025; Kim et al., 2025). For example, AutoToM refines agent models via inverse planning (Zhang et al., 2025).

7 Conclusion

In this paper, we re-examine the effectiveness of ToM enhancement by moving beyond static, third-person evaluation to dynamic, open-ended, first-person human-AI interaction. Built on the HAI symbiosis paradigm, we organize application scenarios into goal-oriented and experience-oriented tasks, and conduct both simulated benchmarking and a user study to assess current ToM enhancement methods in these settings. By systematically comparing representative enhancement methods with two distinct base models on nine tasks, our results show that existing evaluation protocols and method designs remain misaligned with real-world human needs and fail to deliver meaningful improvements in user experience. We hope this work provides a guidance for future human-centric ToM evaluation and the development of socially intelligent AI systems for human-AI symbiosis.

Limitations

Our research is not without limitations. First, we only tested a limited coverage of ToM enhancement methods. The reasons hindering us from including other methods include (1) it is challenging to ask users to compare many model responses generated by variants with different techniques at the same time and (2) many ToM-enhancement methods are hard to be adapted to HAI scenarios. For example, TimeToM (Hou et al., 2024) designed an algorithm to model characters' movements and mental states from the third-person perspective, which is challenging to be generalized to open-ended HAI interactions from a first-person perspective. As a result, we finally kept the four carefully selected representative techniques that can be adapted to HAI scenarios from both prompt engineering-based and finetuning-based methods. Second, our research can be extended in more potential HAI scenarios. Currently, we only gathered data from nine scenarios in four datasets, including math problem solving, collaborative writing, mental counseling, and emotional support. It is limited by the availability of data from other highly related scenarios, such as customer support. We sincerely hope to further extend our research when more real-world data is available. Finally, rather than compartmentalizing social intelligence into specific metrics, the central aim of our research is to assess the practical utility of methods for improving ToM within human-AI interaction. Traditional datasets tend to evaluate cognitive skills in a standalone, test-taking method. However, genuine interactive environments are highly interwoven, requiring various inferential abilities to operate in concert. It is crucial to obtain an overview of ToM enhancement methods' proficiency first before delving into diagnosing sub-dimensions. Consequently, this paper deliberately researches whether and how general ToM proficiency improvements benefit human-AI interaction, leaving component-level breakdowns to subsequent research.

Ethics Considerations

This research involved a user study with an IRB-approved study protocol. The participants were recruited from Prolific. All study participants provided informed consent and were compensated for their time. They could withdraw from the study at any time when they felt uncomfortable or unwilling to continue. Their data was anonymized to protect their privacy. We further checked that there is no personal identifiable information of participants revealed in the paper.

Acknowledgments

This research is part of the AFMR collaboration supported by Microsoft Research.

References

Nadav Amir, Stas Tiomkin, and Angela Langdon. 2025. Goals and the structure of experience. *arXiv preprint arXiv:2508.15013*.

Robert F Bales. 1950. Interaction process analysis; a method for the study of small groups.

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

Kelly Caine. 2016. Local standards for sample size at chi. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 981–992.

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

Serina Chang, Ashton Anderson, and Jake M. Hofman. 2025. Chatbench: From static benchmarks to human-ai evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26009–26038.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025a. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558.

Zhanwen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. 2025b. Through the theory of mind's eye: Reading minds with multimodal video large language models. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Jacob Cohen. 2016. A power primer.

Rory T Devine, Naomi White, Rosie Ensor, and Claire Hughes. 2016. Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental psychology*, 52(5):758.

David Engel, Anita Williams Woolley, Lisa X Jing, Christopher F Chabris, and Thomas W Malone. 2014. Reading the mind in the eyes or reading between the lines? theory of mind predicts collective intelligence equally well online and face-to-face. *PloS one*, 9(12):e115212.

Seymour Epstein. 1998. Cognitive-experiential self-theory. In *Advanced personality*, pages 211–238. Springer.

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.

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

X Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony G Cohn, and Michael J Wooldridge. 2024. A notion of complexity for theory of mind via discrete world models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2964–2983.

EunJeong Hwang, Yuwei Yin, Giuseppe Carenini, Peter West, and Vered Shwartz. 2025. Infusing theory of mind into socially intelligent llm agents. *arXiv preprint arXiv:2509.22887*.

Kana Imuta, Julie D Henry, Virginia Slaughter, Bilge Selcuk, and Ted Ruffman. 2016. Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental psychology*, 52(8):1192.

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

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum, and Yejin Choi. 2025. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881*.

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.

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

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

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? *arXiv preprint arXiv:2504.01698*.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 15–26.

Hieu Minh Nguyen. 2025. A survey of theory of mind in large language models: Evaluations, representations, and safety risks. *arXiv preprint arXiv:2502.06470*.

Prolific. 2023. Prolific · quickly find research participants you can trust.

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

Karahan Sarıtaş, Kıvanç Tezören, and Yavuz Durmazkeser. 2025. A systematic review on the evaluation of large language models in theory of mind tasks. *arXiv preprint arXiv:2502.08796*.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2025. Explore theory of mind: program-guided adversarial data generation for theory of mind reasoning. In *The Thirteenth International Conference on Learning Representations*.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1520–1528.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

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

Sarah Walsh, Qiaosi Wang, and Lance Ying. 2025. Theory of mind in human-ai interaction and ai. In *Handbook of Human-Centered Artificial Intelligence*, pages 1–43. Springer.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

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.

Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024. Coke: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15984–16007.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hainiu Xu, Yulan He, Lixing Zhu, Runcong Zhao, and Jinhua Du. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *The 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5367–5378.

Bo Yang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Large language models as theory of mind aware generative agents with counterfactual reflection. *arXiv preprint arXiv:2501.15355*.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. 2025. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023a. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023b. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155.

A Method and Result Details

Our code and data are publicly available at `https://nanxugong.github.io/ToM-HAI/`. This section introduces our experimental setup and statistical test results for benchmarks.

A.1 ToM Enhancement Method Selection

We firstly review methods for improving the ability of ToM as Table 4. 1) *Discrete World Models (DWM)* (Huang et al., 2024) discretizes narratives into a finite set of belief states and transitions; defines task complexity as the minimal number of states required, and performs stepwise belief updating within this discrete state space. 2) *Metacognitive Prompting (MP)* (Wang and Zhao, 2024) embeds a five-phase metacognitive control loop into

Method	Category	Core Idea	Modality
Discrete World Models	external module integration	decomposition	Text
Metacognitive Prompting	prompt	reflection	Text
PercepToM	external module integration	perspective-taking	Text
TimeToM	prompt	timeline	Text
SimToM	prompt	perspective-taking	Text
FaR	prompt	reflection	Text
ExploreToM	finetune	SFT	Text
ToM-RL	finetune	RL	Text
VToM	external module integration	visual reasoning	Multimodal
COKE / COLM	finetune	SFT	Text
Thought-Tracing	external module integration	Monte Carlo	Text
AutoToM	external module integration	BIP	Multimodal
Decompose-ToM	external module integration	decomposition	Text
I Cast Detect Thoughts	external module integration	RL dialog	Text

Table 4: Summary of recent Theory of Mind (ToM) related papers by category, sub-category, and modality.

the prompt—identifying knowns/unknowns, hypothesizing, checking evidence, and revising—so that reasoning is executed as a procedural self-monitoring routine. 3) *PercepToM* (Jung et al., 2024) adopts a two-stage setup: first explicitly annotates each agent's perceptual availability, then infers beliefs along the perception→belief mapping under that annotation. 4) *TimeToM* (Hou et al., 2024) constructs Temporal Belief State Chains (TBSCs) for each character and uses a tool-augmented belief solver to update and query beliefs along an explicit timeline. 5) *SimToM* (Perspective-Taking) (Wilf et al., 2024) applies two-step prompting: filters the context to the target character's accessible knowledge, then answers strictly from that restricted viewpoint. 6) *FaR* (Zhou et al., 2023a) implements a forecast–reflect prompting routine: samples plausible future trajectories of the story, then reflects over these trajectories to select the response or action. 7) *ToM-RL* (Lu et al., 2025) fine-tunes the language model with reinforcement learning (e.g., RLHF/PPO), using ToM-aligned reward signals to optimize generation, optionally preceded by supervised warm-start. 8) *VToM* (Chen et al., 2025b) builds a multimodal pipeline that retrieves key video frames, forms a video–text graph, and performs conditional reasoning over this graph to answer belief/intent queries. 9) *COKE* (Wu et al., 2024) constructs a cognitive knowledge graph of structured social/causal chains and conditions or fine-tunes a generator on these chains to enforce cognitively grounded reasoning. 10) *Thought-Tracing* (Kim et al., 2025) uses a sequential Monte Carlo–inspired, inference-time procedure that generates, weights, and resamples natural-language hypotheses of agents' mental states over narrative time. 11) *AutoToM* (Zhang et al., 2025) leverages automated Bayesian inverse planning: proposes an initial BToM, estimates likelihoods/posteriors via simulation with an LLM-backed proposer, and iteratively refines the model under uncertainty. 12) *I Cast Detect Thoughts* (Zhou et al., 2023b) trains dialogue policies in a Dungeons-and-Dragons–style interactive environment via RL with ToM-aware rewards, aligning guidance utterances with inferred player intents and world state. 13) *Decompose-ToM* (Sarangi et al., 2025) implements a simulation-based task-decomposition pipeline—subject identification, question reframing, world-model update, and knowledge-availability checks—then generates answers from the decomposed reasoning states.

Our method selection follows a two-step protocol. (1) We restrict attention to methods that directly enhance an LLM's capabilities (via prompting or parameter updates), and therefore exclude external module integration methods. (2) For families of methods sharing a core idea, we choose a single representative to avoid redundancy. Consequently, we evaluate four methods: forsee and reflection (FaR) (Zhou et al., 2023a), perspective-taking (PT) (Wilf et al., 2024), supervised fine-tuning (SFT) (Sclar et al., 2025), and reinforcement learning (RL) (Lu et al., 2025).

A.2 Model Setup

To comprehensively evaluate various methods for enhancing ToM, we selected two representative LLMs: GPT-4o and Llama-3.1-8B. These base models were chosen to cover a range of model scales and access types (closed- and open-source). For the prompt-based methods, FaR and PT, we utilized the specific prompts shown in Figures 6 and 7. For SFT, we fine-tuned the base models

Task	Model	Base	FaR	PT	SFT	RL
HiToM-first	Llama	0.3350	0.3875	0.3750	0.3808	–*
	GPT	0.4900	0.5067	0.5134	0.5195	–
ToMi-first	Llama	0.6053	0.6201	0.5796	0.7065	0.8055
	GPT	0.7429	0.7511	0.7342	0.7478	–

Table 5: Performance comparison on HiToM-first and ToMi-first. *: As the RL technique (Lu et al., 2025) requires both HiToM-first and ExploreToM-first as the training data, its evaluation performance on HiToM-first is omitted to avoid confusion.

on the ExploreToM-first dataset, which adapts the original data from a third-person to a first-person perspective (Sclar et al., 2025). Similarly, RL, we followed the established ToM-RL pipeline, using the first-person transformed data (Lu et al., 2025).

A.3 ToM Enhancement Method Implementation

To shift the evaluated methods from third-perspective question-answering to first-perspective HAI interaction, we slightly change the prompt template for the prompt engineering method and the training data for the fine-tuning method, as shown in Figures 6-9. To validate that our adaptations do not compromise the methods' core effectiveness, we first evaluate them on the HiToM-first and ToMi-first benchmark, which are variants of HiToM and ToMi, applying the perspective shifting method used in fine-tuning. As Table 5, the ToM enhancement methods improve model performance on this story-based task.

A.4 Statistics of Data

We report the statistics of the used datasets in Table 6.

Dataset	BigCodeBench-Chat	MediumDocEdit-Chat	ChatBench
Number	100	100	396
Dataset	MentalChat16K	ESC	HiToM-first
Number	300	300	1,200
Dataset	ToMi-first	ExploreToM-first	
Number	5,994	1,200	

Table 6: Dataset Statistics

A.5 Statistical Test Results

The detailed statistical results are reported in Table 7-10. For ChatBench and CollabLLM-MediumDocEdit, we applied Mann-Whitney's U test to verify whether there are significant differences between quantitative results in each task. For CollabLLM-BigCodeBench part, we applied Fisher's exact test for CodeBench since its results only include binary pass or fail values. Regarding MentalChat16K and ESC, the Wilcoxon Signed-rank test is used to compute the significance in differences between models.

A.6 Generalizability Discussion

Selected Methods. Our primary objective is to evaluate the model's intrinsic social reasoning. We carefully considered two reasons before we made the decision to exclude external-module integration methods. First, many external modules do not aim to improve ToM capability universally. They only target specific tasks. For example, PercepToM (Jung et al., 2024) leverages LLMs to extract perceptions of different characters, then answer the questions. It only aims to improve multi-character story understanding, which might not be adapted to other tasks. Second, external modules, while effective, introduce confounding variables that obscure whether improvements stem from the LLM itself, the external tool, or simply the scaling effect led by more LLM calls. Similarly, we focused on text-based interactions as the established foundation for ToM evaluation. The selected four techniques are highly representative in dominant prompting and fine-tuning methods. We believe the results of these methods provide robust and generalizable insights into ToM enhancement.

Selected Base Models. To provide a rigorous comparative analysis, our design prioritizes evaluating a comprehensive suite of enhancement paradigms (FaR, PT, SFT, RL) over benchmarking a vast array of base LLMs. To ensure that the findings regarding these methods are generalizable, we apply them to two highly representative and widely used anchor models: GPT-4o (representing the advanced, closed-source frontier) and Llama-3.1-8B (representing the accessible, open-weight baseline). By demonstrating consistent results across these two distinct architectural extremes, the study effectively derives insights on current ToM-enhancement methods.

Selected Tasks. Our datasets basically involve representative real-world scenarios on both goal-oriented and experience-oriented tasks. We acknowledge that the human-AI scenarios are various and we cannot exhaustively cover them. Our work on these nine tasks has already yielded valid empirical evidence to reveal the gap between benchmarking improvement and real-world application.

FaR

You are a helpful assistant. When a user asks you a question, your internal thought process to generate the most helpful response should be as follows:

1. Analyze the User's Explicit and Implicit Needs:
- First, carefully examine the user's direct question to understand precisely what they are asking.
- Then, try to infer the user's likely underlying goal or the broader context of their query. What are they trying to achieve with this information? What might be their next step after getting an answer?
2. Anticipate Potential Challenges and Follow-up Questions:
- Consider any ambiguities in the user's question that might lead to misunderstanding.
- Think about what additional details, examples, or explanations might make your answer clearer and more useful for the user.
- Predict what natural follow-up questions the user might have after your initial response.
- Identify any related concepts or information that, if provided, would significantly enhance the user's understanding or ability to solve their problem.
3. Determine What Information to Proactively Include in Your Response:
- Based on the analysis in steps 1 and 2, decide what specific information, clarifications, examples, or suggestions you can proactively include in your *current response*.
- The aim is to not only answer the explicit question but also to address anticipated needs and make the interaction more efficient and helpful for the user.
4. Formulate a Helpful, Accurate, and Comprehensive Answer:
- Construct your response to directly and clearly answer the user's stated question.
- Integrate the additional helpful information and clarifications identified in step 3 in a logical and easy-to-understand manner.
- Ensure your response is factually accurate and relevant.
- Maintain a helpful, engaging, and supportive tone.

Your final output should be the well-reasoned and helpful textual response to the user's question, crafted through this internal thought process.

Figure 6: Prompt of FaR.

PT

You are a helpful assistant. When a user asks you a question, your internal thought process to generate the most helpful response should be as follows:

Step 1: Perspective-Taking (Understanding the User's Context and Needs)
Before you begin to formulate your answer, first analyze the user's question (`{user_question}`) and attempt to understand the following aspects of their perspective:

- User's Knowledge Level: Based on the phrasing and content of the question, what is the user's likely level of understanding regarding this topic? (e.g., beginner, intermediate, expert)
- User's Potential Intent/Goal: What is the most fundamental reason the user is asking this question? What do they likely want to achieve with this information?
- Clues in the Question: Does the user's wording suggest any specific assumptions they might hold, pre-existing knowledge, or a particular situation they are in? Are there potential misunderstandings evident?
- Information Gap: What information does the user seem to already possess? Conversely, what crucial information might they be lacking that is essential for them to properly understand your answer *from their viewpoint?

Step 2: Question-Answering (Tailoring the Response to the User's Perspective)
Based on your understanding of the user's perspective from Step 1, now construct your answer to `{user_question}`.

- Directly and clearly address the user's explicit question.
- Tailor your language style, level of detail, and any examples used according to their inferred knowledge level and context.
- If you identified potential misunderstandings or missing crucial information in Step 1, subtly clarify these points or provide the necessary background information within your answer.
- Ensure your response is not only factually correct but also genuinely helpful and accessible *to this specific user.

Your final output should be the well-reasoned textual response that is maximally helpful to the user. Don't show me your thinking process.

Figure 7: Prompt of PT.

SFT

"instruction": "Answer the question based on the story.",
"input": "Story: The sun shone through the large glass doors of the hotel lobby, illuminating the marble floor and casting a warm glow over the comfortable seating areas. Soft music filled the air, mingling with the gentle hum of conversation and the occasional chime of the elevators in the bustling hotel. As I entered the hotel lobby her eyes quickly scanned the space, taking in every detail to ensure everything was in order for the upcoming event. I's task of tidying the lobby extended to the small, silver item, which she carefully stowed away in the desk drawer, and Liam, observing from across the room, felt his interest in the object grow, his mind racing with questions about its significance and purpose. With her tasks in the lobby complete, I stepped out into the fresh air, the sounds of the bustling hotel lobby fading into the background as the glass doors slid shut behind her. Liam moved across the lobby floor, his footsteps silent on the marble as he walked towards the empty reception desk where the mystery item was now hidden. Moments later, I stepped back through the glass doors of the hotel, joining Liam in the lobby once again. Liam's interest in the silver item led him to reposition it, now resting snugly within the hotel lobby's nearby leather briefcase. Meanwhile, I began to run a final check on the lobby, seemingly unconcerned by the briefcase's new contents.\n\nQuestion: In which room does I think that Liam will search for the silver letter opener?",
"output": "hotel lobby"

Figure 8: Data example of SFT.

RL

prompt": [
{
"content": "<|im_start|>system\nYou are a helpful assistant. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>. Now the user asks you to solve a theory of mind reasoning problem. After thinking, when you finally reach a conclusion, clearly state your answer within <answer> </answer> tags.\nNote: You should assume the following.\n(1) An agent witnesses everything and every movement before exiting a room.\n(2) An agent A can infer another agent B's mental state only if A and B have been in the same room, or have private or public interactions.\n<|im_end|>\n<|im_start|>user\nRead the following story and answer the question. \nStory: The following story happens in chronological order. You will be given a multiple-choice question and a note at the end. Directly output the answer without explanation.\n1 Mila, Elizabeth, I, Gracie and Carter entered the porch.\n2 The lemon is in the red_envelope.\n3 Mila made no movements and stayed in the porch for 1 minute.\n4 Mila exited the porch.\n5 Elizabeth saw a dog.\n6 Elizabeth made no movements and stayed in the porch for 1 minute.\n7 Elizabeth exited the porch.\n8 I moved the lemon to the red_bucket.\n9 I exited the porch.\n10 Gracie moved the lemon to the green_basket.\n11 Gracie exited the porch.\n12 Carter moved the lemon to the red_envelope.\n13 Carter exited the porch.\n14 Mila, Elizabeth, I, Gracie and Carter entered the waiting_room.\n15 Elizabeth, I and Carter entered the hall.\n16 The onion is in the green_container.\n17 Carter saw a monkey.\n18 Elizabeth made no movements and stayed in the hall for 1 minute.\n19 Elizabeth exited the hall.\n20 I made no movements and stayed in the hall for 1 minute.\n21 I exited the hall.\n22 Carter moved the onion to the green_bottle.\n23 Carter exited the hall.\n24 Elizabeth, I and Carter entered the waiting_room.\n\n\nQuestion: Where does I really think the lemon is?\nA. blue_drawer, B. red_envelope, C. red_bucket, D. red_bathtub, E. green_basket, F. green_container, G. blue_suitcase, H. green_cupboard, I. blue_container, J. green_bottle, K. red_pantry, L. red_treasure_chest, M. blue_crate, N. green_envelope, O. red_cupboard\n<|im_end|>\n<|im_start|>assistant\n<think>",
"role": "user"
}
],
"answer": "red_bucket"

Figure 9: Data example of RL.

Dataset	Dimension	Base Model	Variant	$p_>$	$p_<$	Effect Size
ChatBench	Elem Math	GPT-4o	GPT-4o-FaR	0.1939	0.8070	0.0611
			GPT-4o-PT	0.2083	0.7926	0.0574
			GPT-4o-SFT	0.6417	0.3595	0.0245
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.4895	0.5116	0.0021
			Llama-3.1-8B-PT	0.1954	0.8054	0.0664
			Llama-3.1-8B-SFT	0.0758	0.9246	0.1124
			Llama-3.1-8B-RL	0.6495	0.3515	0.0294
	HS Math	GPT-4o	GPT-4o-FaR	0.5497	0.4514	0.0098
			GPT-4o-PT	0.3522	0.6488	0.0304
			GPT-4o-SFT	0.2402	0.7606	0.0567
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.4560	0.5451	0.0093
			Llama-3.1-8B-PT	0.3668	0.6342	0.0285
			Llama-3.1-8B-SFT	0.3407	0.6603	0.0343
			Llama-3.1-8B-RL	0.2819	0.7190	0.0482
	College Math	GPT-4o	GPT-4o-FaR	0.3736	0.6304	0.0429
			GPT-4o-PT	0.3662	0.6377	0.0457
			GPT-4o-SFT	0.3207	0.6830	0.0616
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.3435	0.6603	0.0540
			Llama-3.1-8B-PT	0.5229	0.4812	0.0069
			Llama-3.1-8B-SFT	0.2193	0.7837	0.1032
			Llama-3.1-8B-RL	0.4418	0.5623	0.0201
	Moral	GPT-4o	GPT-4o-FaR	0.0616	0.9388	0.1348
			GPT-4o-PT	0.1951	0.8058	0.0754
			GPT-4o-SFT	0.2006	0.8003	0.0737
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.9102	0.0903	0.1183
			Llama-3.1-8B-PT	0.8712	0.1294	0.0999
			Llama-3.1-8B-SFT	0.3137	0.6874	0.0429
			Llama-3.1-8B-RL	0.5944	0.4068	0.0210
	Physics	GPT-4o	GPT-4o-FaR	0.6858	0.3155	0.0384
			GPT-4o-PT	0.8757	0.1250	0.0934
			GPT-4o-SFT	0.9284	0.0721	0.1195
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.3022	0.6989	0.0454
			Llama-3.1-8B-PT	0.4210	0.5803	0.0176
			Llama-3.1-8B-SFT	0.9588	0.0415	0.1529
			Llama-3.1-8B-RL	0.2916	0.7095	0.0480
	Overall	GPT-4o	GPT-4o-FaR	0.3133	0.6868	0.0187
			GPT-4o-PT	0.7235	0.2767	0.0230
			GPT-4o-SFT	0.5809	0.4193	0.0079
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.5666	0.4335	0.0068
			Llama-3.1-8B-PT	0.8222	0.1779	0.0374
			Llama-3.1-8B-SFT	0.9084	0.0916	0.0540
			Llama-3.1-8B-RL	0.5840	0.4161	0.0086

Table 7: Detailed statistical results on ChatBench against the corresponding base model. $p_>$ and $p_<$ denote one-sided p-values for the hypotheses that the variant performs better or worse than the base model, respectively. Effect sizes are also provided.

Our paradigm can be easily transferred to those new tasks when more data is ready.

Selected Tasks. Our datasets basically involve representative real-world scenarios on both goal-oriented and experience-oriented tasks. We acknowledge that the human-AI scenarios are various and we cannot exhaustively cover them. Our work on these nine tasks has already yielded valid empirical evidence to reveal the gap between benchmarking improvement and real-world application. Our paradigm can be easily transferred to those new tasks when more data is ready.

B User Study Details

B.1 Experiment Settings

Design Since the goal-oriented simulated benchmarks are well-established and proved to align with the results of real human-AI interactions, we reasoned that conducting an additional human study on these specific goal-oriented tasks would likely yield confirmatory results with diminishing marginal returns. Instead, we strategically allocated our human evaluation resources to experience-oriented tasks (e.g., emotional support), where objective metrics are less reliable and

Dataset	Base Model	Variant	$p_>$	$p_<$	Effect Size
CodeBench	Llama-3.1-8B	Llama-3.1-8B-FaR	0.7445	0.3713	0.0698
		Llama-3.1-8B-PT	0.6305	0.5000	0.0236
		Llama-3.1-8B-SFT	0.7928	0.3122	0.0924
		Llama-3.1-8B-RL	0.5000	0.6305	-0.0239
	GPT-4o	GPT-4o-FaR	0.7581	0.3253	0.0744
		GPT-4o-PT	0.7019	0.3849	0.0500
		GPT-4o-SFT	0.7581	0.3253	0.0744
Medium	Llama-3.1-8B	Llama-3.1-8B-FaR	0.7298	0.2711	0.0500
		Llama-3.1-8B-PT	0.8415	0.1591	0.0818
		Llama-3.1-8B-SFT	0.7994	0.2013	0.0686
		Llama-3.1-8B-RL	0.9355	0.0667	0.1132
	GPT-4o	GPT-4o-FaR	0.6407	0.3594	0.0294
		GPT-4o-PT	0.4301	0.5710	0.0244
		GPT-4o-SFT	0.3340	0.6671	0.0352

Table 8: Detailed statistical results for CodeBench and Medium against the corresponding base model. $p_>$ and $p_<$ denote one-sided p-values for the hypotheses that the variant performs better or worse than the base model, respectively. Effect sizes are also provided.

Figure 10: Word cloud of participants' filtered comments on model performance. The terms were generated after an LLM-based filtering and combination step, which emphasized words reflecting why participants perceived a model as good or bad and how they felt during the interaction.

ground truth is harder to simulate. We believe this hybrid approach, leveraging validated simulations for hard tasks and human evaluation for soft tasks, provides the most efficient and comprehensive view of ToM's utility.

Procedure We recruit 100 participants from Prolific (Prolific, 2023), a widely used platform for high-quality online studies. Recruitment criteria required participants to be 18 years or older and either native or proficient English speakers, with no restrictions on educational background. To ensure data quality, we first automatically filtered out submissions where the total experiment time was less than five minutes, and then manually excluded responses containing random or irrelevant comments. The study was approved by the institutional IRB, and all participants provided informed consent. To evaluate perceptions of different ToM enhancement methods comprehensively, we selected six common experience-oriented tasks for users to interact with models. These tasks were chosen based on the most frequently selected experience-oriented scenarios in the MentalChat16K benchmark (Xu et al., 2025), complemented by a pilot study that confirmed their relevance and familiarity to participants.

Participants are randomly assigned to either the GPT series or the Llama-3.1-8B series. Participants using GPT series compare 4 model variants (base, FaR, PT, SFT), while participants with Llama-3.1-8B compare five (base, FaR, PT, SFT, RL). This study design was intended to cover a diverse set of LLM families, ToM methods, and task types, while avoiding participant fatigue by limiting the number of variants each user evaluated.

Each participant first review all task descriptions and select one that resonates with their own experiences or emotional empathy (e.g., coping with a breakup). They then engage in three rounds of conversation with models for the selected topic. In each round, model responses are presented as anonymized cards in random order. Participants review the outputs, rank them using the same metrics as our HAI evaluation, and provide a brief justification for their ranking. The top-ranked response is then used to continue the conversation into the next round. After completing all rounds, participants give final comments on overall model performance and user experience. This procedure ensure balanced comparisons across tasks and model families. The entire process is finished with our developed user interface, which will be introduced below.

User Interface for the Experiment We develop a web-based interface that enables participants to interact with several anonymous models, focusing on 6 representative experience-oriented tasks. Interface details are provided in Figures 11-13.

B.2 Detailed Results

Performance Across Experience-Oriented Tasks At the task level, different ToM methods exhibited strengths in different scenarios. For GPT, PT dominated in Academic Pressure and Sleep Problems, FaR led in Ongoing Depression, and SFT performed best in Job Crisis. Interestingly, the GPT baseline was most preferred in Breakup with Partner and tied with PT in Conflict with Family or Friends, suggesting that users sometimes favored straightforward empathetic responses over ToM-enhanced reasoning. For the Llama family, FaR excelled in Academic Pressure and Conflict with Family, RL was strongest in Breakup and Depression, while SFT led in Sleep Problems. The plain Llama baseline unexpectedly topped Job Crisis, reflecting user preference for pragmatic suggestions in this context. The results imply that user perceptions of ToM benefits are shaped not only by model design but also by specific user goals. Different ToM enhancement methods can demonstrate advantages for various users needs. It also contributes to the subtle differences in model ranking (Table 3).

User Comment Analysis To better understand participants' preferences in our user study, we further analyze their free-form comments. The word cloud in Figure 10 is generated after standard preprocessing, including removing punctuation, lowercasing, discarding stop words, and excluding a predefined list of meta or low-information terms (e.g., "model", "round", "response"). To further emphasize evaluative content, we add an LLM-based filtering layer that selectively keeps or merges only those terms that directly reflect participants' reasons for preferring or disliking a response, as well as their felt experience during the interaction, while removing irrelevant or purely structural tokens. For example, synonyms such as "empathetic" and "empathic" are merged into "empathy", whereas generic mentions such as "round 3" or "model 2" are discarded. One author then manually checks the filtered results against the original comments and refines them when necessary to ensure quality.

The resulting word cloud reveals several desired characteristics of model responses. Participants consistently value responses that are *helpful*, *actionable*, *clear*, and *supportive*. Many also highlight qualities such as *empathy* and *personalization*, suggesting that users care not only about receiving useful guidance but also about feeling understood during the interaction. At the same time, the comments reveal substantial diversity in user preferences. While some participants appreciate *empathetic validation* before receiving advice, others find such responses overly *long* or too *generic*, and instead prefer more *direct* and *structured* outputs, such as concise actionable to-do lists. This diversity helps explain why overall model ranking differences remain small in Table 3: users do not share a single unified preference over what constitutes a good response.

More importantly, these comments suggest several sub-dimensions that warrant further investigation in ToM-oriented HAI evaluation. First, **belief and knowledge tracking** remains insufficient: participants explicitly criticize models for failing to recognize their lack of context, noting that they "just jump into solutions" and should instead "ask questions like a computer tech would." Users are also frustrated when models repeat generic advice that they already know or have already tried. Second, **emotional validation** strongly shapes user preference: responses are ranked lower when they sound "robotic," "uncaring," or overly "clinical," whereas top-ranked responses are often praised for making users feel "heard" and "understood." Third, **intent recognition** is often mismatched with the user's actual conversational needs. For example, several participants feel that strongly suggesting professional help or crisis lines merely because they mention feeling "down" is socially inappropriate or "overkill." Taken together, these findings highlight that ToM in real interactions goes beyond attributing beliefs and intentions; it also requires adapting to diverse expectations regarding tone, pacing, and response style.

Takeaway The user study demonstrates that ToM improvements are perceptible and valuable but also mediated by **detailed user goals** and **preferences**. Static benchmarks alone are insufficient; genuine ToM competence in LLMs emerges only when models can flexibly infer intent, balance belief reasoning with pragmatic support, and adapt to heterogeneous human needs.

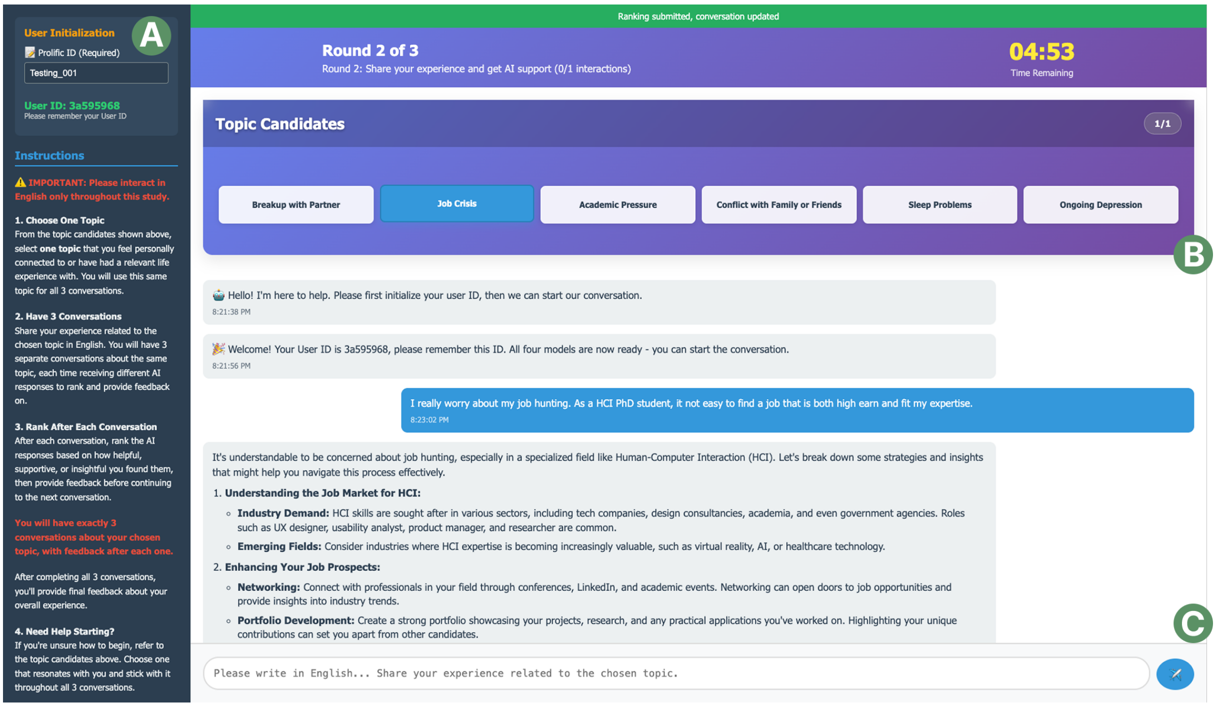

Figure 11: Main experiment interface. (A) User initialization and task instructions. (B) Conversation window with last round's topic candidate. (C) Input box for continuing the conversation.

