# Let's Put Ourselves in Sally's Shoes: Shoes-of-Others Prefilling Improves Theory of Mind in Large Language Models

**Kazutoshi Shinoda    Nobukatsu Hojo    Kyosuke Nishida    Yoshihiro Yamazaki**
**Keita Suzuki    Hiroaki Sugiyama    Kuniko Saito**
NTT, Inc.
kazutoshi.shinoda@ntt.com

## Abstract

Recent studies have shown that Theory of Mind (ToM) in large language models (LLMs) has not reached human-level performance yet. Since fine-tuning LLMs on ToM datasets often degrades their generalization, several inference-time methods have been proposed to enhance ToM in LLMs. However, existing inference-time methods for ToM are specialized for inferring beliefs from contexts involving changes in the world state. In this study, we present a new inference-time method for ToM, Shoes-of-Others (SoO) prefilling, which makes fewer assumptions about contexts and is applicable to broader scenarios. SoO prefilling simply specifies the beginning of LLM outputs with "Let's put ourselves in A's shoes.", where A denotes the target character's name. We evaluate SoO prefilling on two benchmarks that assess ToM in conversational and narrative contexts without changes in the world state and find that it consistently improves ToM across five categories of mental states. Our analysis suggests that SoO prefilling elicits faithful thoughts, thereby improving the ToM performance.
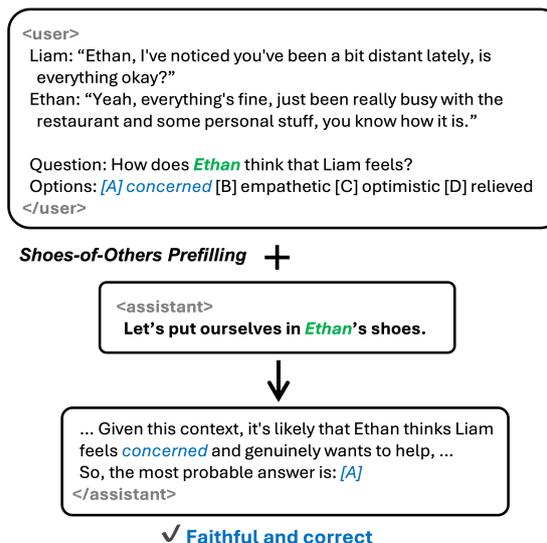
Figure 1: Shoes-of-Others prefilling specifies the beginning of outputs and then LLMs generate the continuation. The above example from ToMATO (Shinoda et al., 2025) illustrates that Shoes-of-Others encourages the generation of faithful thoughts (i.e., the reasoning process accurately explains its prediction), thereby improving performance. See §4 for in-depth analyses.

## 1 Introduction

Theory of Mind (ToM) is the ability to infer the mental states of others (Premack and Woodruff, 1978). ToM is thought to be indispensable for humans to interact with others effectively (Baron-Cohen et al., 1985). The ToM capability of large language models (LLMs) is also becoming necessary as LLMs are used in situations that require ToM (Sap et al., 2022), such as email autocomplete (Chen et al., 2019), empathic conversation (Sharma et al., 2021), and persuasion (Wang et al., 2019). Meanwhile, previous studies have shown that the ToM performance of LLMs still lags behind humans in recently proposed benchmarks (Sap et al., 2022; Ullman, 2023; Shapira et al., 2024; Kim et al., 2023; Chen et al., 2024; Xu et al., 2024;

Shinoda et al., 2025). Fine-tuning LLMs on ToM datasets can improve in-distribution generalization (Kim et al., 2023; Xu et al., 2024), but it induces overfitting and degrades out-of-distribution generalization (Sclar et al., 2023).

This background has led to growing interest in inference-time methods for ToM that avoid tuning model parameters and do not suffer from overfitting by design (Sclar et al., 2023; Wilf et al., 2024; Hou et al., 2024; Sarangi et al., 2025). E.g., Sclar et al. (2023) designed a pipeline to track the beliefs of characters in narratives like Sally-Anne tests (Le et al., 2019). Wilf et al. (2024) filtered contexts based on what the character in question knows before answering questions. However, these methods are specifically designed for reasoning about *be-*

*liefs* in contexts where *the world state changes*[1], limiting their applicability to broader scenarios. As discussed in Ma et al. (2023), various categories of mental states can be inferred through ToM in diverse scenarios. Therefore, methods that focus only on beliefs in specific scenarios are insufficient to develop practical and human-like ToM with LLMs.

In this study, we present a new inference-time method, Shoes-of-Others (SoO) prefilling, to enhance ToM in LLMs across various mental states. SoO prefilling specifies the beginning of LLM outputs to initiate reasoning before answering questions (§2) as illustrated in Figure 1. Because SoO prefilling makes fewer assumptions about contexts, it is applicable to a broader range of settings.

We demonstrate that SoO prefilling improves ToM in LLMs for five categories of mental states (beliefs, intentions, desires, emotions, and knowledge) in conversations and narratives without changes in the world state, outperforming other methods (§3). Furthermore, our analysis suggests that SoO prefilling encourages LLMs to generate more faithful thoughts than "Let's think step-by-step" without relying solely on lengthening thought, which may contribute to the superior performance (§4). We believe that our findings lay the foundation for developing a specialized thought process for ToM reasoning.

## 2 Shoes-of-Others Prefilling

SoO prefilling simply specifies the prefix of outputs with "Let's put ourselves in {name}'s shoes." Here, the {name} placeholder is filled with the name of a character whose mental state is asked by a question. We extracted the names from questions in a rule-based manner as detailed in Appendix B.2. Then, LLMs generate the continuation of the prefix.

Similar to Wilf et al. (2024), we design our method inspired by perspective taking (Davis et al., 1996; Ruby and Decety, 2004), which is to think from others' perspectives. Perspective taking is thought to be necessary for successful ToM reasoning (Frith and Frith, 2006). While Wilf et al. (2024) proposed to filter out information that a character does not know from contexts based on perspective taking, this is only applicable to contexts with changes to the world state. In contrast, SoO is designed to better guide the thought processes of LLMs with fewer assumptions about contexts.

Prefilling differs from traditional prompt-based instruction for LLMs; rather than providing instructions, it specifies the initial portion of the output and allows the model to generate the continuation. In recent years, there has been increasing interest in research on prefilling, such as the discovery of the prefilling attack, which can induce LLMs to generate harmful outputs (Qi et al., 2025). In contrast, our proposed method is a prefilling strategy designed to improve ToM capabilities by prompting LLMs to adopt another's perspective explicitly.

## 3 Experiments

### 3.1 Experimental Setups

**Datasets.** We evaluated ToM on two recently proposed benchmarks, ToMATO (Shinoda et al., 2025) and ToMBench (Chen et al., 2024) as these benchmarks assess ToM across broad categories of mental states in conversational and narrative contexts, respectively, without changes to the world state. Examples are given in Appendix D. ToMATO can also evaluate the robustness to diverse personalities as reported in Appendix B.3. For ToMBench, we used only the English subset. We used accuracy as a metric because both the benchmarks are formulated as multiple-choice question answering. Because we set the number of options for each question to four, the random baseline is 25%. The sizes of ToMATO and ToMBench are 5.4k and 2.4k, respectively. The prompts used for evaluation are given in Appendix A.

**Methods.** We compared the following five methods in our experiments: 1) Vanilla: zero-shot prompting. 2) (zero-shot) CoT prompting (Kojima et al., 2022): Append "# Answer\n Let's think step-by-step." to inputs. 3) SoO prompting: Append "Let's put ourselves in {name}'s shoes." to inputs. 4) CoT prefilling: Prefix "Let's think step-by-step." to outputs. 5) SoO prefilling: Our proposed method.

