On the Role of Linguistic Features in LLM Performance on Theory of Mind Tasks

Ekaterina Kozachenko*
University of Lorraine
ETH Zürich

Gonçalo Guiomar
ETH Zürich

Karolina Stańczak
ETH Zürich

{ekozachenko, gguiomar, kstancza}@ethz.ch

Abstract

Theory of Mind presents a fundamental challenge for Large Language Models (LLMs), revealing gaps in processing intensional contexts where beliefs diverge from reality. We analyze six LLMs across 2,860 annotated stories, measuring factors such as idea density, mental state verb distribution, and perspectival complexity markers. Notably, and in contrast to humans, we find that LLMs show positive correlations with linguistic complexity. In fact, they achieve high accuracy (74-95%) on *high* complexity stories with explicit mental state scaffolding, yet struggle with *low* complexity tasks requiring implicit reasoning (51-77%). Furthermore, we find that linguistic markers systematically influence performance, with contrast markers decreasing accuracy by 5-9% and knowledge verbs increasing it by 4-10%. This inverse relationship between linguistic complexity and performance, contrary to human cognition, may suggest that current LLMs rely on surface-level linguistic cues rather than genuine mental state reasoning.

1 Introduction

While Large Language Models (LLMs) demonstrate remarkable capabilities in code generation (Jiang et al., 2024), multilingual translation (Zhu et al., 2024), and long-context conversational memory (Liu et al., 2024), their performance on social reasoning tasks remains fundamentally unreliable. Although LLMs are approaching human accuracy on simple false-belief tests (Moghaddam and Honey, 2023; Kosinski, 2024), their inconsistent patterns on more sophisticated tasks requiring social reasoning (Sap et al., 2022; Kim et al., 2023), suggest they rely on mechanisms fundamentally different from human cognition.

At the heart of this reasoning lies Theory of Mind (ToM), the human ability to model others' mental states, especially when their beliefs contradict reality (Premack and Woodruff, 1978). Classic false-belief tasks, such as the Sally-Anne test, probe this ability by requiring a model to predict an agent's actions based on their incorrect beliefs. Computationally, this requires processing intensional contexts created by attitude verbs like "believe," where the truth of a proposition is evaluated relative to a subjective perspective rather than objective reality (Montague, 2008). Recent findings reveal that LLMs capable of passing standard false-belief tests often fail on their minor variations (Ullman, 2023). This suggests they lack a robust understanding of how mental state verbs create distinct semantic contexts that block standard entailment (Karttunen, 1973).

In this paper, we empirically analyze six LLMs on ToM tasks to understand their failure patterns on tasks requiring semantic reasoning. We examine 2,860 stories by quantifying linguistic features related to information structure (idea density) and lexical patterns (mental state verb density). We also manually annotate each story for its level of perspectival complexity and linguistic markers. We address three key research questions: (**RQ1**) To what extent do idea density and mental state verb density correlate with LLM performance on mental state reasoning? (**RQ2**) How do linguistic markers of perspectival complexity influence model performance on ToM tasks? (**RQ3**) What systematic failures emerge across different model architectures?

We find that LLMs exhibit opposite correlations to humans in terms of linguistic complexity, yet paradoxically achieve the highest accuracy on high-complexity stories with explicit mental state scaffolding. These findings suggest that LLMs rely on surface linguistic cues rather than genuine perspective-tracking.

*This research was conducted while visiting ETH Zürich.

Proceedings of the 16th International Conference on Computational Semantics, pages 308–316
September 22-23, 2025,

2 A Semantic Framework for Theory of Mind

To formally analyze ToM, the capacity to attribute beliefs, desires, and intentions to oneself and others, and to recognize that these states may diverge from reality (Premack and Woodruff, 1978; Astington, 1993), we ground our analysis in a multi-agent epistemic–doxastic logic (Hintikka, 2005; Fagin and Halpern, 1994). This framework provides a precise language for representing nested perspectives and allows for systematic categorization of the perspectival complexity of social reasoning scenarios (Karttunen, 1971; Giannakidou, 1998).

