Appendix

A Broader Impact

Our work focuses on exploring whether we can improve the faithfulness of the CoT reasoning generated by state-of-the-art LLMs. This has significant positive implications for societal benefit. For instance, if CoT reasoning output by LLMs faithfully captures the underlying model behavior, decision-makers and relevant stakeholders can leverage this to determine if, when, and how much to rely on the recommendations provided by LLMs. Therefore, our exploration itself is very valuable and has a substantial positive societal impact. Our analyses and findings indicate that existing techniques commonly used to steer behavior in LLMs are not effective in enhancing the faithfulness of LLM-generated CoT reasoning. While this finding is not particularly positive, we believe it is a step in the right direction, informing us of the complexity of the problem and underscoring the need for fundamentally different frameworks to address it. As far as we understand, our work does not have any potential negative societal impacts, as it is mainly an exploration to improve the faithfulness of LLM-generated CoT reasoning.

B Related Work

Chain-of-Thought Reasoning Large Language Models (LLMs) produce Chain-of-Thought (CoT) reasoning [23, 2] to help provide end users with a peak into the reasoning process leading up to their response. While the CoT reasoning generated by these models is often appealing to human end users [23, 10], prior research has argued that LLM-generated CoT reasoning does not *faithfully* capture the underlying behavior of these models and that this is a critical challenge particularly in applications involving high-stakes decision making [2]. For instance, as discussed in Agarwal et al. [2], a doctor would benefit from seeing an explanation that faithfully captures why an LLM is recommending a particular diagnosis for a patient, as opposed to seeing some plausible explanation that could lead to the diagnosis at hand. In the former case, the doctor can actually use this faithful explanation to determine if and how much to rely on the model's recommendation.

Evaluating the Faithfulness of CoT Reasoning Despite the criticality of the faithfulness of LLM-generated CoT reasoning, there is very little work on analyzing and measuring this aspect of LLMs. Turpin et al. [22] were the first to demonstrate that CoT explanations may not faithfully capture the behavior of the underlying models. They showed that these explanations can be heavily influenced by biasing model inputs *e.g.,* by reordering multiple-choice options in a few-shot prompt to always make the answer "(A)"—which these models systematically fail to mention in their explanations. Lanham et al. [11] extended the above work and proposed novel metrics to measure the faithfulness of an LLM-generated CoT explanation. 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. Similarly, if *adding mistakes* in a generated CoT causes the model to change its final response, then the original CoT can be considered faithful. Analogously, they proposed other metrics to measure faithfulness based on *paraphrasing* the beginning portions of the original CoT as well as replacing the CoT with *filler* tokens (*e.g.,* ellipses). Using these metrics, they demonstrated that the CoT reasoning produced by state-of-the-art LLMs does not faithfully capture the behavior of the underlying models.

Enhancing the Quality of CoT Reasoning While there are some prior works that tackled the problem of improving the quality of CoT reasoning [18], their focus was on improving its quality vis-a-vis human knowledge or understanding. For example, Lyu et al. [18] focused on generating a reasoning chain that could then be put through a deterministic math solver, and the resulting answer from this solver was compared to the answer produced by the LLM. The reasoning chain was considered to be faithful if the answers of the solver and the LLM matched. Note that this approach does not account for ensuring that the internal computations or the underlying behavior of the LLM was captured in the reasoning chain, which is the focus of our work.

In summary, our work makes one of the initial attempts at exploring the promise of various popular paradigms, namely, activation editing, fine-tuning, and in-context learning, to improve the faithfulness of the CoT reasoning generated by LLMs.

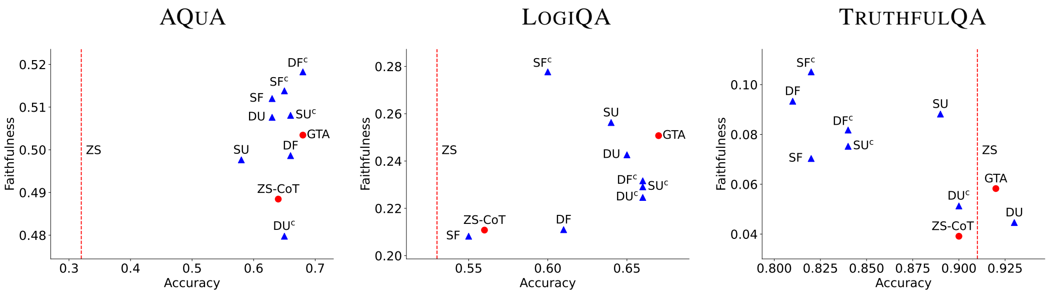

Figure 10: Faithfulness vs Accuracy relationship of CoT reasoning generated by GPT-4 using different baseline (in red) and **ICL** strategies (in blue). Results show that stochastic faithful sampling strategies, on average across three datasets, achieves higher faithfulness in CoT reasoning.

