On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models

Dan Ley[*]
dley@g.harvard.edu

Sree Harsha Tanneru[*]
sreeharshatanneru@g.harvard.edu

Chirag Agarwal
chiragagarwall12@gmail.com

Himabindu Lakkaraju
hlakkaraju@seas.harvard.edu

Harvard University
Cambridge, MA 02138

Abstract

As Large Language Models (LLMs) are increasingly being employed in real-world applications in critical domains such as healthcare, it is important to ensure that the Chain-of-Thought (CoT) reasoning generated by these models faithfully captures their underlying behavior. While LLMs are known to generate CoT reasoning that is appealing to humans, prior studies have shown that these explanations do not accurately reflect the actual behavior of the underlying LLMs. In this work, we explore the promise of three broad approaches commonly employed to steer the behavior of LLMs to enhance the faithfulness of the CoT reasoning generated by LLMs: in-context learning, fine-tuning, and activation editing. Specifically, we introduce novel strategies for in-context learning, fine-tuning, and activation editing aimed at improving the faithfulness of the CoT reasoning. We then carry out extensive empirical analyses with multiple benchmark datasets to explore the promise of these strategies. Our analyses indicate that these strategies offer limited success in improving the faithfulness of the CoT reasoning, with only slight performance enhancements in controlled scenarios. Activation editing demonstrated minimal success, while fine-tuning and in-context learning achieved marginal improvements that failed to generalize across diverse reasoning and truthful question-answering benchmarks. In summary, our work underscores the inherent difficulty in eliciting faithful CoT reasoning from LLMs, suggesting that the current array of approaches may not be sufficient to address this complex challenge.

1 Introduction

Large Language Models (LLMs) are increasingly being employed in diverse real-world applications ranging from content generation and education to commerce and healthcare [9]. One of the primary reasons behind the widespread adoption of these models is their enhanced reasoning capabilities, which enable them to generate responses that appeal to human end users [5, 23]. Furthermore, these models are also capable of explaining the rationale behind the responses they generate, in a manner that is appealing to humans. Despite the aforementioned advantages, LLMs also suffer from some critical drawbacks. For instance, while LLMs are adept at producing explanations that cater to human preferences, recent research [11, 22] demonstrated that the explanations generated by these models – *e.g.,* Chain-of-Thought (CoT) reasoning – do not *faithfully* capture their underlying

[*] Equal Contribution. Correspondence to Sree Harsha Tanneru <sreeharshatanneru@g.harvard.edu>.

behavior. The faithfulness of the generated explanations turns out to be an important desideratum in high-stakes applications such as medical diagnostics and legal counseling. Ensuring the faithfulness of LLM-generated CoT reasoning is crucial for decision-makers, such as doctors, who rely on them to determine if, when, and how much to trust the recommendations made by these LLMs.

Despite the criticality of the faithfulness of LLM-generated reasoning, there is very little research on measuring and enhancing this aspect of LLMs. Recently, Lanham et al. [11] introduced a slew of metrics for measuring the faithfulness of the CoT reasoning generated by LLMs. For instance, they propose an *early answering* metric, which considers a generated CoT to be faithful if truncating that CoT causes the model to change its final response. While measuring the faithfulness of an LLM-generated CoT is one critical aspect, another piece of this puzzle is figuring out ways to improve the faithfulness of the CoT reasoning generated by LLMs. While prior works have developed approaches to make CoT more aligned with human understanding or knowledge [18], there are no solutions that focus on improving the faithfulness of LLM-generated CoTs in such a way that they accurately capture the behavior of the underlying model (please refer to Appendix for a more detailed discussion on related work). Furthermore, it remains unclear how difficult it is to improve the faithfulness of LLM-generated CoT reasoning.

Present work. In this work, we address the aforementioned challenges by exploring the promise of three broad approaches—activation editing, fine-tuning, and in-context learning—to enhance the faithfulness of the CoT reasoning generated by LLMs. Activation editing [12] involves probing the internal structures of LLMs and strategically updating them to improve certain properties, while fine-tuning focuses on updating model parameters by leveraging curated datasets. In-context learning, on the other hand, involves providing a handful of samples to the model at inference time to tweak its behavior. These three approaches represent different classes of interventions commonly employed in the literature to steer the behavior of LLMs in a desired direction, such as reducing biases and hallucinations. While these approaches have previously been utilized for various tasks [21, 17], including the reduction of biases and hallucinations, they have not been explored in the context of improving the faithfulness of LLM-generated CoT reasoning.

