ToM-Synth: Scaling Robust Theory of Mind in LLMs via 6,912 Structured Social Units

Anonymous ACL submission

Abstract

Theory of Mind (ToM), the ability to infer others' mental states from behavior, is pivotal for developing machines with human-level social intelligence. Existing methods endowing LLMs with ToM fall into two paradigms: training-free methods and those repurposing ToM evaluation benchmarks as training data for RL-based fine-tuning. However, training-free methods fail to internalize the augmented ToM into the LLMs. Meanwhile, using evaluation benchmarks as training sources is conceptually problematic and, in practice, results in narrow in-domain overfitting rather than robust ToM. To address the lack of training resources within the ToM community and to empower LLMs with robust ToM, we introduce ToM-Synth, a factorial combinatorial synthesis framework of 6912 social units. This framework enables the systematic synthesis of ToM data, yielding a training dataset of 27,648 instances, termed ToM-Synth-27K. Utilizing ToM-Synth-27K for RL fine-tuning, experimental results demonstrate consistent and significant improvements across models of varying families and scales on ToM, Emotional Intelligence, and Social Commonsense benchmarks. Furthermore, we observe concurrent enhancements in IQ-related tasks (math, science, logic) and effective performance scaling with increasing data scale.

1 Introduction

Theory of Mind (ToM) involves inferring others' mental states (e.g., beliefs, emotions, and intentions) based on their behaviors in complex, real-world social contexts, and is fundamental to human social interaction (Premack and Woodruff, 1978; Wimmer and Perner, 1983). As Large Language Models (LLMs) advance and become increasingly integrated into real-world applications, developing robust and generalizable ToM in LLMs is crucial for ensuring productive and natural human-AI interaction (Liu et al., 2016; Wang et al., 2021; Ying et al., 2024; Zhang et al., 2025b).

Given the critical importance of ToM in LLMs, substantial efforts have been devoted to their evaluation. Beyond assessments based on classic paradigms (Baron-Cohen et al., 1985; Perner et al., 1987) such as the Sally-Anne Test (Unexpected Transfer) and the Smarties Test (Unexpected Contents), Shapira et al. (2024) and Ullman (2023) demonstrate that trivial adversarial perturbations to these classic tests, such as changing an opaque container to a transparent one, lead to significant performance degradation, suggesting that LLMs lack robust ToM. More recently, a growing body of benchmarks such as ToMBench (Chen et al., 2024), OpenToM (Xu et al., 2024), ToMATO (Shinoda et al., 2025), SimpleToM (Gu et al., 2024), and TactfulToM (Liu et al., 2025) have been introduced to evaluate LLMs across richer dimensions of mental states and scenarios that better align with real-world social interactions. Empirical results from these benchmarks consistently reveal a substantial gap between LLMs' ToM and human-level ToM.

The deficiency of ToM in LLMs has spurred significant research efforts aimed at enhancing this capability. As illustrated in Figure 1, these efforts can be broadly categorized into two paradigms: training-free methods and Reinforcement Learning (RL)-based fine-tuning. Among training-free methods, representative works include SimToM (Wilf et al., 2024), which adopts perspective-taking prompting strategies; TimeToM (Hou et al., 2024), which constructs a temporal space and designs a time-aware belief solver; MetaMind (Zhang et al., 2025a), a multi-agent framework inspired by metacognitive theories that decomposes ToM tasks into three collaborative stages; DeL-ToM (Wu et al., 2025), which introduces a process belief model for inference-time scoring and selection among multiple belief traces; and AutoToM (Zhang et al., 2025b), which automatically constructs a suitable agent model and performs automated bayesian inverse planning using an LLM as the computational

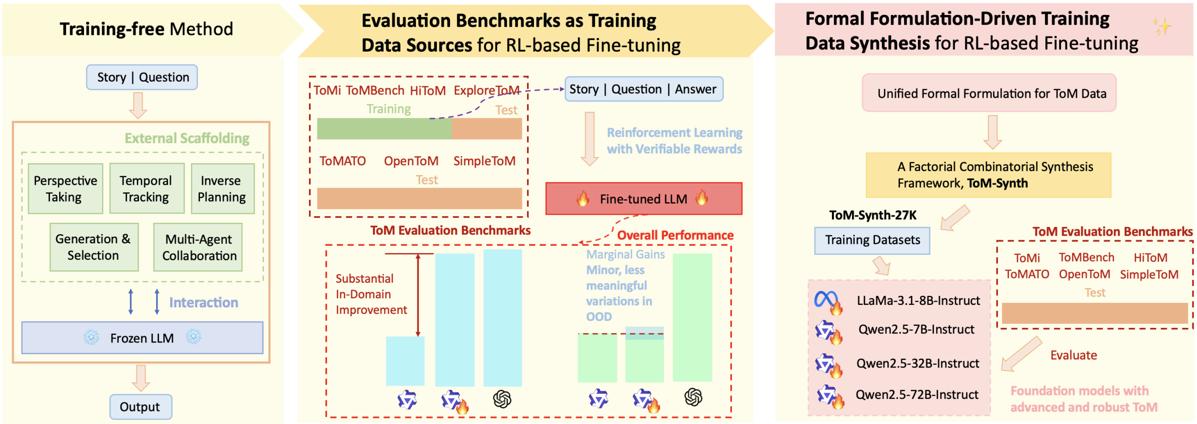

Figure 1: Comparison of paradigms for endowing LLMs with ToM. Unlike training-free methods that fail to internalize ToM due to frozen parameters, or methods that train on evaluation benchmarks, which is conceptually problematic and yields poor robustness, we synthesize training data from a unified formal formulation for ToM data to develop foundation models across diverse families and scales that exhibit advanced and robust ToM.

backend. Despite their demonstrated effectiveness, these methods share a fundamental limitation: **they fail to fundamentally internalize the augmented ToM into the LLMs themselves**.

Most existing works on ToM focus primarily on designing benchmarks for evaluation, leaving a notable gap in the **lack of training resources for developing ToM** in LLMs. Current RL-based fine-tuning methods uniformly rely on data drawn from these **evaluation benchmarks to construct training sets**. For instance, TimeHC-RL (Hou et al., 2025) and ToM-RL (Lu et al., 2025) directly split existing benchmarks into training and test sets, a practice we argue is conceptually fundamentally problematic. Furthermore, Sarangi and Salam (2025) critically demonstrates that LLMs fine-tuned via RL on ToM benchmarks, despite achieving substantial in-domain improvements, show only marginal gains on Out-Of-Distribution (OOD) tasks. This suggests that the learned behavior constitutes a form of **narrow overfitting** rather than the acquisition of **robust ToM**.

In this work, to bridge the notable gap in the lack of training resources within the ToM community and to empower LLMs with robust ToM, we first present a **unified formal formulation for ToM data**. This formulation is anchored in the intrinsic definition of ToM, which essentially involves inferring various dimensions of mental states across diverse real-world social situations. Leveraging the structured dimensions from this formulation, we introduce **ToM-Synth**, a factorial combinatorial synthesis framework comprising 2 presentation formats (Narrative / Dialogue), 36 capabilities spanning 6 mental state dimensions (e.g., Hidden Emotions, Content False Beliefs), and 96 real-world social situations (e.g., Workplace Meetings, Family Dinner), yielding a systematically structured synthesis space of $36 \times 96 \times 2 = 6{,}912$ social units, as illustrated in Figure 2. This multi-factorial combinatorial design ensures comprehensive coverage of the ToM data landscape. Each unit corresponds to a unique combination of factors (e.g., [Hidden Emotions, Workplace Meetings, Dialogue]). We leverage Claude Opus 4.5 (Anthropic, 2025) to synthesize four data entries per unit based on its specific factor combination, yielding a training dataset of 27,648 instances, termed **ToM-Synth-27K**. Data quality is ensured through a rigorous multi-stage pipeline, including cross-model verification using Gemini 3 Pro Preview (DeepMind, 2025).

