RoleToM: Eliciting Theory-of-Mind Abilities through Role-playing in Large Language Models

Anonymous ACL submission

Abstract

The ability to attribute others' mental states, known as Theory-of-Mind (ToM), is a cornerstone of social intelligence. While large language models (LLMs) have exhibited impressive performance at various tasks including role-playing, their ToM reasoning capabilities remain limited and unreliable compared to humans. Meanwhile, the potential of intuitively leveraging role-playing for enhancing social cognition remains largely unexplored. To bridge this gap, we pioneer the investigation into how role-playing influences LLMs' Theory-of-Mind capabilities. We introduce RoleToM, a exploratory approach that integrates step-by-step reasoning with role-playing, demonstrating superior ToM abilities compared to perspective-taking and Chain-of-Thought methods alone. Additional experiments including ablation study of role-playing and fine-tuning Llama 3.1-8B-Instruct on RoleToM-generated data, showed that structured first-person simulation can effectively improve LLMs' ToM capabilities and generalize across different scenarios. We hope the role-playing methodology opens potential avenues for further applications and research in LLMs' social cognition and intelligence.

1 Introduction

Theory-of-Mind (ToM) refers to the ability to understand the mental states of others, including their intentions, beliefs, desires, and emotions. This cognitive capacity is fundamental to social interactions, enabling individuals to predict others' behavior, interpret emotions, and engage in effective communication (Premack and Woodruff, 1978; Wimmer and Perner, 1983). Research on ToM is significant not only for understanding human social cognition but also for advancing fields such as psychology and cognitive science (Williams et al., 2022). With the emergence of large language models (LLMs), this area has gained unprecedented attention, as it intersects with computational linguistics and human-computer interaction, fundamentally shaping how we approach social intelligence mechanisms in artificial systems (Kosinski, 2023; Ullman, 2023; Shapira et al., 2024).

ToM is often implicit in complex social situations and requires careful analysis, including understanding character relationships and events ocurrence (Apperly, 2012), demanding sophisticated approaches that can handle nuanced interpretations and multi-step analysis. Chain-of-Thought (CoT) (Wei et al., 2022) is a versatile technique that has significantly enhanced reasoning abilities in general contexts. Through the lens of Gardner's Multiple Intelligence Theory (Gardner, 2011), this method facilitates the conversion of LLMs' inherent linguistic competence into enhanced logical-mathematical reasoning abilities. However, it still faces issues on vagueness, including shortcut trapping, hallucinations and information loss, particularly when processing complex scenarios with alternating details and belief states (Chu et al., 2023; Feng et al., 2024).

Meanwhile, numerous studies have demonstrated LLMs' remarkable proficiency in role-playing tasks and character simulation (Wilf et al., 2024; Shao et al., 2023). This success has inspired researchers to explore character-based approaches for enhancing LLMs' reasoning capabilities. Notably, SIMTOM (Wilf et al., 2024) introduced a perspective-taking workflow that generates characters' beliefs about story events to answer questions from the character's perspective. While perspective-taking shows promise for enhancing ToM capabilities, psychological research reveals its limitations (Zhang et al., 2012) that although first-person perspective advantages exist in predictive reasoning, perspective-taking neither accelerates reasoning acquisition nor affects cognitive processing efficiency.

To draw an analogy with how humans approach

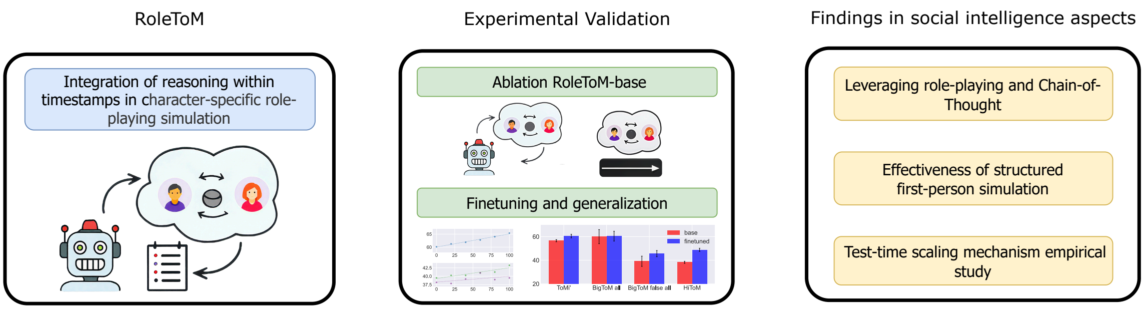

Figure 1: The framework on investigating how role-playing affect on Theory-Of-Mind abilities of LLMs

such complex ToM problems, we would typically immerse ourselves in a role within a scenario, focus on the transitions of events, and distinguish between known and unknown facts (Baron-Cohen et al., 1985). This has inspired us to combine role-playing and reasoning approach, where the LLM plays the role of a character in a story and gradually updates its knowledge through analyzing immediate events. The advantage of this method is that by adopting a first-person perspective, such as "I am xxx," the LLM can intuitively try to appropriately understand and elicit the differing beliefs between characters, observe limitations in its own knowledge, and hopefully detect shifts in key information in time.

Our framework is illustrated in Figure 1. The designed process-oriented role-playing method, named **RoleToM**, outperformed vast majority of existing baselines, which indicates that our exploration of first-person perspective has yielded some positive outcomes. We analyzed some counterintuitive cases and the variance of RoleToM outputs in experiments section and Appendix G. To evaluate to what extend it can simulate well generally, we further conducted ablation studies under settings of RoleToM submodules. Furthermore, we fine-tuned a small model Llama-3.1-8B-Instruct (Dubey et al., 2024) on RoleToM-generated data to validate its effectiveness. It resulted in a slight performance boost and stable out-of-domain generalization, which further highlights the potential of the RoleToM method. Besides, we view this framework as an exploratory trial, given the uncertainties in role-playing design schemes and the lack of standardized evaluation metrics. We point it out that a systematic evaluation of LLMs' role-playing abilities on random stories is essentially needed to gain comprehensive understanding on any simulation workflow (Wang et al., 2024), including our RoleToM. To the best of our knowledge, however, existing benchmarks only evaluate models on predefined stories and questions, which falls short of our intended purpose. We leave this as a direction for future work.