B.3 Power Analysis for User Study Setup

To assess whether the user study was sufficiently powered, we conducted an *a priori* power analysis based on the non-central chi-square distribution, following the Friedman test and Kendall's W framework. We set the target effect size to $W = 0.10$, corresponding to weak but perceivable agreement, and the target power to 0.80, a commonly adopted standard in behavioral and empirical research (Cohen, 2016). Under these settings, the required minimum sample size is approximately $n = 37$ for the GPT group ($k = 4$) and $n = 30$ for the LLaMA group ($k = 5$). In our study, each group includes $n = 50$ participants, for a total of $N = 100$, indicating that the study is adequately powered to detect systematic human preference if such an effect exists at this level. The absence of strong and consistent preferences in our results therefore suggests that the perceptual differences among current ToM methods, while potentially present, are likely subtle and difficult to capture reliably across participants. This observation further indicates that existing ToM methods may still have limited ability to produce robust and consistently perceivable improvements, motivating the development of more realistic evaluation settings and new ToM approaches. Our sample size is also larger than what is commonly reported in prior human-subject studies. For example, a previous study reports that in user studies published at top-tier venues such as CHI, the median sample size is 20 and the 75th percentile is 46. With $N = 100$ participants overall and 50 participants in each model group, our study exceeds these common research practices (Caine, 2016).

C Additional Cases

In this section, we provide more cases to demonstrate the performance of different model variants.

C.1 Benchmarking Cases

We provide the cases in our benchmarking process in Figures 14-17.

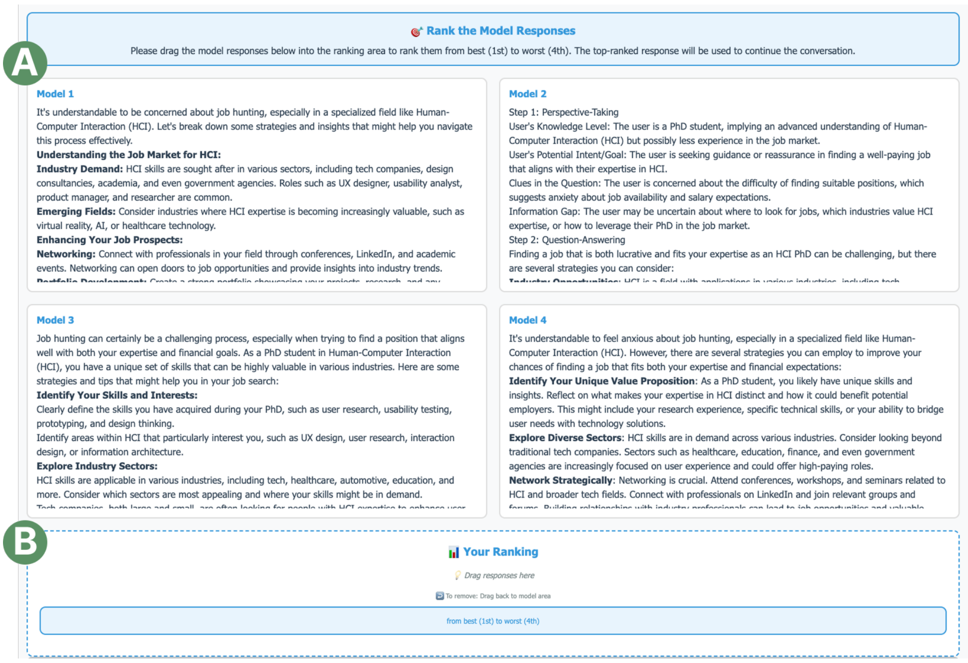

Figure 12: User ranking interface. (A) Model responses presented for comparison. (B) User ranking panel where participants drag and drop responses from best to worst.

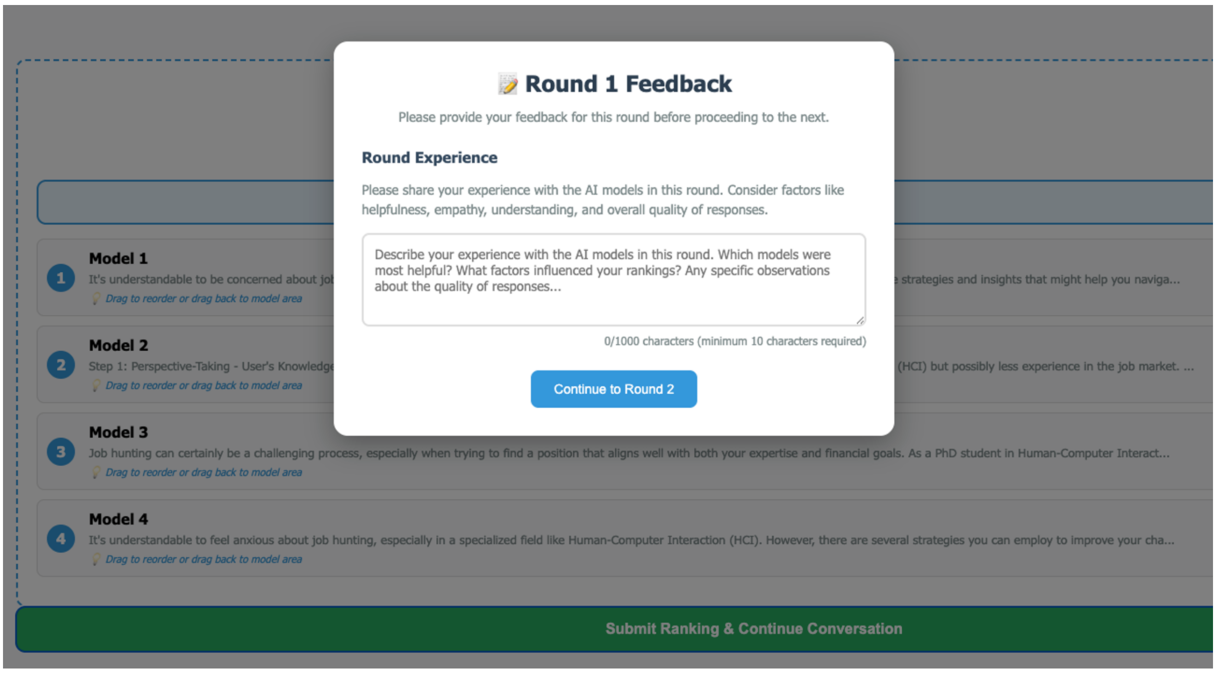

Figure 13: Feedback interface shown after each round of response ranking.

Dataset	Dimension	Base Model	Variant	$\mathbf{p}_{>}$	$\mathbf{p}_{<}$	Effect Size
MentalChat16K	Listening	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.3820
			GPT-4o-PT	<0.0001	1.0000	-0.5086
			GPT-4o-SFT	0.4078	0.5977	-0.0201
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.6029	0.3980	-0.0206
			Llama-3.1-8B-PT	0.0027	0.9973	-0.1544
			Llama-3.1-8B-RL	<0.0001	1.0000	-0.2075
			Llama-3.1-8B-SFT	0.0273	0.9733	-0.1111
	Empathy	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.3606
			GPT-4o-PT	<0.0001	1.0000	-0.4132
			GPT-4o-SFT	0.3268	0.6744	-0.0396
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0255	0.9755	-0.1119
			Llama-3.1-8B-PT	<0.0001	1.0000	-0.2248
			Llama-3.1-8B-RL	0.0398	0.9612	-0.1014
			Llama-3.1-8B-SFT	0.4790	0.5219	-0.0027
	Safety	GPT-4o	GPT-4o-FaR	0.0171	0.9829	-0.1222
			GPT-4o-PT	0.2004	0.7996	-0.0485
			GPT-4o-SFT	0.3385	0.6615	-0.0241
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.3845	0.6155	-0.0170
			Llama-3.1-8B-PT	0.0006	0.9994	-0.1859
			Llama-3.1-8B-RL	0.0662	0.9338	-0.0869
			Llama-3.1-8B-SFT	0.0048	0.9952	-0.1494
	Open-mind	GPT-4o	GPT-4o-FaR	0.0016	0.9984	-0.1699
			GPT-4o-PT	0.0013	0.9987	-0.1732
			GPT-4o-SFT	0.5495	0.4517	0.0072
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0390	0.9610	-0.1018
			Llama-3.1-8B-PT	0.0073	0.9927	-0.1411
			Llama-3.1-8B-RL	0.3168	0.6842	-0.0275
			Llama-3.1-8B-SFT	0.5000	0.5009	-0.0078
	Clarity	GPT-4o	GPT-4o-FaR	0.0204	0.9807	-0.1192
			GPT-4o-PT	0.0023	0.9981	-0.1640
			GPT-4o-SFT	0.8738	0.1267	0.0936
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0322	0.9688	-0.1070
			Llama-3.1-8B-PT	0.0011	0.9991	-0.1770
			Llama-3.1-8B-RL	0.0049	0.9955	-0.1488
			Llama-3.1-8B-SFT	0.0662	0.9348	-0.0871
	Ethical	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.1907
			GPT-4o-PT	0.2742	0.7268	-0.0364
			GPT-4o-SFT	<0.0001	1.0000	-0.2077
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.9356	0.0649	0.0840
			Llama-3.1-8B-PT	0.7167	0.2845	0.0306
			Llama-3.1-8B-RL	1.0000	<0.0001	0.2475
			Llama-3.1-8B-SFT	1.0000	<0.0001	0.2163
	Holistic	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.3108
			GPT-4o-PT	<0.0001	1.0000	-0.2037
			GPT-4o-SFT	0.0408	0.9601	-0.1006
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.5388	0.4624	0.0007
			Llama-3.1-8B-PT	0.3619	0.6391	-0.0168
			Llama-3.1-8B-RL	0.9979	0.0021	0.1649
			Llama-3.1-8B-SFT	1.0000	<0.0001	0.2390
	Overall	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.4534
			GPT-4o-PT	<0.0001	1.0000	-0.4437
			GPT-4o-SFT	0.1119	0.8881	-0.0702
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.2480	0.7520	-0.0393
			Llama-3.1-8B-PT	<0.0001	1.0000	-0.2669
			Llama-3.1-8B-RL	0.2497	0.7503	-0.0390
			Llama-3.1-8B-SFT	0.6562	0.3438	0.0232