A mental-state representation in our framework consists of an agent $a \in \mathcal{A}$ from a set of story participants, an attitude (e.g., knowledge K, belief B), and a content formula φ expressing a proposition about events or states. Our formal language $\mathcal{L}_{[1][2]\ldots[n]}$ extends propositional logic $\mathcal{L}$ with a set of modal operators $[i]$, each corresponding to a mental attitude held by a specific agent. For instance, for agents $a, b \in \mathcal{A}$, the formula $K_a\varphi$ expresses "a knows φ," and $B_b\varphi$ expresses "b believes φ." These operators can be nested to represent higher-order ToM, as in $K_aB_b\neg p$ ("a knows that b believes that p is false").

The semantics are defined using a generalized Kripke model (Voorbraak, 1992), a tuple

$$\begin{aligned} M = \langle & w_0, W, \{\Sigma_1, \Sigma_2, \ldots, \Sigma_m\}, \\ & \langle \sigma_1, \sigma_2, \ldots, \sigma_m \rangle, \\ & \langle F_1, F_2, \ldots, F_m \rangle, \vDash \rangle \end{aligned}$$

where W is a set of possible worlds, Σ_i is a non-empty set of epistemic states for attitude i, $\sigma_i : W \to \Sigma_i$ maps each world to an epistemic state, F_i is a set of projection functions that extract information from an epistemic state, and $\vDash$ is the valuation function, where $\vDash (w, [i]\varphi)$ depends on the epistemic state $\sigma_i(w)$. The key insight of the generalized Kripke models is that epistemic states are explicitly represented as atomic entities, not sets of worlds, with nonstandard valuation for modal operations.

We instantiate this general framework for two attitudes: objective knowledge and rational belief.
Objective Knowledge. Modeled as an S5 modality, objective knowledge corresponds to truthful, introspective information. In an *objective knowledge (OK) model*, the truth condition for $K_a\varphi$ is given as:

$$w \vDash K_a\varphi \text{ iff } \forall w' \in W\big(\kappa(w') = \kappa(w) \Rightarrow w' \vDash \varphi\big)$$

where $\kappa(w)$ is the information state at world w. This states that φ is true in all worlds that are informationally indistinguishable from w.
Rational Belief. Modeled as a KD45 modality, rational belief is not necessarily true but is consistent and introspective. In a *rational belief (RIB) model*, the belief set $\|\beta(w)\|_B$ for a state $\beta(w)$ is non-empty (*consistency*) and constant across all worlds within that set (*introspection*). The truth condition for $B_a\varphi$ is:

$$w \vDash B_a\varphi \quad \text{iff} \quad \forall w' \in \|\beta(w)\|_B \; w' \vDash \varphi.$$

This states that φ is true in all worlds compatible with the agent's beliefs.

Veridicality. Following Karttunen (1971, 1973) and Giannakidou (1998), we classify attitude verbs by their entailment properties. An operator is *veridical* if it entails its complement φ in the actual world (e.g., "know", "realize"), *non-veridical* if it carries no such entailment (e.g., "believe", "suspect"), and *anti-veridical* if it entails $\neg\varphi$ (e.g., "pretend", "imagine"). As a subset of non-veridical operators (Giannakidou, 2013), an operator F is anti-veridical if $F\varphi$ is false in an agent's epistemic model $M(x)$, i.e., $M(x) \cap [\![\varphi]\!] = \varnothing$. We note that this distinction can be modeled within the non-veridical *RIB* framework by adding a constraint that all accessible worlds satisfy $\neg\varphi$ (e.g., for attitudes like "pretend" or "imagine"); however, our analysis focuses on the core attitudes of knowledge and belief.

Perspectival Complexity. We quantify complexity based on the nesting depth of modal operators and the number of distinct agents. Depth 0 (no operators) is *simple*. Depth 1 with a single agent is *low* complexity. Depths 2 with multiple agents are *medium*, and depths of 3+ with at least three agents are *high*. We also annotate linguistic markers, including explicit contrasts ($B_a\varphi \wedge \neg\varphi$) and displacement (a proposition φ appearing only within the scope of an operator).

3 Methodology

3.1 Data

For our analysis, we use the English portion of ToMBench (Chen et al., 2024), a benchmark designed to assess ToM capabilities in LLMs. ToMBench covers 31 distinct aspects of social cognition organized into six categories: *beliefs* (reasoning about divergent or false mental states), *emotions* (understanding situational feelings), *intentions* (recognizing goal-directed actions), *knowledge* (tracking access to information), *non-literal communication* (interpreting indirect meaning), and *desire* (identifying subjective wants). Representative examples from the dataset are provided in the App. A in Tab. 1. Every instance of the ToMBench contains a story, followed by a question, and four plausible options (A, B, C, D) where only one answer is correct and the others are high-quality but misleading wrong answers.

Data annotation. We manually annotated each instance in the dataset for two key properties: *perspectival complexity* and the presence of specific *linguistic markers*. We categorized stories into four levels based on mental state attribution patterns: *simple story* (no explicit mental state attributions), *low* (single agent with mental state), *medium* (multiple agents or belief-reality contrasts), and *high* (nested mental states or three+ agents with contrastive structures). We tracked three types of linguistic markers: (1) contrast markers signaling belief-reality divergence ("but actually," "however"), (2) displacement markers indicating perspective shifts ("from X's perspective"), and (3) verb types distinguishing factive (knows, sees) from non-factive (thinks, believes) mental states. While this surface-level annotation simplifies true intensional complexity, which would require analyzing scope ambiguities, de re/de dicto distinctions, and semantic properties of embedded clauses, it captures identifiable correlates that may proxy for deeper semantic complexity. This approach tests whether LLMs are sensitive to surface markers of perspective complexity, even if we cannot directly assess their handling of formal intensional semantics.

The annotation was performed by a linguistics expert and validated by a second expert, both authors of this work. All discrepancies were resolved through discussion, resulting in a high inter-annotator agreement (Cohen's $\kappa = 0.90$ for complexity and $\kappa = 0.95$ for markers). A detailed guide to our annotation criteria is available in App. A. Additionally, we automatically computed Idea Density and lexical patterns via Mean Syntactic Verb Dependency for each instance using custom scripts based on spaCy.[1]

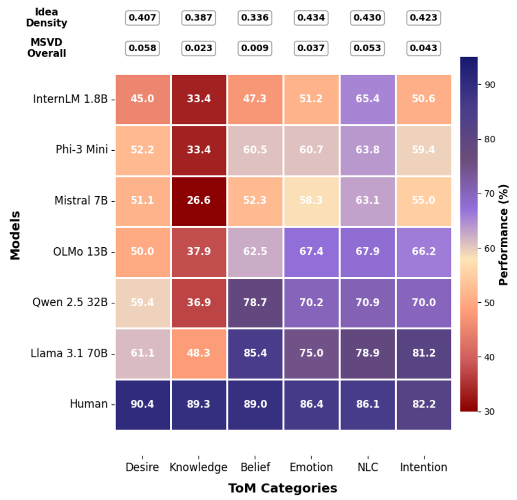

Figure 1: Heatmap of performance (%) for LLMs and humans across six ToM categories with the average idea density and MSVD for stories in each category.

Idea Density (ID). Idea density measures the rate of elementary propositions in a text, normalized by its length. It serves as a metric for informational complexity, where lower density has been linked to cognitive decline and an increased risk of Alzheimer's disease (Sirts et al., 2017). The idea density for a given text is calculated as:

$$\text{Idea Density} = \frac{\text{Number of Propositions}}{\text{Number of Words}} \quad (1)$$

Mean Syntactic Verb Dependency (MSVD). To capture the expression of characters' internal states, which is a key component of ToM (Astington, 1993), we measure the density of state verbs. State verbs (*e.g.*, *think, know, believe, want, feel*) describe cognitive or emotional states rather than physical actions. A higher frequency of these verbs can indicate a greater focus on intentionality and mental representation within a story. For a story S, we calculate MSVD as:

$$MSVD(S) = \frac{|V_{\text{state}}(S)|}{N_{\text{words}}(S)} \quad (2)$$

where $V_{\text{state}}(S)$ is the set of lemmatized state verbs in the text and $N_{\text{words}}(S)$ is the total word count.

3.2 Models

We evaluate several state-of-the-art LLMs on the ToMBench benchmark, ranging from 1.8B to 70B

[1] https://spacy.io/

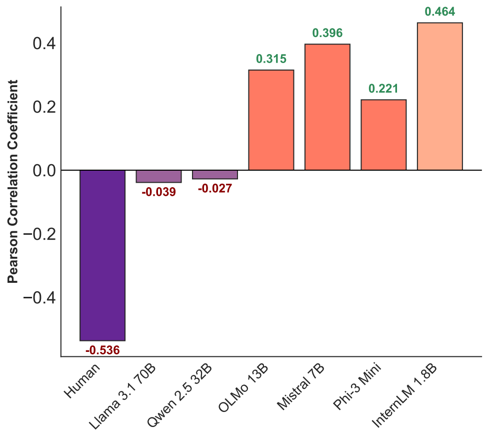

Figure 2: Pearson correlation between idea density and task performance. A strong negative correlation is observed for humans, in contrast to most models.

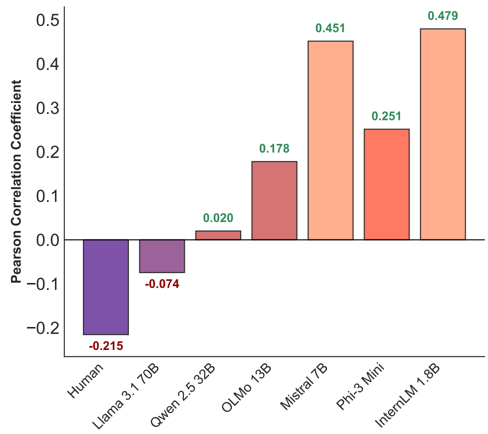

Figure 3: Pearson correlation between MSVD and task performance for humans and LLMs. A negative correlation is observed for humans ($r = -0.215$), while most models exhibit a positive correlation.

parameter count: LLama-3.1-70B (Touvron et al., 2023), Qwen-2.5-32B (Team, 2025), OLMo-2-13B (OLMo et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Phi-3-Mini-4k-Instruct (Abdin et al., 2024), and InternLM-2.5-1.8 (Cai et al., 2024). To this end, we prompt these models to answer the tasks from the dataset discussed above in the multiple-choice setup.

4 Experiments and Results

RQ1: To what extent do idea density and mental state verb density correlate with LLM performance on mental state reasoning? We first examine performance across the six ToM categories shown in Fig. 1, revealing a consistent human advantage across all categories. To investigate the relationship between linguistic features and success on mental state reasoning tasks, we analyze the correlation between performance on ToMBench and two textual features: idea density and MSVD. We compute the Pearson correlation between these features and task performance across both the human baseline and the suite of evaluated LLMs. The human performance data is derived from the original study involving 20 graduate students (Chen et al., 2024). Our analysis reveals a stark, opposing relationship between these linguistic features and performance for humans versus LLMs. In Fig. 2, we observe a negative correlation for human performance with both ID ($r = -0.536$) and MSVD ($r = -0.215$). This indicates that as texts become more informationally dense or contain more explicit mental state verbs, human performance on the ToM tasks tends to decrease. In direct contrast,

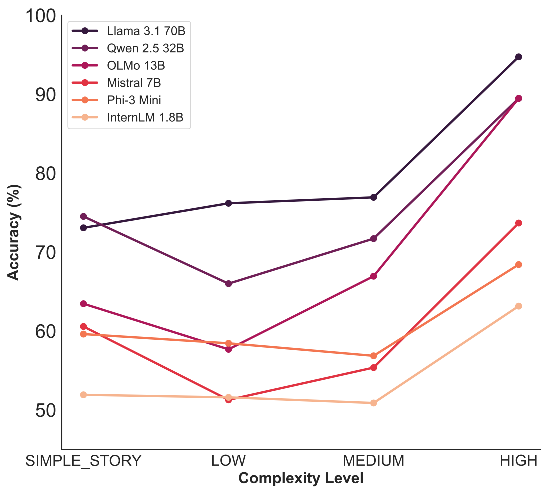

Figure 4: LLMs performance across perspectival complexity categories.

LLMs consistently show a positive correlation with these same features. The correlation between performance and ID is positive across most models, ranging to a moderate $r = 0.464$. A similar positive trend is observed for MSVD (see Fig. 3). This suggests that, unlike humans, LLM performance and comprehension are enhanced through increased linguistic scaffolding.