In summary, the contributions of this paper are three-fold:

- We raised the problem of how role-playing influences the Theory-of-Mind (ToM) abilities of LLMs and pioneered the investigation into this topic. Our framework integrating Chain-of-Thought (CoT) and role-playing showed that structured first-person simulation could pave the way towards advancing social intelligence of LLM-powered agent.
- We proposed RoleToM, an exploratory workflow that splits the events along a timeline, performs multi-stage role-playing and updates character' s perspectives. The result demonstrated the enhancement on ToM benchmarks compared with perspective-taking and CoT alone.
- We leveraged RoleToM-generated data to fine-tune a small model, achieving improved zero-shot ToM reasoning without human annotations or prompt engineering, which ascertains whether role-playing perspective maintains task-specific generalization in resource-constrained settings and consolidates the effectiveness of RoleToM.

2 Related Works

2.1 Theory-of-Mind on Large Language Models

LLMs have demonstrated unprecedented capabilities in natural language processing, driving substantial progress in simulating human-like behaviors and language (Kosinski, 2023; Bubeck et al., 2023; Sclar et al., 2022; Tang and Belle, 2024). Despite their advancements, achieving robust Theory-of-Mind (ToM), social intelligence, and commonsense reasoning in LLMs remains a significant AI challenge (Ullman, 2023).

Additionally, there is debate about the various reasons why LLMs may fail to accurately demonstrate ToM, including the lack of long-term contextual memory, limited understanding of complex social situations, and biases or errors in the model's reasoning processes (Ullman, 2023). Some argue that ToM capabilities are merely advanced pattern recognition and statistical inference (Sap et al., 2022), others suggest that the complexity and adaptability of LLMs may hint at a form of proto-consciousness, raising profound questions about the nature of intelligence and the boundaries between artificial and human cognition (Kosinski, 2023; Strachan et al., 2024; Hagendorff, 2024). (Trott et al., 2023) indicated that human belief reasoning capabilities stem from both language experience (which can be learned by LLMs) and innate social intelligence (which is difficult to model and not inherent in LLMs), while raising cognitive philosophical debates about whether existing evaluation methods can serve as valid benchmarks for proving LLMs' ToM capabilities.

2.2 LLMs-based Agents for Role-Playing

Agent-based Modeling (ABM) has evolved in tandem with neural language systems and their human value alignment. Early work (Zhang et al., 2018) laid the foundation for persona-based dialogue systems, focusing on maintaining consistent personality traits using datasets like Persona-Chat. LaMDA (Thoppilan et al., 2022) sparked the first discussion that consciousness might have emerged in language models. Researchers are now exploring whether LLMs exhibit emergent properties akin to consciousness, such as self-awareness, intentionality, and the ability to reflect on their own outputs (Sclar et al., 2023; Liu et al., 2023). Along with those investigations in different personalities on LLMs (Tan et al., 2024), role-playing with LLMs has evolved significantly, transitioning from basic persona-based dialogue systems to sophisticated, multi-dimensional character simulations (Shanahan et al., 2023). It was revealed that models like GPT-4 and Claude-3 exhibit stronger performance in third-person belief inference compared to first-person scenarios (Suzgun et al., 2024). ToMATO (Shinoda et al., 2025) simulates the formation and expression of different perspectives in human conversations through information asymmetry, introspective cues, and subsets of contradictory beliefs.

3 RoleToM

3.1 Method Overview

Humans possess a natural ability to understand others' mental states, often by mentally "stepping into their shoes" and simulating their perspective given a sequence of events (Wellman et al., 2001). Drawing inspiration from this cognitive process, we introduce RoleToM, a novel method designed to enhance Theory-of-Mind capabilities of LLMs. RoleToM elicits ToM reasoning by prompting LLMs to simulate in story comprehension through a structured, first-person role-playing experience. This approach consists of two core components in functionality: a simulator and a reasoner, which work sequentially to process narrative information and infer mental states.

3.2 Simulator

The simulator component is responsible for constructing a first-person and dynamic experience for the narrative story. It guides the LLM to role-play as the protagonist, meticulously processing the story event by event to build an internal representation of the character's evolving knowledge. This process involves the following key steps:

Temporal Event Segmentation: The input story is first segmented into discrete events $(E_1, E_2, \ldots, E_n)$ based on their temporal order. This ensures that the LLM processes information sequentially and mirrors how events unfold in timestamps. In our experimental setting, event segmentation was conducted by splitting numbered markers within the text beforehand. However, if stories do not have explicit markers, we will integrate this segmentation process into a separate prompt and executing event partitioning in an independent message.

Iterative Belief State Update: For each event E_t in the chronological sequence, the simulator prompts the LLM to role-play as the protagonist in question. When a question pertains to a fact and involves no characters, the LLM then simulates the role of an oracle. This process begins with the LLM identifying key new information introduced, focusing particularly on changes concerning characters,

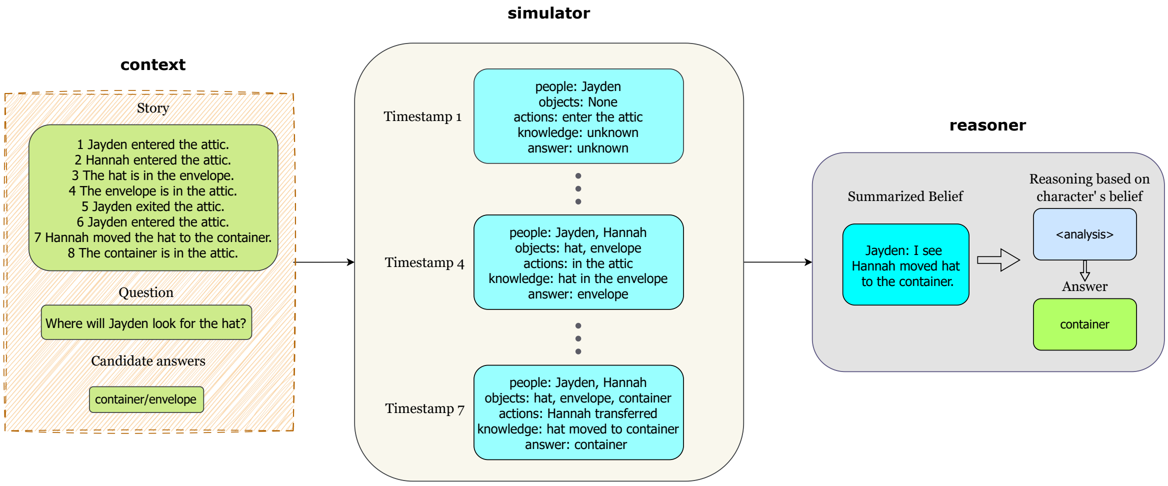

Figure 2: One example for RoleToM workflow

entities, locations, and their actions. Based only on the information available up to and including this current event, the LLM then explicitly articulates the protagonist's understanding and perception of the situation. Finally, this articulated perspective is used to formulate an updated belief state B_t, which reflects the protagonist's cumulative knowledge and perspective at timestamp t. This formulation involves integrating new information from E_t with the previous belief state B_{t-1} (where B_0 represents an initial empty or context-setting state), thereby tracking the crucial shifts in the protagonist's awareness.