Utilizing the ToM-Synth-27K dataset, we apply Group Relative Policy Optimization (GRPO) (Guo et al., 2025) to fine-tune a diverse array of LLMs across varying families and scales, specifically LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and the Qwen2.5 series (7B, 32B, and 72B) (Yang et al., 2024). Experimental results demonstrate consistent performance improvements across all models on six ToM benchmarks, one Emotional Intelligence (EI) benchmark, and one Social CommonSense (CS) benchmark, yielding overall gains of +9.31, +5.95, +2.26, and +4.08, respectively. Notably, the 32B and 72B models achieve performance comparable to the advanced GPT-5 (OpenAI, 2025) on multiple benchmarks. Furthermore, we observe consistent enhancements in IQ-related tasks, spanning mathematics, science, and logic,

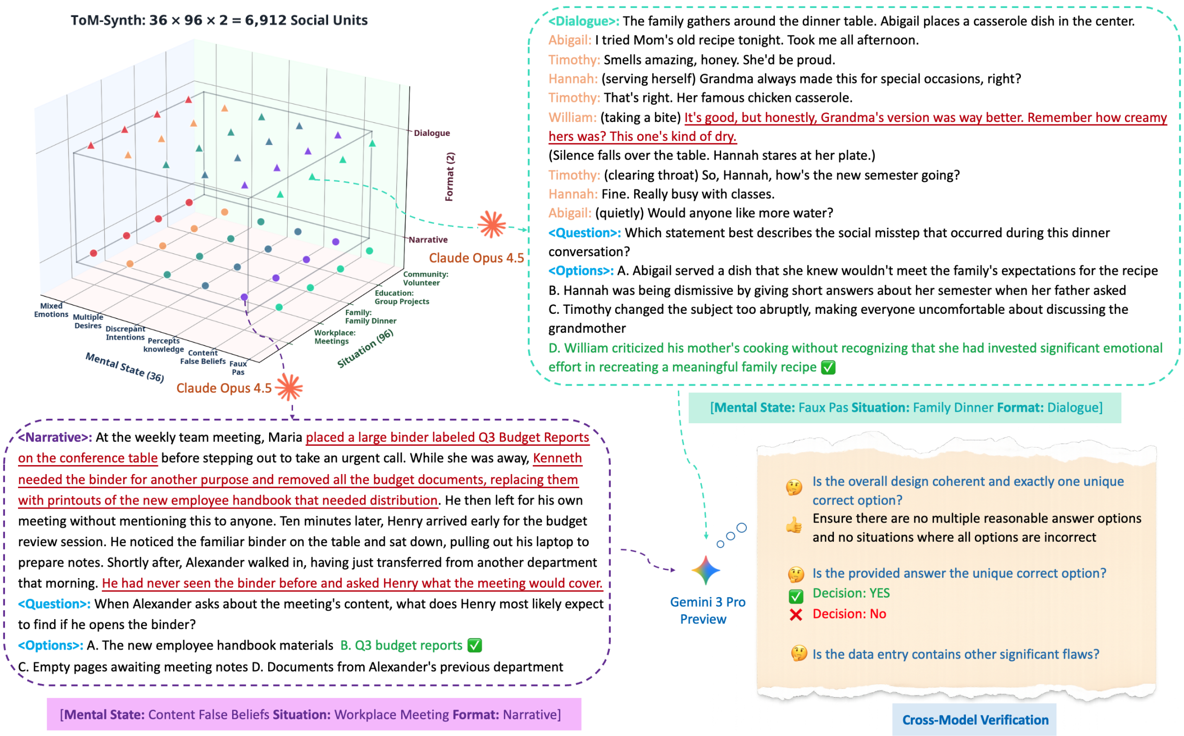

Figure 2: Visualization of the ToM-Synth social units. Each unit within the space corresponds to a unique combination of factors. We utilize Claude Opus 4.5 to generate data based on these combinations, followed by cross-model verification using Gemini 3 Pro Preview to ensure data unambiguity and the reliability of answer.

highlighting the intrinsic value of the ToM-Synth-27K dataset. Beyond standard evaluations, we conduct comprehensive analyses to: (1) investigate the trade-off between increasing unit diversity and data entries per unit under a fixed data budget; (2) explore performance scaling laws with respect to Unit Scaling and Per-Unit Data Scaling; and (3) compare reasoning trajectories before and after RL.

2 Unified Formalization of ToM Data

Fundamentally, ToM data can be characterized by three orthogonal dimensions: the target mental state being inferred (M), the social situation in which the mental state arises (S), and the presentation format (P). Different ToM data vary primarily in their specific instantiations along these axes. Consequently, by systematically exploring the combinatorial space of these dimensions, we can construct a structured and scalable framework that provides comprehensive coverage of the ToM data landscape. Formally:

$$O = \mathcal{F}(m, s, p), \quad \text{where } (m, s, p) \in M \times S \times P \tag{1}$$

where $\mathcal{F}$ denotes the Foundation Model (instantiated here as an LLM). An LLM exhibiting robust ToM is expected to consistently produce contextually accurate outputs (O) across diverse combinatorial inputs of (m, s, p).

3 ToM-Synth Framework

This section details the ToM-Synth Framework, which operates in three steps: (1) constructing structured social units by defining value sets for dimensions M, S, and P; (2) synthesizing data conditioned on these units; and (3) multi-stage quality refinement to ensure dataset validity.

3.1 Constructing Structured Social Units

For dimension M (the target mental state being inferred), we first establish a set of core mental state dimensions grounded in canonical literature (Premack and Woodruff, 1978) and American Psychological Association[1] definitions: $\mathcal{S}_{core}$ = {Belief, Emotion, Intention, Desire, Knowledge}. We also include Non-literal Communication which, while not formally defined as a mental state, is widely acknowledged as a complex pragmatic function that necessitates advanced ToM (Happé, 1994). Leveraging the taxonomy from Ma et al. (2023), we expand these six high-level

[1] https://dictionary.apa.org/theory-of-mind

dimensions into 31 fine-grained sub-dimensions. For instance, Emotion is further decomposed into sub-dimensions such as Mixed Emotion and Hidden Emotion. Furthermore, we introduce five supplementary sub-dimensions by proposing a generalized schema of "Belief about X". Extending standard second-order belief (typically restricted to X = Belief), we argue that inferring beliefs regarding other latent states is equally significant. Accordingly, we define the supplementary set as:

$$\mathcal{S}_{supp} = \{\text{Belief about } X \mid X \in \mathcal{S}_{core} \setminus \{\text{Belief}\}\} \tag{2}$$

This schema accounts for sub-dimensions such as *Belief about Emotion* (i.e., what one believes about another's emotional state). Consequently, the final variable space consists of the union of the 31 fine-grained sub-dimensions and these supplementary sub-dimensions, totaling **36 distinct values**.

For dimension S (the social situation in which the mental state arises), we first prompt the Claude Opus 4.5 model to identify a diverse set of social domains (e.g., Workplace, Family, Education) serving as high-level thematic anchors. Conditioned on each domain, we subsequently generate granular interaction situations (e.g., Meetings, Negotiations, and Performance Reviews within the Workplace domain). From an initial pool of 20 social domains and 15 granular interaction situations per domain, we conduct a manual review to retain those exhibiting **high prevalence and broad representativeness in real-world contexts**. This result in a final selection of 12 social domains and 8 granular interaction situations each, totaling **96 distinct values**.

For dimension P (the presentation format), we consider **2** values: Narrative and Dialogue. We conceptualize the dimensions M, S, and P as three distinct factors. By exhaustively combining them, we form a total of 6,912 unique structured social units (36 × 96 × 2). The value set for each dimension is detailed in Appendix A.

3.2 Data Synthesis Based on Social Units

As illustrated in Figure 2, each social unit corresponds to a specific combination of factors, such as [**Mental State**: Content False Beliefs, **Situation**: Workplace Meeting, **Format**: Narrative]. This factor combination serves as the parameter configuration for data generation using Claude Opus 4.5, where the meaning of each factor is explicitly articulated to the model. We enforce rigorous generation constraints include ensuring that distractors are plausible and non-trivial, eliminating length bias in correct options, and mental states must be inferred from observable behavioral cues rather than stated explicitly, thereby producing sufficiently challenging data. The complete data generation prompt is provided in Appendix B.

3.3 Multi-stage Dataset Quality Refinement

Based on 6,912 structured social units, we generate four entries per unit, forming a dataset of 27,648 samples termed ToM-Synth-27K. To guarantee high data quality, we apply a multi-stage refinement pipeline:

- **Stage 1: Format and Consistency Check**. We screen the generated outputs for adherence to the specified format and check for content duplication among options. This step filters out 89 invalid entries (0.32%).
- **Stage 2: Answer Distribution Balancing**. We address the positional bias where the model disproportionately assigns the correct answer to options "B" or "C" by balancing the distribution of the answer keys.
- **Stage 3: Cross-Model Verification**. We leverage Gemini 3 Pro Preview to perform cross-validation against the generating model (Claude Opus 4.5). We eliminate samples containing ambiguities or disagreements regarding the answer between the two advanced models. This rigorous filtering removes 1,703 entries (6.2%), resulting in a final curated dataset of 25,856 samples.

4 RL on Synthetic Data

Building on the recent advances in Reinforcement Learning with Verifiable Rewards (RLVR), we focus on applying RL to synthetic data using the GRPO algorithm. Following Guo et al. (2025), we adopt a structured training prompt template that encourages the model to first engage in explicit reasoning before providing the final answer. Specifically, the reasoning process and the answer are enclosed within <think> </think> and <answer> </answer> tags, respectively. Our reward function is composed of two additive components: a format reward and an outcome reward. Specifically, let y denote the model response. We define the reward function as:

$$R(y) = R_{\text{format}}(y) + R_{\text{outcome}}(y) \tag{3}$$

Model	ToM Benchmarks						EI & Social CS		Avg
	ToMi	HiToM	ToMBench	SimpleToM	ToMATO	OpenToM	EmoBench	SocialIQA	
Representative and Advanced Foundation Models									
Doubao-1.5-pro-32k	75.53	58.70	69.53	47.11	70.60	56.40	50.92	77.70	63.31
Deepseek-v3-1	73.10	52.50	67.97	53.56	51.50	50.72	52.17	78.25	59.97
Qwen3-max	72.00	48.98	75.04	50.48	81.60	47.32	56.50	84.55	64.56
Qwen3-235B-A22B	71.70	78.15	74.59	65.71	76.70	51.42	52.00	86.70	69.62
GPT-4o	68.50	42.50	75.22	59.37	76.40	57.12	62.25	83.35	65.59
GPT-5	96.80	88.60	77.64	67.45	83.40	72.18	57.50	86.50	78.76
Our Trained Models									
LLaMA-3.1-8B-Instruct	56.90	37.31	59.03	48.93	58.54	55.38	40.67	71.70	53.56
+ ToM-Synth-27K	70.20	47.59	66.32	**72.59**	65.78	61.50	46.00	74.00	62.87
Δ	+13.30	+10.28	+7.29	+23.66	+7.24	+6.12	+5.33	+2.30	+9.31
Qwen2.5-7B-Instruct	64.60	35.92	65.41	41.61	61.50	58.50	39.33	72.30	54.90
+ ToM-Synth-27K	67.30	50.27	65.78	52.16	66.39	63.12	42.50	79.30	60.85
Δ	+2.70	+14.35	+0.37	+10.55	+4.89	+4.62	+3.17	+7.00	+5.95
Qwen2.5-32B-Instruct	74.80	52.22	73.73	58.96	71.66	62.40	52.83	**83.80**	66.30
+ ToM-Synth-27K	75.80	55.37	74.10	**65.38**	**75.16**	64.56	**54.00**	**84.10**	68.56
Δ	+1.00	+3.15	+0.37	+6.42	+3.50	+2.16	+1.17	+0.30	+2.26
Qwen2.5-72B-Instruct	77.20	56.11	73.56	58.26	66.02	65.75	46.50	**81.10**	65.56
+ ToM-Synth-27K	77.80	57.77	**76.12**	**64.57**	**76.75**	**70.62**	50.00	**83.45**	69.64
Δ	+0.60	+1.66	+2.56	+6.31	+10.73	+4.87	+3.50	+2.35	+4.08

Table 1: Results on ToM, EI, and Social CS benchmarks. Avg denotes the mean score across eight benchmarks. Green numbers indicate improvements following RL with ToM-Synth-27K. **Bold orange** highlights our trained foundation models, which achieve performance close to the advanced GPT-5. EI: Emotional Intelligence; CS: Common Sense. Specific version details for the reference models are provided in C.2.

where format reward penalizes malformed outputs:

$$R_{\text{format}}(y) = \begin{cases} 0 & \text{if } y \text{ follows required format} \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

and outcome reward incentivizes correct answers:

$$R_{\text{outcome}}(y) = \begin{cases} 1 & \text{if the answer is correct} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

This formulation yields a total reward $R(y) \in \{-1, 0, 1\}$: the model receives $R(y) = 1$ when both format and answer are correct, $R(y) = 0$ when format is correct but the answer is wrong, and $R(y) = -1$ when the format is malformed.

5 Experiments

5.1 Setup Details

Training Setup. We employ LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and the Qwen2.5 series (7B, 32B, and 72B) (Yang et al., 2024) as our baseline models. We utilize the VeRL framework (Sheng et al., 2025) to implement RL on these models. All experiments are conducted on NVIDIA A100 (80GB) GPUs. Specifically, we use 8 GPUs for the 7B and 8B models, 32 GPUs for the 32B model, and 64 GPUs for the 72B model. The RL parameter configurations are provided in C.

Dataset Name Explanation. **ToM-Synth-27K-Contrast** is constructed to investigate the trade-off between unit diversity and the number of data entries per unit under a fixed data budget. This dataset comprises 2,304 structured social units, where the set of mental state dimension values is reduced from 36 to 12 (specifically retaining six high-level dimensions and six dimensions under the second-order "Belief about X" pattern), with each unit containing 12 data entries. **ToM-Synth-Merged-55K** is obtained by directly merging ToM-Synth-27K and ToM-Synth-27K-Contrast.

