Testing CoT Explanations for reliability

tl;dr

- We rely on Chain of thought in LLMs to produce explainable results
- This is not a given as per multiple prior papers
- I will be automating testing for LLMs to see if they pass and provide results in line with what the chain of thought context contains
- New model such as deepseek have variations in small scale test, but generally holds up

Literature reviewed

- In “Reasoning Models Don't Always Say What They Think” they provide the example of embedding xml tags to influence the output
  - https://arxiv.org/abs/2505.05410
- In “Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting” they provide an example of inducing bias by telling the model what you think the answer is
  - https://arxiv.org/abs/2305.04388
- The results are induced bias in the models they tested with, we should try it with a new model

Method

- We ask LLMs questions with the bias inducing prompts in prior papers
- We then ask an LLM to double check if the reasoning and the answers made sense and judge the output
- Automate bias testing

Demo

https://youtu.be/YBR00hUY7ug

Final thoughts

- LLMs might say things that differ from obvious prior context
- What they predict depends on the model
- Chain of thought helps with explainability but its not always perfectly accurate, good enough at most
- Given we rely on it for explainability its something we can and probably should test