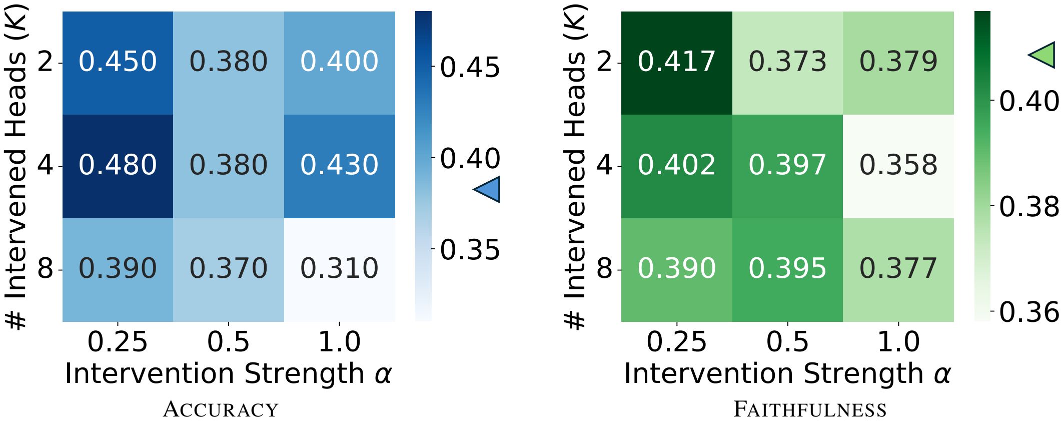

Figure 11: Accuracy and faithfulness of LLM reasoning for different intervention configurations (α, K) for LOGIQA dataset. Activation editing shows different the trade-off between the accuracy and faithfulness performance of LLAMA-3-8B-INSTRUCT and some configuration leads to an increase in accuracy as compared to the zero-shot CoT performance (▲ and ▲ markers) but doesn't improve faithfulness significantly.

C Experiments

Here, we provide additional results of our experiments in tabular format and perform significance testing of all our empirical analysis.

Table 1: GPT-3.5-Turbo Faithfulness for Different Fine-tuning Approaches

Approach	**AQuA**		**LogiQA**		**TruthfulQA**	
	Accuracy	Faithfulness	Accuracy	Faithfulness	Accuracy	Faithfulness
ZS-CoT	0.60 ± 0.05	0.64 ± 0.02	0.47 ± 0.05	0.40 ± 0.03	0.64 ± 0.05	0.30 ± 0.02
GTA	0.60 ± 0.05	0.65 ± 0.02	0.38 ± 0.05	0.38 ± 0.03	**0.83** ± 0.04	0.35 ± 0.03
DU	0.57 ± 0.05	0.62 ± 0.02	0.46 ± 0.05	0.49 ± 0.03	0.65 ± 0.05	0.29 ± 0.03
DU^c	0.61 ± 0.05	0.60 ± 0.03	0.46 ± 0.05	0.43 ± 0.03	0.65 ± 0.05	0.32 ± 0.03
DF	0.58 ± 0.05	0.62 ± 0.02	0.43 ± 0.05	0.51 ± 0.03	0.63 ± 0.05	0.32 ± 0.03
DF^c	**0.63** ± 0.05	0.64 ± 0.02	0.43 ± 0.05	0.44 ± 0.03	0.64 ± 0.05	0.33 ± 0.03
SU	0.61 ± 0.05	**0.66** ± 0.02	0.45 ± 0.05	0.48 ± 0.03	0.61 ± 0.05	**0.36** ± 0.03
SU^c	0.57 ± 0.05	0.65 ± 0.02	0.44 ± 0.05	0.47 ± 0.03	0.67 ± 0.05	0.33 ± 0.02
SF	0.58 ± 0.05	0.64 ± 0.02	0.45 ± 0.05	**0.51** ± 0.02	0.62 ± 0.05	0.35 ± 0.03
SF^c	0.59 ± 0.05	0.64 ± 0.02	**0.49** ± 0.05	0.48 ± 0.03	0.63 ± 0.05	0.34 ± 0.02

Table 2: Llama-3-8B-Instruct Faithfulness for Different Fine-tuning Approaches

Approach	**AQuA**		**LogiQA**		**TruthfulQA**	
	Accuracy	Faithfulness	Accuracy	Faithfulness	Accuracy	Faithfulness
ZS-CoT	0.43 ± 0.05	0.61 ± 0.02	0.38 ± 0.05	0.41 ± 0.03	0.56 ± 0.05	0.23 ± 0.03
GTA	0.43 ± 0.05	**0.68** ± 0.01	0.44 ± 0.05	**0.65** ± 0.01	**0.75** ± 0.04	0.23 ± 0.03
DU	0.45 ± 0.05	0.62 ± 0.02	0.42 ± 0.05	0.48 ± 0.02	0.51 ± 0.05	0.27 ± 0.03
DU^c	0.52 ± 0.05	0.57 ± 0.02	0.41 ± 0.05	0.43 ± 0.03	0.52 ± 0.05	0.24 ± 0.03
DF	0.43 ± 0.05	0.63 ± 0.02	**0.45** ± 0.05	0.44 ± 0.02	0.55 ± 0.05	**0.31** ± 0.03
DF^c	**0.54** ± 0.05	0.62 ± 0.02	0.40 ± 0.05	0.43 ± 0.03	0.54 ± 0.05	0.29 ± 0.03
SU	0.50 ± 0.05	0.67 ± 0.01	0.43 ± 0.05	0.45 ± 0.03	0.53 ± 0.05	0.28 ± 0.03
SU^c	0.47 ± 0.05	0.61 ± 0.02	0.42 ± 0.05	0.45 ± 0.03	0.54 ± 0.05	0.24 ± 0.03
SF	0.39 ± 0.05	0.61 ± 0.02	0.40 ± 0.05	0.44 ± 0.03	0.50 ± 0.05	0.27 ± 0.03
SF^c	0.50 ± 0.05	0.62 ± 0.02	0.39 ± 0.05	0.44 ± 0.03	0.61 ± 0.05	0.27 ± 0.03