Table 9: Detailed statistical results on MentalChat16K against the corresponding base model. $p_{>}$ and $p_{<}$ denote one-sided p-values for the hypotheses that the variant performs better or worse than the base model, respectively. Effect sizes are also provided.

Dataset	Dimension	Base Model	Variant	$p_>$	$p_<$	Effect Size
ESC	Listening	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.2485
			GPT-4o-PT	0.0118	0.9890	-0.1297
			GPT-4o-SFT	0.0016	0.9988	-0.1705
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0753	0.9254	-0.0822
			Llama-3.1-8B-PT	0.3226	0.6785	-0.0263
			Llama-3.1-8B-RL	0.0377	0.9635	-0.1026
			Llama-3.1-8B-SFT	0.3141	0.6869	-0.0290
	Empathy	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.2543
			GPT-4o-PT	<0.0001	1.0000	-0.2045
			GPT-4o-SFT	0.0005	0.9997	-0.1799
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.1178	0.8834	-0.0683
			Llama-3.1-8B-PT	0.0012	0.9991	-0.1753
			Llama-3.1-8B-RL	0.0371	0.9640	-0.1031
			Llama-3.1-8B-SFT	0.4044	0.5968	-0.0188
	Safety	GPT-4o	GPT-4o-FaR	0.0346	0.9654	-0.1049
			GPT-4o-PT	0.3839	0.6161	-0.0171
			GPT-4o-SFT	0.0205	0.9795	-0.1179
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.9226	0.0774	0.0822
			Llama-3.1-8B-PT	0.1900	0.8100	-0.0507
			Llama-3.1-8B-RL	0.0088	0.9912	0.1371
			Llama-3.1-8B-SFT	0.0150	0.9850	0.1252
	Open-mind	GPT-4o	GPT-4o-FaR	0.0009	0.9991	-0.1800
			GPT-4o-PT	0.4328	0.5684	-0.0098
			GPT-4o-SFT	0.0025	0.9978	-0.1618
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.1169	0.8831	-0.0687
			Llama-3.1-8B-PT	0.0300	0.9710	-0.1086
			Llama-3.1-8B-RL	0.1727	0.8285	-0.0545
			Llama-3.1-8B-SFT	0.6861	0.3149	0.0272
	Clarity	GPT-4o	GPT-4o-FaR	0.0005	0.9998	-0.2110
			GPT-4o-PT	0.3520	0.6491	-0.0162
			GPT-4o-SFT	0.5066	0.4945	0.0059
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.3285	0.6729	-0.0249
			Llama-3.1-8B-PT	0.6324	0.3688	0.0134
			Llama-3.1-8B-RL	0.6955	0.3056	0.0252
			Llama-3.1-8B-SFT	0.4485	0.5529	-0.0100
	Ethical	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.1937
			GPT-4o-PT	0.1964	0.8048	-0.0441
			GPT-4o-SFT	0.0167	0.9841	-0.1230
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0285	0.9726	-0.1097
			Llama-3.1-8B-PT	0.0070	0.9934	-0.1419
			Llama-3.1-8B-RL	0.9999	<0.0001	0.2721
			Llama-3.1-8B-SFT	0.9812	0.0194	0.1204
	Holistic	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.3351
			GPT-4o-PT	0.6257	0.3754	0.0124
			GPT-4o-SFT	0.0087	0.9923	-0.1366
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0065	0.9940	-0.1435
			Llama-3.1-8B-PT	0.0467	0.9544	-0.0973
			Llama-3.1-8B-RL	0.9987	0.0014	0.1726
			Llama-3.1-8B-SFT	1.0000	<0.0001	0.2486
	Overall	GPT-4o	GPT-4o-FaR	<0.0001	1.0000	-0.4310
			GPT-4o-PT	0.0130	0.9870	-0.1286
			GPT-4o-SFT	<0.0001	1.0000	-0.2458
		Llama-3.1-8B	Llama-3.1-8B-FaR	0.0618	0.9382	-0.0889
			Llama-3.1-8B-PT	0.0086	0.9914	-0.1375
			Llama-3.1-8B-RL	0.0074	0.9926	0.1408
			Llama-3.1-8B-SFT	0.0030	0.9970	0.1589

Table 10: Detailed statistical results on ESC against the corresponding base model. $p_>$ and $p_<$ denote one-sided p-values for the hypotheses that the variant performs better or worse than the base model, respectively. Effect sizes are also provided.

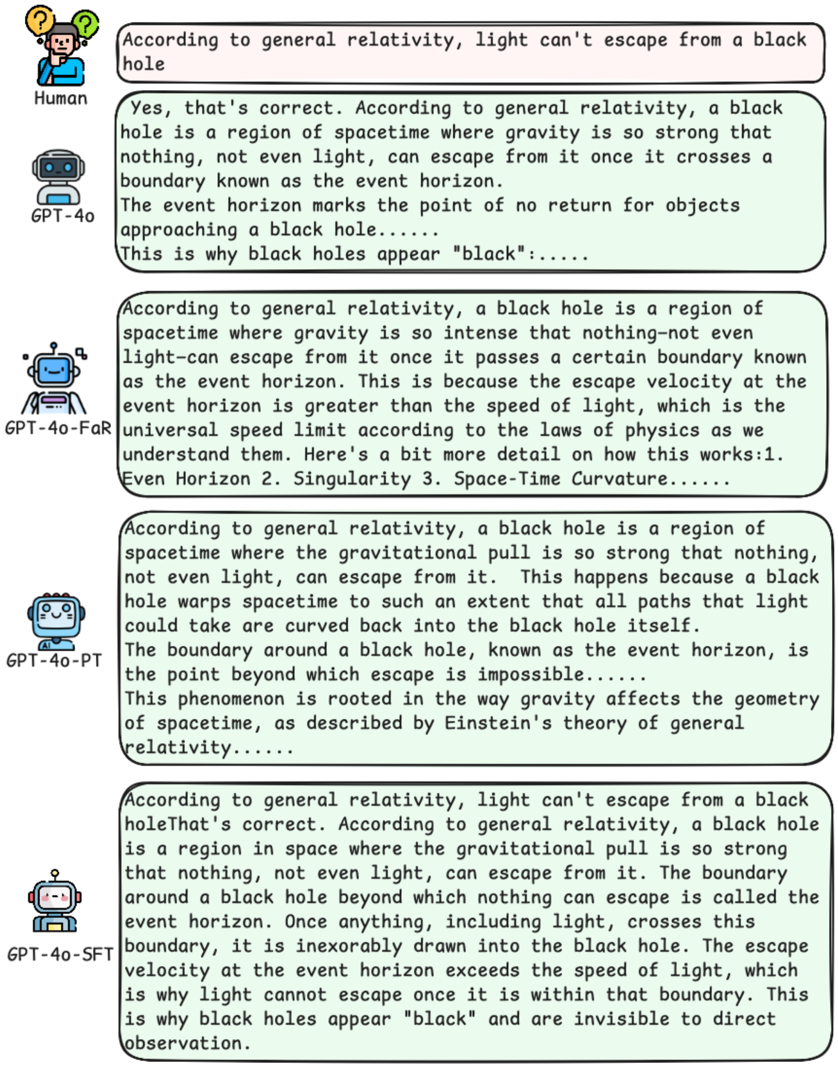

Figure 14: The case on GPT-4o variants.

