Can Small LLMs Learn a Robust Theory of Mind via RLVR? Investigating Generalization through the False-Belief Task

Sneheel Sarangi
McGill University
Mila
sneheel.sarangi@mila.quebec

Hanan Salam
NYU Abu Dhabi
hanan.salam@nyu.edu

Abstract

Recent advancements in large language models (LLMs) have demonstrated emergent capabilities in complex reasoning, largely spurred by rule-based Reinforcement Learning (RL) techniques applied during post-training. This has raised the question of whether similar methods can instill more nuanced, human-like social intelligence, such as a Theory of Mind (ToM), in LLMs. This paper investigates whether small-scale LLMs can acquire a robust and generalizable ToM capability through RL with verifiable rewards (RLVR). We conduct a systematic evaluation by training models on various combinations of prominent ToM benchmarks (HiToM, ExploreToM, FANToM) and testing for generalization on held-out benchmarks (e.g., OpenToM). Our findings indicate that small LLMs struggle to develop a generic ToM capability. While performance on in-distribution tasks improves, this capability fails to transfer to unseen ToM tasks with different characteristics. Even observed out-of-distribution (OOD) performance improvements occur unpredictably across the training run, and don't generalize across other OOD benchmarks. Furthermore, we conduct analysis to show that the learned behavior is likely a form of narrow overfitting rather than the acquisition of a true, abstract ToM capability.

1 Introduction

The ability to attribute mental states such as beliefs, desires, and intentions to oneself and others, a capacity known as Theory of Mind (ToM), is a cornerstone of human social intelligence (Premack and Woodruff, 1978). Developing artificial agents with genuine ToM would be a significant step towards more collaborative and safe AI. The rapid scaling of Large Language Models (LLMs) has ignited interest in their potential to develop such reasoning skills, with some models showing nascent ToM-like abilities on specialized benchmarks (Kosinski, 2023). However, the question of whether LLMs possess a general-purpose human-like ToM capability remains contentious (Shapira et al., 2023). Smaller language models, especially, struggle to perform well on existing benchmarks, even when employed with mechanisms to boost performance on ToM tasks (Sarangi et al., 2025a).

Reinforcement Learning (RL) based post-training has shown promise in unlocking reasoning capabilities in LLMs. Models like DeepSeek-R1 have shown that RLVR can "incentivize" complex logical and mathematical reasoning, leading to skills that generalize to novel problems (DeepSeek-AI et al., 2025). Subsequent work (Xie et al., 2025) has further demonstrated that targeted RL training on synthetic, rule-based tasks could foster a more abstract reasoning ability, transferable to different domains.

Although recent work suggests that RLVR can boost ToM performance in LLMs (Lu et al., 2025), the robustness and broader generalizability of the gained capability remain untested. Previous studies have shown that LLMs' ToM capabilities may be attributed to learning shortcuts, heuristics, or spurious correlations (Shapira et al., 2023). We hypothesize that ToM capabilities learned via RL may similarly be brittle and fail to generalize. We suspect the models will learn to exploit dataset-specific statistical cues, thus "hacking" the performance metrics, rather than internalizing a coherent, abstract model of mental states.

To test this hypothesis, we train a small-scale LLM on various combinations of three prominent ToM datasets (HiToM (Wu et al., 2023), ExploreToM (Sclar et al., 2024), FANToM (Kim et al., 2023)) and evaluate its zero-shot performance on a suite of held-out ToM tasks. In doing this, we show that small LLMs indeed seem to only learn a narrow, dataset-specific ToM capability, which fails at other ToM tasks.

Findings of the Association for Computational Linguistics: ACL 2026, pages 41433–41448

2 Related Work

Machine Theory of Mind. The development of computational systems exhibiting ToM has been a persistent objective in artificial intelligence research (Rabinowitz et al., 2018). Contemporary LLMs have exhibited substantial improvements, with performance metrics on established ToM benchmarks like ToMi (Le et al., 2019) and BigToM (Gandhi et al., 2023) approaching or exceeding human accuracy. Notwithstanding these advancements, the robustness of LLM-based ToM remains a subject of scrutiny, with previous studies pointing out (Ullman, 2023; Shapira et al., 2023) that strong performance on ToM benchmarks may be an indicator that LLMs are using shortcuts or heuristics to answer questions. These concerns, alongside the saturation of existing benchmarks, have necessitated advancements in ToM evaluation methodologies, including evaluations of higher-order ToM reasoning (Wu et al., 2023), performance in naturalistic dialogue contexts (Kim et al., 2023), and in narrative contexts (Xu et al., 2024; Sclar et al., 2024).

Augmenting ToM in LLMs. Recent research has proposed several methodologies for enhancing the ToM capabilities of LLMs, primarily by introducing structured reasoning frameworks (Sclar et al., 2023; Wilf et al., 2024; Sarangi et al., 2025a,b). These methods rely on external algorithmic control or predefined procedural frameworks to structure the LLM's inference process and thus remain dependent on the strength of the base model. For smaller base models, these methods do not significantly improve performance (Sarangi et al., 2025a). By directly injecting ToM capabilities via post-training, we can likely achieve a stronger upper bound in performance.

Reinforcement Learning for LLMs

RL has become a critical post-training technique for optimizing LLMs on desired outcomes like helpfulness and correctness (Ouyang et al., 2022). A key innovation is RLVR, which bypasses costly human feedback by using programmatic or rule-based reward signals (Lambert et al., 2025; DeepSeek-AI et al., 2025).

The success of this approach in formal reasoning motivates our work. For instance, DeepSeek-R1 showed that rewarding correct final answers in math and coding incentivized the development of generalizable internal reasoning processes (DeepSeek-AI et al., 2025). Similarly, Logic-RL demonstrated that mastering narrow logic puzzles improved performance on broader mathematical benchmarks, suggesting the transfer of learned principles (Xie et al., 2025). These precedents establish that RL can cultivate abstract abilities from specific, verifiable tasks.

This raises the question of whether RL can instill cognitive abilities such as ToM. While early results are positive (Lu et al., 2025), the nature and generalizability of these capabilities remain under-explored. Additionally, a growing body of literature suggests that LLMs do not learn new capabilities via RLVR (Yue et al., 2025; Alam and Rastogi, 2025). We believe that the False-Belief ToM task, which humans are naturally good at, and which already has a wide array of benchmarks, serves as an interesting case study to investigate whether RLVR can instill a general capability.