Table 3: GPT-4 Faithfulness for Different In-Context Learning Approaches

Approach	**AQuA**		**LogiQA**		**TruthfulQA**	
	Accuracy	Faithfulness	Accuracy	Faithfulness	Accuracy	Faithfulness
ZS-CoT	0.64 ± 0.05	0.49 ± 0.03	0.56 ± 0.05	0.21 ± 0.03	0.90 ± 0.03	0.04 ± 0.01
GTA	**0.68** ± 0.05	0.50 ± 0.03	**0.67** ± 0.05	0.25 ± 0.03	0.92 ± 0.03	0.06 ± 0.02
DU	0.63 ± 0.05	0.51 ± 0.03	0.65 ± 0.05	0.24 ± 0.03	**0.93** ± 0.03	0.04 ± 0.01
DU^c	0.65 ± 0.05	0.48 ± 0.03	0.66 ± 0.05	0.22 ± 0.02	0.90 ± 0.03	0.05 ± 0.01
DF	0.66 ± 0.05	0.50 ± 0.03	0.61 ± 0.05	0.21 ± 0.02	0.81 ± 0.04	0.09 ± 0.02
DF^c	**0.68** ± 0.05	**0.52** ± 0.03	0.66 ± 0.05	0.23 ± 0.03	0.84 ± 0.04	0.08 ± 0.02
SU	0.58 ± 0.05	0.50 ± 0.03	0.64 ± 0.05	0.26 ± 0.03	0.89 ± 0.03	0.09 ± 0.02
SU^c	0.66 ± 0.05	0.51 ± 0.03	0.66 ± 0.05	0.23 ± 0.03	0.84 ± 0.04	0.08 ± 0.02
SF	0.63 ± 0.05	0.51 ± 0.03	0.55 ± 0.05	0.21 ± 0.02	0.82 ± 0.04	0.07 ± 0.02
SF^c	0.65 ± 0.05	0.51 ± 0.03	0.60 ± 0.05	**0.28** ± 0.03	0.82 ± 0.04	**0.11** ± 0.02

Table 4: GPT-3.5-Turbo Faithfulness for Different In-Context Learning Approaches

Approach	**AQuA**		**LogiQA**		**TruthfulQA**	
	Accuracy	Faithfulness	Accuracy	Faithfulness	Accuracy	Faithfulness
ZS-CoT	0.60 ± 0.05	0.64 ± 0.02	0.47 ± 0.05	0.40 ± 0.03	0.64 ± 0.05	0.30 ± 0.02
GTA	0.56 ± 0.05	0.64 ± 0.02	0.48 ± 0.05	0.44 ± 0.03	0.69 ± 0.05	0.33 ± 0.02
DU	0.55 ± 0.05	0.67 ± 0.02	0.48 ± 0.05	0.40 ± 0.03	0.65 ± 0.05	0.32 ± 0.02
DU^c	**0.64** ± 0.05	0.64 ± 0.02	0.40 ± 0.05	0.44 ± 0.03	0.75 ± 0.04	0.30 ± 0.03
DF	0.57 ± 0.05	0.67 ± 0.02	**0.54** ± 0.05	**0.47** ± 0.03	0.74 ± 0.04	0.30 ± 0.02
DF^c	0.62 ± 0.05	0.65 ± 0.02	0.44 ± 0.05	0.45 ± 0.03	0.73 ± 0.04	**0.34** ± 0.03
SU	0.59 ± 0.05	0.66 ± 0.02	0.43 ± 0.05	0.46 ± 0.03	**0.78** ± 0.04	0.30 ± 0.03
SU^c	0.61 ± 0.05	0.65 ± 0.02	0.44 ± 0.05	0.45 ± 0.03	0.73 ± 0.04	0.32 ± 0.02
SF	0.59 ± 0.05	**0.67** ± 0.02	0.48 ± 0.05	0.46 ± 0.03	0.70 ± 0.05	0.30 ± 0.02
SF^c	0.62 ± 0.05	0.66 ± 0.02	0.39 ± 0.05	0.44 ± 0.03	0.70 ± 0.05	0.31 ± 0.03