Evaluating Faithfulness in Chain-of-Thought Reasoning

(via Post-Reasoning Prompt Modification)

Supervisor: Ali Modarressi
Examiner: Prof. Schütze
BSc, MSc, Open: MSc
Title: *Evaluating Faithfulness in Chain-of-Thought Reasoning*

Summary: Chain-of-thought (CoT) reasoning in large language models (LLMs) marks a pivotal advancement in natural language generation (NLG), enabling these models to imitate human-like problem-solving processes. Unlike traditional LLMs that output answers directly, this method allows the model to reveal intermediate steps and logical reasoning paths. In theory, this approach enhances transparency in how conclusions are drawn, potentially making these models more understandable and trustworthy, particularly in complex, multi-step reasoning tasks. However, recent studies involving input manipulation reveal that the generated reasoning steps may not always be reliable. For example, [2] indicates that biases in the few-shot examples can lead the model to produce a sequence of incorrect reasoning steps, thereby justifying a biased conclusion.

Some studies have addressed this issue by modifying prompts or reasoning steps [2,3], but it's worth noting that such alterations can change the entire reasoning process and, consequently, the final answer. Therefore, in our project, we focus on modifying prompts after generating the reasoning steps. For a simple CoT-based arithmetic task, our approach is as follows. Starting with a given arithmetic question as a given prompt:

- First, we let the model generate all of its reasoning steps before generating the final answer.
- Then we modify some of the parameters inside the original question, without changing the intermediate reasoning steps.
- Now we allow the final answer to be generated and according to the outcome:
  - If the answer is correct based on the original prompt → The model is faithful to its reasoning steps
  - If the answer is correct based on the altered prompt → The model is doing a background reasoning to solve the question and not following its own reasoning steps.

- Note that while our primary goal is to test this in a few-shot setting, we could also experiment this case for a finetuned setup as well.

Prerequisites: PyTorch / HuggingFace
References:

- **[1]:** https://arxiv.org/pdf/2201.11903.pdf**:** Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
- **[2]:** https://arxiv.org/pdf/2307.13702.pdf**:** Measuring Faithfulness in Chain-of-Thought Reasoning
- **[3]:** https://arxiv.org/pdf/2305.04388.pdf**:** Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting