Machine Theory of Mind Needs Machine Validation

Adil Soubki[a][b] and Owen Rambow[c][b]
[a]Department of Computer Science, [c]Department of Linguistics
[b]Institute for Advanced Computational Science, Stony Brook University
asoubki@cs.stonybrook.edu, owen.rambow@stonybrook.edu

Abstract

In the last couple years, there has been a flood of interest in studying the extent to which language models (LMs) have a theory of mind (ToM) – the ability to ascribe mental states to themselves and others. The results provide an unclear picture of the current state of the art, with some finding near-human performance and others near-zero. To make sense of this landscape, we perform a survey of 16 recent studies aimed at measuring ToM in LMs and find that, while almost all perform checks for human identifiable issues, less than half do so for patterns only a machine might exploit. Among those that do perform such validation, which we call machine validation, none identify LMs to exceed human performance. We conclude that the datasets that show high LM performance on ToM tasks are easier than their peers, likely due to the presence of spurious patterns in the data, and we caution against building ToM benchmarks relying solely on human validation of the data.

1 Introduction

In cognitive science, theory of mind (ToM) refers broadly to the capacity to reason about the mental states of oneself and others (e.g., beliefs, intentions, emotions) – especially when they may differ from one's own (Premack and Woodruff, 1978). In recent years there has been an explosion of interest in understanding and quantifying the extent to which language models (LMs) demonstrate this ability. Numerous benchmark datasets have been designed to measure this using narratives (Nematzadeh et al., 2018; Le et al., 2019; Gu et al., 2024), human conversation (Bara et al., 2021; Soubki et al., 2024), and adversarial data generation (Sclar et al., 2024).

Despite, or perhaps due to, the growth of ToM evaluation tools in both diversity and number, the extent to which one can say that LMs display ToM remains unclear. Some evaluation metrics find that ToM is almost non-existent in modern models (Kim et al., 2023), others determine that there is evidence but they lack some sort of robustness (Shapira et al., 2024; Jones et al., 2024), while still others find that they already meet or exceed human performance in some respects (Gu et al., 2024; Street et al., 2024). This contradictory set of results leaves the working scientist wondering – do LMs have ToM?

In this position paper we argue that the variety of results seen across these evaluations is, at least in part, due to a lack of what we refer to as "machine validation", an analysis aimed at identifying patterns in data that neural models (but not humans) might exploit. We begin with a brief history of approaches to measuring ToM prior to 2020, and a discussion of how the data may mislead LM-based studies (§2). We discuss the notion of machine validation (§3), and then perform a meta-analysis of 16 recent papers introducing ToM datasets (§4) and find that those which report strong zero-shot LM performance tend to lack a form of machine validation. We present fine-tuning baselines for a sample of four datasets from our meta-analysis (§5); we find that simple models achieve perfect or near-perfect performance on the datasets that omitted machine validation, leading us to outline a suggested workflow for creating ToM (and other) datasets (§6). We conclude with some final recommendations for the study of LM ToM going forward (§7).

2 Theory of Mind in Language Models

The term *theory of mind* was first introduced by psychologists (Premack and Woodruff, 1978) studying the behavior of chimpanzees. They posit that an agent has a ToM "if [they] impute mental states to [them]self and others". The study of ToM was later extended to examine the behavior of children including the, now famous, Sally-Anne test (Wimmer and Perner, 1983; Baron-Cohen et al., 1985) which presents subjects with a narrated or acted

Findings of the Association for Computational Linguistics: ACL 2025, pages 18495–18505
July 27 - August 1, 2025

scene about two or more agents, and a question to see if the subjects understand the story agents' cognitive state. This style of observer-based probing is especially amenable to the study of ToM in LMs, where question answering is already a well studied capability (Al-Mamari et al., 2024; Yang et al., 2018; Joshi et al., 2017). As a result, a number of datasets inspired by psychological experiments have been adapted for LMs over the years. Nematzadeh et al. (2018) produce a template-based question answering corpus (ToM-bAbi) generated from stories inspired by the Sally-Anne test. Le et al. (2019) note that such formulaic data results in a flawed evaluation, especially when using supervised methods, and produce their own templatic corpus (ToMi) which introduces more noise such as distractor sentences and reorderings. Despite these improvements, Sclar et al. (2023) find ToMi to be vulnerable to similar issues.

While recent approaches (see §4) differ greatly from their predecessors, concerns regarding models exploiting spurious correlations (Gordon and Van Durme, 2013; Aru et al., 2023) in order to display so-called *illusory ToM* have remained. Early work on machine ToM did not necessarily focus on zero-shot performance (Nematzadeh et al., 2018; Chandrasekaran et al., 2017; Grant et al., 2017) or even the inclusion of language as input (Rabinowitz et al., 2018). As zero-shot performance has gained priority, fewer studies seem to provide fine-tuned baselines for comparison.

We argue that one manner of checking for the presence of surface cues is to provide these simple, fine-tuned baselines. As humans are not thought to be exploiting such patterns for ToM, very strong performance of simple models (often prone to relying on these patterns) can be an indicator of undesirable correlations in data or a task that is somehow easier than prior work. We keep these observations in mind in our meta-analysis.

3 Machine Validation

Any step taken to ensure that machine performance on a benchmark is not due to the exploitation of spurious correlations specific to machine systems is a form of what we term *machine validation*, i.e., validation designed specifically for machine "subjects". We find that, despite many ToM datasets being designed for machine subjects, there is an over-reliance on validation techniques more suitable for humans or, in some cases, no discussion of validation at all. We are not the first to call for additional machine validation (Shapira et al., 2024; Ullman, 2023), and several techniques have already been proposed (see §6).

There are two broad approaches to machine validation. The first involves designing the benchmark such that it has features that allow one to validate whether models are using certain surface cues (e.g., Rajpurkar et al. 2018; Kim et al. 2023). This enables additional analysis when using the benchmark that should then be reported on. The second approach is to do a post-hoc analysis of the dataset after creation, checking for the presence of spurious correlations through some (possibly statistical) means (McCoy et al., 2019; Sugawara et al., 2020). Roughly speaking, the former focuses on determining what models use at the time of evaluation while the latter focuses on what is present in the data to begin with.

An advantage of the machine validation techniques above is that they clearly identify the problem in the event of an issue. A disadvantage is that they can be fairly specific to the dataset in question, and can require considerable effort. We consider fine-tuning simple baseline models a form of this post hoc approach to machine validation. Though training on a benchmark is more often thought of as a way to evaluate the model, it can also be used to evaluate the benchmark itself. While it will not pinpoint the exact issue, it can indicate that there is a problem, which can then be diagnosed.

4 Meta-Analysis

To obtain the 16 papers selected for analysis, we searched the ACL Anthology for papers since 2020 matching the keyword "theory of mind" and manually inspected their content. We discarded papers which primarily contributed methods for improving models of ToM, rather than evaluation resources. While a number of datasets in related topics may be relevant (e.g., emotion recognition), we restrict our analysis to those specifically designed for ToM. A similar process was repeated by searching Google Scholar using the term "language model theory of mind". We then read the identified papers, paying special attention to the manner in which their data was created, validated, and used in evaluation. We also reviewed several papers in this citation network which did not meet our recency threshold.

The final collection is a diverse sample. It includes a number of datasets compiled to test

	LM Evals		Data Evals		Superhuman	Metadata			
Study	**Zero-Shot**	**Few-Shot**	**Fine-Tuning**	**Human**	**In 1+ Expt.**	**Perspective**	**Format**	**Source**	**Size**
Common-ToM (Soubki et al., 2024)	60.6	-	64	80	No	Observer	MC (2)	Natural	7,374
FANToM (Kim et al., 2023)	26.6	-	53.7	87.5	No	Observer	FR, MC (2)	LM	10,317
OpenToM (Xu et al., 2024)	52.8	-	72.7	92.2	No	Observer	MC (2/3)	LM	13,708
ToMBench (Chen et al., 2024)	74.7	-	-	86.1	No	Observer	MC (4)	Manual	2,470
Social IQa (Sap et al., 2022)	42	73	83*	87	No	Observer	MC (3)	MTurk	1,954
MindCraft (Bara et al., 2021)	-	-	41.7	56.7	No	Interactant	MC (3, 21)	Natural	1,200
FauxPas-EAI (Shapira et al., 2023)	40	-	-	82	No	Observer	MC (2)	Manual	40
MMToM-QA (Jin et al., 2024)	46.7	-	76.7	93	No	Observer	MC (2)	LM	600
Hi-ToM (Wu et al., 2023)	58.9	-	-	-	No (?)	Observer	MC (15)	Template	600
Adv-CSFB (Shapira et al., 2024)	70	-	-	-	No (?)	Observer	MC (3)	Manual	183
ExploreToM (Sclar et al., 2024)	74	-	-	-	No (?)	Observer	FR, MC (2)	LM	1,000*
EPITOME (Jones et al., 2024)	58.9	-	-	70.6	Yes	Observer	FR, MC (2)	Manual	446
BigToM (Gandhi et al., 2024)	84.5	89.7	-	86	Yes	Observer	MC (2)	LM	5,000
Strachan et al. (2024)	88.2	-	-	89.2	Yes	Observer	MC (2)	Manual	105
MoToMQA (Street et al., 2024) 🔒	88.6	-	-	90.4	Yes	Observer	MC (2)	Manual	70
SimpleToM (Gu et al., 2024)	89.5	97.1*	-	-	Yes (?)	Observer	MC (2)	LM	3,441

Table 1: An overview of the ToM datasets surveyed (🔒 indicates not publicly available). The format of the evaluation is noted as multiple choice (MC) with the number of choices appearing in parenthesis, or free response (FR). Size is based on the number of questions and shading indicates performance relative to human baselines (if available). We make note of if their results find models to exceed human performance by at least one reported metric. For datasets that do not provide a human baseline we guess (?) based on similar tasks. Additional details (∗) are in Appendix A.

higher order ToM (Wu et al., 2023; Street et al., 2024), incorporate more tasks (Chen et al., 2024; Jones et al., 2024; Xu et al., 2024; Strachan et al., 2024), include additional modalities (Soubki et al., 2024; Jin et al., 2024), involve social reasoning (Sap et al., 2022; Shapira et al., 2023), and expand on belief-oriented approaches (Street et al., 2024; Shapira et al., 2024; Gandhi et al., 2024; Kim et al., 2023). Gu et al. (2024) make a distinction between explicit ToM (i.e., inferring mental states) and implicit ToM (i.e., making judgments based on these states). In (Bara et al., 2021), agents are evaluated in their ability to cooperate with humans to complete objectives in MineCraft. Sclar et al. (2024) generate questions adversarially, making the evaluation adaptive.

4.1 Data

We compile summary statistics for the 16 studies reviewed. This includes the performance (where available) of humans and their best models in zero-shot, few-shot, and fine-tuning experiments. Eight of the datasets involve composite scores (i.e., the benchmark evaluated more than one aspect of ToM). In this case we compute the mean of reported performance across these categories. We also identify the type of ToM (Kalbe et al., 2010) the studies focus on, classifying the types as cognitive (e.g. beliefs, thoughts) and/or affective (e.g. emotions, desires), as well as whether non-text modalities are available in the corpus.

Other analyses of ToM benchmarks have called for evaluations which situate models as interactants rather then just passive observers (Shapira et al., 2024; Ma et al., 2023). We make note of this feature. We also record the datasets' answer format (multiple choice or free response), source (e.g., manually created by experts, LM generated), and size. The last thing we collect is whether the evaluation finds models to exceed human performance by at least one of their reported metrics (i.e., "superhuman performance"). For datasets which do not provide a human baseline we make an educated guess based on human performance for similar tasks. For additional details regarding the methods of our survey, see Appendix A.

4.2 Findings

The results from our survey are shown in Table 1.

The Good The use of LMs to generate ToM data has raised some concern due to the possibility of low lexical diversity and other output patterns (Xu et al., 2024; Soubki et al., 2024). However, in our analysis we do not see any indication that model performance is strongly correlated with whether the source was human or synthetic. Prior reviews have also called for ToM benchmarks to broaden their scope (Ma et al., 2023). We find several recent benchmarks answer this call by incorporating a variety of skills beyond false beliefs (Chen et al., 2024; Jones et al., 2024; Gu et al., 2024).

The Bad Only a single benchmark (Bara et al., 2021) places models in the role of an active

Task	Subset	Accuracy
FANToM 📁 Kim et al. (2023)	Y/N	65.8 ± 1.63
	MC	49.3 ± 2.12
Common-ToM 🌱 Soubki et al. (2024)	All	61.3 ± 4.95
SimpleToM 📁 Gu et al. (2024)	State	100 ± 0.00
	Judgment	100 ± 0.00
	Behavior	96.6 ± 5.06
BigToM 📁 Gandhi et al. (2024)	Without Belief	96.7 ± 1.62
	With Belief	92.8 ± 9.41

Table 2: Accuracy of BERT (~110M params), GPT-2 (~137M params), and Flan-T5-base (~248M params) when fine-tuned on various ToM benchmarks. Results are computed over five folds (📁) or three seeds (🌱) for the three models, and then pooled for the mean and standard deviation calculations.

participant – perhaps one of the most common scenarios for humans. The remaining all evaluate models' abilities to make ToM inferences as a passive observer. Additionally, only three of the benchmarks include input data in a form other than text and only four include affective aspects of ToM in their evaluation.

The Ugly Many papers discuss the dangers of models exploiting surface-level patterns and spurious correlations to motivate their data creation methodology. Despite this awareness, only one paper (Xu et al., 2024) performs a validation step aimed at identifying and correcting this. A surprising number of papers provide no human baseline to compare against, making it difficult to situate the source of their dataset's difficulty.

Every benchmark which identifies models with superhuman ToM omits machine validation (e.g., computing lexical overlaps, providing fine-tuned model baselines) of their dataset.

5 Case Study

We hypothesize that datasets which report superhuman performance will likely see strong performance in fine-tuning experiments (i.e., fail machine validation). To investigate this we compare the fine-tuning performance of BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), and Flan-T5-base (Chung et al., 2022) across two datasets which did not find superhuman performance (FANToM and Common-ToM) and two datasets which did (SimpleToM and BigToM). The models were chosen to be relatively small by modern standards and include an encoder-only model (BERT, ~110M params), decoder-only model (GPT-2, ~137M params), and encoder-decoder model (Flan-T5, ~248M params). The datasets were selected somewhat arbitrarily from our set of 16 studies to include datasets which we perceived to report poor, moderate, and strong performance, respectively. For FANToM we discard the free response questions to maintain comparability. We average over three seeds for Common-ToM using the author's splits and, for all other datasets, over five folds using cross-validation. We report the average accuracy of all models across all runs. Further details, including hyperparameters, can be found in Appendix B.

5.1 Results

The results of our fine-tuning experiments, averaged over all runs, can be seen in Table 2. For FANToM and Common-ToM, accuracies roughly replicate those reported by the original authors which also fall broadly in line with fine-tuning performance for the other datasets surveyed. On SimpleToM, our models achieve near perfect performance across both their implicit and applied ToM questions. Similarly high performance is observed on BigToM, both in the case where initial beliefs are and are not provided. All models perform very comparably across the datasets with the exception of BigToM where BERT performs a bit worse than the other models in the condition that included initial beliefs (See Table 3 for details). Across runs, standard deviations were generally no more than a few percent. These results are very unusual and, we argue, likely indicate that either (1) the benchmarks are markedly easier than others or (2) zero-shot models are exploiting spurious correlations in these datasets.

6 Validating Your Benchmark

The results of our survey and case study suggest what we suspected from the start — human validation is not enough for ToM benchmarks. Roughly half of the surveyed papers discuss some form of machine validation in the design phase but only six provide such analysis after construction. We outline a workflow for validating ToM benchmarks here.

When designing tasks, think carefully about the sort of heuristics that a model might use to perform well while avoiding ToM reasoning. Two ways to make this less likely are to introduce noise, and to construct adversarial examples. Some examples of noise include adding distractor sentences (Le et al., 2019) and including (possibly multiple) rephrasings of the same task (Sclar et al., 2023; Kim et al., 2023). This can be complemented by the addition of adversarial examples, such as entries that should be impossible to predict (Rajpurkar et al., 2018), or that vary the scenario to reveal model biases (Ullman, 2023). If this introduces subsets that can be used to detect when models are relying on surface cues, this analysis should be included and made clear to users.

Ideally, there are few spurious correlations in the benchmark for models to exploit in the first place, but some work should be done to estimate their prevalence in the completed benchmark. This gives an idea of how likely strong performance is to be a false positive. After construction, *always train a simple fine-tuned baseline*. This could be a small LM (as we do in §5) or a more classical statistical model. If this baseline performs unexpectedly well, consider searching for lexical overlaps (McCoy et al., 2019), employing one of the growing numbers of interpretability techniques (Rai et al., 2025), or returning to options from the design phase.

7 Where Do We Go From Here?

We have found that less than half of the 16 LM ToM studies we examined evaluate their dataset for patterns only a machine might exploit (i.e., perform machine validation). Among those which do perform such validation, none identify LMs to exceed human performance on any aspect of their benchmark, while all studies that find superhuman performance omit such checks. We then performed machine validation by providing a fine-tuning baseline. We found that a small, fine-tuned system could achieve near perfect accuracy on the datasets which did not perform machine validation. This indicates these datasets are, in some sense, easier than their peers, likely due to the presence of spurious patterns in the data. In the following paragraphs we offer some closing thoughts and suggestions.

How do you interpret LM performance on tests designed for humans? It is notable that ToM was first studied in animals, and the manner of testing underwent significant changes when attention was turned towards humans. It is entirely possible, as others have also noted (Ullman, 2023; Shapira et al., 2024; Markowska et al., 2023), that our methods will need to change further to study this phenomena in LMs. In the case of animals to humans, experimenters had to mind the change in capabilities between these two subjects. When observing the performance of LMs on tests originally used for humans, we can't necessarily take away the same conclusions – the capabilities of the subject have changed again. Models may exploit patterns present in our evaluations, otherwise undetectable by humans, that do not broadly generalize to what we wish to measure.

Other evaluation options Changing our evaluation approach might avoid this situation altogether. Moving away from observer-based ToM evaluations towards ones where the agent is situated (Bara et al., 2021), adaptive evaluations (Sclar et al., 2024; Sap et al., 2022), and simulated environments (Jin et al., 2024) all reduce the chances of measuring primary spurious patterns. In other words, we should couple evaluations more closely to the conditions in which ToM is actually used.

Fine-tuning small models is necessary but not sufficient Fine-tuning small models situates the difficulty of a dataset. Unexpectedly strong performance is likely an indicator of undesirable patterns or relative ease. While this may not directly say what in the data models are exploiting, it will indicate that there is probably an issue. The growing number of interpretability techniques (Zhu et al., 2024) and even more classical approaches like measuring lexical overlap (Xu et al., 2024) can help to track down the culprit. We can borrow from the extensive work on more general QA datasets which has run into similar issues, like shortcutting (Sen and Saffari, 2020; Jiang and Bansal, 2019). This is a sufficient, not a necessary condition. It doesn't mean the dataset is free of spurious patterns but if it fails, then it likely means trouble.