3.3 Reasoner

The reasoner leverages the rich, temporally-grounded belief states generated by simulator to perform explicit ToM reasoning and answer targeted questions. It also employs a role-playing instruction for the LLM, guiding it through two main operational steps.

Core Belief Summarization: After role-playing from simulator, the LLM is prompted to still embody the protagonist' s role and summarize the core belief B_{core}. In the case of factual questions, the reasoner just omits the former step. This summarization serves to achieve contextual condensation by reducing potentially extensive and detailed context into a concise representation of the protagonist's most relevant knowledge and beliefs directly pertinent to the ToM query. Furthermore, this process enhances clarity, mitigating potential confusion that might arise from processing overlapping keywords present in the long raw sequence of belief states.

Question Answering: The reasoner proceeds to perform reasoning and question answering based on the core belief obtained before. Specifically, the LLM is tasked to deduce what the protagonist would know, believe, or feel in response to the query based on the core belief obtained before and generate an answer to the ToM question that is logically consistent with the whole story and simulation.

3.4 Example

Figure 2 provides a step-by-step illustration of RoleToM's application to a sample story. The narrative is first segmented by timestamps (simulator step 1). For each timestamp, the simulator generates a formatted belief description from the protagonist's perspective, reflecting their updated knowledge (simulator step 2) at t. The reasoner synthesizes these evolving beliefs into a core belief summary B_{core} (reasoner step 1). Subsequently, the reasoner derives the answer to the ToM question (reasoner step 2). The specific prompts used to guide the LLM in both the simulator and reasoner components are detailed in Appendix F.

4 Experiments

4.1 Overview

To dive into how Theory-of-Mind ability gain from structured role-playing and demystify the influence behind this method, we conduct RoleToM on three benchmarks with ablations on simulator and reasoner perspectively, and finetuned on Llama-3.1-8B-Instruct with RoleToM-generated data. Moreover, we investigate the prevailing scaling law on ToM scenario by measuring performance on increasing gradually tokens.

4.2 RoleToM Results

4.2.1 Benchmarks

We evaluated RoleToM on three benchmarks.

ToMi (Le et al., 2019) is a dataset inspired by classic Sally-Anne test of false belief created by using a stochastic rule-based algorithm that generates stories involving two participants, an object, and a set of locations or containers where the object is moved. Each story comes with questions that have two different location answers. Here we use an updated version (Sap et al., 2022) with size 1000 which corrects mislabeled second-order questions and clarifies the reference of container locations.

BigToM (Gandhi et al., 2024) is a benchmark designed to assess a more comprehensive Theory-of-Mind reasoning capabilities of LLMs. BigToM encompasses question types in forward and backward, action and belief, and true or false information separately. We only evaluate on forward split with size 800 of the whole dataset as well as half of the false beliefs samples within it.

HiToM (Wu et al., 2023) is a benchmark designed to evaluate the higher-order Theory-of-Mind reasoning capabilities, which involving 600 samples of recursive reasoning about others' beliefs and thereby assessing the ability to understand complex mental states.

4.2.2 Baselines

ZeroShot: The base prompt without any explicit constrains or tools. We added the phrase "do not include any other words" at the end of the question.

Chain-of-Thought (Wei et al., 2022): Chain-of-Thought (CoT) is a well-established technique that enhances LLMs reasoning abilities by encouraging them to generate intermediate reasoning steps before arriving at a final answer.

PercepToM (Jung et al., 2024): It first infer characters' perceptions from an input context and aids in perception-to-belief inference through a perspective context extraction step, which isolates the context perceived by the target character using a simple string-matching algorithm. Finally, LLMs answer ToM questions based on the isolated context.

SIMTOM (Wilf et al., 2024): This method first filters the context to include only information known to the character in their perspective. Subsequently, it prompts the LLM to answer ToM questions based on this filtered context. We provide detailed comparison in Appendix B.

4.2.3 Overall Performance

The evaluated models span different scales and accessibility: Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Dubey et al., 2024), GPT-4 (Achiam et al., 2023), Qwen2.5-7B and Qwen2.5-72B (Yang et al., 2024), and Deepseek-R1 (Guo et al., 2025). We deployed the open source models by vLLM (Kwon et al., 2023) offline and called OpenAI or Aliyun API to access the closed source models separately.

Our experimental results demonstrate that RoleToM consistently outperforms most existing baselines across diverse models and benchmarks. The main experiments affirm RoleToM's superiority and showcase substantial improvements. RoleToM improves performance by approximately 2.1% up to a significant 42.6%. On larger models except for reasoning model R1, like Llama-3.3-70B-Instruct and GPT-4, RoleToM achieves the highest accuracy on most benchmarks, notably improving ToM' performance by approximately 4-10 percentage. The enhanced efficacy observed on all baselines in these models likely stems from their more developed context understanding and instruction-following capabilities. On smaller models like Qwen2.5-7B, RoleToM maintains superior performance with 75.20% on ToMi, while other methods show performance degradation to some extent. The exclusive deep reasoning model, R1, achieved top performance despite marginal improvement, thereby demonstrating the significant advantages conferred by its capacity for profound analysis and sophisticated multi-step reasoning.

While the ablation of RoleToM shows reduced performance, it still outperforms perspective-taking and zero-shot methods in overall, demonstrating that role-playing is inherently valuable for ToM reasoning. Furthermore, we evaluated the reliance on simulator accuracy by having an independent GPT-4o score simulated processes and manually assessing 20% of samples from each benchmark. See Appendix C for details please.