5.2 Evaluation Benchmarks

To rigorously assess the enhancement of ToM in LLMs trained with ToM-Synth-27K, we employ a suite of six ToM benchmarks: ToMi (Nematzadeh et al., 2018), HiToM (Wu et al., 2023), ToMBench, SimpleToM, ToMATO, and OpenToM. Furthermore, to demonstrate the broader utility of our synthetic data, we extend our evaluation to include EmoBench (Sabour et al., 2024) for EI, SocialIQA (Sap et al., 2019) for Social CS reasoning, and three IQ-related benchmarks: AIME 2025 (Mathematical Association of America, 2025), GPQA Diamond (Rein et al.), and AGIEval (Zhong et al., 2024). This compre-

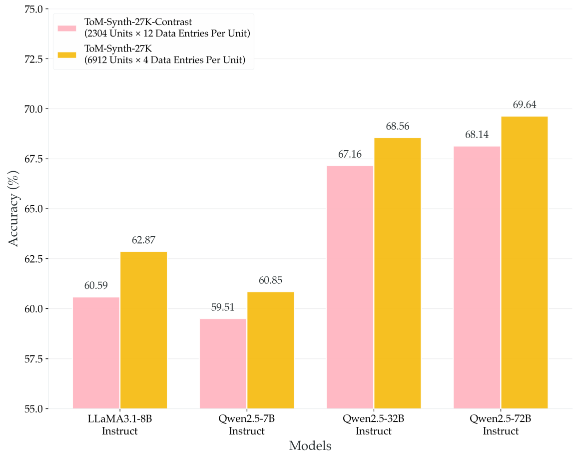

Figure 3: Comparison of average performance on ToM, EI, and Social CS benchmarks across different data strategies under a fixed 27,648 data budget.

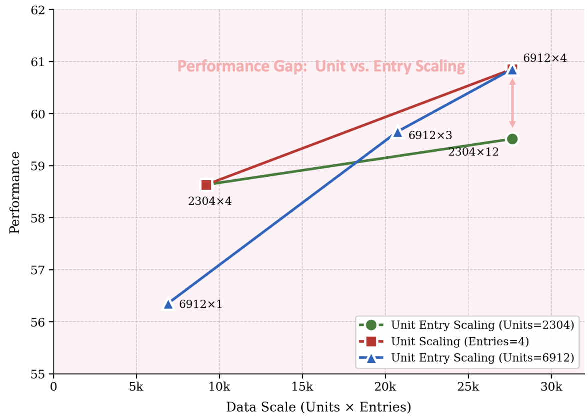

Figure 4: Comparison of scaling strategies: Unit-scaling (diversity) vs. Entry-scaling (depth). Performance of Qwen2.5-7B-Instruct across varying ToM-Synth data scales (Units$\times$ Entries).

hensive evaluation suite comprises a total of **14,015** samples, with the detailed data distribution for each benchmark provided in the Appendix C.3.

5.3 Quantitative Experimental Results

Consistent Improvements on ToM, EI, and Social CS Benchmarks. As shown in Table 1, models across different scales and families exhibit consistent performance gains after RL with ToM-Synth-27K. Notably, the Qwen2.5-72B-Instruct model achieves an overall performance of 69.64. This performance not only surpasses the leading open-source model, Qwen3-235B-A22B-Instruct-2507, but also demonstrates capabilities comparable to the advanced GPT-5 on a majority of benchmarks (five out of eight, 62.5%).

Unit Diversity vs. Data Entries per Unit. As illustrated in Figure 3, we investigate the trade-off between increasing unit diversity and the number of data entries per unit under a fixed data budget (comparing 2,304 units $\times$ 12 entries against 6,912 units $\times$ 4 entries). Experimental results across models of varying scales and families (60.59 vs. 62.87, 59.51 vs. 60.85, 67.16 vs. 68.56, and 68.14 vs. 69.64) demonstrate that, given a fixed 27,648 data budget, higher unit diversity consistently outperforms increasing data entries per unit.

Scaling of Model Performance with Data Scale. As illustrated in Figure 4, with the number of Social Units fixed at 6,912, model performance exhibits a near-linear scaling trend as the number of data entries per unit increases from 1 to 4. Starting from a baseline of $2,304$ units $\times$ 4 entries, we compare scaling data density ($2,304 \times 12$) versus expanding unit count ($6,912 \times 4$). Notably, the latter approach yields superior performance gains. Figure 5 further demonstrates that applying RL to Qwen2.5-72B-Instruct with ToM-Synth-Merged-55K achieves a score of 71.45, surpassing the 69.64 of ToM-Synth-27K, validating effective scaling.

RL on ToM Data Enhances Performance on IQ-related Tasks. As shown in Table 2, models across various scales and families exhibit consistent performance improvements on IQ-related benchmarks (including mathematics, science, and logic) following RL with ToM-Synth-27K. For instance, the Qwen2.5-72B-Instruct model achieves improvements of +1.67, +6.09, and +1.31, respectively. Given that ToM is intrinsically associated with Emotional Quotient (EQ), its capacity to bolster performance in IQ-centric domains is particularly noteworthy. These findings suggest a promising foundation for developing LLMs that possess both advanced IQ and EQ.

5.4 Qualitative Analysis

The Superiority of ToM-Synth: A Data Synthesis Perspective Our data synthesis framework, ToM-Synth, generates novel data through fundamental combinatorial composition of factors rather than paraphrasing or imitating existing examples. This design theoretically minimizes the risk of direct repetition. Figure 6 illustrates data generation using PersonaHub (Ge et al., 2024) with a sample from ToMi as seed data. Although this approach, which uses a persona description as inspirational text, is more likely to produce novel data compared

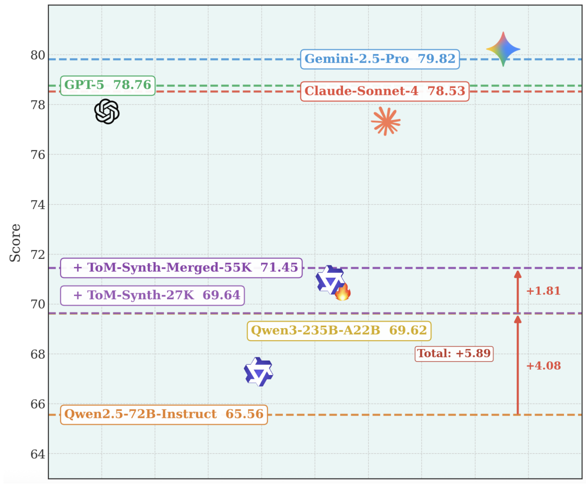

Figure 5: Performance trajectory on ToM, EI, and Social benchmarks. We illustrate the progressive performance gains, achieving a total increase of +5.89 points via our proposed ToM-Synth data scaling. Dashed lines represent the performance ceilings established by leading proprietary models, including Gemini-2.5-Pro (Comanici et al., 2025) and Claude-Sonnet-4 (Anthropic, 2025), highlighting how our method narrows the gap between open-weight models and leading proprietary models.

to direct paraphrasing or imitation, the generated data still exhibits substantial similarity to the seed data. This highlights a fundamental limitation of generation approaches that depend on seed data. The prompt used for PersonaHub-based data generation is provided in Appendix B.

Comparison of Reasoning Trajectories Before and After RL with ToM-Synth-27K for Qwen2.5-7B-Instruct. As illustrated in Figure C.4, the upper panel presents an example from the Mixed Emotions category in ToMBench. Before RL with ToM-Synth-27K, the Qwen2.5-7B-Instruct model fails to effectively identify the mixed emotional state, feeling frustrated about one's own situation while simultaneously experiencing genuine happiness for a friend's achievement. After RL training, the model bridges this gap through more elaborate reasoning, as reflected in increased response length. The lower panel shows a second-order belief example from HiToM. Surprisingly, before RL with ToM-Synth-27K, Qwen2.5-7B-Instruct explicitly states that the question cannot be answered, indicating a complete inability to handle this problem. After RL training, the model systematically deduces each event in the narrative, ultimately arriving at the correct answer. Additional case studies comparing reasoning trajectories are provided in the Appendix C.4.