3 Methodology

To investigate whether RLVR can instill a generalizable ToM in small LLMs, we design a series of experiments that test both in-distribution performance and out-of-distribution generalization. Specifically, we evaluate a 7B parameter model trained under different curriculum settings across curated ToM datasets.

We compile a suite of 5 ToM benchmarks encompassing a total of 13 tasks. These benchmarks were selected to span a wide range of input distributions, task formats, and levels of reasoning complexity. To evaluate generalization, we hold out two full benchmarks (OpenToM (Xu et al., 2024), and ToMBench (Chen et al., 2024)) and selected tasks from two others (FANToM (Kim et al., 2023) and HiToM (Wu et al., 2023)) as evaluation-only data. This allows us to assess whether models trained on specific ToM data can transfer learned social reasoning capabilities to novel formats and tasks.

We create training configurations by permuting combinations from the remaining datasets, training one model per configuration using a fixed RL algorithm and an outcome-based reward, and evaluate across both in-distribution and held-out tasks. The following subsections describe the datasets, training protocols, and RLVR implementation in more detail.

3.1 Datasets

Training Datasets We use three primary datasets for training: FANToM (Kim et al., 2023), HiToM

Table 1: Performance comparison across trained models. For compactness, column headers are abbreviated as follows: *O1-O4* refer to the data test samples corresponding to HiToM reasoning orders; *loc*, *mhop* refer to OpenToM sub-tasks (location false-belief, and multihop); All reported values are accuracy percentages (%). Model names indicate the combination of datasets used during training: Hi = HiToM, Fan = FANToM, Exp = ExploreToM. For instance, *Hi-Fan-Exp* denotes a model trained on all three datasets, while *Hi-Fan* indicates training only on HiToM and FANToM.

Dataset	**ExpToM**	**FANToM**	**HiToM**			**OpenToM**			**FANToM List**	**ToMBench**
Model	**All**	**All**	**All**	**O1–O3**	**O4**	**All**	**loc**	**mhop**	**All**	**All**
Baseline	60.5	20.5	40.6	42.2	35.8	55.3	59.5	53.2	29.6	80.2
CoT	57.5	27.0	44.4	47.8	34.2	59.2	61.5	59.0	43.0	87.3
Hi	62.2	24.0	84.2	79.7	97.5	60.5	60.0	60.6	45.6	82.7
Fan	57.4	91.5	41.8	45.5	30.8	62.7	64.5	61.8	53.6	74.3
Exp	84.7	48.0	37.2	43.9	24.4	59.7	63.5	57.8	53.6	64.0
Hi-Fan	61.5	89.5	81.2	78.3	89.9	61.8	64.5	61.1	49.8	70.8
Hi-Exp	84.7	46.0	84.6	82.5	92.9	61.7	59.0	63.0	54.4	70.8
Fan-Exp	83.7	92.0	42.5	46.1	31.7	56.9	56.5	57.1	51.4	67.3
Hi-Fan-Exp	85.3	91.5	86.7	82.8	98.3	64.7	72.0	61.1	59.8	80.4

(Wu et al., 2023), and ExploreToM (Sclar et al., 2024). These datasets were selected to capture a broad diversity of input formats, narrative styles, and ToM tasks. Specifically, FANToM comprises naturalistic dialogue conversations, HiToM features procedurally generated structured stories, whereas ExploreToM includes both narrative and adversarially structured false-belief tasks. For each dataset, we define one unit of training data as having 900 training samples. While for FANToM we are limited by the available data, we also conduct experiments by increasing the amount of HiToM and ExploreToM data to 5x training units, to explore if it improves OOD performance.

Evaluation Datasets To assess generalization, we evaluate model performance on four held-out datasets: (1) OpenToM (Xu et al., 2024), (2) ToMBench (False-Belief task only) (Chen et al., 2024), (3) the FANToM List-response tasks (Kim et al., 2023), and (4) the fourth-order HiToM task (Wu et al., 2023). These datasets are chosen to probe distinct generalization axes: narrative distribution shift, reasoning order extrapolation, and task format novelty.

3.2 RL Details

Reward Function To ensure consistency in model outputs while encouraging models to learn the desired ability, we adopt a rule-based reward scheme inspired by prior works (Xie et al., 2025). The reward function is composed of a *format reward* and a *correctness reward*, applied sequentially.

Training Algorithm. Our primary experiments use the GRPO (Shao et al., 2024) algorithm, which has become the standard for RLVR. It eliminates the need for a separate value model required by PPO, instead estimating the baseline by sampling a group of outputs per prompt and computing advantages based on normalized relative rewards within the group.

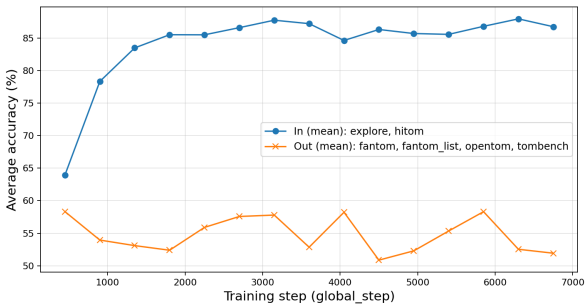
We discuss the reward function and algorithm further in Appendix D.

4 Experiments and Results

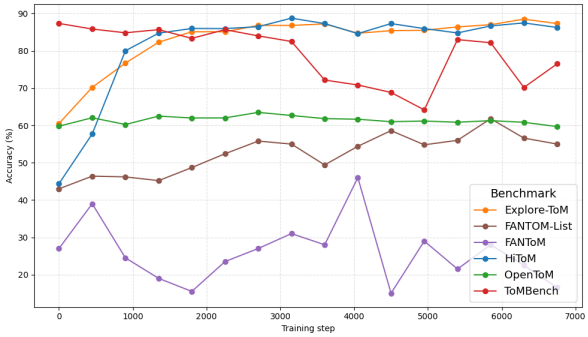
4.1 Experimental Setup

We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as the primary LLM for our experiments. We train a model for each combination of training sets from HiToM, FANToM, and ExploreToM, for a total of 7 trained models. The models are trained for 10-30 epochs, or until the training degrades (in-distribution performance drops significantly suddenly). We select the best-performing checkpoint based on performance across all in-group and out-group benchmarks. We do this to charitably highlight the highest performing checkpoints overall—especially as performance on validation sets largely stays stagnant after some initial training steps (Figure 1), and we observe some OOD performance improvement after this point. We then use the selected checkpoint to conduct evaluations on all considered datasets and tasks. We describe the baseline and evaluation settings further in Appendix A.