However, some unconventional results were observed with ZeroShot baseline, on HiToM benchmark, and Qwen2.5-72B model. For instance, PercepToM underperforms ZeroShot sometimes due to its original, exclusive design for ToMi dataset or Llama' s occasional output in repeating format. Concerning HiToM benchmark, which is character-

Model	Benchmarks	Methods						
		ZeroShot	CoT	PercepToM	SIMTOM	RoleToM	w/o simulator	w/o reasoner
Llama-3.1 -8B-Inst.	ToMi	60.60	64.40	53.30	56.80	**69.90**	64.60	66.40
	BigToM all	59.88	52.38	36.75	46.25	**80.00**	67.50	71.50
	BigToM false all	35.50	40.50	24.25	30.75	**66.75**	45.00	46.00
	HiToM	37.50	50.00	12.67	14.83	**53.50**	51.00	50.50
Llama-3.3 -70B-Inst.	ToMi	71.40	71.50	66.70	77.00	**81.20**	72.50	79.70
	BigToM all	88.00	91.13	69.25	91.75	**93.38**	86.63	89.63
	BigToM false all	79.75	84.75	55.50	86.25	**88.25**	79.00	87.75
	HiToM	38.75	63.33	52.17	62.33	**67.17**	61.33	66.83
GPT-4	ToMi	70.50	76.30	61.50	80.60	**81.70**	71.90	81.90
	BigToM all	88.13	90.00	70.75	91.50	**94.50**	86.00	93.63
	BigToM false all	78.50	82.75	58.50	87.00	**91.00**	75.50	85.25
	HiToM	54.67	11.67	35.00	60.33	**60.50**	60.33	60.17
Qwen2.5-7B	ToMi	62.30	61.70	68.10	59.30	**75.20**	71.80	72.00
	BigToM all	71.00	68.38	50.75	25.38	**72.50**	72.25	72.13
	BigToM false all	48.00	49.50	31.25	9.75	**55.00**	51.50	53.50
	HiToM	47.08	34.50	46.50	27.33	**56.33**	49.00	52.83
Qwen2.5-72B	ToMi	68.70	73.30	73.70	**86.50**	84.80	76.60	81.60
	BigToM all	79.63	90.13	68.38	82.25	**83.63**	78.25	83.25
	BigToM false all	65.75	83.50	50.75	69.50	**70.00**	63.75	64.25
	HiToM	57.17	25.83	48.67	**63.50**	60.83	61.00	60.17
DeepSeek-R1	ToMi	90.60	89.70	83.30	95.90	**96.80**	94.80	96.00
	BigToM all	77.88	84.75	78.63	87.38	**91.75**	90.88	90.38
	BigToM false all	74.75	80.50	61.75	81.50	**89.25**	85.00	79.75
	HiToM	64.00	64.83	67.17	68.50	**70.00**	69.50	66.17

Table 1: Overall performance. Bold indicates the best performance.

ized by its emphasis on specially set high-order beliefs, RoleToM is primarily designed to accurately simulate the unfolding of directly observable facts (*i.e., 0th-order beliefs*) rather than to specialize in processing such massive and confusing entities. This design choice is potentially more effective for benchmarks like ToMi and BigToM in terms of filtering interference noise and discerning key information within intricate scenarios. When performing Chain-of-Thought reasoning on numerous items within a story, a model becomes vulnerable to being led astray by the intricate details. These observations are discussed in detail in Appendix G. It is important to note that these explanations are specific to these anomalous cases and do not represent a general critique of overall efficacy of the methods involved.

Additionally, Figure 3 showed the instability of perspective-taking and CoT compared with our methods on the first five models in our experiments, where error bars indicate the variation. Findings from cognitive psychology align with our observations in LLM experiments, suggesting that perspective-taking approaches alone may yield less stable or incomplete improvements in ToM capabilities, and possibly depend on specific models' underlying capabilities in some cases.

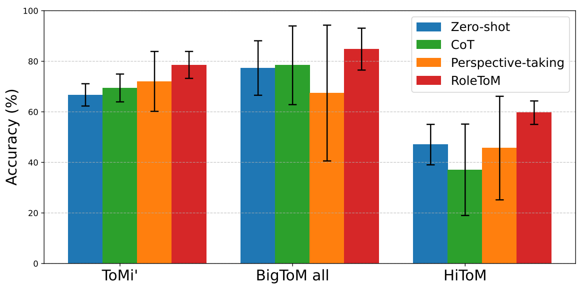

Figure 3: Performance comparison between zero-shot, perspective-taking and CoT across three benchmarks

4.3 Finetuned Experiments

4.3.1 Experiments Overview

In this section, we investigate fine-tuning a small model, Llama-3.1-8B-Instruct, using response data generated by RoleToM from Section 3.1. We collected responses within correct samples pool across four models except Llama-3.1-8B-Instruct and R1. For training data preparation, we specifically utilized the belief summarization from the simulator and responses from the reasoner as our primary dialogue pairs. Training dataset statistics are in Appendix A. We excluded detailed knowledge updates from individual timestamps, ensuring that training data effectively encompasses the story's progression and preserves key information while maintaining temporal consistency.

We conducted three experiments to evaluate the generalization capabilities of models fine-tuned on RoleToM-generated data: (1) the impact of varying proportions of training data from different sources, (2) zero-shot transfer capabilities across unseen benchmarks, and (3) robustness of performance within each benchmark. We finetuned them on $1\times$NVIDIA A100 80G GPUs for 3 epochs using LoRA methods with $rank = 8$ and $\alpha = 16$. The duration of a single fine-tuning process is approximately 10-30 minutes.

4.3.2 Impact of Training Data Composition

This experiment investigates how varying proportions of different data sources affect model performance. We constructed multiple training sets by combining two data categories: correct outputs generated by RoleToM and answers labeled via a ZeroShot method. We set the ratio of two types of data to quintiles from 0 to 100% to create several training dataset configurations. The model was then fine-tuned on each mixed dataset and evaluated across all three benchmarks.

4.3.3 Cross-Benchmark Evaluation

We iteratively selected RoleToM-generated data from two benchmarks for training while reserving the third benchmark for testing. This rotation was performed three times, ensuring each benchmark served as the test set once while the other two provided training data. We set the temperature for LLM inference to 0.3 and ran the process three times to obtain error bars.

4.3.4 Random Split Evaluation within Benchmark

This experiment assesses the reliability and consistency of our fine-tuning approach through within-benchmark evaluation. For each benchmark, we employed a random stratification protocol, allocating 70% of the data for fine-tuning and 30% for testing. We conducted three iterations of random splits individually to compare the base and fine-tuned model performance with temperature at 0, with error bars shown for statistical significance.

4.3.5 Results

The results demonstrate fine-tuning with RoleToM-generated data demonstrates consistent, albeit modest and transferability capabilities.

Figure 4 shows that increasing the proportion of RoleToM data generally leads to better performance, particularly for the BigToM all metric. The cross-benchmark evaluation in Figure 5 reveals that the fine-tuned model consistently outperforms the base model across all benchmarks, with improvements ranging from 2-6 percentage points. Figure 6 further confirms steady performance gains, while showed slightly small enhancement on ToMi and HiToM. While the improvements are relatively small in magnitude, they are notably consistent across different evaluation settings, suggesting that RoleToM-based fine-tuning provides reliable, if incremental, enhancements to model performance.