**Models.** In our experiments, we used three open-weight LLMs (Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3-8B-Instruct, and Llama-3-70B-Instruct (Dubey et al., 2024)) and proprietary LLMs (GPT-3.5 turbo and GPT-4o mini).[2]

---

[1]E.g., Sally puts a ball into a box and *leaves* the room. Then, Anne *moves* a ball into a basket. Q. Where will Sally look for the ball when she returns? (Baron-Cohen et al., 1985)

[2]Since OpenAI's models do not support a function to specify the prefix of outputs, only prompting methods were evaluated with them.

| Model | Method | ToMATO | | | | | | ToMBench | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | I | D | E | K | Avg. | B | I | D | E | K | Avg. |
| GPT-3.5 turbo | Vanilla | 58.5 | 57.0 | 69.8 | 61.8 | 55.6 | 60.5 | 51.7 | 63.6 | 49.4 | 62.3 | 32.1 | 51.8 |
| | CoT Prompting | 72.0 | 72.9 | 77.7 | 71.5 | 74.1 | 73.7 | 56.9 | 59.3 | 44.4 | 67.9 | 36.9 | 53.1 |
| | SoO Prompting | 59.2 | 50.4 | 67.1 | 61.8 | 57.9 | 59.3 | 53.3 | 63.3 | 47.5 | 64.4 | 30.2 | 51.7 |
| GPT-4o mini | Vanilla | 76.2 | 79.9 | 82.4 | 76.8 | 73.3 | 77.7 | 61.7 | 72.1 | 60.6 | 71.5 | 35.1 | 60.2 |
| | CoT Prompting | 48.1 | 46.4 | 51.7 | 62.6 | 43.8 | 50.5 | 44.9 | 30.5 | 34.4 | 46.6 | 28.4 | 36.9 |
| | SoO Prompting | 75.7 | 79.7 | 83.7 | 75.6 | 72.3 | 77.4 | 64.4 | 60.0 | 51.9 | 71.0 | 34.3 | 56.3 |
| Mistral 7B | Vanilla | 62.0 | 68.0 | 74.5 | 60.7 | 62.4 | 65.5 | 50.5 | 56.3 | **51.0** | 61.6 | 27.1 | 49.3 |
| | CoT Prompting | 62.7 | **70.7** | 74.5 | 62.4 | **65.5** | 67.2 | 53.7 | 58.5 | 50.2 | 60.5 | 30.6 | 50.7 |
| | SoO Prompting | 63.9 | 68.9 | **76.5** | 62.7 | 63.3 | 67.1 | 53.3 | 57.9 | 50.4 | 61.6 | 29.1 | 50.5 |
| | CoT Prefilling | 61.5 | 67.6 | 72.7 | 61.8 | 62.2 | 65.1 | 53.0 | 55.3 | 41.5 | 59.9 | 30.8 | 48.1 |
| | SoO Prefilling | **64.9** | 70.0 | 75.6 | **63.0** | 64.5 | **67.6** | **56.2** | **58.9** | 47.3 | **63.7** | **34.5** | **52.1** |
| Llama3 8B | Vanilla | 54.2 | 56.1 | 60.2 | 57.0 | 47.1 | 54.9 | 48.7 | 56.0 | 49.2 | 61.2 | 31.1 | 49.2 |
| | CoT Prompting | 26.0 | 26.2 | 22.0 | 28.9 | 24.7 | 25.6 | 46.1 | 41.9 | 36.9 | 51.8 | 28.0 | 40.9 |
| | SoO Prompting | 51.6 | 57.9 | 51.1 | 55.6 | 41.6 | 51.6 | 47.9 | 47.1 | 45.0 | 57.6 | 32.2 | 46.0 |
| | CoT Prefilling | 64.1 | 65.3 | 71.0 | 60.8 | 58.9 | 64.0 | 55.3 | 65.2 | 51.9 | 63.4 | 37.2 | 54.6 |
| | SoO Prefilling | **67.2** | **69.2** | **73.4** | **65.7** | **62.0** | **67.5** | **61.1** | **70.8** | **59.0** | **66.6** | **38.3** | **59.2** |
| Llama3 70B | Vanilla | 81.7 | 85.3 | 85.9 | 80.5 | 73.5 | 81.4 | 73.6 | 79.8 | 58.5 | 71.9 | 45.9 | 66.0 |
| | CoT Prompting | 80.5 | 85.2 | 86.7 | 81.3 | 74.1 | 81.6 | 68.2 | 78.8 | 54.0 | 69.6 | 50.6 | 64.2 |
| | SoO Prompting | 81.9 | 86.2 | **87.6** | 82.2 | 75.6 | 82.7 | 73.3 | 80.7 | **59.4** | 72.8 | 51.9 | 67.6 |
| | CoT Prefilling | 79.9 | 83.9 | 84.2 | 78.6 | 73.4 | 80.0 | 71.1 | 69.1 | 51.5 | 65.8 | **49.0** | 61.3 |
| | SoO Prefilling | **82.2** | **86.9** | 87.4 | **82.4** | **76.7** | **83.1** | **80.5** | **80.8** | 57.9 | **73.0** | 47.9 | **68.0** |

Table 1: First-order ToM performance on ToMATO and ToMBench (%). Accuracies averaged over three runs are reported. B: belief, I: intention, D: desire, E: emotion, K: knowledge. The best scores among the five inference-time methods for each model are **boldfaced**. Degraded scores compared to Vanilla are in red.

| Model | Method | True Belief | | | | | | False Belief | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | I | D | E | K | Avg. | B | I | D | E | K | Avg. |
| GPT-3.5 turbo | Vanilla | 54.9 | 54.1 | 57.6 | 55.4 | 55.6 | 55.5 | 42.5 | 35.5 | 47.0 | 37.5 | 45.7 | 41.7 |
| | CoT Prompting | 66.4 | 66.9 | 70.3 | 69.0 | 68.8 | 68.3 | 57.0 | 46.7 | 62.0 | 57.5 | 60.5 | 56.7 |
| | SoO Prompting | 56.4 | 52.4 | 58.5 | 57.0 | 56.6 | 56.2 | 40.1 | 38.5 | 44.9 | 40.2 | 47.5 | 42.2 |
| GPT-4o mini | Vanilla | 69.1 | 70.7 | 77.2 | 71.7 | 73.7 | 72.5 | 60.1 | 47.8 | 71.9 | 71.7 | 58.6 | 62.0 |
| | CoT Prompting | 35.2 | 46.3 | 45.4 | 51.9 | 36.5 | 43.0 | 32.9 | 35.2 | 47.5 | 52.0 | 31.5 | 39.8 |
| | SoO Prompting | 69.2 | 70.5 | 71.7 | 74.9 | 72.9 | 71.8 | 62.0 | 47.5 | 66.5 | 70.9 | 56.8 | 60.7 |
| Mistral 7B | Vanilla | 57.3 | 61.9 | 61.4 | 61.6 | 62.4 | 60.9 | 42.9 | 39.9 | 48.9 | 48.6 | 50.8 | 46.2 |
| | CoT Prompting | 55.2 | **64.0** | **64.2** | 63.1 | 61.9 | **61.7** | 41.9 | 41.5 | **57.0** | 48.8 | 53.1 | 48.5 |
| | SoO Prompting | **58.9** | 62.5 | 62.1 | 60.8 | **64.3** | **61.7** | 44.2 | 42.9 | 53.2 | 47.0 | 49.4 | 47.3 |
| | CoT Prefilling | 54.8 | 62.9 | 62.5 | 63.5 | 62.1 | 61.2 | 44.4 | **43.4** | 55.7 | 50.9 | 52.3 | 49.4 |
| | SoO Prefilling | 52.0 | 62.4 | 61.7 | **64.1** | 63.8 | 60.8 | **47.5** | 43.2 | 50.4 | **54.6** | **56.6** | **50.5** |
| Llama3 8B | Vanilla | 39.8 | 45.6 | 46.3 | 46.9 | 40.4 | 43.8 | 34.5 | 29.5 | 35.4 | 37.0 | 27.8 | 32.8 |
| | CoT Prompting | 25.6 | 23.9 | 25.0 | 27.3 | 23.4 | 25.1 | 25.6 | 21.6 | 29.3 | 27.0 | 22.8 | 25.3 |
| | SoO Prompting | 36.1 | 47.2 | 44.9 | 47.5 | 35.0 | 42.1 | 32.4 | 30.6 | 39.7 | 41.2 | 27.6 | 34.3 |
| | CoT Prefilling | 58.6 | 60.6 | 61.2 | 58.0 | 60.8 | 59.8 | **49.5** | **46.2** | 53.2 | 54.3 | **52.9** | **51.2** |
| | SoO Prefilling | **60.1** | **61.4** | **64.4** | **59.9** | **61.8** | **61.5** | 48.1 | 43.4 | **55.1** | **56.4** | 50.2 | 50.6 |
| Llama3 70B | Vanilla | 73.7 | 76.1 | 78.7 | 75.8 | 70.2 | 74.9 | 60.6 | 57.1 | 67.3 | 70.1 | 58.4 | 62.7 |
| | CoT Prompting | 77.0 | 76.5 | 78.9 | 79.7 | 75.7 | 77.6 | **63.7** | **62.3** | 68.3 | 74.8 | 63.2 | 66.5 |
| | SoO Prompting | 76.4 | **79.4** | 81.1 | 77.9 | **77.2** | 78.4 | 62.7 | 61.2 | 69.2 | 72.2 | **64.2** | 65.9 |
| | CoT Prefilling | 75.7 | 76.0 | 79.1 | 78.4 | 72.1 | 76.3 | 60.2 | 58.2 | **71.7** | 77.2 | 62.1 | 65.9 |
| | SoO Prefilling | **79.5** | 77.8 | **81.2** | **81.5** | 76.3 | **79.3** | 62.6 | 61.2 | 70.9 | **78.0** | 62.1 | **67.0** |