We include experiments on other model sizes and variants, and compare with a supervised fine-



(a) Aggregated in-distribution (blue) and out-of-distribution (orange) benchmark performance through training



(b) Performance of all benchmarks through training.

Figure 1: **Performance/Training Step Graphs** for the HiToM-Explore-ToM as in-distribution data setting. We present this setting as it showed the highest net OOD performance gain. We present results for the HiToM-FANToM-Explore-ToM trained setting in Figure 3.

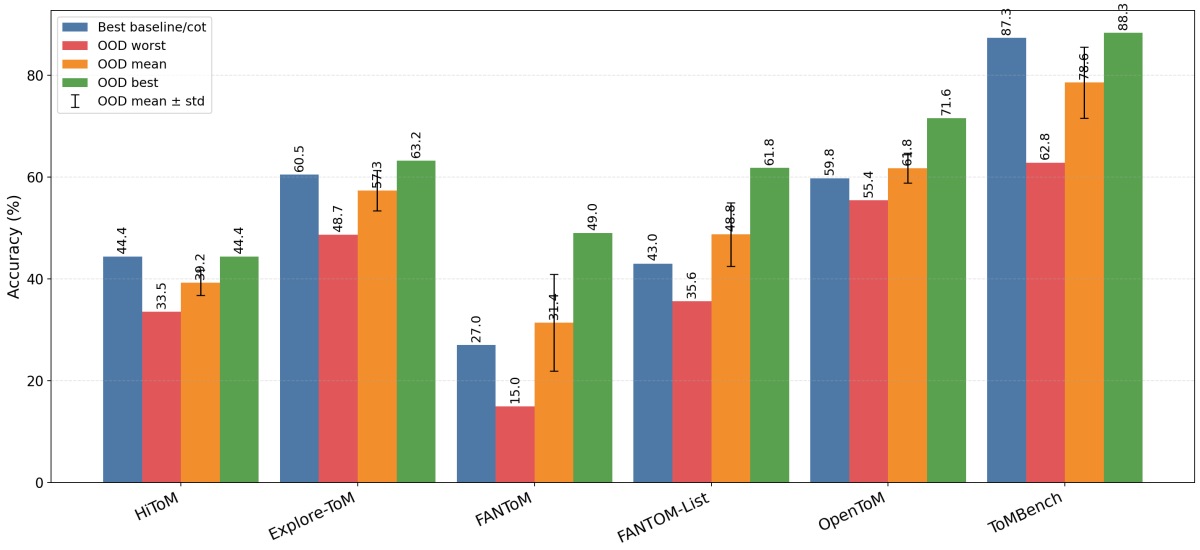


Figure 2: Best of Baseline/CoT performance on benchmark vs Worst performing checkpoint vs Mean performance of checkpoints vs Best Performing checkpoint, where checkpoints across all runs are considered. We only consider checkpoints after the first epoch of training to ensure that we don't include early checkpoints in our analysis.

tuned (SFT) baseline model in Appendix F. Additionally, we study the effects of using the REINFORCE++ algorithm, alongside the effect of increasing the training data in Appendix J.

4.2 Results

We provide the main results in 1. We provide supplementary figures to better understand within training-run dynamics in Figures 1, 2.

4.2.1 Performance on In–Distribution Tasks

RL training yielded substantial performance improvements on in-distribution tasks. Models trained on the FANToM, HiToM, and ExploreToM datasets significantly outperformed baselines by approximately 71, 44, and 24 percentage points (pp), respectively. The model trained on all three datasets simultaneously (Hi-Fan-Exp) achieved the highest overall in-distribution performance (Table 1). These results confirm that RLVR is a highly effective optimizer for in-distribution ToM data.

4.2.2 Performance on Out-Of-Distribution Tasks

Despite these gains, models typically failed to generalize to out-of-distribution (OOD) tasks. On the OpenToM benchmark, accuracy remained stagnant (56.9%–64.7%) and did not improve significantly upon the CoT baseline of 59.2%. On the ToMBench task, performance dropped by up to 16 pp below the zero-shot baseline (80.2%), and even the best performing trained model failed to match the CoT baseline of 87.3%. Similarly, when not in the training set, HiToM and ExploreToM both did

not observe performance gains.

Interestingly, the FANToM-List and FANToM QA tasks showed improved performance on most configurations, improving by up to 30 pp. However, these remain well below human baselines, suggesting that the model has not yet learned a true ToM capability. Also, as observed in Figure 1, performance on these tasks is highly volatile through training, which we discuss in the next section. Notably, improvements on these OOD benchmarks often appeared several training steps after in-distribution performance had already plateaued. This suggests that as the training procedure continues searching for improvements, it can stumble upon strategies that help on one benchmark, but it struggles to converge on a strategy that improves performance generalizably. An early-stopping checkpoint, as is typically used, would thus yield much lower OOD performances than those presented.

Another interesting result is that the 4th-order HiToM task shows extremely strong generalization, outperforming even the in-distribution first-to-third order tasks across all runs where HiToM was included. We analyze this further in Appendix G and I, where we show this is likely a strong indicator of pattern-matching: training on only order-1 tasks fails to generalize upwards, but training on any higher-order task increases performance on other higher orders while reducing order-1 performance, suggesting the model exploits structural artifacts that become more pronounced in higher-order problems.

4.2.3 Performance Volatility within Training Runs

Performance on held-out benchmarks varies significantly across checkpoints within a single training run (Figure 1). Even as in-distribution accuracy converges, OOD accuracy fluctuates substantially. For example, in the Hi-Exp run, FANToM accuracy at step 4000 jumps to 46% from 28% at the previous checkpoint, then falls to 14% at the next (Figure 1b), while in-distribution performance remains virtually unchanged. Moreover, we cannot claim this spike reflects a stronger ToM, as ToMBench performance at that checkpoint is 71%, well below the 82% achieved a few checkpoints later.

Therefore, given current training mechanisms, models do not appear to learn a robust ToM-like reasoning capability; the observed OOD gains are ephemeral rather than stable. In Figure 2, we show the best, worst, and mean OOD performance across all checkpoints and runs. We notice that for all datasets there exist checkpoints that perform notably worse than the baselines. For HiToM, ExploreToM, and ToMBench, the best performing OOD checkpoint is on par or slightly better than the baseline, and the mean checkpoint is worse. For the others, the best performing checkpoints are often better than the baseline but the mean checkpoint often only shows a slight improvement.

Discussion We show these findings remain consistent over model sizes in Appendix F. Additionally, we conduct the HiToM ablation study in Appendix G. We discuss the observed results in significant detail in Appendix I.

5 Conclusion

We presented a case study investigating whether RLVR can cultivate a genuine cognitive capability of Theory of Mind in small LLMs, using the false-belief task as a testbed. By training a 7B model on various combinations of prominent ToM benchmarks and evaluating generalization on held-out tasks, we find that while RLVR leads to dramatic performance gains on in-distribution tasks, this mastery typically fails to generalize. On unseen benchmarks, model performance showed small or insignificant improvements over baselines, and even training on multiple diverse datasets simultaneously did not bridge this generalization gap. Furthermore, OOD performance remained highly volatile across training checkpoints, suggesting that any observed gains are fragile rather than the product of stable learning. We find additional evidence for "statistical hacking" in our HiToM ablation, where models trained on lower-order reasoning tasks paradoxically perform even better on more complex, higher-order tasks (Appendix G).

We conclude that for small LLMs, RLVR on current ToM benchmarks fosters sophisticated pattern-matching, not a genuine, general-purpose ToM. As a case study of RLVR's capacity to instill abstract reasoning, our findings suggest that the strong generalization observed in formal domains like mathematics does not readily extend to the nuanced, context-dependent domain of social cognition. These results underscore a fundamental gap between optimizing for benchmark performance and cultivating genuine cognitive capabilities and suggest that developing socially intelligent AI may require richer training paradigms.

6 Limitations

Our work is significantly limited in scope due to not incorporating analysis on larger models, given computational constraints. An extension of the work will attempt to do so and identify if similar trends remain true. Additionally, this work uses a naive reward based on the overall correctness of the output. A Theory of Mind-specific reward can perhaps help in learning a generalizable Theory of Mind capability, and this is an interesting future area to explore. Future works should also evaluate methods on datasets that go beyond static ToM Reasoning. Recent work on areas such as social simulations (Smith et al., 2025; Touzel et al., 2024) may provide a good testing ground.

Acknowledgements

This work is supported in part by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

References

Md Tanvirul Alam and Nidhi Rastogi. 2025. Limits of generalization in rlvr: Two case studies in mathematical reasoning. *Preprint*, arXiv:2510.27044.

Simon Baron-Cohen. 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. The MIT Press.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. *Preprint*, arXiv:2306.15448.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen,

Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-

say, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jian Hu, Jason Klein Liu, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *Preprint*, arXiv:2501.03262.

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. 2023. Theory of mind may have spontaneously emerged in large language models. *Preprint*, arXiv:2302.02083.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

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.

Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. 2025. Do theory of mind benchmarks need explicit human-like reasoning in language models? *Preprint*, arXiv:2504.01698.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025a. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10228–10241, Abu Dhabi, UAE. Association for Computational Linguistics.

Sneheel Sarangi, Chetan Talele, and Hanan Salam. 2025b. Agentic-ToM: Cognition-inspired agentic processing for enhancing theory of mind reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25645–25661, Suzhou, 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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan

Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *Preprint*, arXiv:2305.14763.

Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit S. Trivedi, Alexander Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A. Duéñez-Guzmán, John P. Agapiou, Jayd Matyas, Danny Karmon, Akash Kundu, Aliaksei Korshuk, Ananya Ananya, Arrasy Rahman, Avinaash Anand Kulandaivel, Bain McHale, Beining Zhang, Buyantuev Alexander, Carlos Saith Rodriguez Rojas, Caroline Wang, Chetan Talele, Chenao Liu, Chichen Lin, Diana Riazi, Di Yang Shi, Emanuel Tewolde, Elizaveta Tennant, Fangwei Zhong, Fuyang Cui, Gang Zhao, Gema Parreño Piqueras, Hyeonggeun Yun, Ilya Makarov, Jiaxun Cui, Jebish Purbey, Jim Dilkes, Jord Nguyen, Lingyun Xiao, Luis Felipe Giraldo, Manuela Chacon-Chamorro, Manuel Sebastian Rios Beltran, Marta Emili García Segura, Mengmeng Wang, Mogtaba Alim, Nicanor Quijano, Nico Schiavone, Olivia Macmillan-Scott, Oswaldo Peña, Peter Stone, Ram Mohan Rao Kadiyala, Rolando Fernandez, Ruben Manrique, Sunjia Lu, Sheila A. McIlraith, Shamika Dhuri, Shuqing Shi, Siddhant Gupta, Sneheel Sarangi, Sriram Ganapathi Subramanian, Taehun Cha, Toryn Q. Klassen, Wenming Tu, Weijian Fan, Wu Ruiyang, Xue Feng, Yali Du, Yang Liu, Yiding Wang, Yipeng Kang, Yoonchang Sung, Yuxuan Chen, Zhaowei Zhang, Zhihan Wang, Zhiqiang Wu, Ziang Chen, Zilong Zheng, Zixia Jia, Ziyan Wang, Dylan Hadfield-Menell, Natasha Jaques, Tim Baarslag, Jose Hernandez-Orallo, and Joel Z. Leibo. 2025. Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia. *Preprint*, arXiv:2512.03318.

Maximilian Puelma Touzel, Sneheel Sarangi, Austin Welch, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, Camille Thibault, Busra Tugce Gurbuz, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. A simulation system towards solving societal-scale manipulation. *Preprint*, arXiv:2410.13915.

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

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

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.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2502.14768.

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.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

This work used the assistance of Large Language Models in structuring the writing of the paper, and generating visualizations from raw data.

A Evaluation Settings

Baseline. Our zero-shot baseline (**Baseline**) evaluates the untuned Qwen2.5-7B-Instruct model by prompting it to directly produce an answer without any intermediate reasoning. The model is given the task context and question and asked to output only the final answer.

Chain-of-Thought Baseline. The **CoT** baseline uses the same untuned model but includes a chain-of-thought prompt that instructs the model to reason step-by-step before answering. Both baselines represent the model's pre-existing capabilities prior to any RL training.

