Understanding Theory of Mind in Large Language Models: The Role of Perturbations and Chain-of-Thought Prompting

Anonymous submission

Abstract

Recent advancements in large language models (LLMs) have sparked debate over their Theory of Mind (ToM) capabilities, raising questions about whether they exhibit genuine ToM capabilities or merely rely on linguistic pattern recognition. Our study investigates the robustness of ToM in LLMs using perturbations on false-belief tasks and examines the potential of Chain-of-Thought prompting (CoT) to enhance performance and explain the LLM's decision. We generate richly annotated ToM data, including spaces of valid reasoning chains, and propose metrics to evaluate to what extent final answers are faithful to the generated CoT. In line with previous work, we show a steep drop in ToM capabilities under perturbations for all LLMs. CoT prompting surprisingly degrades accuracy for some perturbation classes and control tasks, indicating selective application is necessary. However, our structured approach is limited to classical Sally-Anne Tests. We highlight the need for more advanced evaluation methods to gain finer insights into models' broader social reasoning capabilities and advocate for rigorous definitions of capabilities in LLMs to leverage interdisciplinary research potential and build a clear, commonly understood taxonomy. Moreover we propose new approaches to synthetic data generation of ToM data, including CoT annotations.

Introduction

The emergence of Theory of Mind (ToM)-like behavior in large language models (LLMs) has sparked significant interest. (Kosinski 2023) ToM, the ability to infer the mental states of others, is foundational for social reasoning and has broad applications in AI, e.g. from autonomous agents to empathetic customer service systems (Premack and Woodruff 1978). While promising, ToM capabilities in LLMs are not well understood, particularly their lacking robustness to task perturbations (Ullman 2023). Chain-of-Thought Prompting (CoT), which structures intermediate reasoning steps, has shown potential for improving reasoning tasks (Suzgun et al. 2022; Wei et al. 2022), but its impact on ToM remains under-explored. Analyzing reasoning chains, given models' final answers are faithful, might indicate where and why models go wrong in social reasoning, indicating the way to make ToM capabilities more resilient and trustworthy.

Thus we intend to answer the following questions:

RQ1 How robust are to-date LLMs to ToM task perturbations in syntax, semantics, and common knowledge?

RQ2 To what extent does CoT reasoning enhance the performance of LLMs on original and perturbed tasks?

RQ3 What are the underlying mechanisms by which CoT reasoning impacts the models' final answers?

RQ4 How to best use CoT step explanations to evaluate, interpret, and improve ToM abilities in an automated way?

Related Work

Measuring the effect of task perturbations, Shapira et al. (2023) employ an approach similar to ours, using false belief tasks among others. In accordance with our work they show ToM capabilities vanishing to varying degrees with most types of perturbations applied to tasks. (Shapira et al. 2023; Ullman 2023). ToMBench (Chen et al. 2024) offers a comprehensive ToM evaluation framework employing several task types, similarly finding mixed results with regards to improving accuracy using CoT. Similar to some of our perturbed tasks (Xu et al. 2024) create tasks that are especially challenging by integrating personality traits and intentions, noting some significant performance gains employing CoT, while there is degradation as well. Steps towards more real world challenges in social cognition are taken in Holterman and van Deemter (2023), where the authors use faux-pas-detection tasks derived from Kahneman's cognition biases to assess ToM in ChatGPT-3 and -4, showcasing that the models struggle with achieving performance of a 9 year old human child. This is in accordance with the results on faux-pas-tests of Shapira, Zwirn, and Goldberg (2023). FANToM (Kim et al. 2023) makes use of dynamic social interaction tasks to measure ToM. Fighting the data drought researchers increasingly turn to synthetic data for training and evaluation. Abdullin et al. (2024) introduce an approach generating synthetic data for training a conversational agent, intended to interview a user, understanding and elicitng its goals and possible constraints in the realization of this goal. They employ LLM-as-a-judge to evaluate the generated data. However comparison with human evaluation on a subset shows that the automated evaluation still needs refinement. The lack of scenario diversity and complexity has been more recently addressed by agentic settings such as AgentSense Mou et al. (2024) or AgentPro (Zhang et al.

2024).

Approach

We base our experiments on our novel, human-created dataset, inspired by the works of Kosinski (2023) and Ullman (2023). It comprises 68 false-belief tasks in total (42 unexpected content and 26 unexpected transfer). From each classic task we derive perturbed versions according to our 10 perturbation classes, including the introduction of concepts like the knowledge of automatic state changes, where understanding environmental dynamics is crucial, or "untrustworthy" testimony, in which the protagonist has to evaluate the credibility of information provided by others. Moreover each task is augmented by 15 additional sub-tasks - sanity checks that can be used to ensure that the models are not using simple statistical heuristics to answer the main task, totaling 1088 subtasks, each in the form of open-ended questions. Crucial to our approach is the rich CoT annotation of each task with all valid reasoning paths (space of correct reasoning chains). We introduce evaluation metrics for CoT correctness [1], model faithfulness [2] and impact of CoT reasoning on accuracy on all classes of (un-)perturbed tasks [3]. To understand the presence or size of placebo effects we also compute and compare the ATEs based on the partition of the tasks into those to which a correct ($ATE_{CoT=T}$) respectively incorrect COT ($ATE_{CoT=F}$) has been generated. A substantial increase in performance without correct CoTs justifying that increase would be called a placebo effect. Ongoing work comprises improvement to our metrics. We seek to make them even more informative, providing more precise knowledge on where and why the CoT models make mistakes, and more generally applicable, including non-structured, NL CoTs. Moreover, agentic method to generate synthetic ToM data with reliable CoT augmentations are desirable for more fine grained yet statistically significant investigations of the TOM phenomenon.