Model	AIME 2025	GPQA Diamond	AGIEval
LLaMA-3.1-8B-Instruct	0.83	9.14	16.08
+ ToM-Synth-27K	0.83	13.70	20.43
Δ	0.00	**+4.56**	**+4.35**
Qwen2.5-7B-Instruct	2.50	24.36	24.34
+ ToM-Synth-27K	5.42	31.97	24.78
Δ	**+2.92**	**+7.61**	**+0.44**
Qwen2.5-32B-Instruct	0.83	40.60	25.65
+ ToM-Synth-27K	4.17	43.65	32.17
Δ	**+3.34**	**+3.05**	**+6.52**
Qwen2.5-72B-Instruct	0.00	35.53	31.73
+ ToM-Synth-27K	1.67	41.62	33.04
Δ	**+1.67**	**+6.09**	**+1.31**

Table 2: Evaluation on IQ-related benchmarks across different models. Purple numbers denote improvements yielded by RL with ToM-Synth-27K. All values represent percentage accuracy.

6 Related Work

Data Synthesis Existing data synthesis methods can be broadly categorized into three paradigms. The first involves rewriting seed data from multiple perspectives. MetaMath (Yu et al., 2024) augments mathematical reasoning data through diverse reformulations of original problems. The second paradigm synthesizes data from scratch without relying on seed data. GRIP (Wang et al., 2025a) constructs a relational graph over key concepts to enable synthesis based on diverse concept com-

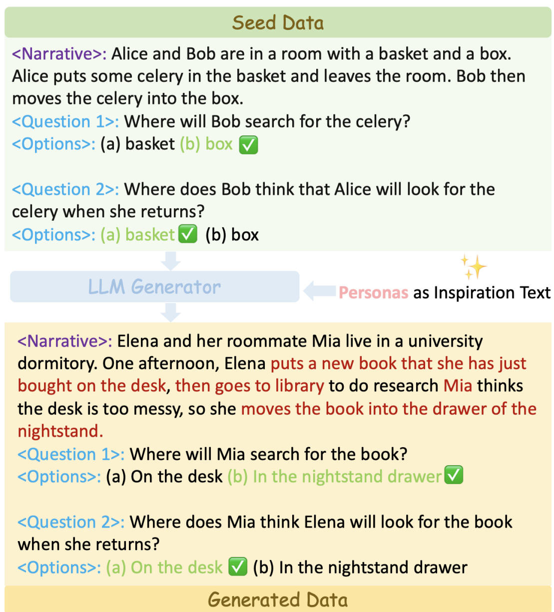

Figure 6: The top panel shows a ToMi sample used as seed data, while the bottom panel displays data generated using PersonaHub with persona descriptions as inspiration text.

Figure 7: Comparison of Qwen2.5-7B-Instruct's reasoning trajectories before and after RL with ToM-Synth-27K. Upper: A Mixed Emotions example from ToMBench where after RL training enables the model to identify concurrent emotional states. Lower: A second-order belief example from HiToM where after RL training transforms complete reasoning failure into systematic narrative deduction.

binations, while TreeSynth (Wang et al., 2025b) employs a hierarchical tree structure. The third paradigm performs online data synthesis during RL (Liang et al., 2025), where problems that the policy model struggles to solve are identified, evolved into challenging variants, and then combined with the original data for subsequent policy updates. These methodologies have been extensively applied across domains, including mathematical reasoning, logical inference, and code generation.

Methods for Enhancing ToM in LLMs Beyond the methods for enhancing LLMs' ToM discussed in the introduction, several notable approaches have been proposed. SymbolicToM (Sclar et al., 2023) constructs an explicit belief tracker to maintain mental state representations. PercepToM (Jung et al., 2024) improves perception-to-belief inference by extracting relevant contextual information from the input. Agentic-ToM (Sarangi et al., 2025) enables LLMs to autonomously determine when to invoke cognitive tools for solving ToM problems. BIP-ALM (Jin et al., 2024) adopts a Bayesian inverse planning framework and, distinctively, employs a language model fine-tuned on human activity data to evaluate the likelihood of hypotheses regarding an agent's beliefs and goals. DWM-ToM (Huang et al., 2024) utilizes an LLM as a world model to track environmental dynamics and iteratively refine prompts.

7 Conclusion

In this paper, we introduce ToM-Synth, a factorial combinatorial synthesis framework comprising 6,912 social units, which we use to synthesize the ToM-Synth-27K training dataset. By applying RL with ToM-Synth-27K, models across different scales, families achieve consistent improvements on ToM, EI, and Social CS benchmarks. Notably, we also observe transfer effects to IQ-related task benchmarks, suggesting that structured social cognition training may enhance broader cognitive capabilities. Theoretically, our framework enables infinitely scalable data generation; both the number of social units and the data entries per unit can be scaled up to achieve performance levels beyond those reported in this paper.

Limitations

To our knowledge, this work has the following limitations:

- In our current data synthesis pipeline, each social unit contributes an equal number of data entries. Investigating optimal data allocation ratios across social units may help cultivate stronger ToM in LLMs while potentially improving data efficiency.
- As we continue to scale the data volume, it remains an open empirical question whether a performance plateau will emerge at a certain threshold. In our current experiments, we have scaled the dataset to 55K entries across the 6,912 social units. Our findings indicate that up to this 55K limit, model performance scales monotonically with data size without signs of saturation. Should a bottleneck occur in future scaling, it may necessitate a further expansion of the underlying social units to maintain performance gains.

References

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

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

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

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google DeepMind. 2025. Gemini 3 pro. https://deepmind.google/models/gemini/pro/.

DeepSeek. 2025. Deepseek-v3.1 release. https://api-docs.deepseek.com/news/news250821.

Doubao Team. 2025. Doubao-1.5-pro. https://seed.bytedance.com/en/special/doubao_1_5_pro.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

Guiyang Hou, Xing Gao, Yuchuan Wu, Xiang Huang, Wenqi Zhang, Zhe Zheng, Yongliang Shen, Jialu Du, Fei Huang, Yongbin Li, and 1 others. 2025. Timehcrl: Temporal-aware hierarchical cognitive reinforcement learning for enhancing llms' social intelligence. *arXiv preprint arXiv:2505.24500*.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

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

Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. 2025. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. *arXiv preprint arXiv:2508.14029*.

Chang Liu, Jessica B Hamrick, Jaime F Fisac, Anca D Dragan, J Karl Hedrick, S Shankar Sastry, and Thomas L Griffiths. 2016. Goal inference improves objective and perceived performance in human-robot collaboration. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 940–948.

Yiwei Liu, Emma Jane Pretty, Jiahao Huang, and Saku Sugawara. 2025. Tactfultom: Do llms have the theory of mind ability to understand white lies? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25054–25072.

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*.

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

Mathematical Association of America. 2025. American invitational mathematics examination (AIME) 2025. https://maa.org/math-competitions/.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

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

Josef Perner, Susan R Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

Sneheel Sarangi and Hanan Salam. 2025. Small llms do not learn a generalizable theory of mind via reinforcement learning. *arXiv preprint arXiv:2507.15788*.

Sneheel Sarangi, Chetan Talele, and Hanan Salam. 2025. Agentic-tom: Cognition-inspired agentic processing for enhancing theory of mind reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25645–25661.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

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.

