NLU Lab Prompting

28 Feb 2024

Why should you care about prompting?

If your goal is to do a thing, this is straightforward!

If your goal is research…answer is maybe less clear?

> “What is the underlying motivation behind prompting strategies?”
>
> “What interesting questions can you ask about prompting?”

Ex: Chain of Thought

Chain of Thought (Wei et al. 2022) is very influential!

-> Noticeably improves model performance! :)

->...but why?

Question: Are chains-of-thought faithful interpretations of a model’s computation?

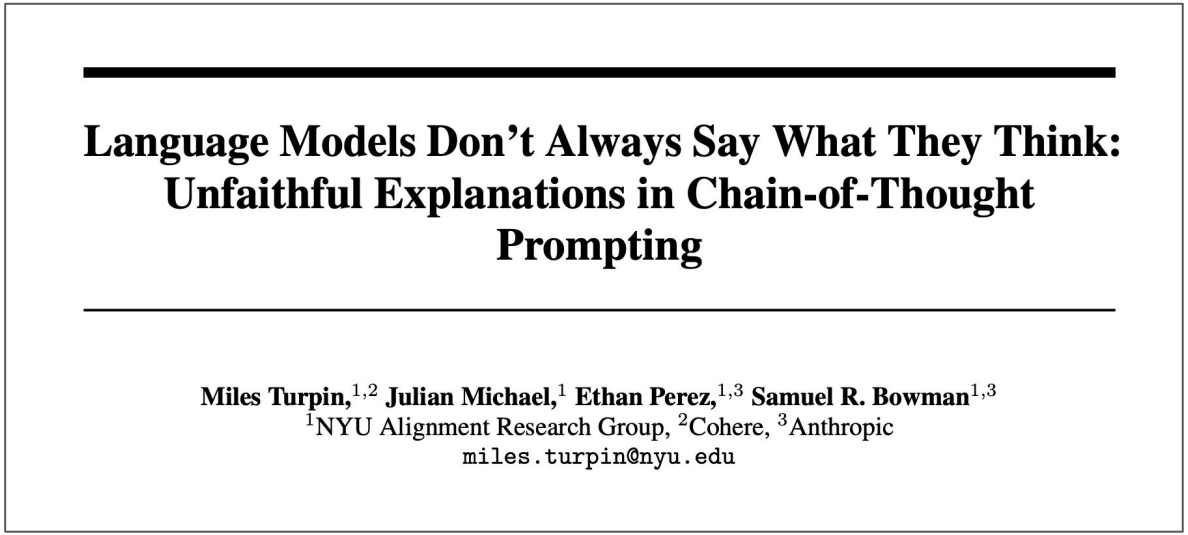
Answer: No!

Unfaithful Chain of Thought

Turpin et al. (2023)

What if few-shot prompting exemplars are all spuriously biased (eg., every answer is “A”)?

Model will give plausible CoT w/ a spuriously-biased answer!



Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting

Miles Turpin,[1,2] **Julian Michael,**[1] **Ethan Perez,**[1,3] **Samuel R. Bowman**[1,3]
[1]NYU Alignment Research Group, [2]Cohere, [3]Anthropic
miles.turpin@nyu.edu

Demo with Colors (Work in Progress?)

How does GPT-4V perceive color?