But do LMs have ToM? This is a tricky question. If the question is simply "Can they infer mental states?", as described in (Premack and Woodruff, 1978), the answer is plainly, *yes*. However, this has never been the problem. The trouble has always been making sense of the inconsistencies in their performance across seemingly similar contexts. Evaluation tools should not be aimed at measuring the presence of ToM but the *robustness* of ToM (Shapira et al., 2024; Chen et al., 2024). With few common goalposts to situate the difficulty of so many ToM datasets, it's hard to say if models are improving, but it seems clear that performance is not yet robust.

Limitations

We acknowledge that the study of ToM in LMs is progressing rapidly and, while we did our best to include as much work as possible, that our survey may not be comprehensive. We understand that our case study presented in Section 5 could be improved by including additional baselines (e.g., logistic regression on word embedding features) for more datasets and that this lends some uncertainty to our conclusions.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under the CCU (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034) program. Soubki has received additional support from the National Science Foundation (NSF) under No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or DARPA.

We thank both the Institute for Advanced Computational Science and the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources needed for this work. These resources were made possible by NSF grant No. 1531492 (SeaWulf HPC cluster maintained by Research Computing and Cyberinfrastructure) and NSF grant No. 1919752 (Major Research Infrastructure program), respectively.

We would also like to thank our anonymous reviewers for their perceptive comments, which improved this work.

References

Asmahan Al-Mamari, Fatma Al-Farsi, and Najma Zidjaly. 2024. SQUad at FIGNEWS 2024 shared task: Unmasking bias in social media through data analysis and annotation. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 646–650, Bangkok, Thailand. Association for Computational Linguistics.

Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. Mind the gap: Challenges of deep learning approaches to theory of mind. *Artificial Intelligence Review*, 56(9):9141–9156.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Simon Baron-Cohen, Michelle O’riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29:407–418.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of ai’s mind. *Preprint*, arXiv:1704.00717.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC ’13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Erin Grant, Aida Nematzadeh, and Thomas L Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task? In *CogSci*.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *Preprint*, arXiv:2410.13648.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

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.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Elke Kalbe, Marius Schlegel, Alexander Thomas Sack, Dennis A. Nowak, Manuel Dafotakis, Christopher Bangard, Matthias Brand, Simone G. Shamay-Tsoory, Oezguer A Onur, and Josef Kessler. 2010. Dissociating cognitive from affective theory of mind: A tms study. *Cortex*, 46:769–780.

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.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Preprint*, arXiv:2103.13009.

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 the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

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

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2025. A practical review of mechanistic interpretability for transformer-based language models. *Preprint*, arXiv:2407.02646.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2024. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. *Preprint*, arXiv:2412.12175.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online. Association for Computational Linguistics.

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 *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

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.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2024. Views are my own, but also yours: Benchmarking theory of mind using common ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14815–14823, Bangkok, Thailand. Association for Computational Linguistics.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *Preprint*, arXiv:2405.18870.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

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

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. In *Forty-first International Conference on Machine Learning*.

A Survey Details

We provide more information on the source of the performance scores for each dataset, as summarized in Table 1. Table 4 is an extended version of Table 1 with additional columns.

- **Common-ToM** - See Table 3 in Soubki et al. (2024) which reports Mistral-7B-Instruct zero-shot results and Mistral-7B fine-tuning results.

- **FANToM** - See Table 9 from Kim et al. (2023). We take the best results from the “All Question Types” column which includes GPT-4-0613 (June) with CoT for zero-shot performance and Flan-T5-XL for fine-tuning performance.
- **OpenToM** - See Table 2 from Xu et al. (2024) which reports macro-averaged F1 scores. We average over all rows for GPT-4-turbo for zero-shot and Llama2-13B for fine-tuning.
- **ToMBench** - See Table 2 from Chen et al. (2024). We use GPT-4-1106 zero-shot results averaged over English and Chinese performance.
- **Social IQa** - For zero-shot, few-shot, and human performance see Figure 7 from Sap et al. (2022). We use their results for PALM-535B. For fine-tuning performance see Table 8 from Lourie et al. (2021). As the result comes from another paper we note this with an asterisk. The dataset is originally from Sap et al. (2019).
- **MindCraft** - See Figure 5 from Bara et al. (2021) which reports F1. We average V. Tran. performance over all three prediction tasks for fine-tuning performance.
- **FauxPas-EAI** - See Table 1 from Shapira et al. (2023). We take the final accuracy (requiring correct answers on all four questions) of Flan-T5-xxl for zero-shot performance. Human performance cites a study of children aged 9-11 (Baron-Cohen et al., 1999).
- **Hi-ToM** - See Table 5 from Wu et al. (2023). We use the overall performance of GPT-4-32k for the zero-shot results.
- **Adv-CSFB** - See Table 2 from Shapira et al. (2024). We average the zero-shot accuracy of text-davinci-003 over the question and story levels.
- **ExploreToM** - See Table 2 from Sclar et al. (2024). For zero-shot performance we use the accuracy report for GPT-4o when Mixtral 7x8B Inst. was used for question generation. This was computed over a sample of 1000 question pairs and this is what we report in the size column, however note that the “size” of this dataset is ambiguous since the tool can be used for generation. The authors release a set of 13,300 questions to demonstrate this.
- **EPITOME** - See Table 1 from Jones et al. (2024). We use the average zero-shot performance of text-davinci-002 over all tasks.
- **BigToM** - See Table 2 from Gandhi et al. (2024) for model results. We use GPT-4 accuracy without initial beliefs and average over all conditions. We take their 0-shot-CoT results for zero-shot performance and 1-shot-CoT for few-shot. Human performance is taken from Figure 3 and averaged over the same conditions.
- **MoToMQA** - See Table 7 from Street et al. (2024). We average over task types and use results reported with GPT-4 for zero-shot performance.
- **SimpleToM** - See Table 5 from Gu et al. (2024). For zero-shot performance, we use accuracy averaged over belief, behavior and judgment prediction tasks reported for Claude-Sonnet-3.5 with their CoT* prompt. For few-shot performance we take the same information from their MS-remind prompt-chaining experiments. We denote this value with an asterisk to acknowledge that few-shot approaches and prompt-chaining are similar but not equivalent.
- **MMToM-QA** - See Table 1 from Jin et al. (2024). We use multimodal accuracy across all question types. We categorized the BIP-ALM models as fine-tuned and the remaining models as zero-shot, though the distinction is somewhat complicated in this case.

B Experimental Details

All experiments were performed on Tesla V100 or A100 GPUs. We fine-tune `bert-base-uncased`, `gpt2`, and `google/flan-t5-base` for classification for a fixed 10 epochs and record the accuracy at the last epoch. All experiments use cross-entropy loss, the AdamW optimizer with a learning rate of 2e-5 and linear schedule, and a batch size of 1 (GPT-2 and Flan-T5) or 16 (BERT). We pad input text to the maximum sequence length of 512 and manually inspect training loss curves to ensure that models were converging.

We report average accuracy over three seeds (42, 0, 1) for Common-ToM using the provided splits. For corpora without established splits (FANToM, SimpleToM, and BigToM), we perform five-fold cross-validation and report the average over all five folds. Training times typically ranged from 4 to 6 hours for all runs on a given dataset.

SimpleToM asks multiple questions regarding specific scenarios. When splitting we ensure that no scenario appears in both the train and test data.

Task	Subset	BERT	GPT-2	Flan-T5
FANToM ■ Kim et al. (2023)	Y/N	64.9 ± 1.66	66.1 ± 1.63	66.4 ± 1.52
	MC	48.8 ± 1.74	49.7 ± 2.41	49.5 ± 2.50
Common-ToM ❦ Soubki et al. (2024)	All	58.2 ± 2.15	58.1 ± 1.68	67.5 ± 2.67
SimpleToM ■ Gu et al. (2024)	State	100 ± 0.00	100 ± 0.00	100 ± 0.00
	Judgment	100 ± 0.00	100 ± 0.00	100 ± 0.00
	Behavior	96.0 ± 5.95	96.8 ± 4.51	97.1 ± 5.76
BigToM ■ Gandhi et al. (2024)	Without Belief	94.9 ± 1.42	97.8 ± 0.73	97.5 ± 0.50
	With Belief	83.3 ± 11.8	97.1 ± 0.52	97.9 ± 0.59

Table 3: Accuracy of BERT (~110M params), GPT-2 (~137M params), and Flan-T5-base (~248M params) when fine-tuned on various ToM benchmarks. Results (mean and standard deviation) are calculated over five folds (■) or three seeds (❦).