4.4 Inference-time Scaling trial

We investigate the relationship between inference-time tokens consumption (Snell et al., 2024) and model performance, comparing the scaling behavior of RoleToM against traditional CoT approaches on Llama-3.1-8B-Instruct. We conducted experiments on ToMi with token lengths of approximately 128, 256, 512 and 1024. We set 4 tokens for zero-shot answers only. Given that full RoleToM produced an average of 1029 tokens, we considered this output length representative of a full-scale generation (conceptually a 1024 token point). For subsequent experiments, we then set maximum token lengths to 256 and 512 via vLLM inference for rerunning. Figure 7 show that increasing the number of inference tokens on Llama-3.1-8B-Instruct correlates positively with performance improvements. In contrast, similar increases in token allocation for CoT methods yield relatively minimal performance gains, suggesting that merely extending the reasoning chain length does not enhance ToM capabilities as effectively as the character-specific simulation. Results for four Llama and Qwen models presented in Appendix D exhibit a similar tendency.

5 Discussion

Many studies debate whether Theory-of-Mind can spontaneously emerge in LLMs (Kosinski, 2023; Ullman, 2023; Sap et al., 2022), but the underlying mechanisms remain largely unexplained due to LLMs' black-box nature and diverse yet complex training methods (Yao et al., 2024).The most advanced reasoning LLMs, such as OpenAI-o1 and Deepseek-R1, have made remarkable strides in handling complex text tasks and mathematical prob-

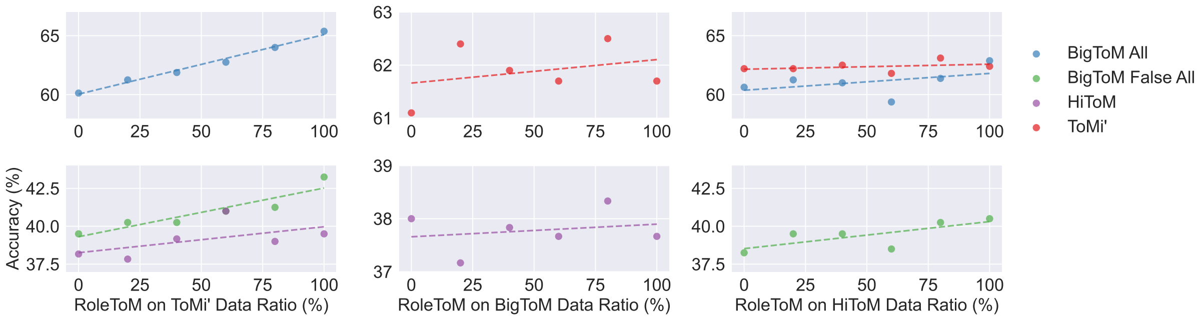

Figure 4: Impact of Training Data Composition

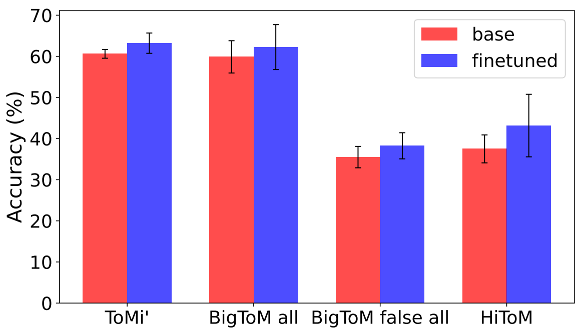

Figure 5: Cross-Benchmark Evaluation

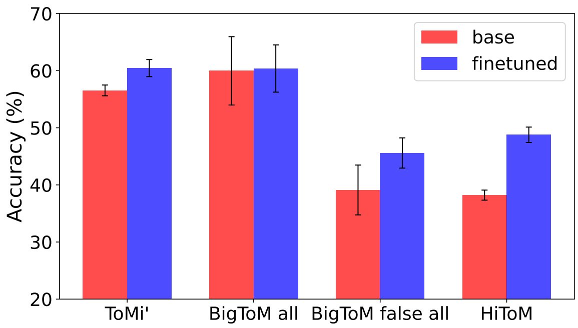

Figure 6: Random Split Evaluation within Benchmark

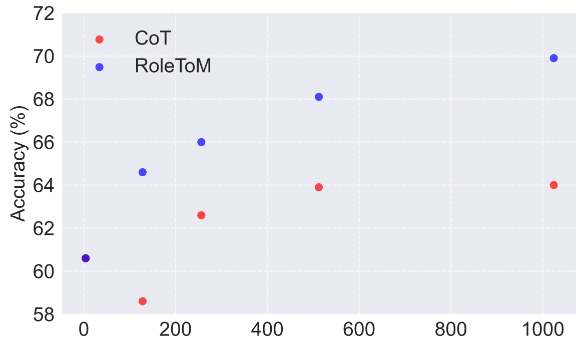

Figure 7: Tokens scaling impact on CoT and RoleToM for Llama-3.1-8B-Instruct

lems, even surpassing most human experts (Didolkar et al., 2024). Yet, these models still require carefully curated and evaluated data, highly technical training schemes, external tools, reasoning strategies, etc. The excellent performance of these models in engineering, mathematical reasoning, and coding does not robustly generalize to social intelligence (Ullman, 2023). Additionally, no research has proven that scaling up during such training can enable LLMs to spontaneously develop robust ToM capabilities. In specific scenarios, LLMs fail at critical deduction steps in long narrative texts even with correct information (Sap et al., 2022). For example, they might acknowledge that a character is in a room while denying a publicly known event's occurrence. And the exact point at which the failure occurs, leading to deviations in ToM performance, has not been thoroughly investigated (Jung et al., 2024). Besides, more discussion concerning the impact of biases introduced by training methods and especially overlooked data composition on LLMs' social intelligence should be given greater priority. While a comprehensive account of their underlying principles and mechanisms may be exceedingly difficult to provide, offering insights upon this area would be invaluable for researchers in constructing new benchmarks and designing novel methodologies and techniques.

6 Conclusion

In this paper, we explore how role-playing influences LLMs' Theory-of-Mind capabilities. Through RoleToM approach, ablation studies and fine-tuning experiments, we demonstrate the advantages of our approach in inference time over standalone role-playing and third-person reasoning methods. Though challenges remain in developing universal role-playing frameworks, we provide valuable insights into how role-playing influences LLMs' cognitive abilities and suggest promising directions for future AI development in social intelligence.