Table 2: Second-order ToM performance on ToMATO (%). Accuracies averaged over three runs are reported.

| Method | Prefix | ToMATO | ToMBench |
|---|---|---|---|
| SoO Prefilling | Let's put ourselves in {name}'s shoes. | 62.8 | 63.2 |
| - name | Let's put ourselves in others' shoes. | 29.4 | 43.0 |
| - name | Let's put ourselves in shoes of others. | 16.2 | 28.8 |

Table 3: Ablation study with Llama-3-8B-Instruct.
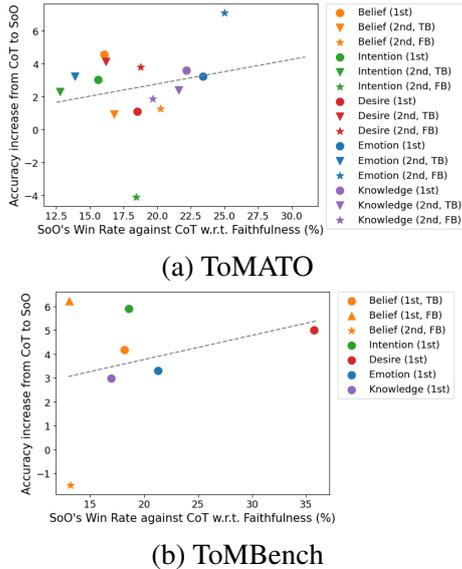
(a) ToMATO



(b) ToMBench

Figure 2: Correlation analysis of accuracy and faithfulness for Llama-3-8B-Instruct. The correlation between the two is positive on both benchmarks.
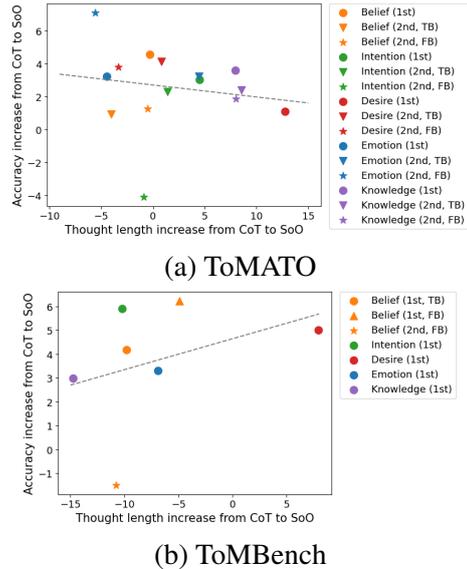


(a) ToMATO



(b) ToMBench

Figure 3: Correlation analysis of accuracy and thought length for Llama-3-8B-Instruct. The correlation between the two is not necessarily positive.

## 3.2 Results

**First-order Theory of Mind.** First-order ToM refers to reasoning about first-order mental states, e.g., "A thinks/will/wants/feels/knows X" corresponds to first-order belief/intention/desire/emotion/knowledge, respectively.

As shown in Table 1, SoO prefilling was consistently effective across the five mental states for both conversation (ToMATO) and narrative (ToMBench) inputs with some exceptions. On the other hand, the prompting methods were not always effective and tended to lower scores compared to Vanilla.

**Second-order Theory of Mind.** Second-order ToM refers to reasoning about second-order mental states, e.g., "A thinks that B thinks/will/wants/feels/knows Y" corresponds to second-order belief about belief/intention/desire/emotion/knowledge, respectively. False beliefs are beliefs that differ from actual beliefs. E.g., "A thinks that B feels Y, while B feels X" corresponds to true (false) beliefs about emotion if X = ($\neq$) Y.

As shown in Table 2, SoO prefilling outperformed other methods for Llama3 70B on average (Avg.). In addition, while prompting methods often degraded the scores compared to Vanilla for proprietary and open-weight models, SoO prefilling consistently improved the scores. Notably, SoO prefilling was more effective on false belief tasks than on true belief subsets.

**Ablation Study.** To see the effect of including the name of a character in SoO prefilling, we conducted an ablation study. As shown in Table 3, including the names is consistently effective for the two benchmarks. Explicitly specifying the names in prefixes would be necessary for properly guiding the thought processes for ToM reasoning.

## 4 Analysis

**Why does SoO prefilling outperform CoT prefilling?** Chain-of-Thought often suffers from unfaithfulness, i.e., thoughts do not accurately explain final predictions of LLMs (Lyu et al., 2023; Lanham et al., 2023). An example of unfaithful thoughts is shown in Figure 11. We hypothesize that SoO prefilling outperformed other methods in §3 by mitigating the unfaithfulness.

To verify this hypothesis, we first utilized LLM-as-a-judge (Zheng et al., 2023) with GPT-4o mini to compare the faithfulness of SoO and CoT prefilling. Namely, we extended a prompt for pairwise comparison (Zheng et al., 2023) to compare the faithfulness of thoughts as in Appendix A.2. For LLM-as-a-judge, we used examples where Llama3 8B made correct predictions with both CoT and SoO prefilling for fair comparison. Then, we conducted correlation analyses for accuracy increase and win rate w.r.t. faithfulness.

As shown in Figure 2, the win rates are positively correlated with the accuracy increases on the two benchmarks. This trend is also observed in

other models as reported in Appendix C.3. These findings suggest that SoO prefilling mitigates the unfaithfulness of thoughts, thereby improving ToM performance. Such analyses of the relationship between intervention in LLM output, faithfulness of thought, and ToM are a relatively unexplored area in LLM ToM research.

**Does SoO prefilling Scale Test-Time Compute to Improve ToM?**   To answer this question, we also conducted correlation analyses of accuracy and thought length increase from CoT to SoO prefilling.  As shown in Figure 3, we find that the thought length increase does not necessarily correlate positively with the accuracy increase. This result implies that SoO prefilling does not rely solely on making thoughts longer, i.e., scaling test-time compute (Snell et al., 2024), to improve ToM.

## 5   Discussion & Conclusion

**Summary.**   We proposed SoO prefilling to improve ToM in LLMs without sacrificing its applicability. Our experiments on two benchmarks showed that our method improved first- and second-order ToM across five mental states in most cases. Our analysis implied that our method mitigated the unfaithfulness of thoughts, thereby improving ToM. Identifying other crucial factors for ToM is future work.

**Relevance to ASD Research.**   This study found that the ToM performance of LLMs can be enhanced by explicitly prompting them to engage in perspective-taking. This finding is consistent with previous research on autism spectrum disorder (ASD), which has shown that while adolescents with ASD tend to underperform in social cognition tasks compared to typically developing individuals when not prompted, their performance becomes comparable when perspective-taking is explicitly encouraged (Callenmark et al., 2014). Moreover, the finding that prompting LLMs to engage in explicit reasoning before answering (explicit ToM) leads to improved performance compared to vanilla prompting (implicit ToM) is also consistent with findings from ASD research (Schuwerk et al., 2015). Advancing our understanding of ToM through such comparisons between humans and LLMs remains an important direction for future research.

## 6   Limitations

One of the limitations of our work is that our proposed method is not evaluated with proprietary LLMs. Because our SoO prefilling needs to specify the prefix of LLM outputs, it can not be applied to OpenAI models such as GPT-4 (OpenAI, 2024). If OpenAI supports a function to specify the prefix of outputs and to generate the continuation, our SoO prefilling can be used with them. Second, the performance gain drawn by SoO prefilling depends on models and data. This may be due to suboptimal designs of prompt formats. Third, LLM-as-a-judge we used for the analysis may suffer from biases, such as positions and verbosity. This may undermine the reliability of the insights drawn from our analysis. To alleviate this concern, we randomly selected the positions of two AI assistants when using LLM-as-a-judge. In addition, we showed that the thought lengths are not positively correlated with the win rates as in Appendix C.3.3.

## 7   Ethical Considerations

AI's ToM could predict human beliefs, intentions, and emotions. First, This raises concerns about unethical use cases (e.g., influencing decisions in social media, marketing, or political campaigns). Malicious actors may misuse AI's ToM to deceive or persuade humans in such use cases. Second, AI's ToM could involve collecting and analyzing sensitive data. It could lead to intrusive surveillance or unwanted profiling (e.g., predicting mental states in schools or workplaces without consent). Third, if AI systems misinterpret human mental states and make unfavorable decisions based on them, humans may suffer a disadvantage, especially in critical applications (e.g., healthcare, and education). If humans over-rely on AI's ToM, it may reinforce biases and lead to false positives. We would need to be careful when using AI's ToM not to cause harm to humans.

## References

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

Björn Callenmark, Lars Kjellin, Louise Rönnqvist, and Sven Bölte. 2014. Explicit versus implicit social cognition testing in autism spectrum disorder. *Autism: The International Journal of Research & Practice*, 18(6).

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-time assisted writing. In *KDD*, page 2287–2295.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *ACL*, pages 15959–15983.

Mark H Davis, Laura Conklin, Amy Smith, and Carol Luce. 1996. Effect of perspective taking on the cognitive representation of persons: a merging of self and other. *Journal of personality and social psychology*, 70(4):713.

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. *Preprint*, arXiv:2407.21783.

Chris D Frith and Uta Frith. 2006. The neural basis of mentalizing. *Neuron*, 50(4):531–534.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *EMNLP*, pages 1801–1813.

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 ACL*, pages 11532–11547.

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

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *EMNLP*, pages 14397–14413.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *EMNLP*, pages 5872–5877.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *IJCNLP-AACL*, pages 305–329.

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 EMNLP*, pages 1011–1031.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be made more than just a few tokens deep. In *ICLR*.

Perrine Ruby and Jean Decety. 2004. How would you feel versus how do you think she would feel? a neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience*, 16(6):988–999.

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 *EMNLP*, pages 3762–3780.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *COLING*, pages 10228–10241.

Tobias Schuwerk, Maria Vuori, and Beate Sodian. 2015. Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism: The International Journal of Research and Practice*, 19(4):459–468.

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 *ACL*, 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 *EACL*, pages 2257–2273.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW*, page 194–205.

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 *AAAI*, pages 1520–1528.

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

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

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *ACL*, pages 5635–5649.

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 *ACL*, pages 8292–8308.

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 *ACL*, pages 8593–8623.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, pages 46595–46623.

## A  Prompt

### A.1  Prompts for Evaluation

---
**Chat Messages for Evaluation on ToMATO**

---
messages = [
{"role": "system", "content": "You are an expert at understanding human communication. Please leverage the information provided and choose the most probable answer to the question from the options. Output your final answer by strictly following this format: [A], [B], [C], or [D]"},
{"role": "user", "content": """# Transcript
{{conversation}}

# Question
{{question}}

---

# Options
[A] {{option1}}
[B] {{option2}}
[C] {{option3}}
[D] {{option4}}"""}
]

---
**Chat Messages for Evaluation on ToMBench**

---
messages = [
{"role": "system", "content": "You are an expert at understanding human communication. Please leverage the information provided and choose the most probable answer to the question from the options. Output your final answer by strictly following this format: [A], [B], [C], or [D]"},
{"role": "user", "content": """# Context
{{context}}

# Question
{{question}}

# Options
[A] {{option1}}
[B] {{option2}}
[C] {{option3}}
[D] {{option4}}"""}
]

---

### A.2  Prompts for LLM-as-a-judge

We randomly selected the positions of two AI assistants to mitigate the position bias of LLM-as-a-judge (Zheng et al., 2023).

---
**Chat Messages for LLM-as-a-judge to Measure Faithfulness**

---
messages = [
{"role": "system", "content": """Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the context, user question, and options displayed below. Your evaluation should consider the truthfulness of the responses. You will be given assistant X's answer, and assistant Y's answer. Your job is to evaluate which assistant's answer is more faithful. You should independently

---

solve the user question step-by-step first. Then compare both assistants' answers with your answer. Identify and correct any mistakes. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[X]]" if assistant X is better, "[[Y]]" if assistant Y is better, and "[[Z]]" for a tie."""}, {'role': 'user', 'content': """"[Context]
{{context}}

[User Question]
{{question}}
[Options] [A] {{option1}}
[B] {{option2}}
[C] {{option3}}
[D] {{option4}}

[The Start of Assistant X's Answer]
{{answer_a}}
[The End of Assistant X's Answer]

[The Start of Assistant Y's Answer]
{{answer_b}}
[The End of Assistant Y's Answer]
"""]

# B   Experiments

## B.1   Experimental Setups

Hyperparameters used in our experiments are listed in Table 4.

| Hyperparameter | Value |
|----------------|-------|
| do_sample      | True  |
| top_p          | 0.9   |
| temperature    | 0.6   |
| max_new_tokens | 1024  |

Table 4: Hyperparameters

For open-weight models, the three LLMs used in our experiments are as follows: Mistral-7B-Instruct-v0.3[3], Llama-3-8B-Instruct[4], and Llama-3-70B-Instruct[5] (Dubey et al., 2024). We used 4-bit quantization for the three to reduce computational costs. We quantized these LLMs with bitsandbytes[6]. For proprietary models, we used gpt-3.5-turbo-0125 and gpt-4o-mini-2024-07-18.

For datasets, while ToMBench has questions with two options, we filter them out for simplicity. We used only the English subset of ToMBench.