For FANToM we use only the "multiple-choice" and "binary" answer types, as the free response questions are not amenable to classification models. When generating input sequences for BigToM, we shuffle the order of answer choices.

C Expanded Case Study Results

Table 3 shows the results of our case study experiments aggregated per model. As discussed in Section 5, performance across the three models is fairly consistent with one exception. BERT did not train consistently across the folds for the version of BigToM that included initial beliefs in the context, resulting in a lower mean accuracy with higher standard deviation. This could instability could be addressed with additional hyperparameter tuning, but maximizing performance was not the purpose of this study.

Study	LM Evals		Data Evals		Superhuman	Task Type		Modality		Metadata			
	Zero-Shot	Few-Shot	Fine-Tuning	Human	In 1+ Expt.	Cogn.	Affect.	Text	Other	Perspective	Format	Source	Size
Common-ToM (Soubki et al., 2024)	60.6	-	64	80	No	✓	✗	✓	✓	Observer	MC (2)	Natural	7,374
FANToM (Kim et al., 2023)	26.6	-	53.7	87.5	No	✓	✗	✓	✗	Observer	FR, MC (2)	LM	10,317
OpenToM (Xu et al., 2024)	52.8	-	72.7	92.2	No	✓	✗	✓	✗	Observer	MC (2/3)	LM	13,708
ToMBench (Chen et al., 2024)	74.7	-	-	86.1	No	✓	✓	✓	✗	Observer	MC (4)	Manual	2,470
Social IQa (Sap et al., 2022)	42	73	83*	87	No	✓	✓	✓	✗	Observer	MC (3)	MTurk	1,954
MindCraft (Bara et al., 2021)	-	-	41.7	56.7	No	✓	✗	✓	✓	Interactant	MC (3, 21)	Natural	1,200
FauxPas-EAI (Shapira et al., 2023)	40	-	-	82	No	✓	✗	✓	✗	Observer	MC (2)	Manual	40
MMToM-QA (Jin et al., 2024)	46.7	-	76.7	93	No	✓	✗	✓	✓	Observer	MC (2)	LM	600
Hi-ToM (Wu et al., 2023)	58.9	-	-	-	No (?)	✓	✗	✓	✗	Observer	MC (15)	Template	600
Adv-CSFB (Shapira et al., 2024)	70	-	-	-	No (?)	✓	✗	✓	✗	Observer	MC (3)	Manual	183
ExploreToM (Sclar et al., 2024)	74	-	-	-	No (?)	✓	✗	✓	✗	Observer	FR, MC (2)	LM	1,000*
EPITOME (Jones et al., 2024)	58.9	-	-	70.6	Yes	✓	✓	✓	✗	Observer	FR, MC (2)	Manual	446
BigToM (Gandhi et al., 2024)	84.5	89.7	-	86	Yes	✓	✗	✓	✗	Observer	MC (2)	LM	5,000
Strachan et al. (2024)	88.2	-	-	89.2	Yes	✓	✓	✓	✗	Observer	MC (2)	Manual	105
MoToMQA (Street et al., 2024) 🔒	88.6	-	-	90.4	Yes	✓	✗	✓	✗	Observer	MC (2)	Manual	70
SimpleToM (Gu et al., 2024)	89.5	97.1*	-	-	Yes (?)	✓	✗	✓	✗	Observer	MC (2)	LM	3,441

Table 4: Expanded overview of ToM datasets surveyed.