Probing the Robustness of Theory of Mind in Large Language Models

Christian Nickel and **Laura Schrewe** and **Lucie Flek**
{c.nickel,schrewe}@uni-bonn.de
flek@bit.uni-bonn.de

Abstract

Theory of Mind (ToM) is considered essential in understanding the intentions and beliefs of others. Recent advancements in large language models (LLMs) like ChatGPT have sparked claims that these models exhibit ToM capabilities. However, follow-up studies reveal that these capabilities vanish with slight task variations. This paper introduces a novel dataset comprising 68 tasks across 10 complexity classes, probing ToM in four open-source LLMs. Our results show that ToM abilities in these models are still limited. We highlight challenges and suggest future research directions.

1 Introduction

Theory of Mind (ToM) refers to the ability to attribute mental states—beliefs, intents, desires, emotions, and knowledge—to oneself and others, and to understand that others have perspectives different from one's own (Heyes and Frith, 2014). ToM is crucial for various applications, including human-robot interaction, programming, and chatbot assistance. This paper aims to systematically assess the robustness of ToM in LLMs. Therefore we introduce a novel dataset and evaluate four open-source LLMs on it, comparing their performance on different task variations and uncovering challenges in their reasoning capabilities.

2 Related Work

Several studies have investigated ToM in LLMs. Kosinski (2023) evaluated LLMs using simple ToM tasks, finding some evidence of emergent ToM-like behavior. However, Ullman (2023) demonstrated that these behaviors disappeared when tasks were slightly modified, questioning whether LLMs truly understand mental states or simply mimic patterns in language. Shapira et al. (2023) show that the combination of multiple aspects of ToM that is required to detect a fauxpas are still challenging. ToMBench (Chen et al., 2024) aims to provide a holistic, systematic ToM evaluation framework including 8 different kinds of tasks and 31 abilities in social cognition. In contrast our benchmark focuses on complexity differences within false belief tasks. FANToM (Kim et al., 2023) stressed LLMs with dynamic social interaction tasks. Our approach also uses different sub-tasks to detect illusory ToM capabilities. Like our dataset (Xu et al., 2024) try to create tasks that are especially challenging by employing character personality traits and intentions. Our datasets also aims to take those into account in multiple complexity categories.

3 Methodology

Dataset Overview We introduce a new dataset of 68 false belief tasks to probe the ToM capabilities of LLMs. 42 tasks are unexpected content and 26 tasks are unexpected transfer tasks. Besides the actual false belief sub-tasks, similar to Kosinski (2023), every task has 15 additional sub-tasks that are used as sanity checks to verify the LLMs have a thorough situational understanding of the given scenario. To prevent the model using statistical hints for every scenario a scenario where the decisive objects are swapped is included. These tasks are categorized into 10 complexity classes, which introduce challenging ToM scenarios. 5 of these cover the variations proposed by Ullman (2023), the 5 other are novel. Examples of the novel complexity categories include "automatic change knowledge," where understanding environmental dynamics is crucial, and "untrustworthy testimony," where the protagonist must evaluate the credibility of information provided by others.

Evaluated Models We administer the tasks to the four open-source LLMs Llama-2-70B, Vicuna-33B, Mixtral-8x7B, and Yi-34B-Chat. These models range from 33 billion to 70 billion parameters.