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

Jiankang Wang, Jianjun Xu, Xiaorui Wang, Yuxin Wang, Mengting Xing, Shancheng Fang, and Hongtao Xie. 2025a. Grip: A graph-based reasoning instruction producer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Sheng Wang, Pengan Chen, Jingqi Zhou, Qintong Li, Jingwei Dong, Jiahui Gao, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2025b. Treesynth: Synthesizing diverse data from scratch via tree-guided subspace partitioning. *arXiv preprint arXiv:2503.17195*.

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.

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

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Yuheng Wu, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. 2025. Del-tom: Inference-time scaling for theory-of-mind reasoning via dynamic epistemic logic. *arXiv preprint arXiv:2505.17348*.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Lance Ying, Kunal Jha, Shivam Aarya, Joshua B Tenenbaum, Antonio Torralba, and Tianmin Shu. 2024. Goma: Proactive embodied cooperative communication via goal-oriented mental alignment. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7099–7106. IEEE.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. 2025a. Metamind: Modeling human social thoughts with metacognitive multi-agent systems. *arXiv preprint arXiv:2505.18943*.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. 2025b. Autotom: Scaling model-based mental inference via automated agent modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Value Sets

We present the complete set of possible values for character roles in Figure 8, and the full sets of possible values for the three dimensions M, S, and P in Figure 9.

Character

James, Michael, Robert, John, David, William, Richard, Joseph, Thomas, Christopher, Charles, Daniel, Matthew, Anthony, Mark, Donald, Steven, Andrew, Paul, Joshua, Kenneth, Kevin, Brian, Timothy, Ronald, George, Jason, Edward, Jeffrey, Ryan, Jacob, Nicholas, Gary, Eric, Jonathan, Stephen, Larry, Justin, Scott, Brandon, Benjamin, Samuel, Gregory, Alexander, Patrick, Frank, Raymond, Jack, Dennis, Jerry, Tyler, Aaron, Jose, Adam, Nathan, Henry, Zachary, Douglas, Peter, Kyle, Noah, Ethan, Jeremy, Christian, Walter, Keith, Austin, Roger, Terry, Sean, Gerald, Carl, Dylan, Harold, Jordan, Jesse, Bryan, Lawrence, Arthur, Gabriel, Bruce, Logan, Billy, Joe, Alan, Juan, Elijah, Willie, Albert, Wayne, Randy, Mason, Vincent, Liam, Roy, Bobby, Caleb, Bradley, Russell, Lucas, Mary, Patricia, Jennifer, Linda, Elizabeth, Barbara, Susan, Jessica, Karen, Sarah, Lisa, Nancy, Sandra, Betty, Ashley, Emily, Kimberly, Margaret, Donna, Michelle, Carol, Amanda, Melissa, Deborah, Stephanie, Rebecca, Sharon, Laura, Cynthia, Dorothy, Amy, Kathleen, Angela, Shirley, Emma, Brenda, Pamela, Nicole, Anna, Samantha, Katherine, Christine, Debra, Rachel, Carolyn, Janet, Maria, Olivia, Heather, Helen, Catherine, Diane, Julie, Victoria, Joyce, Lauren, Kelly, Christina, Ruth, Joan, Virginia, Judith, Evelyn, Hannah, Andrea, Megan, Cheryl, Jacqueline, Madison, Teresa, Abigail, Sophia, Martha, Sara, Gloria, Janice, Kathryn, Ann, Isabella, Judy, Charlotte, Julia, Grace, Amber, Alice, Jean, Denise, Frances, Danielle, Marilyn, Natalie, Beverly, Diana, Brittany, Theresa, Kayla, Alexis, Doris, Lori, Tiffany

Figure 8: Complete set of possible values for character roles.

Mental State

Emotions
Discrepant Emotions
Mixed Emotions
Hidden Emotions
Moral Emotions
Emotion Regulation
Typical Emotional Reactions
Atypical Emotional Reactions

Desires
Discrepant Desires
Desire-action Contradiction
Multiple Desires
Desires influence on emotions and actions

Non-literal Communication
Irony/sarcasm
Involuntary Lies
White Lies
Egocentric Lies
Humor
Faux Pas

Intentions
Completion of Failed Actions
Discrepant Intentions
Prediction of Actions
Intentions Explanations

Beliefs
Content False Beliefs
Location False Beliefs
Identity False Beliefs
Second-order Beliefs
Beliefs Based Action/Emotions
Sequence False Beliefs
Second-order Beliefs (Belief about emotion)
Second-order Beliefs (Belief about non-literal communication)
Second-order Beliefs (Belief about intention)
Second-order Beliefs (Belief about desire)
Second-order Beliefs (Belief about knowledge)

Knowledge
Knowledge-pretend Play Links
Percepts-knowledge Links
Information-knowledge Links
Knowledge-attention Links

Social Situations

Workplace: Meetings, Negotiations, Performance Reviews, Workplace Conflicts, Team Collaboration, Promotion Competition, Cross-Departmental Communication, Resignation Handover

Family: Parent-Child Interaction, Sibling Rivalry, Family Dinner, Inheritance Disputes, Elderly Care, Parenting Disagreements, In-law Relations, Family Financial Decisions

Education:Classroom Discussions, Teacher-Student Interaction, Group Projects, Academic Competitions, Parent-Teacher Conferences, School Bullying, College Counseling, Thesis Defense

Healthcare: Doctor-Patient Consultations, Delivering Bad News, Treatment Decisions, Psychological Counseling, End-of-Life Care Discussions, Medical Fee Negotiations, Surgical Informed Consent, Rehabilitation Training

Business: Sales Interactions, Customer Complaints, Bargaining, Contract Signing, Returns and Refunds, Business Negotiations, Supplier Communication, Warranty Claims/After-Sales Rights

Social Gatherings: Parties, Weddings, Funerals, Class Reunions, Birthday Banquets, Holiday Dinners, Graduation Ceremonies, Awards Galas

Public Services: Government Services, Police Enforcement, Court Litigation, Immigration Interviews, Banking Services, Emergency Assistance, Pettiions and Complaints, Public Transportation

Romantic Relationships: First Dates, Couples' Conflicts, Proposals, Breakups, Meeting the Parents, Long-Distance Communication, Fidelity Questions, Discussing Future Plans

Community: Neighborhood Disputes, Homeowners Meetings, Volunteer Activities, Local Elections, Noise Complaints, Pet Disputes, Parking Space Conflicts, Property Management Communication

Friendship: Making New Friends, Borrowing/Lending Money, Keeping Secrets, Resolving Misunderstandings, Drifting Apart, Competition and Jealousy, Travel Companions, Crisis Support

Stranger Interactions: Asking for Directions, Waiting in Line, Carpooling/Ridesharing, Accidental Collisions, Seat-Offering Etiquette, Asking for Help in Public, Witnessing Incidents, Impromptu Cooperation

Crisis Response: Sudden Accidents, Medical Emergencies, Financial Crisis, Family Upheaval, Coping with Unemployment, Bereavement, Diagnosis of Major Illness, Handling Legal Disputes

Presentation Format

Narrative: Third-person story description
Dialogue: Conversational interaction between characters

Figure 9: Complete set of possible values for mental state (M), social situation in which the mental state arises (S), and presentation format (P) dimension.