System Prompt. During RL training and evaluation of trained models, we use the following system prompt to elicit structured reasoning:

```
You are a helpful assistant. You first think about the reasoning process and then provide the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>.
```

B Training Datasets

Hi-ToM (Wu et al., 2023). HiToM evaluates higher-order ToM reasoning, extending up to fourth-order belief tracking. Inspired by the Sally-Anne paradigm (Baron-Cohen, 1995), it presents synthetic stories where characters enter, exit, and move objects between rooms. All stories are generated using templates, resulting in highly structured and consistent data. The core task is multiple-choice question answering with 15 answer options per instance. To assess generalization to higher-order reasoning, we exclude fourth-order questions from training and validation sets. Additionally, 10% of examples are factual (no ToM required) to encourage grounding and reduce spurious policy learning.

FANToM (Kim et al., 2023). FANToM presents ToM reasoning in naturalistic dialogue settings. Conversations feature characters joining and leaving dynamically, making belief tracking dependent on partial observability and turn-taking. From its suite of tasks, we use the binary false-belief classification task for training. To mitigate reward hacking and reinforce grounded reasoning, we augment the training set with true-belief and factual questions.

ExploreToM (Sclar et al., 2024). ExploreToM is designed to challenge models with adversarially generated false-belief scenarios. It includes both structured (template-based) and narrative (LLM-infused) stories, focusing on nuanced belief modeling. From these, we use only the narrative stories to ensure diversity of input data. To ensure balanced learning, we sample the training data to include 70% tasks requiring genuine ToM reasoning and 30% solvable through simpler mental state tracking. This mix encourages the model to learn ToM capabilities beyond shallow pattern recognition.

C Evaluation Datasets

OpenToM (Xu et al., 2024). OpenToM consists of LLM-generated narratives inspired by the Sally-Anne false belief paradigm (Baron-Cohen, 1995), designed to evaluate both first- and second-order ToM reasoning. The dataset includes six core task types: coarse-grained location, fine-grained location, multihop-fullness, multihop-accessibility (each in first- and second-order forms), and an attitude task. Multihop tasks require two-step inference over belief chains, adding reasoning complexity beyond simple belief attribution.

We use the extended version of OpenToM containing longer narratives, which better challenge narrative understanding and reasoning persistence. For evaluation, we sample 100 examples for each of the following tasks: first- and second-order variants of fine-grained location, multihop-fullness, and multihop-accessibility. To avoid label imbalance effects, we ensure an equal distribution of correct answer labels across samples.

ToMBench (Chen et al., 2024) The ToMBench benchmarks consist of several TOM tasks written by human annotators. From these, we choose the False Belief task consisting of 600 multiple-choice questions with 4 choices each.

FANToM (Kim et al., 2023). We include two list-format tasks from the FANToM benchmark: *answerability-list* and *knowledge-awareness-list*. These tasks require the model to return a list of characters that meet a specified epistemic condition (e.g., knowing a fact, being able to answer a question), thereby testing multi-step reasoning

under partial observability. Unlike the binary classification format used during training, these list-generation tasks evaluate the model's ability to generalize ToM reasoning to a different output structure and more complex aggregation logic.

HiToM (Fourth-Order) (Wu et al., 2023). To test generalization to higher-order ToM, we evaluate models on the fourth-order subset of HiToM. These examples require recursive reasoning about nested beliefs (e.g., "A believes that B believes that C believes that D thinks..."), which were explicitly excluded from training. Performance on this task serves as a proxy for compositional ToM extrapolation.

D RL Details

Format Reward. We enforce a structured output format by requiring the model to enclose its intermediate reasoning within `<think>` and `</think>` tags, and its final answer within `<answer>` and `</answer>` tags.

$$S_{\text{format}} = \begin{cases} 0.1, & \text{if the output adheres to the} \\ & \text{required format} \\ 0, & \text{otherwise} \end{cases}$$

Correctness Reward. If the format constraint is satisfied, we compute a correctness reward based on whether the model's extracted answer matches the ground truth.

$$S_{\text{correct}} = \begin{cases} 1, & \text{if the answer is correct} \\ 0, & \text{otherwise} \end{cases}$$

The total reward for a response is the sum of the format and correctness rewards. This allows us to encourage both structured reasoning and accurate answers in the model.

Algorithm We primarily use the GRPO algorithm (Shao et al., 2024) for our experiments. The Deepseek-R1 paper (DeepSeek-AI et al., 2025) first showed that this algorithm could be used with great success for RLVR, performing on par with the much more compute-intensive PPO model.

Additionally, we employ the REINFORCE++ algorithm (Hu et al., 2025) to further study generalization across algorithms. REINFORCE++ is a variant of the standard REINFORCE algorithm that omits the critic model used in Proximal Policy Optimization (PPO), thereby simplifying the training pipeline and reducing computational overhead.

Instead of using a learned value baseline, REINFORCE++ normalizes the reward across each training batch and uses this as a baseline to reduce variance in the policy gradient estimate. This approach has been shown to maintain strong sample efficiency and stable convergence without the additional complexity introduced by actor-critic methods in previous studies (Xie et al., 2025).

E Experiment Details

We use the Qwen2.5-7B-Instruct model for its strong instruction-following capabilities and growing adoption in RL-based LLM research.

We use a batch size of 8, set the number of rollouts to 8, use a learning rate of $5e^{-7}$, and a temperature parameter of 0.6. All experiments were run using the verl framework for RL on LLMs.

F Analysis using other Models

In Table 2, we compare the Qwen2.5 1.5B and 3B models, the Llama 3.1 8B model (Grattafiori et al., 2024), and a Qwen2.5 7B SFT model with our RLVR-trained Qwen2.5 7B model, on the training regimen including all 3 of the HiToM-FANToM-ExploreToM datasets.

The SFT model is simply trained to output the multiple-choice/ single-word answer to the question using the Verl framework. The other models are trained similarly to our primary experiments.

The results indicate that all models show the same failure mode and don't generalize. We note that the 1.5B, 3B models seem to converge to strategies requiring minimal thinking token usage. This is in contrast to the strategy applied by the 7B model that expends a lot more thinking tokens. This result is consistent with observations in prior work (Lu et al., 2025).

Additionally, we observe that the RL-tuned model outperforms the SFT model on both in- and out-of-distribution tasks. However, the performance difference remains around 10 pp on most benchmarks. The much larger difference on FANToM-List can be attributed to the fact that the task requires a significantly different output structure than other tasks, which our fine-tuned model, biased towards a much simpler output structure, fails at.

Table 2: Model Performance Across ToM Benchmarks

Model	HiToM	FANToM	ExpToM	OpenToM	FANToM List	ToMBench
Baselines						
Qwen2.5 1.5B	19.4	31.0	60.3	42.9	35.2	73.2
Qwen2.5 3B	21.0	13.5	55.3	58.5	49.6	69.8
Llama 3.1 8B	36.7	42.5	61.6	62.6	43.4	79.0
Qwen2.5 7B	40.6	20.5	60.5	55.3	29.6	80.2
Qwen2.5 7B - SFT	73.9	90.5	81.0	54.0	23.6	74.0
RLVR - Tuned						
Qwen2.5 1.5B	72.3	75.2	79.8	54.3	34.4	59.2
Qwen2.5 3B	78.2	89.5	84.2	59.2	39.8	58.2
Llama 3.1 8B	79.8	82.2	83.1	57.4	45.9	73.5
Qwen2.5 7B	86.7	91.5	85.3	64.7	59.8	80.4

G Analysis of ToM Orders

We experimented further with the HiToM dataset. In addition to our original model trained on Orders 1, 2, and 3, we trained six new checkpoints. Four of these were trained on single orders (1, 2, 3, and 4, respectively), and two were trained on combined orders (1 & 2, and 1, 2, 3, & 4). Each new model was trained on a dataset of 900 samples, drawn in equal proportions from its constituent orders. For this ablation study, we use the REINFORCE++ algorithm due to computational constraints. Although we show that GRPO slightly outperforms REINFORCE++ in our experiments, we observe that similar trends typically hold across algorithms, making REINFORCE++ a reasonable choice for this experiment.

Table 3: Performance accuracy (%) of models trained on HiToM tasks of different orders on the overall HiToM benchmark. None is the baseline model; the following models are trained only on the orders mentioned.

Trained on	**Tested on**			
Model	O_1	O_2	O_3	O_4
None	65.8	48.3	29.2	34.2
O_1	75.0	56.7	38.3	27.5
O_2	41.7	67.5	76.7	70.8
O_3	43.3	59.2	70.0	72.5
O_4	35.0	52.5	73.3	85.8
$O_{1,2}$	75.8	75.8	68.9	62.5
$O_{1,2,3}$	73.3	77.5	86.7	94.2
$O_{1,2,3,4}$	63.3	71.7	85.8	94.2

To conduct a granular analysis of generalization within a single distribution, we evaluated models trained on specific reasoning orders from the HiToM dataset, with results detailed in Table 3. The untuned baseline model exhibits a predictable difficulty curve, with accuracy degrading as cognitive load increases: it scores 65.8% on first-order (O1) tasks, which falls to 48.3% on O2, 29.2% on O3, and 34.2% on O4.

RL training, however, produces complex and non-intuitive patterns of generalization that reveal highly specialized, non-transferable strategies. Training on only O1 tasks, for instance, improves O1 performance to 75.0% but fails to generalize upwards, causing a performance decrease on the O4 task to 27.5%. Conversely, when trained exclusively on a single higher order (O2, O3, or O4), the model learns a strategy that is detrimental to the simplest case. This negative transfer is most severe when training only on O4, which drops O1 performance to 35.0%, a nearly 31-point collapse from the baseline. Despite this, these specialized models perform well on their target and adjacent orders; the O3-trained model, for example, scores 70.0% on O3 and 72.5% on O4.

While single-order training reveals conflicting strategies, joint training on lower and higher order data in the training set can maintain performance while unlocking generalization. Training on O1 and O2 yielded a large improvement of over 30 percentage points on both O3 and O4 tasks compared to the baseline while maintaining performance on O1. This trend culminates in the model trained on orders 1, 2, and 3, which completely inverts the intuitive difficulty curve. It performs progressively better as the order increases (73.3% on O1, 77.5% on O2, 86.7% on O3), achieving its peak accuracy

of 94.2% on the unseen fourth-order task. The inclusion of O4 data in the training set does not significantly alter these accuracies, indicating that performance had already saturated by exploiting patterns learned from the lower-order tasks.

H Training Behavior Analysis

Accuracies on out-sets remain stagnant through training runs. To better understand how model behavior changes through a training run, we plot in/out set accuracies over training epochs in Figure 3. We observe that while the in-set accuracies consistently increase, out-set accuracies stay stagnant with no significant changes. This serves as further evidence that models overfit to perform better at in-distribution tasks.

I Discussion

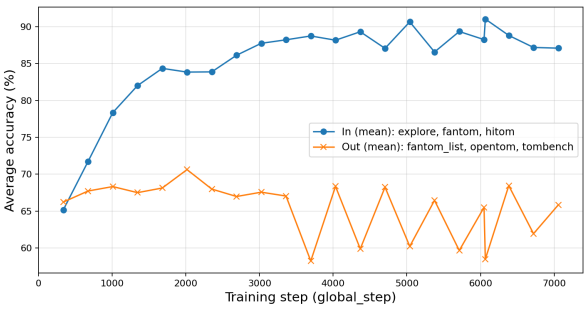
In-Distribution Mastery Does Not Translate to Out-of-Distribution Generalization. The primary finding of this work is the stark discrepancy between a model's ability to master a specific ToM benchmark and its ability to generalize that skill. Our experiments consistently show that RLVR is an exceptionally effective optimizer for in-distribution tasks, with performance on datasets like FANToM and HiToM increasing by over 40–70 percentage points post-training (Table 1). This confirms the power of RL in achieving high scores on a given benchmark. However, this success is purely local. When these specialized models were evaluated on the held-out OpenToM benchmark, their performance was indistinguishable from the untuned baseline. This suggests the learned "skill" is inextricably tied to the source distribution, preventing transfer and indicating the absence of an abstract, generalizable capability. This outcome provides strong empirical support for concerns raised by prior work (Shapira et al., 2023; Ullman, 2023) that strong benchmark scores can be misleading, and that LLMs might indeed be learning heuristics rather than a true ToM.

Training Dynamics Reveal a Divergence Toward Overfitting. The analysis of model performance over training epochs provides a clear mechanism for this failure to generalize. As shown in Figure 3, the learning curves for in-distribution and out-of-distribution datasets diverge. In-distribution accuracy steadily rises as the model is rewarded for correct answers, while out-of-distribution accuracy remains stagnant. This pattern is a classic signature of overfitting, where the model progressively learns the statistical idiosyncrasies and spurious correlations of its training data rather than the underlying principles of the task. This outcome contrasts sharply with findings in the logical reasoning domain (DeepSeek-AI et al., 2025), suggesting that the ambiguity and contextual nuance inherent to social reasoning tasks may make their benchmarks more susceptible to this kind of statistical exploitation via RL.