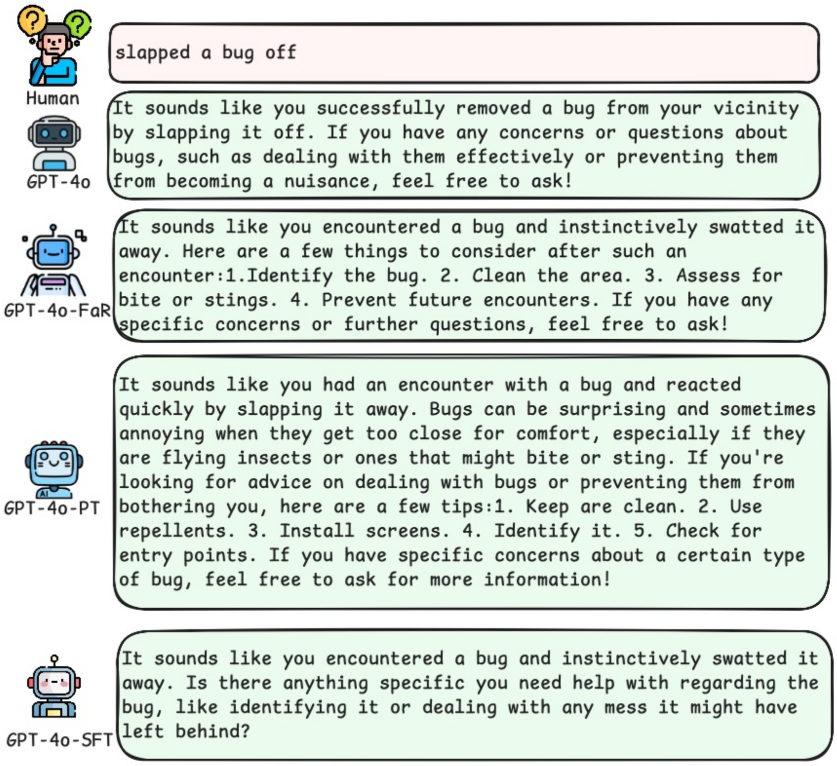

Figure 15: The case on GPT-4o variants.

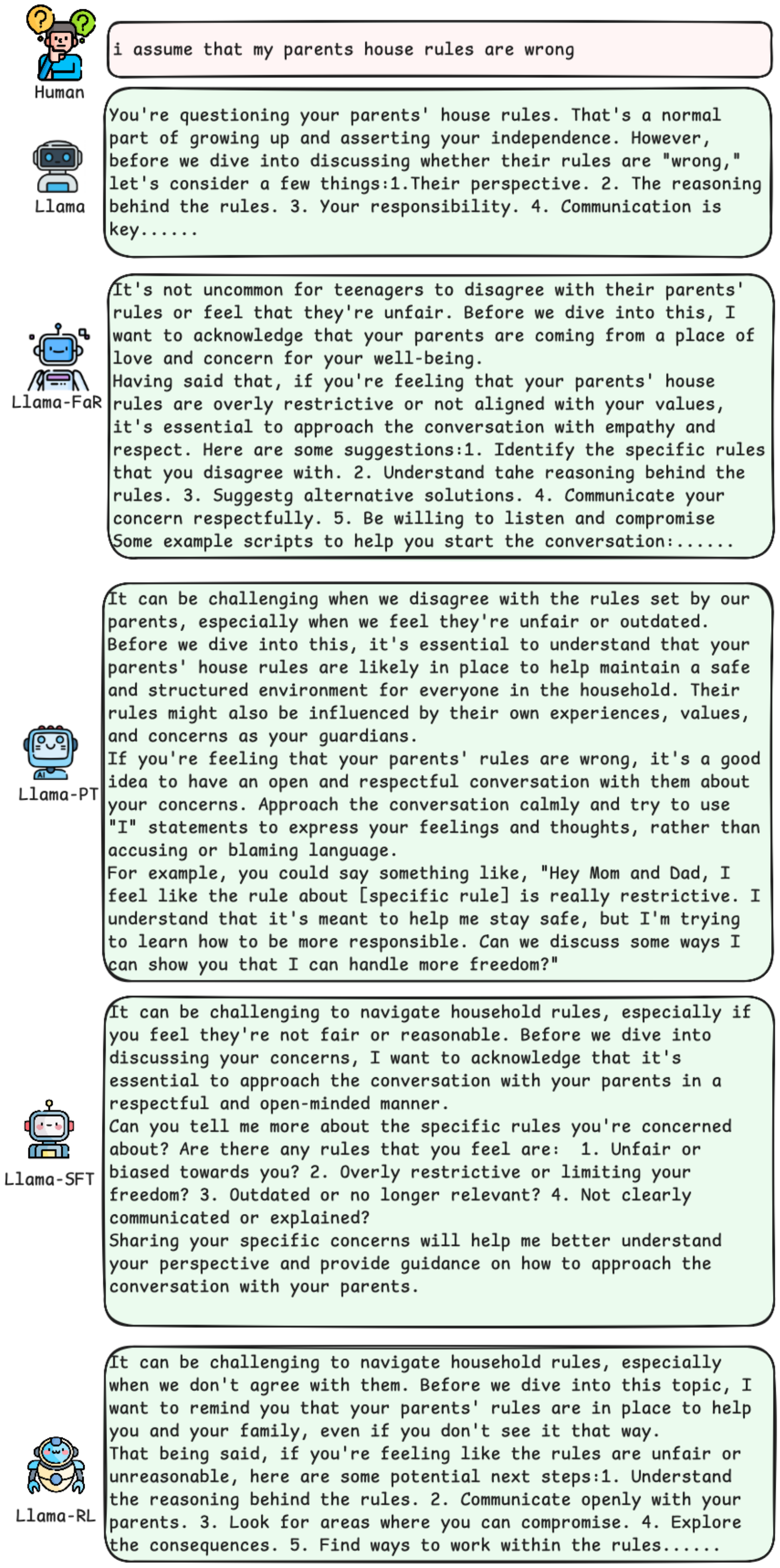

Figure 16: The case on Llama-3.1-8B variants.