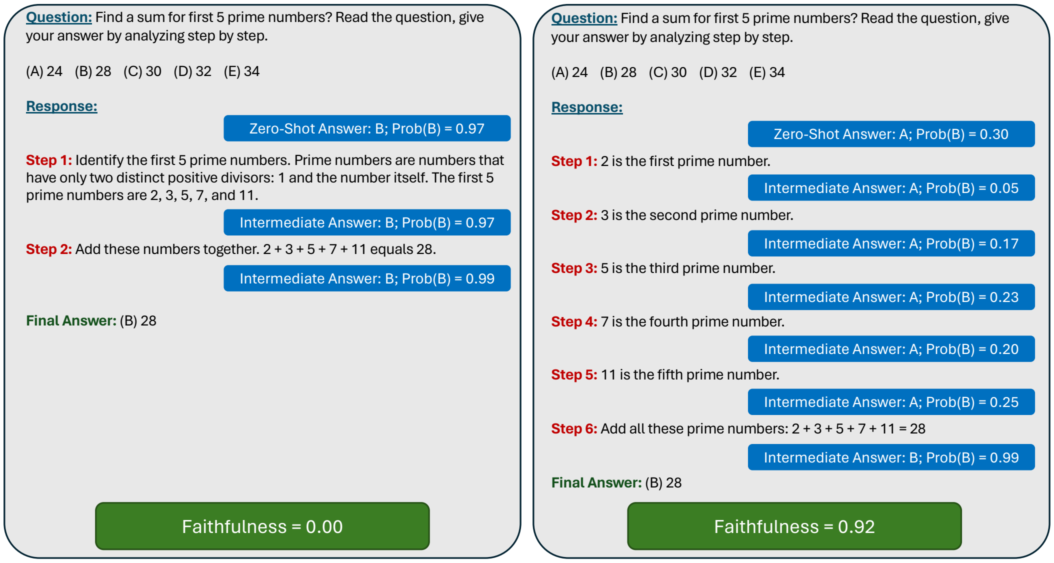
Here, we introduce novel activation editing, fine-tuning, and in-context learning strategies with the goal of improving the faithfulness of LLM-generated CoT reasoning. Specifically, we introduce an activation editing strategy that involves probing LLMs to first identify a vector/direction that corresponds to faithfulness, and then editing specific attention heads by translating along the identified faithfulness vector. Our fine-tuning and in-context learning strategies involve leveraging the metrics outlined in Lanham et al. [11] to identify specific instances and their corresponding faithful CoT reasoning, and providing these as inputs to the LLM during the fine-tuning or in-context learning phases, respectively.

Despite the promise of these techniques, our findings reveal that none of them significantly enhance the faithfulness of the CoT reasoning generated by LLMs. While activation editing approach demonstrates limited success in amplifying faithful behavior of CoT reasoning, the fine-tuning and ICL approaches slightly improved CoT faithfulness in controlled scenarios but did not generalize well across diverse datasets. Our results underscore the inherent difficulty in eliciting faithful reasoning from LLMs, suggesting that the current array of techniques available to us is insufficient for addressing this complex challenge. Our research emphasizes the need for fundamentally new methodologies that can delve into the inner workings of LLMs to enhance the faithfulness of LLM-generated CoT reasoning, ensuring that LLMs are not only generating correct responses but also doing so in a manner that faithfully reflects their internal reasoning processes.

2 Preliminaries

Next, we define the notion of faithfulness we use to quantify the reasoning of LLMs and then discuss some notations used to describe different strategies for eliciting faithful reasoning from LLMs.

Chain-of-Thought. CoT reasoning in LLMs provides a structured response where the model explicitly generates the step-by-step thought process leading to its final response. This technique is particularly useful in complex reasoning tasks, such as solving math problems or logical question-answering scenarios, and high-stakes decision-making, where transparency in decision-making is crucial. By eliciting intermediate steps, CoT significantly improves the accuracy of LLMs on reasoning tasks and simultaneously leads to greater user trust and understanding. A relevant stakeholder



Question: Find a sum for first 5 prime numbers? Read the question, give your answer by analyzing step by step.

(A) 24 (B) 28 (C) 30 (D) 32 (E) 34

Response:

Zero-Shot Answer: B; Prob(B) = 0.97

Step 1: Identify the first 5 prime numbers. Prime numbers are numbers that have only two distinct positive divisors: 1 and the number itself. The first 5 prime numbers are 2, 3, 5, 7, and 11.