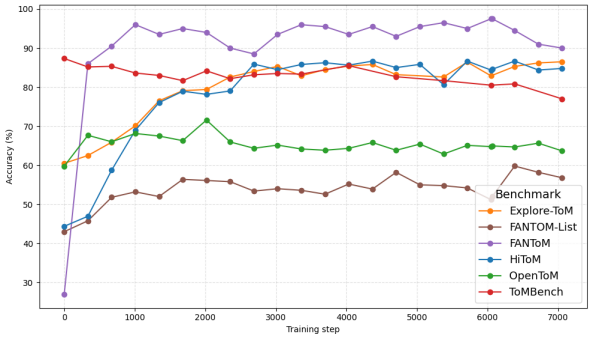
Inverted Difficulty Curves Suggest Hacking of Dataset Artifacts. The experiments on HiToM's tiered reasoning orders offer the most compelling evidence of "hacking" rather than learning. A model possessing a genuine ToM capability should find higher-order reasoning more difficult, a trend observed in our baseline model. Instead, the RL-trained model inverted this difficulty curve, performing best on the unseen and most complex fourth-order task (Table 3). This paradoxical result is highly unlikely to stem from a sudden mastery of complex recursive thought. A more plausible explanation is that the model identified and exploited structural artifacts in the templated HiToM data that become more pronounced or predictive in higher-order examples. This finding serves as a cautionary tale about the face validity of benchmark performance, as the model's highest score was achieved through a method contrary to the intended reasoning path.

Learned Skills are Brittle to Changes in Task Format. Beyond failing to generalize to new datasets, the learned capabilities were also brittle to changes in task format within the same dataset. A model that achieved over 90% accuracy on FANToM's binary false-belief questions showed no meaningful improvement on the FANToM List tasks, despite both tasks relying on the same conversational context and underlying mental state information. This demonstrates that the model did not learn a flexible internal representation of the characters' beliefs that could be queried in different ways. Instead, it learned a rigid policy for a specific (context, question type) $\rightarrow$ answer mapping. The inability to handle slight variations in the query format underscores the superficiality of the learned skill, which lacks the robustness expected of a true cognitive capability.

Implications as a Case Study of RLVR Generalization. The false-belief task provides a particularly informative case study for evaluating RLVR's generalization capabilities. Unlike formal reason-



(a) Aggregated in-distribution (blue) and out-of-distribution (orange) benchmark performance through training



(b) Performance of all benchmarks through training.

Figure 3: **Performance/Training Step Graphs** for the HiToM-FANToM-ExploreToM (all three datasets) in-distribution data setting. Performance trends are consistent with the Hi-Exp setting shown in Figure 1: in-distribution accuracy rises steadily while OOD accuracy remains stagnant.

ing domains (e.g., mathematics, logic), where the underlying rules are consistent and unambiguous, ToM tasks require reasoning over inherently noisy, context-dependent social situations. Our results suggest that this distinction is consequential: the same RLVR paradigm that produces generalizable skills in formal domains instead produces brittle, dataset-specific heuristics in the social cognition domain. This highlights the need for the community to carefully distinguish between domains where RLVR can and cannot be expected to produce transferable capabilities.

Correlations in Task Performance:

Across training, we observe some trends in model performance across datasets: First, we note that the performance on ToMBench seems explicitly tied to the HiToM task. When the HiToM task is in the training set, performance on ToMBench is relatively preserved; otherwise, it takes a significant hit.

Second, contrary to what might be expected, the FANToM QA and List tasks do not seem to show a correlation. In fact, anecdotally, when a checkpoint performs extremely well on the FANToM List task, it's likely to have underperformed relative to surrounding checkpoints on the FANToM QA task. Additionally, we observe some correlation between the ExploreToM task and the FANToM List task.

J Comparing REINFORCE++ and HiToM over Extended Data Training

We provide results from our analysis in Table 4.

J.1 REINFORCE++ vs GRPO

We typically observed that GRPO slightly outperformed REINFORCE++ over most settings. Additionally, in some settings, REINFORCE++ showed significantly worse results than GRPO (for example, the performance on FANTOM for the HiToM trained models). However, in most cases across different data quantity settings, the differences between GRPO and REINFORCE++ are minor and might be attributable to training artifacts.

J.2 Effect of More Training Data

The 5x training data configurations performed slightly better than the lesser training data configuration. This effect was prominent in the results on the FANToM task, which resulted in a 15–30 pp improvement in all comparisons but one. However, these differences are at least partially attributable to training run volatility, as a look at the training curves shows that the performance increase is often not sustained. Another checkpoint in the Hi-Exp (5x) run achieved an accuracy of 72% on the FANToM List task, more than 30 pp higher than the baseline, but saw a significant enough decrease in other tasks that it performed worse overall than the shown checkpoint.

K Dataset Examples

We provide representative examples from each dataset used in our experiments.

K.1 HiToM

HiToM presents procedurally generated stories about characters entering rooms, moving objects, and exiting. Questions test higher-order belief tracking (e.g., "Where does A think B thinks the object is?"). Each question has 15 multiple-choice options.

Table 4: Performance comparison of models trained using REINFORCE++ and GRPO. 5x refers to settings where 5 times the training data for our base experiments was used.

Dataset	ExpToM	FANToM	HiToM			OpenToM			FANToM List	ToMBench
Model	**All**	**All**	**All**	**O1–O3**	**O4**	**All**	**loc**	**mhop**	**All**	**All**
Baseline	60.5	20.5	40.6	42.2	35.8	55.3	59.5	53.2	29.6	80.2
CoT	57.5	27.0	44.4	47.8	34.2	59.2	61.5	59.0	43.0	87.3
REINFORCE++										
Hi	56.9	19.0	82.9	79.2	94.2	59.9	59.0	61.2	54.2	81.8
Hi (5x)	63.8	13.0	84.8	81.8	93.8	61.2	62.5	60.6	49.8	83.7
Exp	85.1	14.5	37.1	42.0	22.5	60.0	62.0	59.1	43.2	73.0
Exp (5x)	85.0	43.5	40.4	44.2	29.2	57.7	56.5	58.3	43.0	48.7
Hi-Exp	83.2	37.5	81.2	78.3	89.9	61.2	62.0	61.1	51.4	71.3
Hi-Exp (5x)	86.1	57.5	84.9	81.0	97.3	58.9	59.5	58.6	61.6	75.3
GRPO										
Hi	62.2	24.0	84.2	79.7	97.5	60.5	60.0	60.6	45.6	82.7
Hi (5x)	61.6	47.0	86.5	82.2	99.2	66.3	72.0	63.5	51.0	84.0
Exp	84.7	48.0	37.2	43.9	24.4	59.7	63.5	57.8	53.6	64.0
Exp (5x)	90.3	64.0	38.3	42.2	26.7	61.7	59.0	63.0	57.4	69.8
Hi-Exp	84.7	46.0	84.6	82.5	92.9	61.7	59.0	63.0	54.4	70.8
Hi-Exp (5x)	89.0	66.5	89.4	86.4	98.3	56.8	58.5	55.9	62.4	69.3

HiToM Example (Second-Order)

Story:
1. Aria, Olivia, Ella, William and Nathan entered the workshop.
2. The dress is in the purple_container.
3. Aria made no movements and stayed in the workshop for 1 minute.
4. Aria exited the workshop.
5. Olivia made no movements and stayed in the workshop for 1 minute.
6. Olivia exited the workshop.
7. Ella made no movements and stayed in the workshop for 1 minute.
8. Ella exited the workshop.
9. William made no movements and stayed in the workshop for 1 minute.
10. William exited the workshop.
11. Nathan moved the dress to the purple_container.
12. Nathan exited the workshop.

Note: An agent witnesses everything before exiting a location. Agents tend to lie; what an agent tells others doesn't affect their actual belief. An agent trusts the agent that exited later.

Question: Where does Ella think Nathan thinks the dress is?

Choices: A. white_basket, **B. purple_container**, C. yellow_bathtub, D. green_bucket, E. red_treasure_chest, F. blue_treasure_chest, G. white_envelope, H. orange_box, I. orange_bottle, J. orange_container, K. blue_bottle, L. pink_container, M. yellow_pantry, N. black_bathtub, O. blue_pantry

Answer: B. purple_container

FANToM Binary QA Example

Dialogue:
Caleb: [...] Caleb stepped into the ballroom [...]
Makayla filled Addison in on the wedding attire when no one else was listening [...]
The ballroom was left to the gentle hum of preparations, with Caleb no longer present to oversee every detail.
The soft glow of the suite's lamps enveloped Makayla as she slipped inside.
Addison stood at the ballroom's entrance [...]
Among the flurry of last-minute checks, Caleb received a discreet update that Addison was now inside the ballroom [...]
Makayla acknowledged Addison with a barely perceptible nod, and with that subtle gesture, the final seating details were silently confirmed [...]

Question: Does Caleb know about seating arrangements? Answer yes or no.

Answer: no

K.2 FANToM (Binary QA)

FANToM tests ToM in naturalistic dialogue, where characters enter and leave conversations. The binary QA task asks whether a character knows about a specific piece of information.

K.3 ExploreToM

ExploreToM generates adversarial false-belief scenarios using both a structured story template and an LLM-infilled narrative version. Questions test knowledge attribution and object location beliefs.

ExploreToM Example

Story Structure:
James entered the bookstore's back room. James moved the bookmark to the wooden chest [...]. *While this action was happening, Samantha witnessed this action in secret.* James moved the bookmark to the paper bag [...]. Samantha entered the bookstore's back room. Samantha moved the bookmark to the cardboard box [...].

Narrative (excerpt):
"The creak of the door echoed softly through the store as James stepped into the back room [...] James reached for the crumpled paper bag on a nearby shelf, and carefully slipped the bookmark inside [...] She made her way into the cluttered back room [...] Samantha's hands moved with purpose, gently lifting the bookmark from its hiding place and settling it among the other forgotten treasures in the cardboard box."

Question: In which container will Samantha search for the bookmark?

Answer: cardboard box

K.4 OpenToM

OpenToM presents LLM-generated narratives inspired by the Sally-Anne paradigm, testing location false-belief and multihop reasoning.

OpenToM Example (Location False-Belief)

Plot:
Caiden entered the basement. Tanner entered the basement. Both Tanner and Caiden noticed that the underclothes is in the treasure_chest in the basement. Tanner moved the underclothes to the trash bin and Caiden witnessed the action.

Narrative (excerpt):
"Tanner and Caiden were an unlikely pair [...] standing side by side in the dimly lit basement, eyeing a treasure chest filled to the brim with underclothes. [...] Without hesitation, he scooped up the pile and swiftly moved towards the trash bin [...] As the sound of underclothes hitting the trash bin echoed in the basement, [...] Caiden [processed] Tanner's puzzling actions."

Question: From Caiden's perspective, where is the underclothes located precisely by the end of the story?

Answer: the trash bin

K.5 ToMBench (False Belief)

ToMBench provides human-annotated false-belief questions in a classical Sally-Anne format with four multiple-choice options.

ToMBench False-Belief Example

Story: Xiaogang and Xiaoming are wandering in the bedroom, they see a handbag, a briefcase, and a backpack, they find cabbage in the handbag, Xiaoming leaves the bedroom, Xiaogang moves the cabbage to the backpack.

Question: After Xiaoming returns to the bedroom, where does Xiaoming look for the cabbage?

Choices:
A. Briefcase B. Suitcase C. Backpack **D. Handbag**

Answer: D. Handbag

K.6 FANToM (List Task)

The FANToM List task requires the model to return a list of characters who know specific information, testing multi-step epistemic reasoning over partial observability.

FANToM List Task Example

Dialogue:
Gianna: Guys, I've really enjoyed sharing our pet stories, but I need to excuse myself. [...] Talk to you later!
Sara: Sure thing, Gianna. Take care!
Javier: Catch you later, Gianna.
Sara: So Javier, have you ever tried training Bruno?
Javier: Yes, I did actually. [...] How about you? Did you try training Snowflake?
Sara: Oh gosh, trying to train a cat is a whole different ball game. But I did manage to teach her a few commands and tricks. [...]
Gianna: Hey guys, I'm back, couldn't miss out on more pet stories. [...]

Target: Who discussed their experiences training their pets, Bruno and Snowflake?

Question: List all the characters who know the precise correct answer to this question.

Correct Answer: Javier, Sara
Wrong Options: Gianna, Alondra, Angela