ing with only three agents, and we use $K=2$ for the same reason. We use knowledge predicate negations 80% of the time to encourage richer inferences (as the fact that an agent does not know something conveys information to others) in announcements and 50% of the time otherwise.

4.2 Controlling for example difficulty

Shortcuts, like hypothesis only bias (Gururangan et al., 2018; Zhang et al., 2023), can lead to the answer without correct reasoning. To control for shortcuts, we trained a relatively shallow supervised model (deberta-small (He et al., 2021), 6 layers, 44M backbone parameters) on a training set combining all setups (ensuring that there was no duplicate and no example that was also in the test set). We used 11.2k training examples for 3 epochs and a learning rate of 3e-5 and 3.73k test and validation examples. Overall validation accuracy was 83%. We also experimented with simpler lexical baselines like TF-IDF which did not capture negations well enough. We assumed that examples correctly predicted by deberta-small with high confidence contained shortcut cues. We used these deberta-small predictions and confidence as additional metadata. We found that the evaluated language models already failed on easy examples. So we used a random subset of the validation and test subsets for our experiments, but our dataset can be filtered by difficulty using the provided confidence level and the discrepancy between deberta-small prediction and ground truth.

We limit the number of agents to 3 and deduplicate then undersample the problems to generate 400 test cases with a perfect balance of True/False labels per setup. We refer to the resulting dataset as MindGames.

4.3 Scaling experiments

We conduct zero-shot experiments and few-shots with a range of language models. We use standard prompting to follow Kosinski (2023) setup. We use the `lm-eval-harness` software (Gao et al., 2021) to measure whether a language model perplexity favors the correct reasoning in a multiple-choice setting, with a natural language inference prompt from Brown et al. (2020): <PREMISE> *Question:* <HYPOTHESIS> *True or False ?"* with two possible continuation choices, *True* and *False*. We evaluate two families of language models:

Human evaluation We present 50 test samples per setup to two NLP researchers only instructed to perform entailment detection. Inter-annotator agreement is 0.89, and average accuracy is 94%[6].

Pythia language models We select the Pythia (Biderman et al., 2023) language models for our open-source scaling experiments. We use the checkpoints trained on the deduplicated corpus (`deduped`) with checkpoint sizes of 70M, 160M, 410M, 1B, 1.4B, 2.8B, and 6.9B.

OpenAI API We evaluate the OpenAI GPT-3 (Brown et al., 2020) models, specifically the ada, babbage, curie, and davinci checkpoints, through the public API. We assume that their model sizes are respectively 350M, 1.3B, 6.7B, and 174B and we use the default temperature.

Figure 1 displays the results for various Pythia model sizes. We observe that scaling improves 5-shot[7] reasoning, but it has no impact on zero-shot reasoning. In contrast to the emergence results reported by Kosinski (2023), Figure 2 does not show a clear scaling trend for GPT-3 models on MindGames data, which suggests that the emergent behavior they observed was not due to robust epistemic logic capabilities.

4.4 Qualitative analysis with ChatGPT

We also run brief qualitative analyses with GPT-3.5 and GPT-4 (OpenAI, 2023), as of May 2023. On 20 randomly sampled problems, we found that GPT3 was 60% correct and GPT-4 70% correct. We present a brief qualitative analysis of the respective models.

As shown in Appendix A, GPT3.5 tends to answer that there is not enough information and to perform correct inferences only when it requires very shallow reasoning. GPT-4 can solve this particular example. However, some problems are still challenging, as shown in Figure 3. GPT-4 rarely answers that there is not enough information and its reasoning looks has the surface form of epistemic reasoning, but occasionally contains glaring mistakes.

5 Conclusion

We developed a novel dataset aimed at evaluating epistemic logic reasoning, addressing a particular

[6]Most errors arose from failing to distinguish between *know whether* and *know that*.

[7]Increasing number of examples did not improve validation accuracy.

User:
[Premise] There are two persons. Everyone is visible to others. There is a mirror in the room. It is publicly announced that someone's forehead is muddy. It is publicly announced that not everyone's forehead is muddy. It is publicly announced that not everyone's forehead is muddy. [Hypothesis] Robert can now know whether or not everyone's forehead is muddy. Is the hypothesis true or false ?

GPT-4:
False. Based on the information provided, Robert cannot know for certain whether or not everyone's forehead is muddy. The statements provided are: There are two persons. Everyone is visible to others. There is a mirror in the room. It is publicly announced that someone's forehead is muddy. It is publicly announced that not everyone's forehead is muddy. Since it is publicly announced that not everyone's forehead is muddy, Robert can deduce that at least one person does not have a muddy forehead. However, this information alone does not allow Robert to definitively determine whether or not his own forehead is muddy.

Figure 3: GPT4 conversation

aspect of ToM. Our results reveal that this task continues to pose challenges for contemporary large-scale language models. When future models can solve MindGames for 2-3 agents, the difficulty can be easily scaled up with more agents. Future studies could better explore human performance on our dataset, taking into account factors such as age and educational background. Additionally, further investigation can examine the impact of fine-tuning on other downstream tasks and assess how well Transformer circuits model Kripke structures that represent modal logic problems.

6 Limitations

Theory of mind is a complex subject, and our study takes a deliberately specific angle, leaving multiple open problems:

Language Our work is centered on English, the method could be adapted to other languages using a subject-verb-object structure. Besides, we restricted our study to templates that do not cover the full variety of the English language.

Prompt structure and models scaling We focused on zero-shot and few-shot prompting, which were sufficient to (Kosinski, 2023), however, Moghaddam and Honey (2023) recently showed that more advanced prompting schemes made significant differences. In addition, we did not explore the full range of Pythia models due to computational limitations.

Task complexity, annotators variation The task we proposed is relatively complex, and raises questions about the profiles of annotators that would match the results of a symbolic reasoner. The framework of DEL itself can also provide insights on theory of mind, as a DEL solver perfectly solves this task, even though we could feel uncomfortable attributing ToM to the solver. We might argue that failing on simple DEL examples disproves ToM, but proving failure is difficult, as mentioned in the previous paragraph.

7 Ethical considerations

This work involves human annotations. However, we used procedurally generated data, ensuring no confidential or harmful content. Besides, annotations were carried out during the researchers' working hours. For these reasons, our Institutional Review Board has determined that it was exempted from formal review according to internal guidelines.

References

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.

Johan Benthem, Jan van Eijck, Malvin Gattinger, and Kaile Su. 2018. Symbolic model checking for dynamic epistemic logic — s5 and beyond*. *Journal of Logic and Computation*, 28:367–402.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.

Thomas Bolander. 2018. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 207–236.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Michael Cohen. 2021. Exploring roberta's theory of mind through textual entailment. *philarchive*.

F. de Waal. 2016. *Are We Smart Enough to Know How Smart Animals Are?* W. W. Norton.

Lasse Dissing and Thomas Bolander. 2020. Implementing theory of mind on a robot using dynamic epistemic logic. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1615–1621. International Joint Conferences on Artificial Intelligence Organization. Main track.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2022. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, pages 1–42.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2022. Logitorch: A pytorch-based library for logical reasoning on natural language. In *The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

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.

Clarence Irving Lewis, Cooper Harold Langford, and P Lamprecht. 1959. *Symbolic logic*, volume 170. Dover publications New York.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*.

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

Ester Navarro, Sara Anne Goring, and Andrew R. A. Conway. 2020. The relationship between theory of mind and intelligence: A formative g approach. *Journal of Intelligence*, 9.

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.

OpenAI. 2023. Gpt-4 technical report.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Kyle Richardson and Ashish Sabharwal. 2022. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11209–11219.

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.