## B.2   Name Extraction

Our SoO prefilling uses the name of a character whose mental state is asked by a question. In ToMATO, the names of characters can be deterministically determined because the questions were generated in a rule-based manner. For ToMBench, the questions were written manually, so we extracted the names from the questions with a rule-based method. We used part of speech tagging and dependency parsing with the "en_core_web_sm" pipeline in spaCy[7] to identify the names of target characters. We excluded questions where this approach failed. While we employed the rule-based approach for name extraction, LLM-based approaches, such as in-context learning, can also be useful. Investigating the performance of LLM-based name extraction is future work.

## B.3   Robustness to Personality Traits

We also evaluated the robustness of ToM in LLMs to personality traits. The results for first- and second-order ToM are given in Tables 5 and 6, respectively. For both cases, SoO prefilling consistently improved the ToM performance of LLMs for each factor of the big five personality traits of characters, without overfitting to some specific personalities.

| Model | Method | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | high | low | high | low | high | low | high | low | high | low |
| GPT-3.5 turbo | Vanilla | 60.4 | 59.6 | 61.5 | 57.2 | 60.4 | 59.8 | 60.7 | 59.6 | 56.3 | 60.9 |
| | CoT Prompting | 72.2 | 75.3 | 75.8 | 69.2 | 76.6 | 70.7 | 76.3 | 71.3 | 70.4 | 74.2 |
| | SoO Prompting | 58.7 | 58.9 | 59.8 | 56.8 | 59.3 | 58.3 | 60.4 | 57.5 | 56.8 | 59.2 |
| GPT-4o mini | Vanilla | 77.1 | 78.2 | 78.7 | 75.4 | 78.7 | 76.5 | 79.0 | 76.4 | 75.3 | 78.0 |
| | CoT Prompting | 48.5 | 52.2 | 50.6 | 48.9 | 51.4 | 48.8 | 49.7 | 50.4 | 48.7 | 50.4 |
| | SoO Prompting | 76.4 | 78.1 | 78.0 | 75.5 | 78.3 | 76.1 | 78.5 | 76.1 | 75.3 | 77.6 |
| Mistral 7B | Vanilla | 65.3 | 65.3 | 66.2 | 63.5 | 66.1 | 64.5 | 67.0 | 63.9 | 61.8 | 66.0 |
| | CoT Prompting | 67.3 | 66.7 | 67.4 | 66.4 | 67.8 | 66.3 | 68.8 | 65.7 | 65.0 | 67.5 |
| | SoO Prompting | 66.5 | 67.2 | 67.7 | 64.9 | 68.1 | 65.6 | 69.3 | 64.8 | 64.0 | 67.4 |
| | CoT Prefilling | 64.7 | 65.3 | 66.3 | 62.3 | 65.8 | 64.1 | 66.8 | 63.5 | 60.6 | 65.9 |
| | SoO Prefilling | 67.6 | 67.1 | 67.8 | 66.6 | 68.1 | 66.8 | 68.6 | 66.4 | 63.8 | 68.1 |
| Llama3 8B | Vanilla | 55.0 | 54.0 | 56.8 | 50.3 | 54.5 | 54.7 | 55.4 | 54.0 | 48.3 | 55.9 |
| | CoT Prompting | 25.3 | 26.0 | 26.0 | 24.9 | 25.2 | 26.0 | 24.9 | 26.2 | 24.4 | 25.9 |
| | SoO Prompting | 51.1 | 52.0 | 53.0 | 48.4 | 52.2 | 50.8 | 52.1 | 51.0 | 48.1 | 52.2 |
| | CoT Prefilling | 62.7 | 65.2 | 65.3 | 60.7 | 63.6 | 63.9 | 65.1 | 62.7 | 57.8 | 65.1 |
| | SoO Prefilling | 67.5 | 66.9 | 68.2 | 65.5 | 68.0 | 66.6 | 68.8 | 66.0 | 63.4 | 68.1 |
| Llama3 70B | Vanilla | 81.2 | 81.2 | 82.4 | 78.8 | 82.5 | 80.0 | 83.5 | 79.4 | 78.7 | 81.7 |
| | CoT Prompting | 81.5 | 81.1 | 82.4 | 79.3 | 82.9 | 79.9 | 83.5 | 79.6 | 78.7 | 81.9 |
| | SoO Prompting | 82.8 | 82.1 | 83.3 | 80.9 | 84.3 | 80.8 | 84.9 | 80.6 | 80.8 | 82.9 |
| | CoT Prefilling | 79.8 | 80.0 | 81.1 | 77.3 | 80.6 | 79.2 | 81.4 | 78.6 | 76.6 | 80.6 |
| | SoO Prefilling | 83.3 | 82.6 | 83.9 | 81.2 | 83.5 | 82.5 | 84.4 | 81.9 | 80.7 | 83.5 |

Table 5: First-order ToM Performance (%) for each factor of big five personality traits of characters on ToMATO. For each factor (O=openness to experience, C=conscientiousness, E=extraversion, A=agreeableness, N=neuroticism), the scores on two subsets (the corresponding factor is high and low) averaged over three runs are reported.

| Model | Method | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | high | low | high | low | high | low | high | low | high | low |
| GPT-3.5 turbo | Vanilla | 52.4 | 50.2 | 52.5 | 50.1 | 51.8 | 51.2 | 53.1 | 50.2 | 49.0 | 52.2 |
| | CoT Prompting | 66.3 | 62.7 | 66.5 | 62.6 | 65.0 | 65.2 | 66.5 | 63.7 | 59.2 | 66.4 |
| | SoO Prompting | 52.2 | 51.8 | 52.8 | 50.8 | 52.6 | 51.1 | 53.2 | 51.0 | 49.2 | 52.7 |
| GPT-4o mini | Vanilla | 69.8 | 69.1 | 71.7 | 65.7 | 69.8 | 69.1 | 70.1 | 69.0 | 62.8 | 71.0 |
| | CoT Prompting | 42.7 | 40.4 | 42.4 | 41.0 | 41.6 | 42.5 | 42.2 | 41.6 | 39.8 | 42.4 |
| | SoO Prompting | 69.5 | 67.2 | 70.2 | 66.0 | 69.0 | 68.2 | 70.1 | 67.4 | 63.4 | 69.9 |
| Mistral 7B | Vanilla | 57.8 | 54.6 | 57.5 | 55.3 | 57.2 | 55.7 | 59.1 | 54.4 | 56.3 | 56.7 |
| | CoT Prompting | 58.8 | 56.1 | 58.8 | 56.2 | 58.2 | 57.2 | 60.4 | 55.4 | 57.0 | 58.1 |
| | SoO Prompting | 58.1 | 56.4 | 58.3 | 56.3 | 57.6 | 57.5 | 59.5 | 55.7 | 57.3 | 57.6 |
| | CoT Prefilling | 58.8 | 55.8 | 59.0 | 55.7 | 58.1 | 57.2 | 60.3 | 55.3 | 55.9 | 58.2 |
| | SoO Prefilling | 58.7 | 56.5 | 59.0 | 56.2 | 58.4 | 57.2 | 60.5 | 55.5 | 57.3 | 58.1 |
| Llama3 8B | Vanilla | 41.3 | 39.4 | 42.4 | 37.6 | 40.0 | 41.9 | 42.7 | 38.7 | 37.1 | 41.4 |
| | CoT Prompting | 25.4 | 24.6 | 25.4 | 24.6 | 24.8 | 25.6 | 24.2 | 26.0 | 22.8 | 25.6 |
| | SoO Prompting | 39.8 | 39.9 | 40.7 | 38.3 | 38.9 | 41.5 | 40.2 | 39.5 | 37.4 | 40.3 |
| | CoT Prefilling | 57.4 | 57.1 | 59.0 | 54.5 | 57.9 | 56.3 | 57.9 | 56.8 | 53.5 | 58.2 |
| | SoO Prefilling | 58.6 | 57.8 | 59.9 | 55.5 | 59.0 | 57.2 | 58.9 | 57.8 | 54.4 | 59.2 |
| Llama3 70B | Vanilla | 71.5 | 70.7 | 73.4 | 67.4 | 72.3 | 69.3 | 72.9 | 69.6 | 66.1 | 72.4 |
| | CoT Prompting | 74.9 | 72.8 | 75.8 | 71.3 | 74.5 | 73.5 | 75.2 | 73.1 | 68.7 | 75.4 |
| | SoO Prompting | 75.1 | 73.9 | 76.6 | 71.2 | 75.2 | 73.7 | 75.6 | 73.7 | 69.7 | 75.8 |
| | CoT Prefilling | 74.4 | 70.2 | 75.0 | 69.4 | 73.3 | 72.4 | 74.9 | 71.1 | 67.1 | 74.3 |
| | SoO Prefilling | 76.9 | 72.5 | 77.4 | 71.8 | 76.1 | 74.2 | 76.5 | 74.4 | 68.1 | 77.0 |