7 Limitations

First, a framework that specifically evaluates the simulation of LLM in non-preset or arbitrary social scenarios should be prior to quantitatively analyze the effect of our method. As far as we know, the existing benchmarks focus on evaluating preset story templates or scoring based on game strategies. These works are limited by the manually designed simulation templates and application scenarios, and fail to evaluate the simulation effect of arbitrary stories. In particular, there is no benchmark for evaluating simulation in three story scenarios of our evaluation datasets. Multiple variants of RoleToM can be designed based on ours to observe the performance of Theory-of-Mind, but a more comprehensive and general framework needs to be introduced to the community for reliable and systematic role-playing evaluation. In our approach, we emphasized that developing a universal role-playing framework applicable to all scenarios for current generative LLMs is very challenging (Chen et al., 2024b,a; Xu et al., 2025; Hao et al., 2024). Current limitations exist at both ends of the spectrum, for instance, smaller language models often struggle with basic instruction following (Kasneci et al., 2023), while larger models face issues related to training data contamination (Zhao et al., 2024). While achieving human-level performance remains beyond our current capabilities, our objective is to raise and highlight this topic and, as an example, introduce an intuitive, heuristic role-playing approach to investigate its influence the ToM abilities of LLMs.

Second, our approach consumes more tokens (list vs. tokens per baseline), which means it may be a challenge for applications with high concurrency requirements. In addition, RoleToM does not involve explicit LLMs agent dialogue, although some data samples may involve dialogue, but this is not the focus of the dataset construction, as well as the experimental practice of mental ability does not explicitly distinguish between dialogue and non-dialogue if needed. The tradition is to align dialogue with public events and assume that every involved character can grasp this fact.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ian A Apperly. 2012. What is “theory of mind”? concepts, cognitive processes and individual differences. *Quarterly journal of experimental psychology*, 65(5):825–839.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024a. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *arXiv preprint arXiv:2405.12205*.

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

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36.

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

Howard E Gardner. 2011. *Frames of mind: The theory of multiple intelligences*. Basic books.

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

Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121.

Qianyue Hao, Fengli Xu, Yong Li, and James Evans. 2024. Ai expands scientists' impact but contracts science's focus. *arXiv preprint arXiv:2412.07727*.

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*, Miami, Florida, USA.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

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

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

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

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Melanie Sclar, Graham Neubig, and Yonatan Bisk. 2022. Symmetric machine theory of mind. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19450–19466. PMLR.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

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

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. *arXiv preprint arXiv:2501.08838*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

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*, pages 1–11.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2024. Belief in the machine: Investigating epistemological blind spots of language models. *arXiv preprint arXiv:2410.21195*.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil

Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

Weizhi Tang and Vaishak Belle. 2024. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models. *CoRR*, abs/2404.15515.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and 1 others. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

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

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Henry M Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763.

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, Singapore. Association for Computational Linguistics.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

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*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Jun Zhang, Trey Hedden, and Adrian Chia. 2012. Perspective-taking and depth of theory-of-mind reasoning in sequential-move games. *Cognitive science*, 36(3):560–573.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Dora Zhao, Jerone TA Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: measure dataset diversity, don't just claim it. *arXiv preprint arXiv:2407.08188*.

A Training dataset statistics

Model	ToMi	BigToM	HiToM
Llama-3.3-70B-Inst.	812	747	403
GPT-4	817	756	363
Qwen2.5-7B	752	580	338
Qwen2.5-72B	848	669	365
Total	3229	2752	1469

Table 2: Training dataset statistics

B Differences between RoleToM and SIMTOM

The fundamental differences between RoleToM and SIMTOM lie in their core design philosophies, operational complexities, and adaptive capabilities for reasoning tasks. SIMTOM employs a relatively straightforward, two-stage process: it first establishes a character's perspective and then generates an answer based on that viewpoint. This method relies on a fixed two-round prompt structure, attempts to capture all narrative developments in one go, and does not incorporate additional reasoning in its answering phase. In stark contrast, RoleToM implements a more sophisticated and dynamic methodology.

Design:

- SIMTOM only has two rounds of prompts, namely perspective-taking and answering. First, the perspective information of a certain role is obtained, and then the answer is given based on this. There is no additional reasoning technology designed for the second step of answering.

- The method of RoleToM is to combine role-playing and step-by-step reasoning. First, the story scene is split by timestamps, and then reasoning is performed at each time node, including the transformation of key objects, the characters involved, different scene information and instant answers (in order to track the connection with the original question). Finally, the plot of the whole story is summarized and key information is retained. The step-by-step reasoning here not only includes the update of each timestamp, but also includes orderly disassembly of problems, sorting out the analysis process and generating the final answer.

Prompt Structure:

- SIMTOM has only two fixed rounds, the key to SIMTOM is a single perspective-taking and attempts to capture all plot changes of the story at one time.

- RoleToM will change according to the complexity of the story, atomically perform multi-step role-playing, and additionally summarize the step-by-step process for reasoning.

The motivation behind RoleToM's more elaborate architecture is to better equip language models, particularly those with fewer parameters, to navigate and interpret complex information and high-order beliefs—challenges where simpler approaches might prove insufficient. Moreover, RoleToM's strategy of summarizing its detailed simulation process effectively mitigates potential issues such as excessive context length and information redundancy.

C Reliance on Simulator Accuracy

Since RoleToM's reasoning depends on the simulator's ability to correctly segment events and track belief updates, we conducted evaluation study on simulator' s accuracy. Specifically, we used an additional independent GPT-4o to score the processes simulated by the simulator and manually scored the accuracy of 20% of the random samples from each benchmark. Scores range from 1 to 5, and the human annotation procedure was done by one of the authors. We compared results of the two, showing that the accuracy and stability of the simulation process are close to the level of human scoring, but there are still differences. This shows that there is room for improvement within the simulator, but the conclusion that the model of step-by-step role-playing and then reasoning is effective still holds.

D Tokens scaling on Llama and Qwen series

Please see Figure 8 below.

