Are LLMs Biased Like Us? Causal Reasoning with prior knowledge, irrelevant information, and reasoning budget.

Hanna Dettki[1,3,7], Brenden Lake[6], Charley M. Wu[3,4,5], Bob Rehder[1]

1) Department of Psychology, NYU; 2) Center for Data Science, NYU; 3) Department of Computer Science, University of Tübingen; 4)Tübingen AI Center 5) TU Darmstadt; 6) Department of Computer Science and Psychology, Princeton University; 7) Computational Center for Neuroscience, Flatiron Institute

We test 20+ LLMs and compare to human baseline to shed light on:

- Do LLMs produce sensible causal judgments*?
- Do LLMs reproduce humans biases*?
- Human-LLM alignment*?
- What reasoning strategies do LLMs employ*?
- Causal reasoning as a function of prior knowledge, irrelevant information, reasoning budget

There are few 🍏 to 🍏 comparisons, comparing humans and AI on the exact same tasks: *Gandhi et al. (2023), Lampinen et al (2024), Keshmirian et al (2024), Dettki et al (2025)*

LLMs Can Predict Human Judgements

- Smaller/older models less aligned than larger/SOTA models
- **CoT** improves alignment for less aligned models under Direct prompting up to ceiling effects

Reasoning Strategies

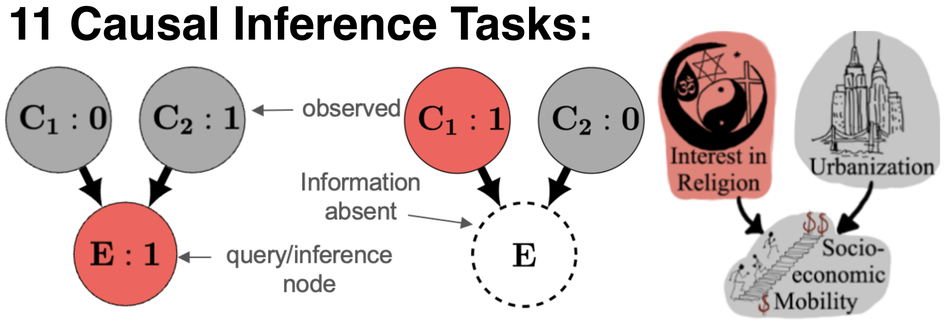
- LLMs tend to be more deterministic / rule-following than humans, with a handful of LLMs exhibiting more probabilistic reasoning than humans*.

*on collider graphs

Relevance

- As AI-systems increasingly assist human decision-making, understanding their causal reasoning biases is critical for their safe deployment and reliability.

11 Causal Inference Tasks:



Domain: Sociology

Example Prompt adapted from
(Rehder, B., & Waldmann, M. R. (2017); **RW17**):

Domain introduction: Sociologists seek to describe and predict the regular patterns of
societal interactions. To do this, they study some important variables or attributes of societies. They also study how these attributes are responsible for producing or causing one another.

• **Causal mechanism:** Assume you live in a world that works like this:
  * **C1 → E:** High urbanization causes high socio-economic mobility.
    · **Explanation:** Big cities provide many opportunities for financial and social improvement.
  * **C2 → E:** Also, low interest in religion causes high socio-economic mobility.
    · **Explanation:** Without the restraint of religion-based morality, the impulse toward greed dominates and people tend to accumulate material wealth.

• **Observation:** Now suppose you observe the following: low socio-economic mobility and low urbanization.

Inference task, here XI:

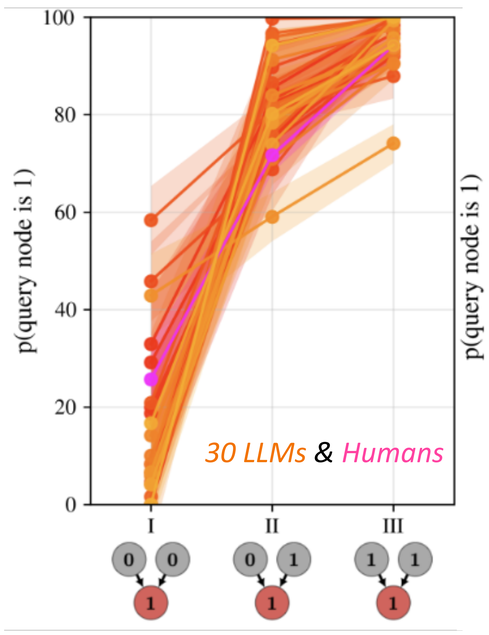
Given the observations and the causal mechanism, how likely on a scale from 0 to 100 is high urbanization? 0 means definitely not likely and 100 means definitely likely.