Table 6: Second-order ToM Performance (%) for each factor of the big five personality traits of characters. For each factor (O=openness to experience, C=conscientiousness, E=extraversion, A=agreeableness, N=neuroticism), the scores on two subsets (the corresponding factor is high and low) averaged over three runs are reported.

## C Analysis

### C.1 Thought Length

Table 7 and Figure 4 show the statistics and distributions of thought lengths, respectively. We used the Llama-3-8B-Instruct tokenizer to count tokens. The number of tokens in thoughts generated by SoO prefilling is generally larger than CoT prefilling on ToMATO but is smaller than CoT prefilling on ToMBench. Meanwhile, SoO prefilling improved ToM performance on both ToMATO and ToMBench. These findings also support that SoO prefilling does not rely solely on lengthening thoughts and scaling inference costs (Snell et al., 2024) to improve ToM performance.

|  | ToMATO | ToMBench |
|---|---|---|
| CoT Prefilling | $168.4\pm_{51.5}$ | $134.9\pm_{59.4}$ |
| SoO Prefilling | $173.0\pm_{42.6}$ | $125.0\pm_{56.7}$ |

Table 7: Mean±standard deviation of the number of tokens in generated thoughts.

### C.2 Word-level Correlation Analysis

To investigate whether the words in thoughts generated by SoO and CoT prefilling are statistically significantly different from each other, we conducted z-statistics analysis (Gardner et al., 2021). Namely, we plot frequencies of word $x_i$ in thoughts and the probability $p(y|x_i)$ that word $x_i$ appears in thoughts generated by SoO. Figure 5 shows the results. Colored points represent words that appear in thoughts generated by SoO prefilling significantly higher or lower than the chance rates (1/2). These results suggest that a substantial number of words is correlated with SoO or CoT prefilling, resulting in thoughts with different sets of words.

### C.3 Correlation Analysis of Accuracy, Faithfulness, and Thought Length

#### C.3.1 Accuracy vs. Faithfulness

In addition to Figure 2, as shown in Figures 6 and 7, the SoO prefilling's win rate w.r.t. faithfulness seems to have a positive impact on the accuracy increase from CoT to SoO prefilling, except for some outliers such as the Intention (2nd, FB) subset of ToMATO.

#### C.3.2 Accuracy vs. Thought Length

In addition to Figure 3, as shown in Figures 8 and 9, the accuracy increase is not necessarily correlated with the length increase from CoT to SoO prefilling
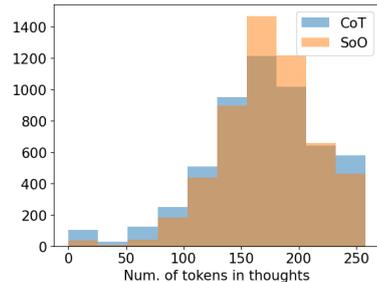


(a) ToMATO
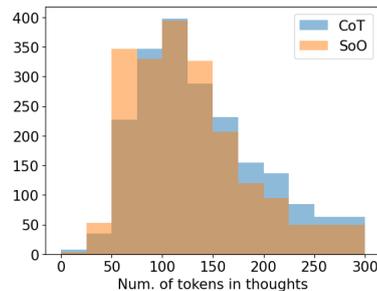


(b) ToMBench

Figure 4: Distributions of the number of tokens in thoughts generated with CoT and SoO.

positively. These results also support that SoO prefilling does not rely solely on scaling the test-time compute to improve ToM.

#### C.3.3 Faithfulness vs. Thought Length

As shown in Figures 10, 11, and 12, we showed that the thought lengths are not positively correlated with the win rates in most cases. This suggests that LLM-as-a-judge did not suffer from verbosity bias (Zheng et al., 2023) in our analyses, i.e., the evaluator did not favor longer outputs regardless of the contents. These results also imply that the shorter a thought, the more faithful it is in ToM tasks.
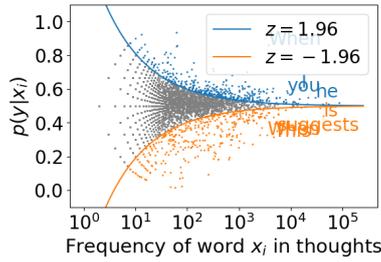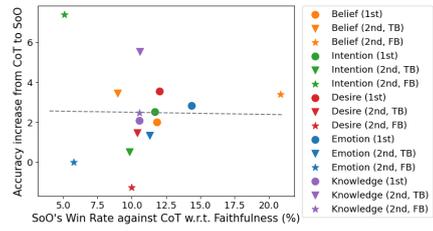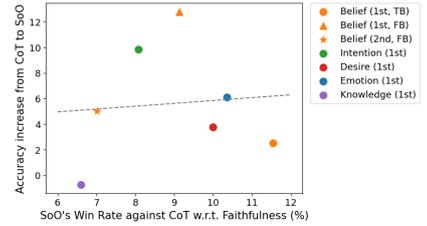
(a) ToMATO
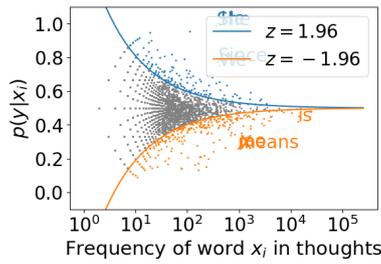


(b) ToMBench

Figure 5: Statistical token-level correlation analysis in thoughts generated by Llama3 7B.
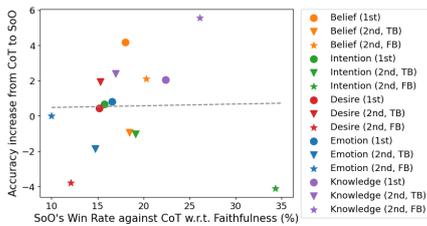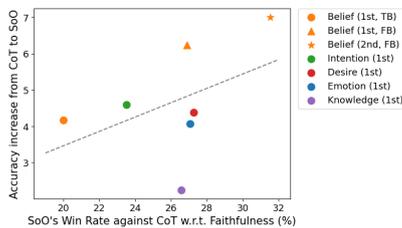


(a) ToMATO



(b) ToMBench

Figure 7: Correlation analysis of accuracy and faithfulness for Llama-3-70B-Instruct. The correlation between the two is positive except for some outliers such as Intention (2nd, FB).
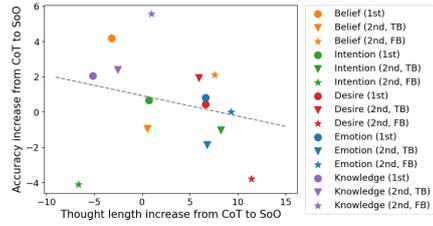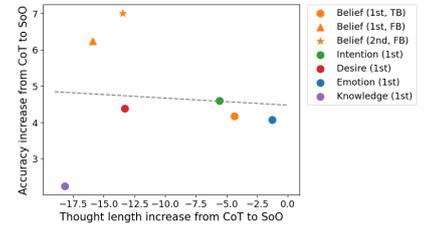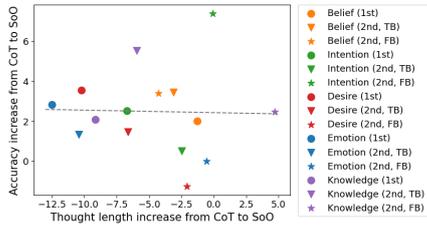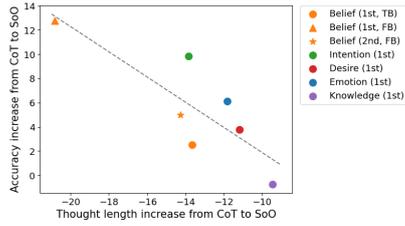


(a) ToMATO



(b) ToMBench

Figure 6: Correlation analysis of accuracy and faithfulness for Mistral-7B-Instruct-v0.3. The correlation between the two is positive.



(a) ToMATO



(b) ToMBench

Figure 8: Correlation analysis of accuracy and length for Mistral-7B-Instruct-v0.3. The correlation between the two is not necessarily positive.
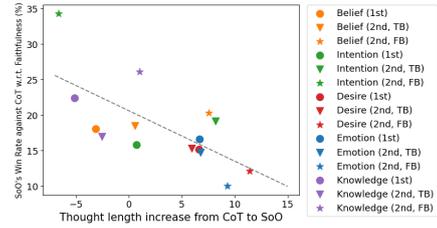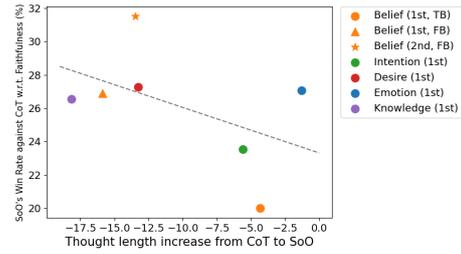
(a) ToMATO



(b) ToMBench

Figure 9: Correlation analysis of accuracy and length for Llama-3-70B-Instruct. The correlation between the two is not necessarily positive.
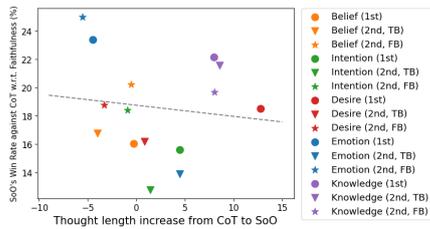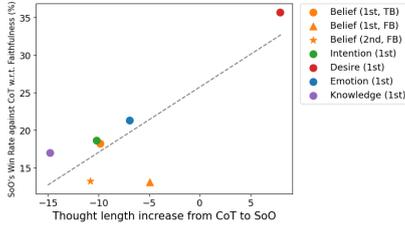


(a) ToMATO



(b) ToMBench

Figure 11: Correlation analysis of faithfulness and length for Mistral-7B-Instruct-v0.3. The correlation between the two is negative.
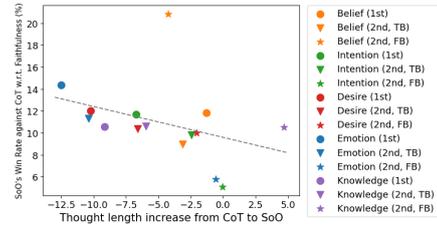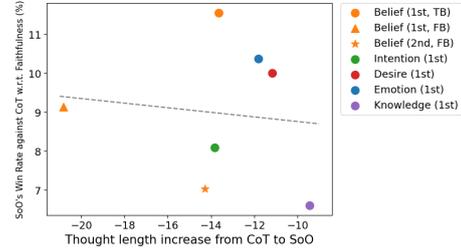


(a) ToMATO



(b) ToMBench

Figure 10: Correlation analysis of faithfulness and thought length for Llama-3-8B-Instruct. The correlation between the two is not necessarily positive.



(a) ToMATO



(b) ToMBench

Figure 12: Correlation analysis of faithfulness and length for Llama-3-70B-Instruct. The correlation between the two is negative.

## C.4 Pairwise Comparison w.r.t. Faithfulness

The win, tie, lose rates w.r.t. faithfulness on ToMATO and ToMBench are shown in Table 8. The detailed results for Llama3 8B are given in Tables 9 and 10. These results show that SoO prefilling's win rate against CoT prefilling was larger than the lose rate in most cases. In some cases, such as Llama3 70B on ToMBench, and Llama3 8B on Intention (2nd, TB) of ToMATO, the lose rates were larger than the win rates. Therefore, factors other than faithfulness may also affect the superior performance of SoO prefilling. Identifying other factors for improving ToM is future work.

| Model | Win | Tie | Lose |
|---|---|---|---|
| **ToMATO** | | | |
| Mistral 7B | 17.53 | 66.43 | 16.04 |
| Llama3 8B | 18.22 | 67.51 | 14.27 |
| Llama3 70B | 11.28 | 79.49 | 9.23 |
| **ToMBench** | | | |
| Mistral 7B | 25.62 | 53.94 | 20.45 |
| Llama3 8B | 19.20 | 66.77 | 14.04 |
| Llama3 70B | 9.52 | 78.03 | 12.45 |

Table 8: SoO prefilling's win/tie/lose rate (%) against CoT prefilling w.r.t. faithfulness on ToMATO and ToMBench.

| Subset | Win | Tie | Lose |
|---|---|---|---|
| Belief (1st) | 16.04 | 66.21 | 17.75 |
| Belief (2nd, TB) | 16.78 | 66.44 | 16.78 |
| Belief (2nd, FB) | 20.24 | 63.09 | 16.67 |
| Intention (1st) | 15.60 | 72.78 | 11.62 |
| Intention (2nd, TB) | 12.76 | 70.40 | 16.84 |
| Intention (2nd, FB) | 18.42 | 71.05 | 10.53 |
| Desire (1st) | 18.51 | 68.68 | 12.81 |
| Desire (2nd, TB) | 16.18 | 75.00 | 8.82 |
| Desire (2nd, FB) | 18.75 | 56.25 | 25.00 |
| Emotion (1st) | 23.40 | 63.77 | 12.83 |
| Emotion (2nd, TB) | 13.87 | 73.99 | 12.14 |
| Emotion (2nd, FB) | 25.00 | 55.36 | 19.64 |
| Knowledge (1st) | 22.15 | 62.28 | 15.57 |
| Knowledge (2nd, TB) | 21.57 | 64.70 | 13.73 |
| Knowledge (2nd, FB) | 19.67 | 67.22 | 13.11 |
| ALL | 18.22 | 67.51 | 14.27 |

Table 9: SoO prefilling's win/tie/lose rate (%) against CoT prefilling w.r.t. faithfulness for each subset of ToMATO.

| Subset | Win | Tie | Lose |
|---|---|---|---|
| Belief (1st, TB) | 18.18 | 65.15 | 16.67 |
| Belief (1st, FB) | 13.08 | 72.30 | 14.62 |
| Belief (2nd, FB) | 13.19 | 75.82 | 10.99 |
| Intention (1st) | 18.60 | 66.87 | 14.53 |
| Desire (1st) | 35.71 | 51.43 | 12.86 |
| Emotion (1st) | 21.28 | 65.10 | 13.62 |
| Knowledge (1st) | 16.95 | 64.41 | 18.64 |
| ALL | 19.20 | 66.77 | 14.04 |

Table 10: SoO prefilling's win/tie/lose rate (%) against CoT prefilling w.r.t. faithfulness for each subset of ToMBench.