RQ2: How do linguistic markers of perspectival complexity influence model performance on ToM tasks To investigate the impact of narrative structure, we evaluated LLM accuracy across four levels of perspectival complexity (Fig. 4). Our results reveal a "complexity paradox:" contrary to expectations, models achieve peak performance (74-95% accuracy) on *high* complexity stories with

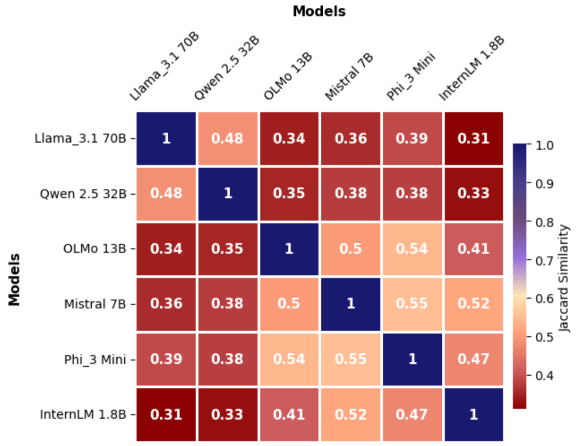

Figure 5: A Jaccard similarity matrix illustrating the degree of overlap in errors between model pairs, where higher values signify more similar failure modes.

nested mental states, while struggling most with the *low* complexity category (51-77%). This suggests that explicitly complex narrative structures may provide a form of linguistic scaffolding that aids model reasoning more than the subtler challenges of medium-complexity texts.

RQ3: What systematic failures emerge across different model architectures? To identify systematic failures across architectures, we computed the Jaccard similarity of incorrect responses for all model pairs, as shown in Fig. 5. The results reveal a clear cluster of smaller models (Mistral 7B, Phi-3 Mini, OLMo 13B) that exhibit high error overlap ($J \approx 0.5 - 0.55$), suggesting they share a common failure mode. In contrast, the largest models show more idiosyncratic errors, indicating they may overcome some specific systematic challenges. Moreover, we identified 245 stories (8.6%), where all models fail universally, concentrated in *low* (9.7%) and *medium* (6.9%) complexity levels. These systematic failures occur despite the presence of linguistic markers: stories with contrast markers, knowledge verbs, or moderate MSVD still cause universal failure when they require reasoning beyond surface cues. This pattern reinforces our finding that LLMs rely on explicit linguistic scaffolding: they fail systematically when answering correctly requires inference rather than pattern-matching on mental state markers.

This aligns with Ross and Pavlick (2019), who showed NLI models like BERT fail on non-veridical verbs (e.g., "think", "believe") due to pattern-matching biases rather than true inference. In our universal failure cases, similar non-veridical mental state verbs dominate low-complexity stories requiring implicit reasoning, while veridical "knowledge verbs" provide insufficient scaffolding, extending their veridicality bias to ToM contexts.

5 Related Work

Early ToM evaluations revealed superficial success on classic false-belief tasks, such as the Sally-Anne test (van Duijn et al., 2023), prompting more rigorous benchmarks. Recent work like ToMBench (Chen et al., 2024) and EPITOME (Jones et al., 2024) benchmarks show a recurring pattern of models handling basic belief-tracking but failing on tasks requiring pragmatic or social inference.

This weakness in compositional reasoning, also probed by procedurally generated narratives in ExploreToM (Sclar et al., 2025), suggests models exploit statistical shortcuts rather than genuinely tracking mental states. Other work reveals failures in more fundamental capabilities, such as the Two Word Test study (Riccardi and Desai, 2023). A common finding across these methods is that models often succeed by exploiting statistical patterns rather than by genuinely tracking mental states. However, prior work has not systematically distinguished between tasks with low and high intentionality (*i.e.*, simple belief attribution versus complex deception) or investigated how specific linguistic features influence LLM performance on ToM tasks. Our work aims to address these gaps.

6 Conclusions

We analyzed linguistic features in LLM performance on ToM tasks, revealing surprising patterns: (1) LLMs show positive correlations with idea density and MSVD, opposite to humans' negative correlations, (2) Models paradoxically excel on *high* complexity stories (74-95%) while struggling with *low* complexity (51-77%), and (3) All models fail systematically when implicit reasoning is required. These patterns suggest LLMs may leverage explicit linguistic markers rather than genuine mental state reasoning, though our correlational analysis cannot prove causation. The complexity paradox, where explicit mental state scaffolding aids performance, warrants further causal investigation to understand whether models truly rely on surface cues or develop deeper representations.