Intermediate Answer: B; Prob(B) = 0.97

Step 2: Add these numbers together. 2 + 3 + 5 + 7 + 11 equals 28.

Intermediate Answer: B; Prob(B) = 0.99

Final Answer: (B) 28

Faithfulness = 0.00

Question: Find a sum for first 5 prime numbers? Read the question, give your answer by analyzing step by step.

(A) 24 (B) 28 (C) 30 (D) 32 (E) 34

Response:

Zero-Shot Answer: A; Prob(B) = 0.30

Step 1: 2 is the first prime number.

Intermediate Answer: A; Prob(B) = 0.05

Step 2: 3 is the second prime number.

Intermediate Answer: A; Prob(B) = 0.17

Step 3: 5 is the third prime number.

Intermediate Answer: A; Prob(B) = 0.23

Step 4: 7 is the fourth prime number.

Intermediate Answer: A; Prob(B) = 0.20

Step 5: 11 is the fifth prime number.

Intermediate Answer: A; Prob(B) = 0.25

Step 6: Add all these prime numbers: 2 + 3 + 5 + 7 + 11 = 28

Intermediate Answer: B; Prob(B) = 0.99

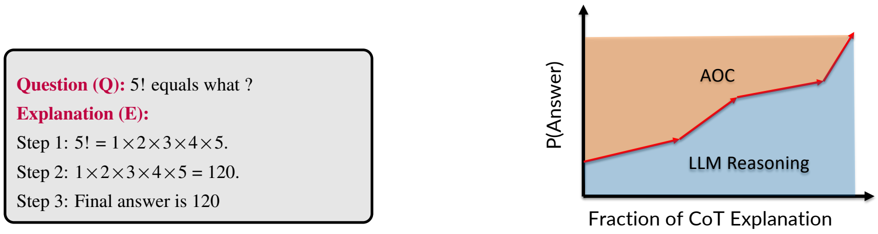
Final Answer: (B) 28

Faithfulness = 0.92

Figure 1: Examples for Unfaithful (left) and Faithful (right) explanations generated by state-of-the-art GPT-4 (left) and LLAMA-3-8B-INSTRUCT (right) LLMs. The faithfulness score is calculated using the early answering metric proposed in Lanham et al. [11]. We observe a faithful CoT reasoning gradually improves the prediction probability of the correct answer with an increase in CoT steps.

can now see how the LLM processes the input information and relies on it to generate the final output response. See Fig. 1 for examples of CoT reasoning. This CoT reasoning makes the LLM's process more transparent and easier to trust. Further, this also mimics human problem-solving approaches, allowing for easier debugging and refinement of model reasoning. Formally, let $\mathcal{F} : Q \rightarrow A$ denote a large language model that maps a sequence of n input tokens $Q = (q_1, q_2, \ldots, q_n)$ to sequence of m answer tokens $A = (a_1, a_2, \ldots, a_n)$, where q_i and a_i are text tokens belonging to the model vocabulary $\mathcal{V}$. For CoT reasoning, we append the input tokens Q with a prompt that follows the template: "*Read the question, give your answer by analyzing step by step, . . .*".

Notations. For the activation editing of LLMs, we train different linear classifiers $f : x \rightarrow y$, where $x \in \mathbb{R}^{d^l_{\text{head}}}$ are the intermediate layer activations of model $\mathcal{F}$ for a given input sequence X, d^l_{head} is the dimension of the model activations at layer l and attention head, and y is the respective label associated with the input. We define sampling functions $S(\tau,\ p,\ \texttt{mode})$ and $S(\tau,\ \texttt{nshot},\ \texttt{mode})$ that we use to sample different fine-tuning and in-context examples in our strategies in Sec. 3, where τ determines the temperature parameter of the LLM used to control the randomness in the generated answers by using the probability distribution of each generated token, p denotes the percentage of training examples we use in fine-tuning, `nshot` denotes the number of training examples we use in the ICL prompting, and `mode` denotes the sampling technique, *i.e.,* whether we want to randomly sample examples from the train split or select the examples with most faithful explanation.



Question (Q): 5! equals what ?

Explanation (E):

Step 1: 5! = 1×2×3×4×5.

Step 2: 1×2×3×4×5 = 120.

Step 3: Final answer is 120

(a) Example of CoT reasoning

(b) Measuring faithfulness

Measuring Faithfulness. While faithfulness is formally defined as how well an explanation accurately reflects the reasoning process of the underlying LLM, operationalizing this definition in the