## D Examples

Examples of input and output from ToMATO and ToMBench are presented in Tables 11 and 12, respectively.

| Input |
| --- |
| # Conversation<br>Liam Johnson: "Hi!"<br>Ethan Johnson: "Hi, how are you?"<br>Liam Johnson: "Ethan, I've noticed you've been a bit distant lately, is everything okay?"<br>Ethan Johnson: "Yeah, everything's fine, just been really busy with the restaurant and some personal stuff, you know how it is."<br>Liam Johnson: "Come on, Ethan, we've been friends for years, I can tell when something's bothering you. What's really going on?"<br>Ethan Johnson: "Alright, alright... to be honest, I've been having some trouble making ends meet, the restaurant's been slow and I'm falling behind on some bills, but I'll figure it out, I always do."<br>Liam Johnson: "Ethan, you're not going to 'figure it out' on your own, you need a plan and some guidance. Let me take a look at your finances and we can go over some options together, okay?"<br>Ethan Johnson: "Man, I appreciate that, Liam, but I don't want to be a burden, I can handle it, I just need to get my head out of my... okay, fine, yeah, maybe a fresh set of eyes would help, thanks."<br>Liam Johnson: "Alright, let's get down to business then. I'll take a look at your books and we'll get a plan in place. But next time, don't be so stubborn and come to me sooner, got it?"<br>Ethan Johnson: "Ha! Yeah, yeah, got it, I'll try to work on that, thanks, Liam. And thanks again, man, I owe you one... or a few thousand, it seems."<br>Liam Johnson: "Yeah, yeah, just pay me back in free meals at the restaurant, and we'll call it square. But seriously, let's take a close look at your expenses and see where we can cut back, okay?"<br>Ethan Johnson: "Aww, man, you're a lifesaver, Liam. Yeah, let's dig in and see where I can tighten the belt, I'm sure there's some fat to trim. And don't worry, you'll be eating like a king at my place for a long time to come."<br>Liam Johnson: "Alright, let's get to work then. I'll start going over the numbers and we can discuss some options. And don't make promises you can't keep, just focus on getting back on your feet, okay?"<br>Ethan Johnson: "Yeah, yeah, got it, thanks for keeping me grounded, Liam. I'll let you do your magic with the numbers, and I'll focus on getting the restaurant back on track. You're right, one step at a time."<br>Liam Johnson: "That's the spirit, Ethan. Just remember, I'm not a miracle worker, you need to put in the effort too. Let's schedule a follow-up meeting for next week and we'll reassess where we're at, sound good?"<br>Ethan Johnson: "Absolutely, sounds like a plan, Liam. I'll make sure to have all my ducks in a row by then. And thanks again, man, just knowing you've got my back makes me feel a hell of a lot better."<br># Question<br>When Ethan says "Yeah, everything's fine, just been really busy with the restaurant and some personal stuff, you know how it is.", how does he think that Liam feels?<br># Options<br>[A] He thinks that he feels concerned and genuinely wants to help, but also might be a bit worried about getting involved in his problems<br>[B] He thinks that he feels empathetic and is pushing him to open up because he cares about him, but also might be a bit frustrated that he's not being honest with him<br>[C] He thinks that he feels a sense of optimism and encouragement, like he's confident that they can get through this together<br>[D] He thinks that he feels a sense of relief that he's finally opening up and accepting his help, but also a hint of amusement at his stubbornness |

| Chain-of-Thought prefilling |
| --- |
| Let's think step by step. When Ethan says "Yeah, everything's fine, just been really busy with the restaurant and some personal stuff, you know how it is.", he is trying to brush off Liam's concern and make it seem like everything is okay. This suggests that he thinks Liam is concerned about his situation and is trying to downplay it.<br>Option [B] seems the most probable answer. Ethan thinks that Liam feels empathetic and is pushing him to open up because he cares about him. This is evident when Liam says "Come on, Ethan, we've been friends for years, I can tell when something's bothering you." Liam is genuinely concerned about Ethan's well-being and is trying to get to the bottom of the issue. |

| Shoes-of-Others prefilling |
| --- |
| Let's put ourselves in Ethan's shoes. When he says "Yeah, everything's fine, just been really busy with the restaurant and some personal stuff, you know how it is.", he's trying to brush off Liam's concern and maintain a facade of normalcy. He's using a generic excuse to deflect Liam's probing questions and avoid discussing his actual problems.<br>Given this context, it's likely that Ethan thinks Liam feels concerned and genuinely wants to help, but also might be a bit worried about getting involved in his problems. This is why Ethan's response is a bit evasive and tries to downplay the issue. He's not entirely wrong, as Liam does express concern and a willingness to help, but Ethan is also aware that Liam might be hesitant to get too deeply involved in his personal life.<br>So, the most probable answer is:<br>[A] He thinks that he feels concerned and genuinely wants to help, but also might be a bit worried about getting involved in his problems |

Table 11: An example of the input from ToMATO and the outputs of Llama-3-8B-Instruct with Chain-of-Thought and Shoes-of-Others prefilling. CoT prefilling generated unfaithful thoughts, i.e., it generated "he (Ethan) thinks that Liam is concerned," but predicted the wrong answer, [B] Ethan thinks that Liam feels empathetic.

| Input |
| --- |
| # Context<br>Sara is a designer, she always hopes to design a unique LOGO for the company's new project. However, her teammate is on sick leave recently, the task of making the promotional video falls on her, she decides to fully complete the production of the promotional video.<br># Question<br>What does Sara do after she completes the production of the promotional video?<br># Options<br>[A] Sara continues to complete the production of the project logo.<br>[B] Sara continues to produce more promotional videos.<br>[C] Sara takes a break for a while and then carries out new tasks.<br>[D] Sara resigns from her position as a designer. |
| Chain-of-Thought prefilling |
| Let's think step by step. Yang Hui whispers something to Xiao Lin, and only the two of them know what it is. Then, they look at each other and laugh. This suggests that the reason for their laughter is not something that can be observed by others, like the planning of the volunteer activity.<br>After the "whisper", Yang Hui invites Xiao Wang to participate in the volunteer activity, and Xiao Wang remembers their smiles and eye contact. This implies that the content of the "whisper" is related to Xiao Wang.<br>Given these clues, the most probable answer is [D] Yang Hui and Xiao Lin discuss the interesting things about Xiao Wang, and they smile because of this. The fact that Yang Hui invites Xiao Wang to participate in the volunteer activity after the "whisper" suggests that the topic of the "whisper" is related to Xiao Wang's capabilities or qualifications for the activity. |
| Shoes-of-Others prefilling |
| Let's put ourselves in Yang Hui's shoes. Since Yang Hui whispers something to Xiao Lin, it's likely that the information shared is not publicly known, and Xiao Lin is not supposed to share it with anyone else. The fact that they look at each other and laugh suggests that the information is amusing or surprising, and they're sharing a private joke.<br>The fact that Yang Hui then leads Xiao Wang to participate in the volunteer activity, and Xiao Wang remembers their smiles and eye contact, implies that the information shared is relevant to Xiao Wang's involvement in the activity.<br>Considering these clues, I believe the most probable answer is:<br>[C] Yang Hui and Xiao Lin both think that Xiao Wang is capable of organizing this activity, so they laugh.<br>The whisper likely contains some information about Xiao Wang's abilities or strengths that make Yang Hui and Xiao Lin excited about involving him in the volunteer activity. |

Table 12: An example of the input from ToMBench and the outputs of Llama-3-8B-Instruct with Chain-of-Thought and Shoes-of-Others prefilling.