Are LLMs Biased Like Humans? Causal Reasoning as a Function of Prior Knowledge, Irrelevant Information, and Reasoning Budget

Hanna M. Dettki[1,2], **Charley M. Wu**[3,4] **& Bob Rehder**[1]
[1]Department of Psychology, New York University, NY, USA
[2]Department of Computer Science, University of Tübingen, Germany
[3]Center for Cognitive Science, Technical University Darmstadt, Germany
[4]Hessian.AI

Abstract

Large language models (LLMs) are increasingly used in domains where causal reasoning matters, yet it remains unclear whether their judgments reflect normative causal computation, human-like shortcuts, or brittle pattern matching. We benchmark 20+ LLMs against a matched human baseline on 11 causal judgment tasks formalized by a collider structure ($C_1 \rightarrow E \leftarrow C_2$). We find that a small interpretable model compresses LLMs' causal judgments well and that most LLMs exhibit more rule-like reasoning strategies than humans who seem to account for unmentioned latent factors in their probability judgments. Furthermore, most LLMs do not mirror the characteristic human collider biases of weak explaining away and Markov violations. We probe LLMs' causal judgment robustness under (i) semantic abstraction and (ii) prompt overloading (injecting irrelevant text), and find that chain-of-thought (CoT) increases robustness for many LLMs. Together, this divergence suggests LLMs can complement humans when known biases are undesirable, but their rule-like reasoning may break down when uncertainty is intrinsic—highlighting the need to characterize LLM reasoning strategies for safe, effective deployment.

Keywords: Large Language Models; Causal Inference; Human and Machine Reasoning; Bayesian Modeling

References

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*