E Probing Rectification

We reproduce the probing methodology to extract answers from SIMTOM framework. However, it required modifications due to several factors. While we ran evaluations on ToMi and BigToM, we encountered challenges when applying original

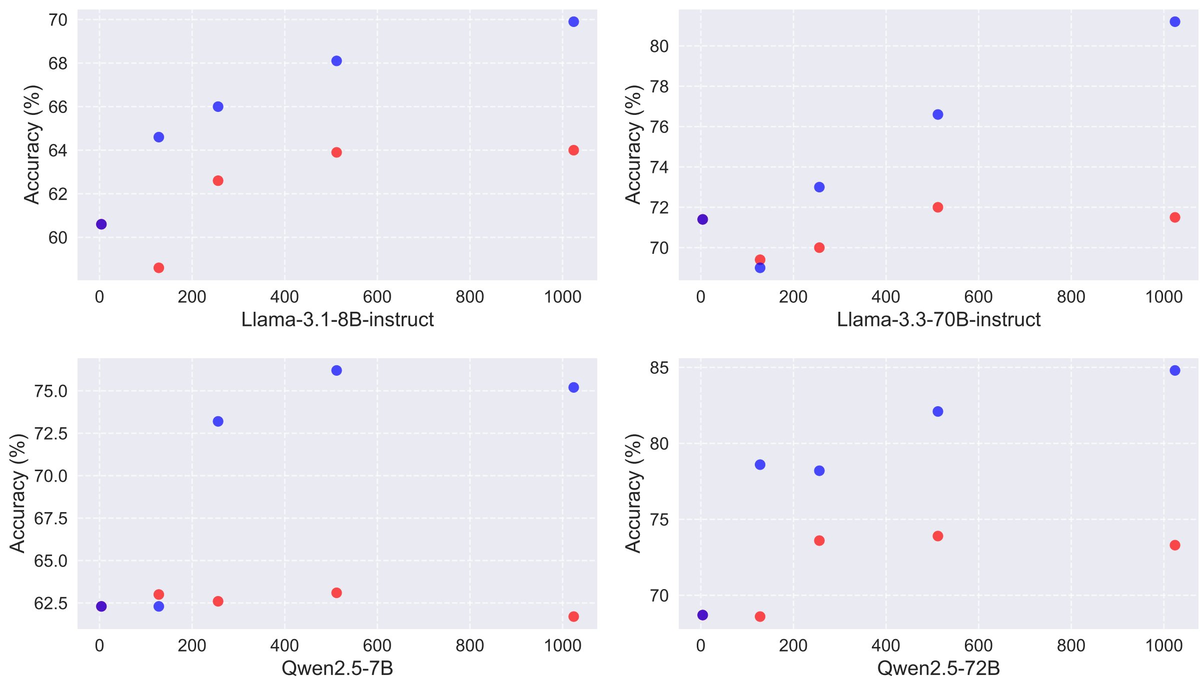

Figure 8: Tokens scaling impact on CoT and RoleToM

Model	Benchmarks	Judge	
		GPT-4o	Human
Llama-3.1-8B-Inst.	ToMi	3.99	4.05
	BigToM all	4.00	4.30
	false all	3.84	3.85
	HiToM	3.68	3.98
Llama-3.3-70B-Inst.	ToMi	4.56	4.65
	BigToM all	4.67	4.97
	BigToM false all	4.41	4.42
	HiToM	4.36	4.66
GPT-4	ToMi	4.58	4.65
	BigToM all	4.73	4.83
	BigToM false all	4.55	4.56
	HiToM	4.03	4.33
Qwen2.5-7B	ToMi	4.26	4.30
	BigToM all	3.62	3.92
	BigToM false all	2.75	2.76
	HiToM	3.32	3.62
Qwen2.5-72B	ToMi	4.74	4.84
	BigToM all	4.18	4.48
	BigToM false all	3.50	3.51
	HiToM	3.54	3.84
Deepseek-R1	ToMi	4.78	4.96
	BigToM all	4.61	4.98
	BigToM false all	4.56	4.86
	HiToM	4.34	4.78

Table 3: Accuracy score of simulated process in RoleToM

prompting methods to newer models. The original approach, which used "choose from the following" prompts and judged whether correct answers appeared in the model's complete output, proved inadequate for all models we evaluated. Some models tend to generate detailed analytical responses rather than direct answers, which included both choices appear in the long context and made the original detection method unreliable. For instance, when evaluating GPT-4 using ZeroShot, we observed an unusually high accuracy of 94% on ToMi with detailed analyses of each response. To address these issues, we adopted different strategies based on model accessibility. For closed-source models, we implemented a two-stage approach: first, obtaining the initial response to the question, and then requesting a concise final answer without additional commentary. For open-source models, we utilized the vLLM inference engine's GuidedDecodingParams method to constrain output formats, ensuring consistent and comparable responses. We mainly tested them between November 2024 and January 2025.

F Full Prompt of RoleToM

The note part is adapted from HiToM (Wu et al., 2023) prompt.

F.1 Template

Simulator:

- You are <name>.
You are presented with a record of events, described in time order by a bystander. This bystander has observed all events, including details that you may or may not know.
<Story>
<Question>
To minimize bias and prevent misleading conclusions, your task is to reconstruct your knowledge step by step before addressing the question directly. Follow the steps below:
(1) Divide the story by timestamps or clear time markers: Break the story into discrete events, noting the specific timestamp or sequence when each occurred.
(2) For each timestamp or event: Explain your knowledge available at that point in the story, while identifying any potentially irrelevant information.
Determine what your choice would be regarding the question at that specific time and explain your reasoning.
Note: Each agent only knows what they have observed directly or through communication and direct interaction. If not explicitly stated that something was unnoticed or unseen, an agent is assumed to know everything that occurs and every objects at a location from the time they enter until the time they leave.
By completing these steps, ensure your reasoning and conclusions are transparent and rooted in the temporal sequence of the events.
Please answer in the following format, completing all items ending with "_1" and adding more if necessary. Keep your words concise, but sufficient to restore the whole story:
{
    "timestamp_1":{
        "people involved": "person_1, ",
        "objects": "object_1, ",
        "actions": "action_1",
        "my knowledge": "belief_1",
        "answer and why": "reason_1, Answer:<option_1>"
    }
}
- <agent_full_knowledge>

Reasoner:

- Based on your knowledge, what' s your answer and reason? Please respond in the following structured format:
{{
    "question": "Restate the full question context here to ensure accuracy and avoid misunder-standings.",
    "reason": "Provide a clear and concise explanation for your answer, rooted in the knowledge and reasoning available."
    "answer": "Based on my perspective, the answer is <option>.",
}}
- Based on the previous conversation, summarize your response to the following question:
<Question>
Instructions:
Output one and only one of the candidate answers.
Do not include any additional words, explanations, or formatting. Only provide the chosen answer in plain text.