Results

ToM robustness We test six open-source LLMs listed in Table 1. While four models demonstrate apparent ToM capabilities in default task settings, under perturbations only two maintain these abilities (Llama-3-70B-Instruct and dbrx-instruct).

Effects of CoT Prompted for CoT, accuracy averaged across all perturbed classes stagnates. No additional models achieve robustness against perturbations. On a per-class basis we see some improvements, but surprisingly also some steep drops in accuracy. Accuracy under CoT rises on unperturbed tasks and on perturbed tasks requiring spatial reasoning or filtering of information. A degradation in performance by employing CoT can be witnessed in tasks, where an early modifier changes the protagonist's information processing (e.g. the protagonist can not read or decipher symbols, that would convey information to the standard protagonist). In classes that don't fall into either of these two categories we see mixed results. When including controls (i.e. non-false belief), the overall performance even diminishes under CoT.

Faithfulness Significant positive correlations across all measures and models underline the importance of CoT fidelity. The most precise measure here turned out to be the Φ-Coefficient between subsequence-based CoT correctness and Final Answer correctness. The ROUGE-based correlations exhibit similar trends, although with more outliers. More recent models are in tendency more faithful.

Placebo Effect Generally we only see substantial overall improvements when correct CoTs had been generated. Only in Mixtral we see a "placebo" effect with any CoT.

Model	$ATE_{CoT=Tr}$	**Φ-Coeff.**
Vicuna 33b v1.3	11,8%	0.342
Mixtral-8x7B-Instruct-v0.1	7%	0.429
Yi-34B-Chat	10,7%	0.306
Meta-Llama-3-70B-Instruct	15,7%	0.584
dbrx-instruct	11,2%	0.489
Llama 2 70B	4,9%	0.549

Table 1: Faithfulness and Placebo Effect: The faithfulness of the models' final answers is indicated by the strong correlations between CoT- and final-answer-correctness calculated as Φ-Coefficient. The presence of a Placebo Effect is judged by comparing the effect strengths of employing CoT partitioned into cases with correct (CoT=T) or incorrect (CoT=F) calculated as Average Treatment Effect (ATE).

Discussion

The strong variance in CoT impact on accuracy indicates its selective use is necessary until underlying reasoning problems are resolved. While our evaluation reveals which perturbation classes see degraded accuracy under CoT, we cannot yet pinpoint where and why along the reasoning chain models fail. This would provide crucial insights into models' demonstrated faithfulness to generated CoTs. Our current limitation to structured reasoning chains demands careful prompting and extensive benchmarking data annotation, making such data costly and rare, constraining statistically significant fine-grained analyses.

To address these challenges, we suggest exploring advanced evaluation methods for diverse reasoning types and natural language, respectively unstructured ToM data. Moreover we propose an agentic approach for generating synthetic interaction data augmented with internal states, enabling broader ToM evaluation through methods capable of analyzing unstructured reasoning chains.

[1] Based on subsequences of valid reasoning states, as well as summarization scores as ROUGE-L and state transition overlaps

[2] Using the correlation of CoT-correctness and final answer correctness, most importantly employing the Φ-coefficient

[3] By computing the Average Treatment Effect (ATE) on each perturbation class of tasks, unperturbed tasks and the overall effect.

References

Abdullin, Y.; Molla-Aliod, D.; Ofoghi, B.; Yearwood, J.; and Li, Q. 2024. Synthetic Dialogue Dataset Generation Using LLM Agents. https://arxiv.org/abs/2401.17461v1.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; and Huang, M. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15959–15983. Bangkok, Thailand: Association for Computational Linguistics.

Holterman, B.; and van Deemter, K. 2023. Does ChatGPT Have Theory of Mind? arXiv:2305.14020.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14397–14413. Singapore: Association for Computational Linguistics.

Kosinski, M. 2023. Theory of Mind Might Have Spontaneously Emerged in Large Language Models. *CoRR*, abs/2302.02083.

Mou, X.; Liang, J.; Lin, J.; Zhang, X.; Liu, X.; Yang, S.; Ye, R.; Chen, L.; Kuang, H.; Huang, X.; and Wei, Z. 2024. AgentSense: Benchmarking Social Intelligence of Language Agents through Interactive Scenarios. arXiv:2410.19346.

Premack, D.; and Woodruff, G. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*, 1(4): 515–526.

Shapira, N.; Levy, M.; Alavi, S. H.; Zhou, X.; Choi, Y.; Goldberg, Y.; Sap, M.; and Shwartz, V. 2023. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. arXiv:2305.14763.

Shapira, N.; Zwirn, G.; and Goldberg, Y. 2023. How Well Do Large Language Models Perform on Faux Pas Tests? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 10438–10451. Toronto, Canada: Association for Computational Linguistics.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; and Wei, J. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv:2210.09261.

Ullman, T. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 2024. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8593–8623. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, W.; Tang, K.; Wu, H.; Wang, M.; Shen, Y.; Hou, G.; Tan, Z.; Li, P.; Zhuang, Y.; and Lu, W. 2024. Agent-Pro: Learning to Evolve via Policy-Level Reflection and Optimization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5348–5375. Bangkok, Thailand: Association for Computational Linguistics.