B Prompt

Cross-Model Verification Prompt

```
You are an expert evaluator for Theory of
Mind (ToM).
Your task is to evaluate the quality of a
ToM data entry.

## Data Entry to Evaluate

### {key1}:
{val1}

### {key2}:
{val2}

### {key3}:
{val3}

Please evaluate the entry based on the
following strict criteria logic:

1. Ambiguity Check:
   * Is the overall design coherent?
     Is the Question logically based on the {
     key1}?
   * Is there exactly one unique correct
   option?
     (Check to ensure there are no multiple
     reasonable answer
     options and no situations where all
     options are incorrect).
   * Decision: If any ambiguity exists, the
   Judgment is AMBIGUOUS.

2. Answer Verification (If not ambiguous):
   * Is the labeled answer {val4} the unique
    correct option?
   * Decision:
      * If YES, the Judgment is PASS.
      * If NO (another option is the correct
       one),
        the Judgment is FALSE.

3. Other Defects:
   * If the entry contains other significant
    flaws
     (e.g., severe logical gaps not covered
     above),
     the Judgment is OTHER.

## Output Format
```

```
You must output a single JSON object.
Do not output any other text.

{
  "judgment": "AMBIGUOUS" | "PASS" | "FALSE"
   | "OTHER",
  "correct_answer": "Only fill this if
  judgment is FALSE, otherwise null",
  "reason": "A concise explanation for your
  judgment"
}
```

Data Synthesis Prompt

```
You are an expert in Theory of Mind (ToM).
ToM involves reasoning
about others' mental states. Your task is to
 generate a high
quality ToM data entry.

## Generation Configuration

- Format: {format_key}
- Social Scenario: {scenario_key}
- Specific Situation: {situation}
- Target Mental State Dimension: {
ability_key}
- Target Concept (Key Focus): {xilidu_key}
- {xilidu_key} Concept Definition: {
xilidu_value}

## Generation Requirements

1. {format_key} Construction:
   - {format_value}
   - In {scenario_key} scenario, a {
   situation} situation.
     {format_key} should be natural and
     plausible.
   - Select the characters from {selected}
   that you will use to
     construct {format_key}. The number of
     characters in
     {format_key} should not exceed 4.
   - Do NOT explicitly state the character's
    mental state (e.g.,
     do not say "John think X" or "Mary feel
      Y"). Instead, provide
     observable cues (context, actions,
     information access) that
     allow a reader to infer the mental
     state.
   - Ensure all necessary information to
   answer the question is
     embedded in the {format_key}, but
     requires synthesis (not
     just distinct recall).
   - The constructed {format_key} should not
    exceed 800 characters.

2. Question Design:
   - For the {ability_key} mental state
   dimension.
   - The question must strictly target the {
   xilidu_key}.

3. Answer Options Requirements:
   - There should be 4 options (1 correct, 3
    distractors) in total.
     The distractors options should be
     designed to appear
     plausible, avoiding obvious outliers
     that can be easily
     eliminated.
   - The position of the correct answer
   should be randomized
     among A, B, C, and D to avoid
     positional bias.
   - Option lengths should not exhibit
   obvious bias, such as the
     correct option being noticeably longer
     or shorter than others.

4. The overall design of the {format_key},
Question, and Answer
   options should be coherent and reasonable.

   - Avoid the question phrasing hinting at
   the correct answer,
     the correct answer being explicitly
     stated in the
     {format_key}, or spurious correlations
     between {format_key}
     elements and the correct answer that
     could enable shortcut
     reasoning without genuine comprehension.

   - Avoid overly detailed descriptions of
   body language or facial
     expressions in {format_key}. These
     factors should not
     dominate the questions and answers.
   - The question and corresponding answer
   options should be based
     on {format_key}, maintaining complete
     and rigorous logic.
     The correct answer should not be based
     on partial speculation
     or conjecture.
   - The overall design must be unambiguous;
    there should be no
     cases where two options both seem
     reasonable answers to the
     question.

5. Difficulty Requirement:
   - The question should be of a high level
   of difficulty and
     present a non-trivial challenge.
   - Increase the challenge by ensuring the
   correct answer relies
     on implicit inference rather than
     explicit statements.
   - Elevate difficulty by constructing
   highly deceptive
     distractors.

## Output Format

Return the result as a single valid JSON
object contained within
a list. Do not output markdown code blocks.
Just the raw JSON.
```

```
Example Output:
[
  {
    "{format_key}": "{format_value}...",
    "Question": "The specific question...",
    "Options": {
      "A": "Option text...",
      "B": "Option text...",
      "C": "Option text...",
      "D": "Option text..."
    },
    "Answer": "one of A, B, C, D",
    "Analysis": "explanation for the
    selected answer"
  }
]
```

Prompt for PersonaHub-based Synthesis

```
You are an expert in Theory of Mind, skilled
 at analyzing human
psychological activities. You are adept at
capturing and
interpreting various mental states that
humans display during
social interactions.

Task Description:
Please imitate the case below (including the
 story, questions,
and options) to generate one similar case in
 Chinese for the
given scenario, to serve as a test sample
for evaluating the
Theory of Mind capabilities of large
language models.

I will provide you with:

Reference case: I will provide you with a
sample that tests the
same task capability for your reference

Inspiration text: This serves as a basis and
 direction for this
generation, similar to a prompt for a
composition

Input

[A reference case]:
Story:
Alice and Bob are in a room with a basket
and a box.
Alice puts some celery in the basket and
leaves the room.
Bob then moves the celery into the box.

First-order question:
Where will Bob search for the celery?
Options: (a) basket (b) box
Answer: box

Second-order question:
Where does Bob think that Alice will look
for the celery
when she returns?
Options: (a) basket (b) box
Answer: basket

[Inspiration text]:
Name: Elena Martinez
Age: 22
Gender: Female
Race: Hispanic
Born Place: San Diego, California

Appearance:
Elena has shoulder-length, wavy dark brown
hair that she often
ties back when she's deep in study. Her
expressive hazel eyes
are framed by a pair of stylish, yet
practical, glasses. She
stands at 5'6" with a slender build, often
seen in casual,
comfortable clothing that allows her to move
 freely between
classes and study sessions.

General Experience:
Elena is currently in her final year of a
Political Science
degree at a prestigious university. She has
interned with
several non-profit organizations focused on
conflict resolution
and has participated in numerous academic
conferences where
she presented her research on political
communication. Her
work has been recognized by her professors
for its depth and
originality.

Personality:
Elena is naturally curious, always eager to
ask questions and
dig deeper into the subjects that fascinate
her. Her ambition
drives her to excel academically and seek
out opportunities to
apply her knowledge in real-world settings.
She is empathetic
and has a strong sense of justice, which
fuels her passion for
understanding and resolving conflicts.
Despite her serious
academic pursuits, Elena has a warm and
approachable demeanor,
making her a beloved figure among her peers.
```

C Experiment Settings

C.1 Training Parameter Configuration

For training hyperparameters, we set the maximum prompt and response lengths to 1,280 and 768, respectively. The learning rate is initialized at 3×10^{-7}, and the KL coefficient is set to 0.001 to