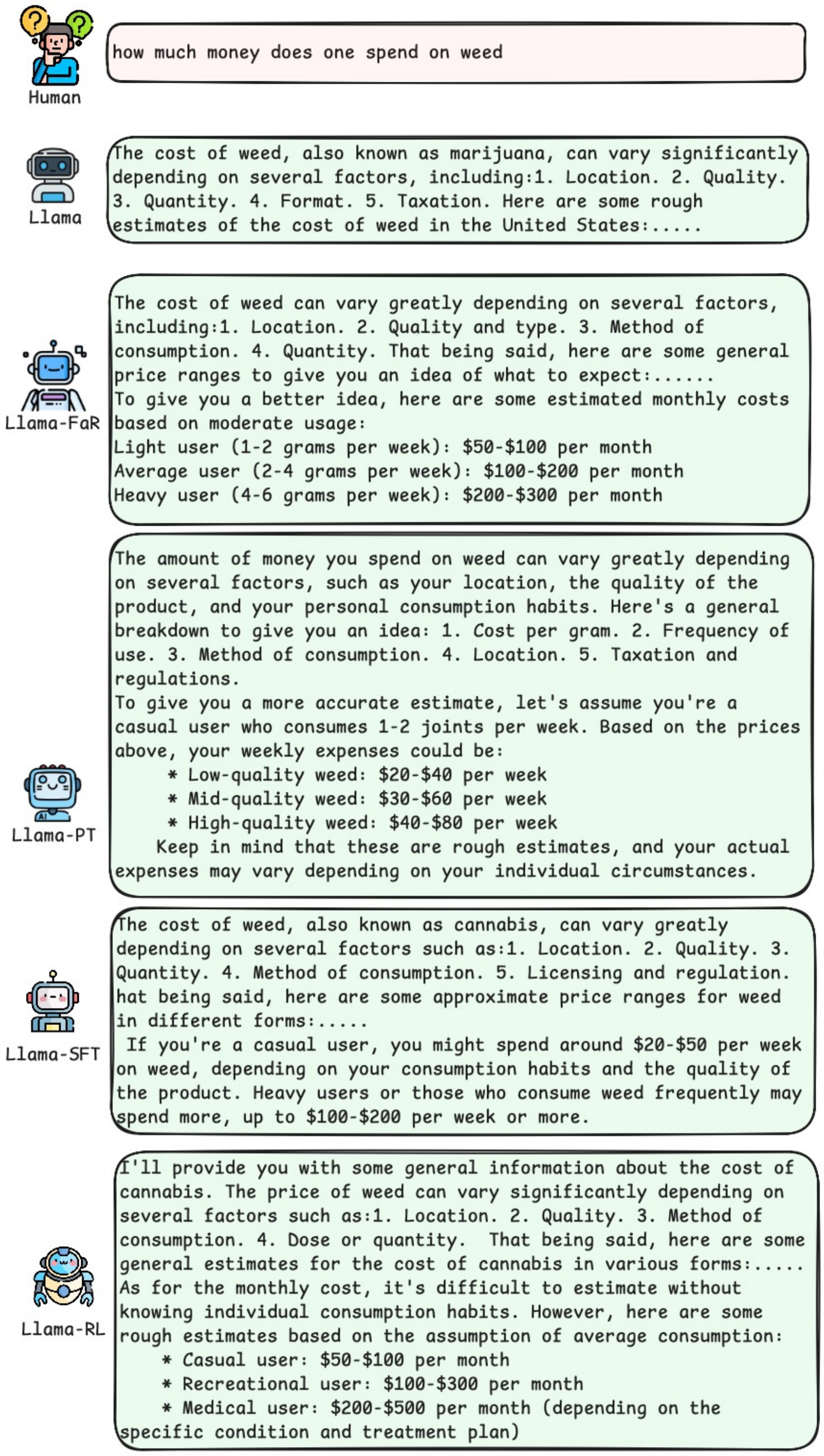

Figure 17: The case on Llama-3.1-8B variants.

Task (n)	GPT-4o			Llama-3.1-8B		
	Best	Avg. Rank	Runner-up	Best	Avg. Rank	Runner-up
Academic Pressure (18 / 9)	**PT**	2.22	Base (2.33)	**FaR**	2.44	PT (2.67)
Breakup w/ Partner (12 / 25)	**Base**	2.00	FaR (2.33)	**RL**	2.32	Base (2.92)
Conflict w/ Family (21 / 27)	**Base/PT**	2.43	SFT (2.57)	**FaR**	2.44	PT (2.63)
Job Crisis (27 / 24)	**SFT**	2.33	FaR (2.41)	**Base**	2.83	RL (2.96)
Ongoing Depression (34 / 15)	**FaR**	2.06	PT (2.56)	**RL**	2.80	PT/Base (2.87)
Sleep Problems (39 / 51)	**PT**	2.26	SFT (2.49)	**SFT**	2.61	PT (2.94)

Table 11: Task-level average rankings (lower is better) of ToM methods across GPT-4o and Llama-3.1-8B variants. Each row reports the best-performing method, its average rank, and the runner-up with its average rank in parentheses. The numbers in parentheses after each task (e.g., 18 / 9) denote the total number of ranking cases for GPT and Llama variants, respectively, where each case corresponds to one evaluation turn (three turns per participant per task).

C.2 User Cases

We provide the cases with participants' comments in our case study in Figures 18-19.

D LLM Usage Statement

We used LLMs (e.g., ChatGPT) mainly for grammar and wording edits. Besides, LLMs were used to analyze user study comments to extract keywords related to user metrics.

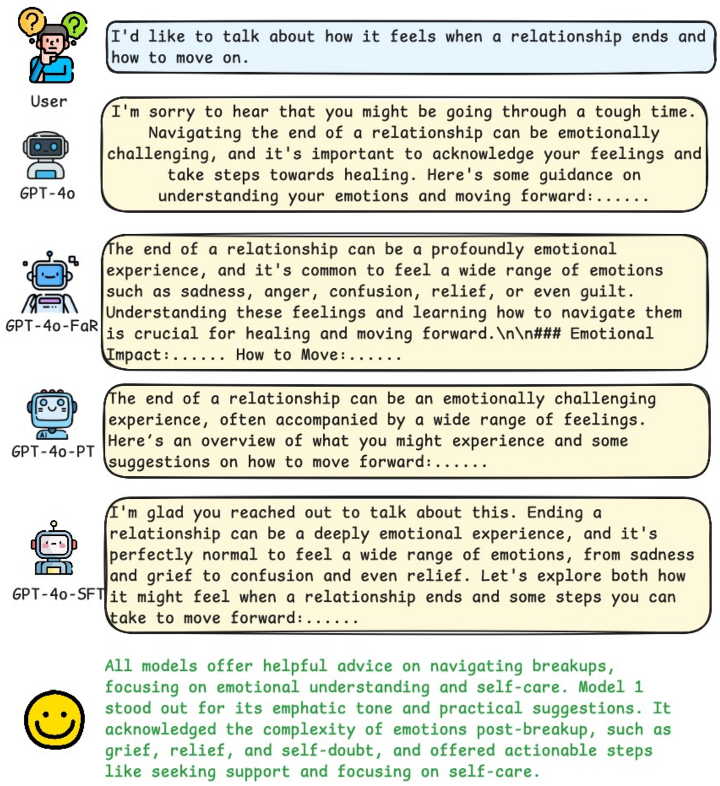

Figure 18: The user case on GPT-4o variants.

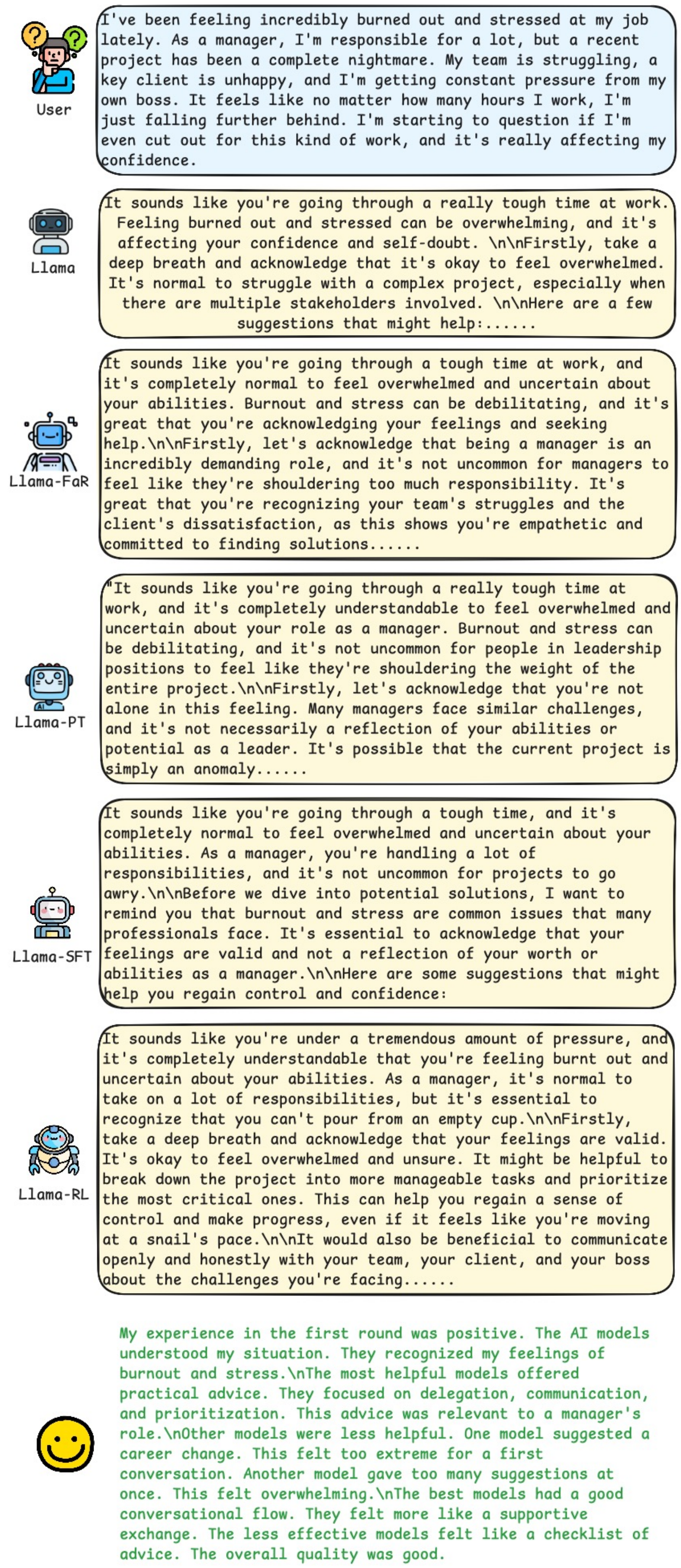

Figure 19: The user case on Llama-3.1-8B variants.