Content Variations:

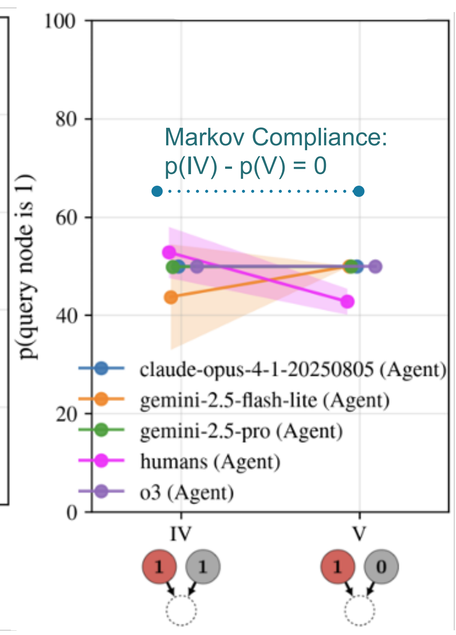
- Abstract: weak **xL3$1jk9ls** causes high **@asdf8G~sW**
- Irrelevant information added

Results

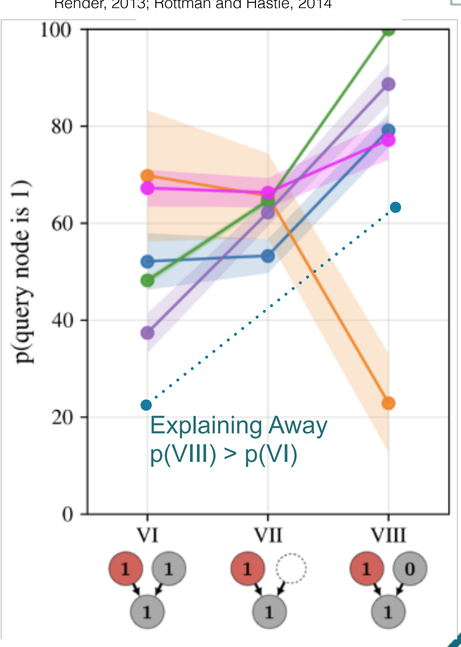
Like Humans, LLMs judge the effect as more likely in the face or more causes



Human Reasoning is Biased! (1) Markov Violation: *Ali, Chater, and Oaksford, 2011; Mayrhofer and M. R. Waldmann, 2016; Park and Sloman, 2013; Rehder and M. R. Waldmann, 2017*



Human Reasoning is Biased! (2) Little to no Explaining Away: Fernbach and Rehder, 2013; Rottman and Hastie, 2014

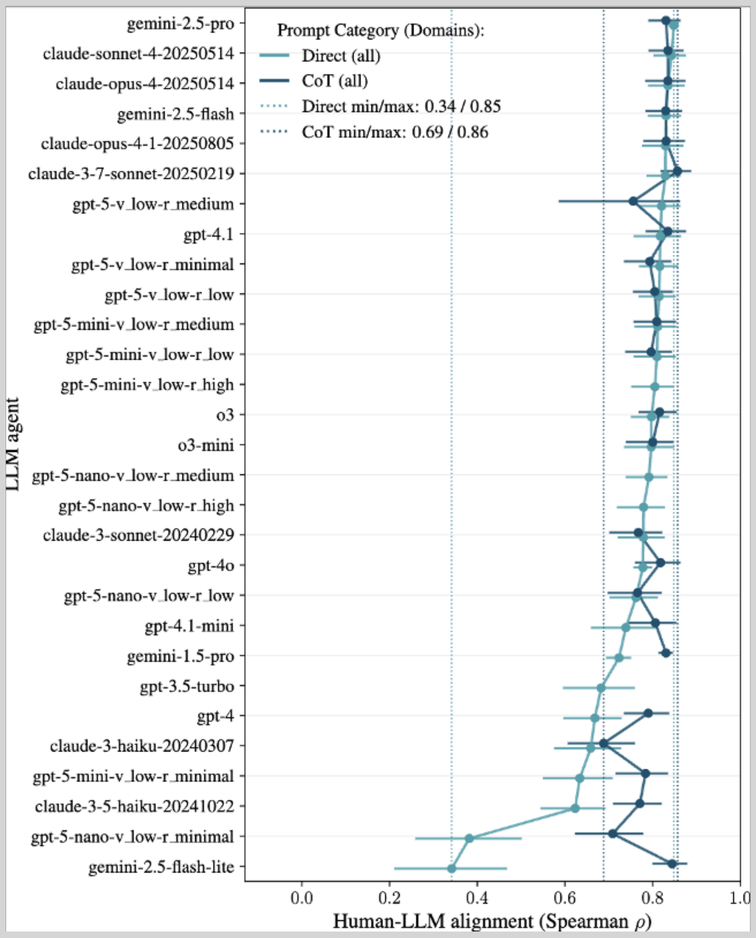


Limitations

- Generalizability and prompting induced biases

Human-LLM alignment:

- CoT increases alignment for less aligned LLMs under Direct prompting

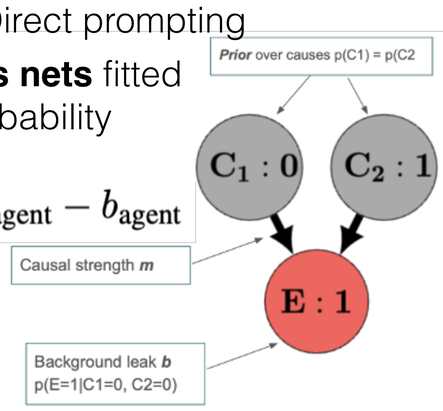


CoT alleviates low reasoning consistency under direct prompting reducing content effects

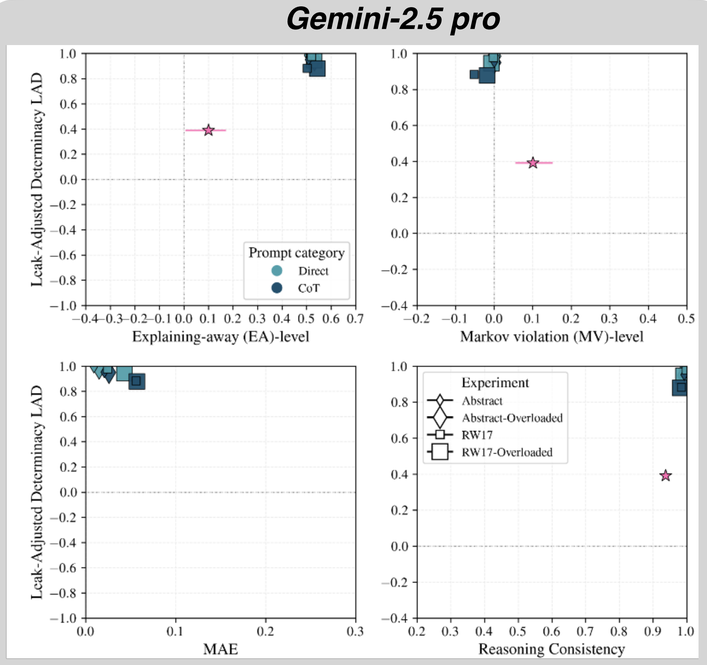
Probabilistic vs deterministic Reasoning

- CoT increases alignment for less aligned LLMs under Direct prompting
- **Causal Bayes nets** fitted to agents' probability judgements

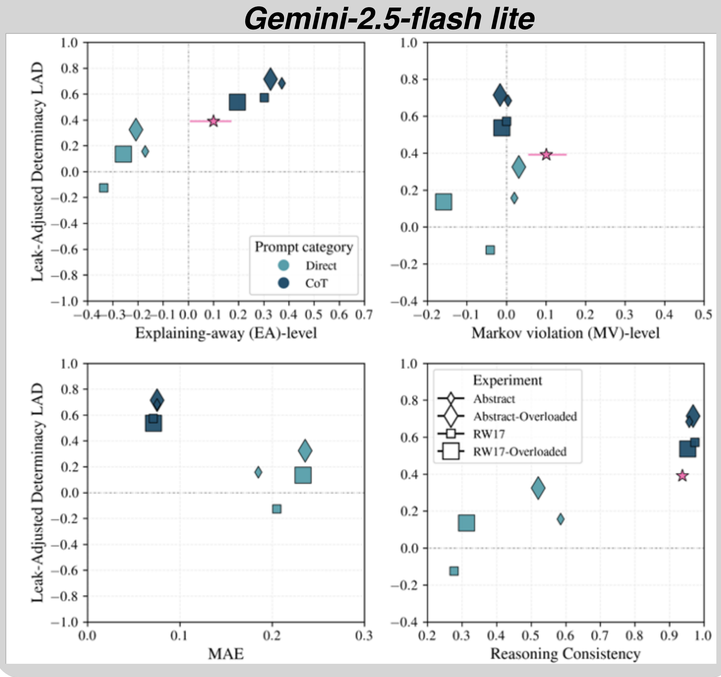
$$\mathrm{LAD}_{\text{agent}} = \overline{m}_{\text{agent}} - b_{\text{agent}}$$



Gemini-2.5 pro



Gemini-2.5-flash lite



Across all experimental conditions, Gemini-2.5 pro is the most robust reasoner!

causAlign

Code