7 Limitations

While this study provides novel insights into the relationship between linguistic features and LLM performance on ToM tasks, we acknowledge several limitations that frame avenues for future research. First, our primary metrics, Idea Density and MSVD, are by design surface-level proxies for informational and perspectival complexity. While effective for establishing high-level correlation, these features do not capture the fine-grained syntactic and semantic structures that underpin intensional reasoning. Future work should augment this analysis with more structurally aware features. Second, our four-level classification of perspectival complexity may simplify a multifaceted phenomenon into discrete categories. However, this operationalization was necessary to analyze performance trends. A more fine-grained, continuous complexity score could enable a more nuanced regression analysis in future studies. Finally, our conclusion that LLMs rely on "linguistic scaffolding" and heuristics is drawn from the observed performance patterns and correlations. This study demonstrates that models behave in a way consistent with heuristic-based processing, but does not isolate the precise nature of these heuristics. A crucial next step, is to move from correlation to causation.

Acknowledgments

Gonçalo Guiomar and Karolina Stańczak were supported by ETH AI Center postdoctoral fellowships.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft.

Janet W. Astington. 1993. *The Child's Discovery of the Mind*. The Developing Child. Harvard University Press, Cambridge, MA.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. InternLM2 technical report. *arXiv preprint arXiv:2403.17297*.

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.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Ronald Fagin and Joseph Y. Halpern. 1994. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2):340–367.

Anastasia Giannakidou. 1998. *Polarity Sensitivity as (Non)Veridical Dependency*. Linguistik Aktuell/Linguistics Today. John Benjamins Publishing Company.

Anastasia Giannakidou. 2013. (non)veridicality, evaluation, and event actualization: evidence from the subjunctive in relative clauses. In Maite Taboada and Ljiljana Tvranc, editors, *Nonveridicality, Perspective, and Discourse Coherence*, Studies in Pragmatics Series. Brill.

Jaakko Hintikka. 2005. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Texts in Philosophy. King's College Publications.

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. *arXiv preprint arXiv:2310.06825*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

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. 1971. Some observations on factivity. *Paper in Linguistics*, 4(1):55–69.

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).

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. 2024. From LLM to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Richard Montague. 2008. *The Proper Treatment of Quntification in Ordinary English*, volume 49. Springer, Dordrecht.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

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.

Alexis Ross and Ellie Pavlick. 2019. How well do nli models capture verb veridicality? 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 2230–2240.

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, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2025. Explore theory of mind: program-guided adversarial data generation for theory of mind reasoning. In *The Thirteenth International Conference on Learning Representations*.

Kairit Sirts, Olivier Piguet, and Mark Johnson. 2017. Idea density for predicting Alzheimer's disease from transcribed speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 322–332, Vancouver, Canada. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Frans Voorbraak. 1992. Generalized Kripke models for epistemic logic. In *Proceedings of the Fourth Conference on Theoretical Aspects of Reasoning about Knowledge*, TARK '92, page 214–228, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Story: Xiao Ming receives a bicycle on his birthday. **Ability:** Emotion **Question-1:** What is Xiao Ming's emotion? (A) Embarrassed **(B) Happy** (C) Disappointed (D) Regretful **Ability:** Belief **Question-2:** He should be very happy, but he is very disappointed, why? (A) Xiao Ming worries that riding a bicycle affects his studies. (B) Xiao Ming fears that riding a bicycle to school makes his classmates laugh at him. (C) Xiao Ming thinks the color of the bicycle does not match his clothes. **(D) Xiao Ming hopes for a computer as a gift, not a bicycle.** **Ability:** Emotion **Question-3:** Xiao Ming is having a birthday, he hopes for a computer or a new game as a birthday gift, on his birthday he receives a bicycle. What is Xiao Ming's emotion at this time? (A) Embarrassed (B) Happy **(C) Disappointed** (D) Regretful
Story: Almost every letter to Laura Company contains a check. Today, Laura receives 5 letters. Laura tells you on the phone "I look at 3 out of 5 letters. There are checks in 2 of the letters." **Ability:** Knowledge **Question-1:** Before Laura calls you, how many of these 5 letters do you think contain checks? (A) 0 (B) 1 (C) 2 **(D) 4** **Question-2:** After Laura calls you, how many of these 5 letters do you think contain checks? (A) 0 (B) 1 (C) 2 **(D) 4**

Table 1: Example of the theory of mind questions from the ToMBench.

A Additional Data Details

In Tab. 1, we show a few examples from the ToMBench that we use for the analysis.