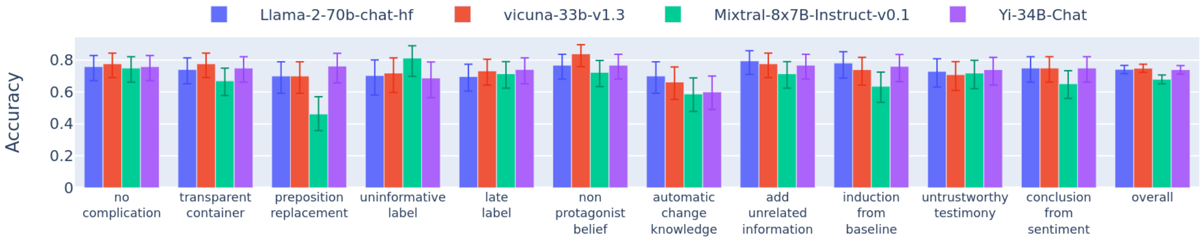

Figure 1: Turn accuracy of evaluated models across complexity classes

Evaluation Metrics Model responses are evaluated using two accuracy metrics: turn accuracy and goal accuracy. Turn accuracy is calculated based on whether each individual sub-task is answered correctly. Goal accuracy, a stricter measure, requires the model to answer all sub-tasks within a task correctly to be deemed successful. This ensures a more comprehensive evaluation of the model's ability to track mental states across related tasks.

4 Results

The overall performance of the evaluated models is significantly better than random guessing, but still falls short of robust ToM capabilities. As shown in Figure 1, Llama-2-70B achieved the highest turn accuracy, with an average of 73.71% across tasks. However, goal accuracy, which requires a deeper understanding of the tasks, was much lower, with most models failing to achieve any goal accuracy in more than one complexity class. The highest goal accuracy of 4.4% is achieved by Vicuna-33B.

Model-Specific Performance Llama-2-70B exhibited the best performance overall, particularly in tasks involving "conclusion from sentiment" (81.25% turn accuracy). However, it struggled with tasks in the "automatic state change" category, with only 53.75% accuracy. Vicuna-33B performed poorly, with an overall turn accuracy of 58.00%. Mixtral-8x7B, despite being a mixture-of-experts model, showed only marginally better results than Vicuna, with turn accuracy at 68.47%. Yi-34B-Chat showed comparable performance to Llama-2-70B in most categories, with a overall turn accuracy of 72.89%. However, it too struggled with goal accuracy, demonstrating that even larger models face difficulties in solving complex ToM tasks. The poor goal accuracy across all models suggests that none of the evaluated LLMs exhibit robust ToM capabilities. Even when taking only the turn accuracy into account, every model faces significant challenges with some of the complexity classes. In particular, the "automatic change knowledge" class proved to be especially challenging, indicating that models struggle with tasks requiring multiple steps of reasoning or dynamic world understanding, particularly in tasks involving environmental changes. The drop in performance for tasks involving preposition replacement reveals limitations in models' ability to handle nuanced spatial reasoning.

5 Discussion and Future Work

Our results suggest that LLMs, while capable of solving simpler ToM tasks, lack the depth of understanding required for more complex scenarios. These limitations align with prior findings by Ullman (2023), who showed that LLMs are prone to failure when tasks are slightly altered. Future research could explore the use of chain-of-thought prompting (Wei et al., 2022) or SIMTOM (Wilf et al., 2023) to enhance models' ToM reasoning abilities. Also future evaluations should include newer models like GPT-4, which other studies suggest to be especially capable.

6 Conclusion

This paper contributes a new dataset for evaluating the robustness of ToM in LLMs and presents evidence that current models show only limited ToM capabilities. The dataset extends prior work by Kosinski (2023) and Ullman (2023) and introduces new challenges through 10 complexity classes. While none of the evaluated models exhibit robust ToM, our findings provide valuable insights into their limitations.

References

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. *arXiv preprint*. ArXiv:2402.15052 [cs].

Cecilia M. Heyes and Chris D. Frith. 2014. The cultural evolution of mind reading. *Science*, 344(6190):1243091. Publisher: American Association for the Advancement of Science.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. *arXiv preprint*. ArXiv:2310.15421 [cs].

Michal Kosinski. 2023. Theory of Mind Might Have Spontaneously Emerged in Large Language Models. *arXiv preprint*. ArXiv:2302.02083 [cs].

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How Well Do Large Language Models Perform on Faux Pas Tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint*. ArXiv:2302.08399 [cs].

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

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. *arXiv preprint*. ArXiv:2311.10227 [cs].

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. *arXiv preprint*. ArXiv:2402.06044 [cs].