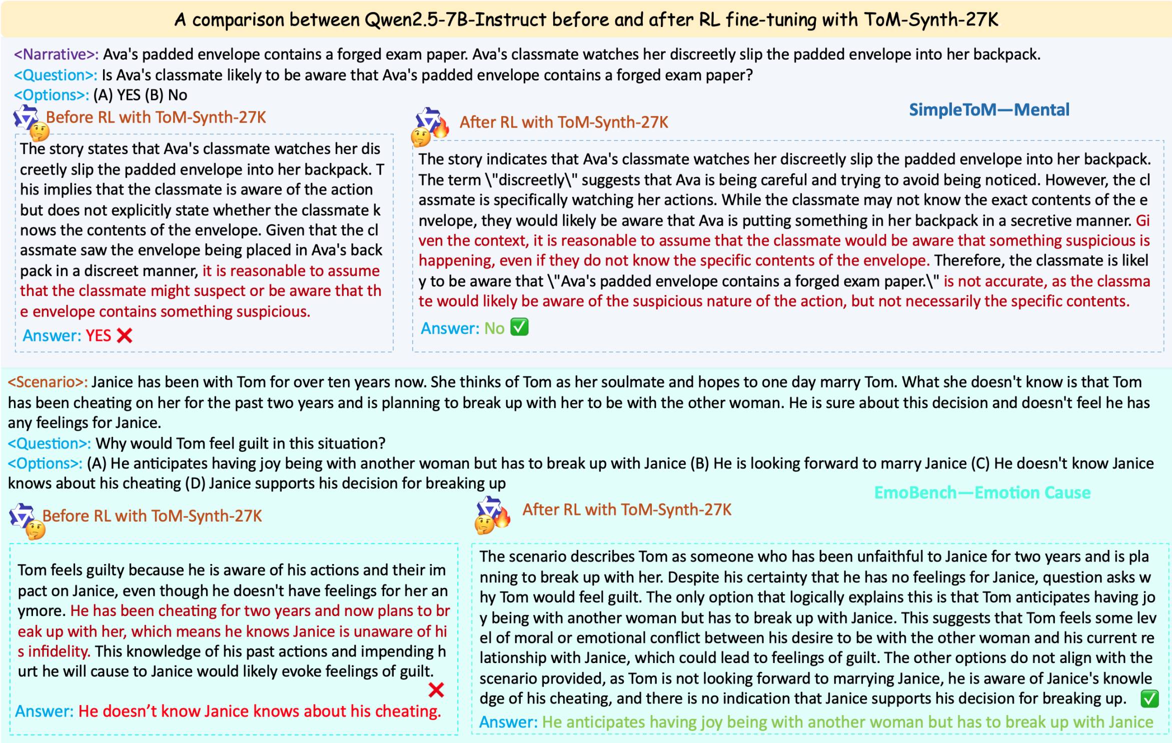

Figure 10: Comparison of Qwen2.5-7B-Instruct's reasoning trajectories before and after RL with ToM-Synth-27K. Upper: A Mental state example from SimpleToM where after RL training the model correctly distinguishes between observing an action and knowing its specific contents. Lower: An Emotion Cause example from EmoBench where after RL training transforms superficial causal attribution into systematic analysis of conflicting emotional motivations.

ensure sufficient optimization of the policy model. The sampling temperature is fixed at 1.0. Regarding the GRPO algorithm, the group size is set to 8. For the 7B, 8B, and 32B models, we set the global batch size to 64. For the 72B model, we set the global batch size to 128.

C.2 Reference Models

To contextualize the capabilities of our trained foundation models, we select a diverse set of representative and advanced models to serve as performance references. Specifically, we include four open-source models: Doubao-1.5-pro-32k-character-250228 (Doubao Team, 2025), Deepseek-v3-1-250821 (DeepSeek, 2025), Qwen3-max, and Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025); as well as two proprietary models: GPT-4o-2024-11-20 (Hurst et al., 2024) and GPT-5-2025-08-07 (OpenAI, 2025).

C.3 Evaluation Benchmark Data Distribution

As shown in Table 3, we present the number of data instances for each benchmark used in our evaluation.

C.4 Case Study

As shown in Figure 10, we present additional reasoning trajectory analysis for the Qwen2.5-7B-Instruct model before and after RL training with ToM-Synth-27K.

D Synthetic Data Examples

We present two additional visualizations of the synthesized data. Figure 11 illustrates a case involving mixed emotions in a **school bullying** social situation, while Figure 12 illustrates a case involving the link between knowledge access and pretend play in a **teacher-student interaction**.

<Dialogue>: Jean: Amanda, I saw what happened in the hallway. Are you okay?
Amanda: I... yeah. I mean, Jeffrey finally got suspended for what he did to me. The principal called my mom.
Jean: That's good news, right? He can't bother you anymore.
Amanda: I guess. I just... I keep thinking about how everyone was watching when he pushed my books. And now they'll all know I'm the one who reported him.
Jean: You did the right thing though.
Amanda: I know. I wanted this. I've wanted him to face consequences for months. But now I can't stop shaking, and I almost wish I could just disappear. What if his friends come after me next?
Jean: We'll stick together. You're not alone in this.
Amanda: Thanks. I just didn't expect to feel this way after finally getting what I wanted.
<Question>: Based on Amanda's responses about Jeffrey's suspension, what best describes her emotional experience?
<Options>: A. She feels conflicted, experiencing both satisfaction about the outcome and anxiety about potential consequences ✅
B. She feels purely relieved that the bullying situation has been resolved
C. She feels regretful about reporting Jeffrey and wishes she hadn't done it
D. She feels angry that the school took so long to take action against Jeffrey

[**Mental State:** Mixed Emotions **Situation:** Education School Bullying **Format:** Dialogue]

Figure 11: Visualization of a synthesized data based on: **[Mental State: Mixed Emotions | Situation: Education School Bullying | Format: Dialogue]**. The red text highlights the complex behavioral cues (shaking, fear of retaliation) and metacognitive reflection that justify the inference of mixed emotions (satisfaction vs. anxiety) following a bullying incident.

<Dialogue>: Lisa (teacher): Today we'll do creative drama activities. Jordan, you'll play a scientist discovering something amazing.
Jordan: Cool! What am I discovering?
Lisa: Actually, let's keep it a surprise. Emily, come whisper to me what Jordan should discover.
Emily (whispering to Lisa): A new element that makes things float!
Lisa: Perfect. Emily, you'll be Jordan's lab assistant who already knows about this discovery. Jordan, just improvise - pretend you're in a lab making a breakthrough.
Jordan (acting): I'm mixing chemicals... something is happening... the beaker is... um... changing colors?
Emily (acting): Dr. Jordan! Look at the objects around the beaker!
Jordan (confused, breaking character): Wait, what am I supposed to notice? I don't know what I discovered.
Lisa: Interesting. Emily, why do you think Jordan is struggling while you're not?
<Question>: Based on the interaction, why is Jordan unable to effectively pretend-play the scientist discovering the floating element while Emily can play along?
<Options>: A. Jordan was not told what the discovery is, so cannot incorporate that specific knowledge into the pretend scenario ✅
B. Jordan prefers to follow scripts rather than improvise during drama activities
C. Jordan feels uncomfortable being the lead role and wants Emily to take over
D. Jordan lacks acting experience compared to Emily who has practiced drama before

[**Mental State:** Knowledge-Pretend Play Links **Situation:** Education Teacher-Student **Format:** Dialogue]

Figure 12: Visualization of a synthesized data based on: **[Mental State: Knowledge-Pretend Play Links | Situation: Education Teacher-Student | Format: Dialogue]**. Red text highlights the causal link between the lack of access to specific information ("Jordan was not told") and the inability to perform the corresponding pretend action.

Theory of Mind (Total: 11,157)					
ToMi	1,000	Hi-ToM	600	ExploreToM	1,000
ToMbench	2,816	SimpleToM-Mental	1,147	SimpleToM-Behavior	1,147
SimpleToM-Judgement	1,147	ToMATO	1,000	OpenToM	1,300
Social & Emotional Intelligence (Total: 2,400)				*IQ-Related Reasoning* (Total: 458)	
SocialIQA	2,000	EmoBench	400	AIME 2025	30
				GPQA Diamond	198
				AGIEval	230
					Grand Total: 14,015

Table 3: Evaluation Benchmarks Overview