F.2 Example

Simulator:

- You are Jayden.
You are presented with a record of events, described in time order by a bystander. This bystander has observed all events, including details that

```
you may or may not know.
1 Jayden entered the attic.
2 Hannah entered the attic.
3 The hat is in the envelope.
4 The envelope is in the attic.
5 Jayden exited the attic.
6 Jayden entered the attic.
7 Hannah moved the hat to the
container.
8 The container is in the attic.
There is a Quesiton:
Where will Jayden look for the hat?
Choose from the following:
container, drawer
...
Please answer in the following format,
completing all items ending with "_1"
and adding more if necessary. Keep
your words concise, but sufficient
to restore the whole story:
{
    "timestamp_1":{
      "people involved": "person_1, ",
        "objects": "object_1, ",
        "actions": "action_1",
        "my knowledge": "belief_1",
        "answer and why": "reason_1,
        Answer:<option_1>"
    }
}
```

- ```
  {
      "timestamp_1":{
          "people involved": "Jayden",
          "objects": "attic",
          "actions": "Jayden entered the
          attic",
          "my knowledge": "Jayden is in
          the attic",
          "answer and why": "No information
          about the hat yet, Answer: N/A"
      },
      "timestamp_2":{
          "people involved": "Hannah",
          "objects": "attic",
          "actions": "Hannah entered the
          attic",
          "my knowledge": "Both Jayden and
          Hannah are in the attic",
          "answer and why": "Still no
          information about the hat,
          Answer: N/A"
      },
      "timestamp_3":{
          "people involved": "None
          directly",
          "objects": "hat, envelope",
          "actions": "The hat is in the
          envelope",
          "my knowledge": "The hat is
          inside the envelope, but location
          of envelope unknown to Jayden",
          "answer and why": "Jayden does
          not know the location of the
          envelope yet, Answer: N/A"
      },
      "timestamp_4":{
          "people involved": "None
          directly",
          "objects": "envelope, attic",
          "actions": "The envelope is in
          the attic",
          "my knowledge": "The envelope,
          and thus the hat, is in the attic.
          Jayden is aware since he is in
          the attic.",
          "answer and why": "Jayden knows
          the hat is in the
          envelope in the attic, Answer:
          envelope"
      },
      "timestamp_5":{
          "people involved": "Jayden",
          "objects": "attic",
          "actions": "Jayden exited the
          attic",
          "my knowledge": "Jayden left the
          attic, last known location of
          the hat was in the envelope in
          the attic",
          "answer and why": "Jayden last
          saw the hat in the envelope in
          the attic, Answer: envelope"
      },
      "timestamp_6":{
          "people involved": "Jayden",
          "objects": "attic",
          "actions": "Jayden entered the
          attic again",
          "my knowledge": "Jayden re-
          entered the attic, still
  ```
  ```

```
        believes the hat is in the
        envelope",
      "answer and why": "Jayden still
        believes the hat is in the
      envelope in the attic, Answer:
        envelope"
    },
    "timestamp_7":{
      "people involved": "Hannah",
      "objects": "hat, container",
      "actions": "Hannah moved the hat
        to the container",
      "my knowledge": "Jayden does not
        see this action; he still
        believes the hat is in the
        envelope",
      "answer and why": "Jayden is
      unaware of the hat's new location,
        still believes it's in the
        envelope,
        Answer: envelope"
    },
    "timestamp_8":{
        "people involved": "None
        directly",
      "objects": "container, attic",
      "actions": "The container is in
        the attic",
      "my knowledge": "Jayden does not
      know the hat has been moved to
        the container",
      "answer and why": "Jayden still
        believes the hat is in the
        envelope, Answer: envelope"
    }
}
```

Reasoner:

- Based on your knowledge, what' s your answer and reason? Please respond in the following structured format:

```
{{
    "question": "Restate the full
    question context here to ensure
    accuracy and avoid
    misunderstandings.",
    "reason": "Provide a clear and
    concise explanation for your
    answer, rooted in the knowledge
    and reasoning available."
    "answer": "Based on my perspective,
     the answer is <option>.",
}}
```

-
```
{
    "question": <Question>,
    "reason": "Based on the sequence of
    events, Jayden last knew that the
    hat was in the envelope when he was
     in the attic. He exited and re-
    entered the attic without observing
     Hannah moving the hat to the
    container. Therefore, Jayden's last
    known location of the hat is still
     the envelope.",
    "answer": "Based on my perspective,
     the answer is envelope."
}
```

- Based on the previous conversation, summarize your response to the following question:
  Where will Jayden look for the hat?
  Instructions:
  Output one and only one of the candidate answers. Do not include any additional words, explanations, or formatting. Only provide the chosen answer in plain text.

- Based on the full context of the story and the given beliefs, Jayden will look for the hat in the container. Here's the reasoning:

  1. Jayden entered the attic (Step 1) and knew the envelope was in the attic (Step 4).
  2. Jayden exited the attic (Step 5) and then re-entered (Step 6).
  3. While Jayden was not in the attic, Hannah moved the hat from the envelope to the container (Step 7).
  4. The container is in the attic (Step 8).

  Since it is not explicitly stated that Jayden did not notice the movement of the hat by Hannah, and given that agents are assumed to know everything that occurs in a location from the time they enter until the time they

```
leave, Jayden would be aware that the
hat is now in the container. Therefore,
Jayden will look for the hat in the
container.
```

G Counterintuitive cases in main experiments

Some cases in main experiments have undergone several counterintuitive results. It's noted that certain results in Table 1 depict ZeroShot, the simplest baseline, occasionally outperforming other baseline methods. This phenomenon, while seemingly counterintuitive, could be attributed to several factors specific to model behaviors and baseline characteristics, rather than indicating a general superiority of ZeroShot approach. We observed instances of out-of-distribution outputs from certain models. For example, Llama-series sometimes tend to continue completing options presented in the question rather than providing a direct answer. Smaller models, in some cases, produced answers with internal logical inconsistencies that were not rectified by re-prompting or slight modifications to the prompt wording.

As to Qwen-2.5-72B case, which have undergone prior training of reasoning pattern, was observed thet it spontaneously incorporated CoT-like reasoning steps even when interfacing with simpler baselines. This spontaneous combination of methods did not consistently lead to improved outcomes, as the integration of different approaches does not always ensure stable performance enhancements; indeed, individual methods inherently possess certain error rates.

The PercepToM baseline, in particular, sometimes performed abnormally lower than ZeroShot. This might be contextualized by its original design, which was specifically tailored for the ToMi dataset. Given that the original ToMi dataset is known to contain few systematic errors and incorrect answers, applying PercepToM (which is sensitive to this dataset's nuances) can lead to performance deviations when compared to a more general approach like ZeroShot.