

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

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023b. 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.

Damien Sileo and Marie-Francine Moens. 2022. Probing neural language models for understanding of words of estimative probability. *arXiv preprint arXiv:2211.03358*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Iris Van De Pol, Iris Van Rooij, and Jakub Szymanik. 2018. Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *Journal of Logic, Language and Information*, 27:255–294.

Hans Van Ditmarsch and Willem Labuschagne. 2007. My beliefs about your beliefs: a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209.

Jan van Eijck. 2014. Dynamic epistemic logics. *Johan van Benthem on logic and information dynamics*, pages 175–202.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2023. On the paradox of learning to reason from data. In *IJCAI 23*, pages 3365–3373. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Qualitative example with GPT3.5

User:

[Premise] There are three persons. Everyone is visible to others. It is publicly announced that someone's forehead is muddy. It is publicly announced that Herbert knows whether someone's forehead is muddy. [Hypothesis] Herbert can now know that Herbert's forehead is muddy. Is the hypothesis true or false ?

GPT-3.5:

It is not possible to determine the truth value of the hypothesis based on the given premise. While it is publicly announced that someone's forehead is muddy, it is not clear who the muddy forehead belongs to. Therefore, it is not possible to determine whether Herbert's forehead is muddy or not based on this information alone. However, it is stated that Herbert knows whether someone's forehead is muddy, so Herbert may know whether their own forehead is muddy or not based on additional information not provided in the premise.

Figure 4: GPT-3.5 conversation