Instructions for IWCS 2021 Proceedings

Anonymous IWCS submission
The style files are borrowed from ACL-IJCNLP 2021

Abstract

This document contains the instructions for preparing a manuscript for the proceedings of ACL-IJCNLP 2021. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

Today's Large Language Models (LLMs) can write production code (Jiang et al., 2024), translate fluently across 100+ languages (Zhu et al., 2024), and remember conversational nuances for hours-long chat (cite), but somehow misplace a marble when Sally leaves the room. Understanding meaning, especially the meaning of mental state attributions, remains fundamentally problematic for LLMs. GPT-4's performance on simple false-belief tasks is 75-80% (humans get 87%) (Moghaddam and Honey, 2023; Kosinski, 2024), but it struggles with complex scenario tasks. LLMs' performance 30-60% less, when tasks involve understanding why people lie, tracking multiple perspectives, or reading social cues, which is worse than humans' performance on the same tasks (Sap et al., 2022; Kim et al., 2023). These failures reveal a fundamental limitation in semantic understanding.

Theory of Mind (ToM) allows humans to predict behavior by modeling others' mental states, even when their beliefs contradict reality (Premack and Woodruff, 1978). It presents a core challenge in computational semantics to process intensional contexts where agents' beliefs diverge from reality. The sentence "Sally believes the marble is in the basket" creates an opaque context, where the truth of the embedded clause is evaluated relative to Sally's belief state, but not the actual world (Montague, 1973). This semantic complexity becomes evident in LLMs' failure patterns, where models that can solve standard false-belief tasks fail when containers become transparent or locations change (Ullman, 2023). These failures suggest LLMs lack compositional understanding of how mental state verbs like "believe," "know" and "think" can create distinct modal contexts that block standard entailment relations (Karttunen, 1973).

In this paper, we empirically analyze six LLMs on Theory of Mind tasks to understand their semantic failures. We quantify features at multiple linguistic levels, including information structure (idea density), discourse relations (RST parsing), lexical patterns (mental state verb distribution), and distributional semantics (semantic similarity), to address three key research questions: (1) To what extent do these features predict LLM success on mental state reasoning? (2) Do different models reveal distinct patterns in processing intensional contexts? (3) What systematic failures emerge across model architectures? We find that the complexity metrics show surprisingly weak correlations with performance on the ToM tasks, revealing that current LLMs neither leverage linguistic complexity cues nor employ compositional semantic processing, suggesting reliance on task-specific patterns or spurious correlations.

2 Related Work

Modern ToM testing goes much further than classic false-belief scenarios. ToMBench dataset (Chen et al., 2024) measures 31 aspects of social cognition using 8 different tasks and 6 categories, showing persistent model weaknesses. The benchmark uses two languages and multiple-choice questions to prevent memorization and allow automatic scoring. EPITOME (Jones et al., 2024) applies psychology

research methods to categorize ToM errors into seven types. Models fail more frequent on pragmatic reasoning and social inference tasks, while performing better on basic belief questions. Researchers now intensionally use difficult test cases to uncover more limitations. ExploreToM generates difficult story structures using A* search algorithms to test how well models handle complex compositions. The Two Word Test study (Riccardi and Desai, 2023) finds that models fail at simple noun-noun combinations. These new evaluation methods show that models often succeed by memorizing common patterns rather than actually tracking beliefs.

3 Methodology

3.1 Data

We used the ToMBench dataset (Chen et al., 2024), a benchmark designed to evaluate Theory of Mind capabilities in LLMs. We focused exclusively on the English version of the dataset. ToMBench consists of 31 distinct aspects of social cognition organized into 6 categories: beliefs, emotions, intentions, knowledge, non-literal communication, and desire. The dataset uses multiple-choice question (A, B, C, D) with underlying text answer formulation to facilitate automatic scoring and reduce ambiguity during evaluation.

3.2 Models

3.3 Experimental Setup

4 Results

4.1 RQ1

4.2 RQ2

4.3 RQ3

5 Discussion

6 Conclusion

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation.

Cameron R. Jones, Sean Trott, and Benjamin Bergen. 2024. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*, 12:803–819.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic Inquiry*, 4(2):167–193.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In Patrick Suppes, Julius Moravcsik, and Jaakko Hintikka, editors, *Approaches to Natural Language*, pages 221–242. Dordrecht.

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

Nicholas Riccardi and Rutvik H. Desai. 2023. The two word test: A semantic benchmark for large language models.

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.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.

Table 1: Model Performance by Theory of Mind Ability (%)

Model	Emotion	Desire	Intention	Knowledge	Belief	NL Comm.	AVG
Human	86.4	78.2	90.4	82.2	89.3	89.0	85.9
Llama 70B	75.0	61.1	81.2	48.3	85.4	78.9	71.6
Qwen 32B	70.2	59.4	70.0	36.9	78.7	70.9	64.4
OLMo 13B	67.4	50.0	66.2	37.9	62.5	67.9	58.6
Mistral 7B	58.3	51.1	55.0	26.6	52.3	63.1	51.1
Phi-3 Mini	60.7	52.2	59.4	33.4	60.5	63.8	55.0
InternLM 1.8B	51.2	45.0	50.6	33.4	47.3	65.4	48.8

Table 2: Linguistic Metrics by Theory of Mind Ability

Metric	Emotion	Desire	Intention	Knowledge	Belief	NL Comm.	AVG
Idea Density	0.434	0.407	0.423	0.387	0.336	0.430	0.403
Num EDUs	8.138	7.511	13.615	9.493	8.713	15.652	10.520
RST Tree Depth	4.274	4.128	5.997	5.148	4.016	6.992	5.092
Rel Attribution	0.950	1.228	1.526	1.486	0.422	3.531	1.524
Rel Causal	0.640	0.439	1.062	0.376	0.643	0.652	0.635
Rel Explanation	0.057	0.117	0.176	0.086	0.181	0.551	0.195