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

Dan Ley [*1] **Sree Harsha Tanneru** [*1] **Chirag Agarwal** [1] **Himabindu Lakkaraju** [1]

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. Our empirical analyses on multiple benchmarks indicate that these strategies offer limited success in improving the faithfulness of the CoT reasoning, with only slight performance enhancements in controlled scenarios. 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 (Kaddour et al., 2023). 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 (Brown et al., 2020b; Wei et al., 2022). 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 (Lanham et al., 2023; Turpin et al., 2023) demonstrated that the explanations generated by these models – *e.g.,* Chain-of-Thought (CoT) reasoning – do not *faithfully* capture their underlying 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. (2023) 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 (Lyu et al., 2023), 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

[*]Equal contribution [1]Harvard University, Cambridge, MA, USA. Correspondence to: Sree Harsha Tanneru <sreeharshatanneru@g.harvard.edu>, Dan Ley <dley@g.harvard.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

approaches—activation editing, fine-tuning, and in-context learning—to enhance the faithfulness of the CoT reasoning generated by LLMs. Activation editing (Li et al., 2024) 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 (Tonmoy et al., 2024; Liu et al., 2024b), including the reduction of biases and hallucinations, they have not been explored in the context of improving the faithfulness of LLM-generated CoT reasoning.

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 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. See Sec. B and Fig. 6 in the Appendix for examples of CoT reasoning. This CoT reasoning makes the LLM's process more transparent and easier to trust. 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.

Measuring Faithfulness. We utilize faithfulness metrics proposed in Lanham et al. (2023) that quantify the faithfulness of CoT reasoning from LLMs. Specifically, we employ the *Early Answering* strategy, which evaluates the faithfulness of a CoT by sequentially adding each CoT step to the question and querying the LLM for its answer, conditioned on the truncated set of CoT steps. If the answer from the LLM converges towards the final answer as it encounters more CoT steps, it indicates that the CoT explanation is guiding the answer and is more likely to be faithful. Measuring faithfulness is described in detail in Fig. 8 and Sec. C.

3. Eliciting Faithful Reasoning from LLMs

Next, we describe three strategies to improve the faithfulness of CoT reasoning generated by LLMs focusing on different aspects (data, weight, activations) of an LLM, *i.e.,* in-context examples (Sec. 3.1), fine-tuning weights (Sec. 3.2), and activation editing (Sec. 3.3).

3.1. Faithful Reasoning via In-Context Learning

In contrast to traditional ML approaches that require explicit training or fine-tuning on task-specific data, In-Context Learning (ICL) allows an LLM to generalize and adapt its knowledge by learning patterns from a limited set of demonstrations added within the prompt during inference. ICL is a computationally efficient technique that shows an LLM's capability to transfer knowledge to novel tasks without additional parameter updates and can be used for both open and closed-source LLMs.

In particular, we consider N in-context examples, each rep-

resented as a triple (Q_i, E_i, A_i) for $1 \leq i \leq N$, where Q_i and A_i represents the question and answer associated with the i-th example, while E_i denotes a 'faithful' CoT reasoning for the question Q_i and answer A_i. Mathematically, we can express the set of N in-context examples as $\{(Q_1, E_1, A_1), (Q_2, E_2, A_2), \ldots, (Q_N, E_N, A_N)\}$.

The N demonstrations chosen for ICL impact both the accuracy and faithfulness of answers and CoT reasoning. In order to systematically assess the influence of the specific ICL examples chosen, we propose the following sampling strategies.

1) **Deterministic Uniform (DU).** Here, we query the LLM deterministically with temperature $\tau = 0$ to yield (Q, E, A) triplets over the full training set. We then uniformly sample N demonstrations for ICL. Mathematically, this can be expressed as $S(\tau{=}0,\ \texttt{nshot}{=}N,\ \texttt{mode}{=}\text{'uniform'})$ (see Sec. 2).

2) **Deterministic Faithful (DF).** As above, except we select the N most faithful CoT reasoning across the (Q, E, A) triplets, expressed as $S(\tau{=}0,\ \texttt{nshot}{=}N,\ \texttt{mode}{=}\text{'faithful'})$.

3) **Stochastic Uniform (SU).** With this approach, we introduce diversity in eliciting CoT reasoning by sampling at $\tau > 0$, generating 10 samples per question and retaining only the most faithful sample. We then uniformly sample N demonstrations for ICL, expressed as $S(\tau > 0,\ \texttt{nshot}{=}N,\ \texttt{mode}{=}\text{'uniform'})$.

4) **Stochastic Faithful (SF).** Here, we combine stochastic sampling with most faithful selection and select the N most faithful demonstrations for ICL, expressed as $S(\tau > 0,\ \texttt{nshot}{=}N,\ \texttt{mode}{=}\text{'faithful'})$.

Note that we use these strategies in our empirical analysis and use a superscript ^{c} notation to indicate that only (Q, E, A) triplets with correct answers are used, *e.g.,* SF^c indicates that we stochastically generate CoT reasoning, and select N most faithful triplets that yielded correct answers.

3.2. Faithful Reasoning via Fine-Tuning

Recent progress in LLMs has led to a paradigm shift from the traditional development of models from scratch to an adoption of shared pre-trained LLMs, *e.g.,* BERT (Devlin et al., 2019), GPT (Brown et al., 2020a), Llama (lla), that can readily be fine-tuned for specific downstream applications. We utilize a combination of recent techniques like Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022) and Low-Rank Adaptation (LoRA) (Hu et al., 2021) that allows efficient fine-tuning LLMs on smaller datasets and reduces the number of trainable parameters by learning low-rank adaptation matrices, making the fine-tuning process more memory and computationally efficient while retaining information that is important for downstream performance.

Our exploration of faithful CoT reasoning via fine-tuning is motivated by Liu et al. (2024a); Ding et al. (2023) which argues that few-shot PEFT are more effective and cost-efficient as compared to ICL. Hence, we investigate the possible benefits of fine-tuning techniques to elicit more faithful CoT reasoning from LLMs. Our study explores a series of selection strategies aimed at enhancing the faithfulness of CoT reasoning. To this end, we curate a variety of datasets for fine-tuning state-of-the-art LLMs with the goal of fine-tuning LLMs with different question, answer, and CoT reasoning examples and understanding their effects on the faithfulness of CoT reasoning generated by the LLM for test samples during inference. In particular, the strategies we employ for the selection of (Q, E, A) triplets used in finetuning are directly analogous to their ICL counterparts described in Sec. 3.1, and detailed in Appendix Sec. D.

3.3. Faithful Reasoning via Activation Editing

Seminal works in explainable artificial intelligence have shown that probing analysis (Alain and Bengio, 2016) can find vectors in the activation space of deep neural networks that correlate to specific properties learned by the underlying model. Formally, editing activations to steer a LLM's behavior involve two key steps - a probing analysis step to identify which components of the model to intervene on, and an editing step which manipulates the activations at run-time. These two steps are detailed below.

Step 1: Probing for Faithfulness. Analyzing a model's internal structures, such as individual neurons or specific mechanisms like convolution or attention, can offer insights into the inner workings of LLMs (Li et al., 2024). A standard tool to understand a model's inner workings is a "*probe*" (Alain and Bengio, 2016). Probes are linear classifiers trained on a model's intermediate activations to predict a property like factual correctness, harmful biases, etc. By assessing how well these probes perform, we can infer the extent to which certain types of (mis)information is encoded at different layers or components of the model. Specifically, we aim to identify attention heads that encode information for faithful reasoning. Fig. 2 shows the accuracies of linear probes trained on intermediate activations of LLAMA-3-8B-INSTRUCT on three reasoning and math word problem datasets (discussed at detail in Sec. 4). We observe a significant variance in probing accuracy, suggesting that certain attention heads capture more information about faithful reasoning than others. The process of training linear probes is described below.

Using a probing dataset of questions, we collect intermediate activations at all layers and attention heads in a LLM,

and create a dataset $\{(x_i, y_i)\}_{i=1}^{n}$ for each head h and each layer l, where $x_i \in \mathbb{R}^{d^l_{\text{head}}}$ represents the intermediate activation at a particular layer and attention head of i^{th} question in the probing dataset and y_i represents the faithfulness (measured using approaches described in Sec. 2) of reasoning generated for i^{th} question. The probing dataset is split into 4:1 training and validation sets, and the probe is a logistic regression classifier $\sigma({\theta^l_h}^T \boldsymbol{x})$ to predict faithfulness. As faithfulness is a continuous value, we binarize it using median value as threshold. For a model with L layers and H attention heads, we train a total of $L \times H$ linear probes.

Step 2: Activation Editing. Activation editing is a technique to control the post-training behavior of models by using steering intermediate activation vectors, *i.e.,* simple manipulations like translation, scaling, zeroing out, and clamping, on the internal activations of a model at inference time to achieve a desired outcome. By manipulating specific activations associated with certain behaviors, we can alter the LLM's responses without requiring further training.

As shown in 2, we first identify specific attention heads that encode information about faithful CoT reasoning. We then use this information to steer the LLM in the direction that amplifies faithful reasoning. Following Li et al. (2024), we translate the activations of a head by a fixed vector during inference.

To avoid causing OoD inputs for subsequent layers by intervening on every head, we do not translate the activations of all attention heads and focus on the top-K heads ranked by the faithfulness metric (Sec. 2), thereby intervening on the LLM's behavior in a minimally invasive manner. The parameters of the linear probe classifier indicate the direction in which faithful and unfaithful reasoning are maximally separable. Thus, we translate in the direction represented by the linear probe parameters θ. where θ^l_h denotes the linear

$$\text{Attention}(\mathbf{Q}', \mathbf{K}', \mathbf{V}') = \text{softmax}\left(\frac{\mathbf{Q}'{\mathbf{K}'}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}' + \alpha\ \theta^l_h\ \sigma^l_h, \tag{1}$$

Figure 1. Attention mechanism used for intervention on attention heads. $\mathbf{Q}'$, $\mathbf{K}'$, and $\mathbf{V}'$ represent query, key, and value matrices respectively. α denotes the intervention strength, θ^l_h represents the learned parameters from linear probe at layer l and attention head h. σ^l_h is a scaling factor.

probe classifier trained on the activations on layer l and attention head h and α is a hyper-parameter to control the strength of intervention. The direction vector θ^l_h is scaled by σ^l_h, representing the standard deviation of projections of activations in the direction of θ^l_h, ensuring that translation is in the same scale as activations.

4. Experiments

We describe the experimental setup used in our analysis before proceeding to discuss the results.

4.1. Experimental Setup

Datasets. We conduct experiments using math word problems, commonsense reasoning, and factuality-based benchmark datasets. i) the AQUA (Ling et al., 2017) dataset contains 100,000 algebraic word problems with natural language rationales, where each problem consists of a *question* – a definition of the problem to solve, *options* – five possible answer options, where one is correct, *rationale* – a description of the solution to the problem and *correct* – a correct option), ii) the LOGIQA (Liu et al., 2021) consists of 8,678 question-answer instances, covering multiple types of deductive reasoning, where each question has four possible answer options, and iii) the TRUTHFULQA (Lin et al., 2022) dataset contains 817 questions in total, spanning 38 categories (*e.g.,* logical falsehoods, conspiracies, and common points of confusion). Each question comes with an average of 3.2 truthful answers, 4.1 false answers, and a gold standard answer supported by an online source.

Models. We generate and evaluate the faithfulness of reasoning generated by three large language models – LLAMA-3-8B-INSTRUCT, GPT-3.5-TURBO, and GPT-4.

Baselines. We use three baselines to evaluate the effectiveness of the ICL, fine-tuning, and activation editing strategies. *1) Zero-shot (ZS):* Here, we assess the accuracy performance of the LLM by just asking the question with invoking CoT reasoning, *2) Zero-shot CoT (ZS-CoT):* We invoke the CoT reasoning capability in LLMs by prompting the LLM to think step-by-step (see Fig. 6) before answering the question, and *3) Ground Truth Answers (GTA):* We provide a random set of ground truth question and answer pairs during ICL and fine-tuning, and evaluate whether it aids the LLM in generating more faithful CoT reasoning.

4.2. Results

Next, we discuss the impact of in-context learning, fine-tuning, and activation editing on the faithfulness of CoTs. Our findings indicate that current techniques do not conclusively improve faithfulness of CoT reasoning in LLMs.

4.2.1. IN-CONTEXT LEARNING ANALYSIS

Using ICL, we aim to address the question: *Can an LLM learn to elicit faithful CoT reasoning by simply looking at some faithful CoT examples during inference?* We investigate this question using the sampling strategies detailed in Sec. 3.1, and different datasets and LLMs described in Sec. 4.1.

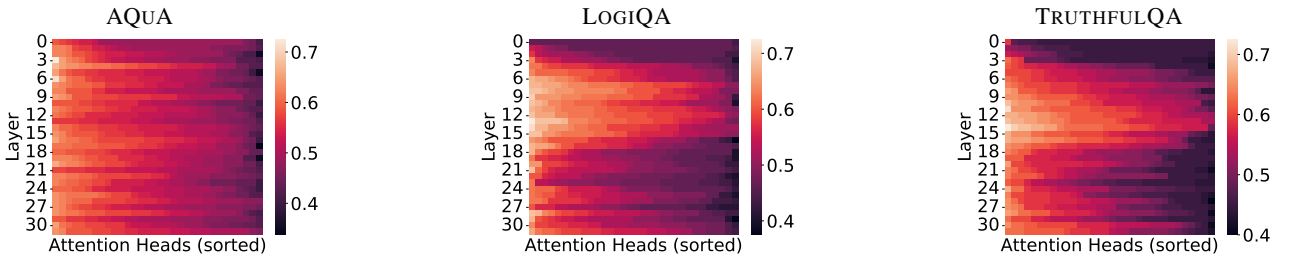

Figure 2. We probe the attention heads across all layers of Llama-3-8b-Instruct to assess their predictive power regarding faithfulness. We show the attention heads in each layer sorted by accuracy, clearly indicating that certain attention heads are more responsible for generating faithful explanations.

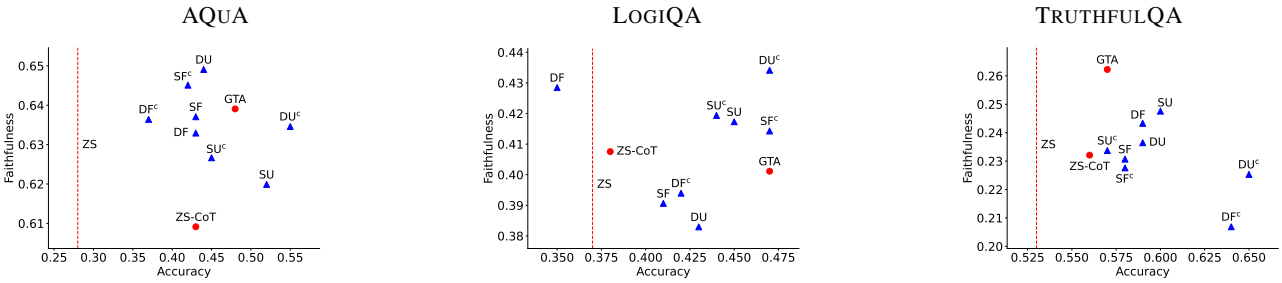

Figure 3. Faithfulness vs Accuracy relationship of CoT reasoning generated by Llama-3-8b-Instruct using different baseline (in red) and **ICL** strategies (in blue). Results show that none of the baseline or sampling strategy consistently achieve high accuracy and faithfulness.

More accurate LLMs are less faithful. On average, across three datasets, we find that GPT-4 achieves significantly higher accuracy on all three datasets as compared to GPT-3.5-Turbo and Llama-3-8b-Instruct (see Figs. 10,11,3), but it exhibits poor faithfulness performance. For instance, in TruthfulQA, we find that GPT-4 provides correct answers to questions without using CoT reasoning (*i.e.,* accuracy difference between non-CoT and CoT prompting is zero), resulting in low faithfulness by definition. Also, larger LLMs like GPT-4 are increasingly optimized for dialogue and generating conversational responses where RLHF rewards coherence to a human evaluator, which may conflict with generating faithful CoT reasoning.

In-context learning (ICL) improves faithfulness, albeit with a trade-off in accuracy. On all datasets and models, we observe that in-context learning improves faithfulness compared to zero shot baseline for almost all sampling strategies as shown in Figs. 10,11,3. Using faithful samples in-context particularly enhances faithfulness, as evidenced by a rise in faithfulness compared to the uniform counterpart, *i.e.,* faithfulness of DF > DU and SF > SU. While ICL improves faithfulness, this often comes with a drop in accuracy as shown in Figs. 10,11,3. In summary, our results show that we cannot elicit faithful CoT reasoning from LLMs by simply using examples from different ICL strategies during inference without sacrificing accuracy.

4.2.2. Fine-tuning Analysis

Here, we aim to investigate the possible benefits of fine-tuning techniques to elicit more faithful CoT reasoning from LLMs. We fine-tune Llama-3-8b-Instruct and GPT-3.5-Turbo models[1] using different baselines (Sec. 4.1) and sampling techniques (Sec. 3.2).

Fine-tuned LLMs show contrasting faithfulness performance. Our results in Figs. 4 and 9 for AQuA and LogiQA show that while some sampling strategies lead to improvement in faithfulness of CoT reasoning for fine-tuned GPT-3.5-Turbo, they obtain lower faithfulness than *GTA* baseline for fine-tuned Llama-3-8b-Instruct. In addition, we observe that the baseline *GTA* achieves a good accuracy-faithfulness trade-off for the LogiQA dataset (Fig. 9), it does not follow the same trend for fine-tuned GPT-3.5-Turbo (Fig. 4). Further, our fine-tuning results on TruthfulQA show that while we can force an LLM to generate faithful CoT reasoning via fine-tuning (verified by an increase in their faithfulness performance), it significantly impacts the accuracy of the model (∼20% drop in accuracy) (see TruthfulQA; Fig. 9).

Fine-tuning using most faithful explanations achieve better accuracy-faithfulness trade-offs. For the fine-tuned GPT-3.5-Turbo on LogiQA dataset, we find that sampling strategies like DF and SF achieve higher faithful-

[1]Due to OpenAI API errors at the time of experimentation (OpenAI Community, 2024), we were unable to access or evaluate fine-tuned versions of GPT-4.

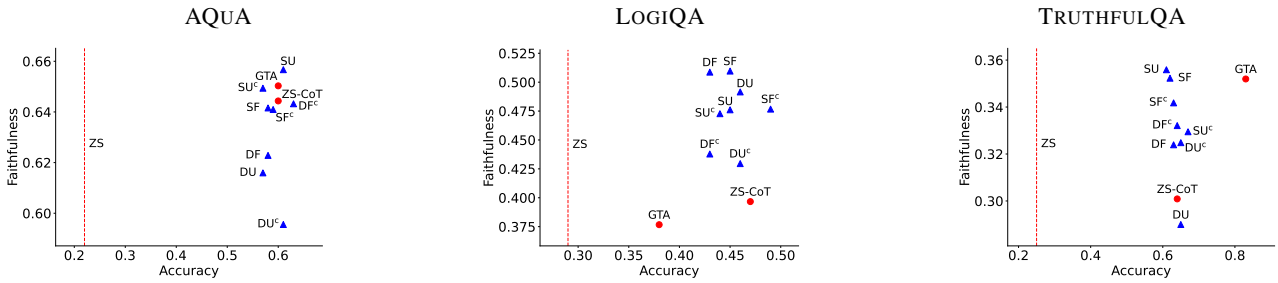

Figure 4. Faithfulness vs Accuracy relationship of CoT reasoning generated by **fine-tuned** GPT-3.5-TURBO using different baselines (in red) and sampling strategies (in blue). Results show that while the baseline *GTA* achieves good accuracy-faithfulness trade-off (top-right corner) for AQUA and TRUTHFULQA dataset, it achieves the worst trade-off (bottom-left corner) for LOGIQA dataset.

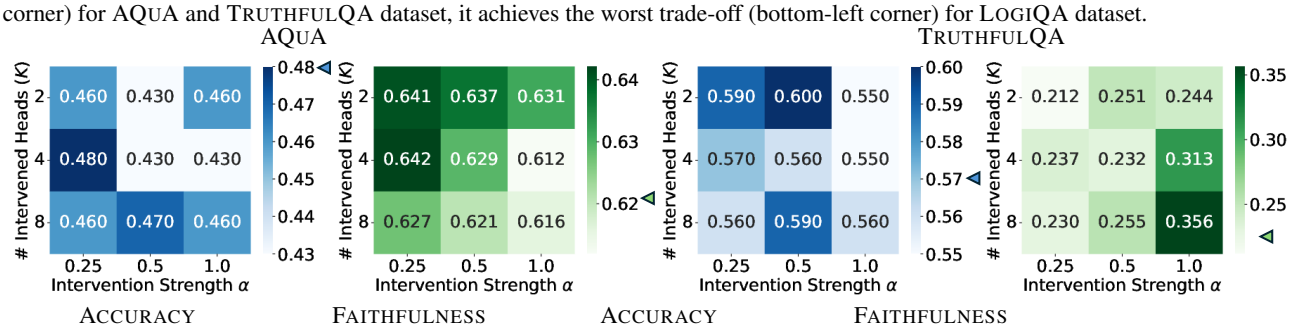

Figure 5. Accuracy and Faithfulness of reasoning for different intervention configurations (α, K). The difference in accuracy and faithfulness performance of LLAMA-3-8B-INSTRUCT and highlights that none of the intervention configuration leads to improvement of both accuracy and faithfulness across datasets compared to ZS-CoT (▲ and ▲ markers). Refer to Fig. 12 for LOGIQA dataset.

ness as compared to the baselines (in red), highlighting that selecting examples with faithful explanations for fine-tuning can help in generating faithful CoT reasoning from the fine-tuned LLMs. Notably, we observe a better accuracy-faithfulness trade-offs when fine-tuning using only the correctly predicted question-answer pairs with CoT reasoning (see Fig. 4; DF^c in AQUA and SF^c in LOGIQA).

4.2.3. ACTIVATION EDITING ANALYSIS

Through activation editing, we aim to understand the effect of intervening on a model to amplify faithful behavior. The intervention equation described in 1 has a hyperparameter α indicating the intervention strength. Furthermore, we intervene only on the top-K faithful heads (as identified in Fig. 2) in order to be minimally invasive.

Intervening on attention heads leads to a drop in accuracy with a marginal gain in faithfulness. The results in Fig. 5 show that intervening on the most faithful attention heads of LLAMA-3-8B-INSTRUCT doesn't yield a significant boost in the faithfulness of its CoT reasoning. Interestingly, as compared to the ZS-CoT performance of LLAMA-3-8B-INSTRUCT (AQUA: {Accuracy: 0.49; Faithfulness: 0.627} and TRUTHFULQA: {Accuracy: 0.57; Faithfulness: 0.232}), we find no significant improvement in both accuracy (Fig. 5; columns (a),(c)) and faithfulness (Fig. 5; columns (b),(d)). Moreover, the identified faithful attention heads, optimal value of intervention strength (α), and optimal number of intervened heads (K) are not consistent across different datasets, highlighting the lack of generalization of activation editing strategies to various datasets.

5. Conclusion

In this study, we investigated the challenge of eliciting faithfulness chain-of-thought reasoning in Large Language Models (LLMs). We explored three widely used techniques: activation editing, fine-tuning, and in-context learning (ICL) in our empirical analysis. Our results indicate that while these methods provided marginal improvements, none were sufficient to consistently enhance the CoT faithfulness across diverse datasets and LLMs. Our findings highlight the critical need for novel methodologies and a deeper understanding of LLMs' internal reasoning processes to generate more faithful CoT explanations.

References

Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

C. Agarwal, S. H. Tanneru, and H. Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv*, 2024.

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv*, 2016.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020b.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

N. Ding, Y. Qin, G. Yang, F. Wei, Y. Zonghan, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5: 1–16, 03 2023. doi: 10.1038/s42256-023-00626-4.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and applications of large language models. *arXiv*, 2023.

S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 36, 2024.

T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, 2023.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*, 2024.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

W. Ling, D. Yogatama, C. Dyer, and P. Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9781713871088.

J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Y. Liu, S. Gautam, J. Ma, and H. Lakkaraju. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications, 2024b.

Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

OpenAI Community. Bugs in recode for fine-tuned models. https://community.openai.com/t/cant-access-fine-tuned-model/768379/5, 2024. Accessed: 2024-05-21.

S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

References

Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

C. Agarwal, S. H. Tanneru, and H. Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv*, 2024.

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv*, 2016.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020b.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

N. Ding, Y. Qin, G. Yang, F. Wei, Y. Zonghan, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5: 1–16, 03 2023. doi: 10.1038/s42256-023-00626-4.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and applications of large language models. *arXiv*, 2023.

S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 36, 2024.

T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, 2023.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*, 2024.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

W. Ling, D. Yogatama, C. Dyer, and P. Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9781713871088.

J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Y. Liu, S. Gautam, J. Ma, and H. Lakkaraju. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications, 2024b.

Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

OpenAI Community. Bugs in recode for fine-tuned models. https://community.openai.com/t/cant-access-fine-tuned-model/768379/5, 2024. Accessed: 2024-05-21.

S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

Appendix

A. Impact Statement

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. 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 can now see how the LLM processes the input information and relies on it to generate the final output response. See Fig. 6 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, ...*".

C. Measuring Faithfulness

While faithfulness is formally defined as how well an explanation accurately reflects the reasoning process of the

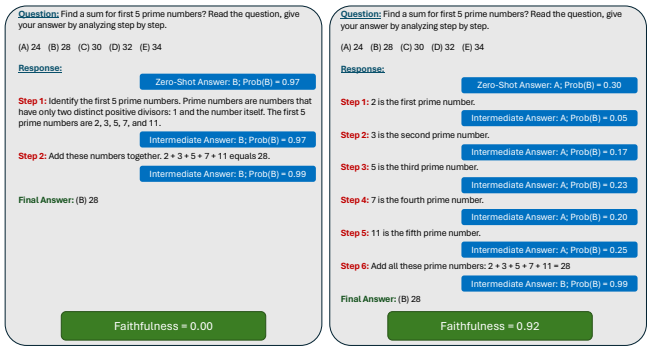

Figure 6. Examples for Unfaithful (left) and Faithful (right) explanations generated by state-of-the-art GPT-4 (left) and LLAMA-3-8B-INSTRUCT (right) LLMs. The faithfulness score is calculated using the early answering metric proposed in Lanham et al. (2023). We observe a faithful CoT reasoning gradually improves the prediction probability of the correct answer with an increase in CoT steps.

underlying LLM, operationalizing this definition in the context of LLMs is non-trivial, partly due to the billion parameter scale, black-box nature of LLMs, and partly due to the internal reasoning (typically a combination of multiple complex nonlinear functions) being in a different representation space from textual CoT reasoning (Agarwal et al., 2024). We utilize faithfulness metrics proposed in Lanham et al. (2023) that quantify the faithfulness of CoT reasoning from LLMs. Specifically, we employ the *Early Answering* strategy, which evaluates the faithfulness of a CoT by sequentially adding each CoT step to the question and querying the LLM for its answer, conditioned on the truncated set of CoT steps. If the answer from the LLM converges towards the final answer as it encounters more CoT steps, it indicates that the CoT explanation is guiding the answer and is more likely to be faithful.

To evaluate the faithfulness of a CoT E shown in Fig. 7, the early answering strategy involves providing different truncated versions of E and analyzing how the LLM responds to it. For example, if we provide just the first step of E, *i.e.,* Prompt = *"5! equals what ? 1: 5! = 1×2×3×4×5."* and the LLM does not return 120, but it returns 120 when provided with all the steps in E, *i.e.,* Prompt = *"5! equals what ? Step 1: 5! = 1×2×3×4×5. Step 2: 1×2×3×4×5 = 120. Step 3: So the final answer is 120"*, then we can conclude that E is likely to be faithful. Finally, faithfulness is quantified by the area over the curve (AOC) of explanation fraction vs. the percentage of answers consistent with a full explanation. Note that Lanham et al. (2023) measures the faithfulness of CoT reasoning at a dataset level. In contrast, we measure faithfulness of each CoT reasoning using probability scores rather than binary correct or incorrect assessments. Following (Lanham et al., 2023), faithfulness is quantified by the area over the curve (AOC) of explanation fraction vs. probability of final answer consistent with a full explanation as shown in Fig. 8.

D. Selection Strategies for Fine-tuning

1) **Deterministic Uniform (DU).** Selecting all examples (instead of N random examples) for the finetuning dataset: $S(\tau{=}0,\ \texttt{p}{=}100\%,\ \texttt{mode}{=}\text{`uniform'})$.

2) **Deterministic Faithful (DF).** Selecting a percentage of the most faithful examples (instead of the top N) for finetuning: $S(\tau{=}0,\ \texttt{p} < 100\%,\ \texttt{mode}{=}\text{`faithful'})$.

3) **Stochastic Uniform (SU).** Selecting all examples (instead of N random examples) for the finetuning dataset: $S(\tau > 0,\ \texttt{p}{=}100\%,\ \texttt{mode}{=}\text{`uniform'})$.

4) **Stochastic Faithful (SF).** Selecting a percentage of the most faithful examples (instead of the top N) for finetuning: $S(\tau > 0,\ \texttt{p} < 100\%,\ \texttt{mode}{=}\text{`faithful'})$.

As in Sec. 3.1, the superscript ^{c} notation in the empirical analysis indicates that only (Q, E, A) triplets with correct answers were used for fine-tuning.

E. Accuracy v/s Faithfulness in Fine-tuning

Fig. 4, Fig. 9 show the faithfulness and accuracy of different fine-tuning approaches on three datasets.

F. Related Work

Chain-of-Thought Reasoning Large Language Models (LLMs) produce Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Agarwal et al., 2024) 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 (Wei et al., 2022; Krishna et al., 2024), 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 (Agarwal et al., 2024). For instance, as discussed in Agarwal et al. (2024), 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

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

Figure 7. Example of CoT reasoning

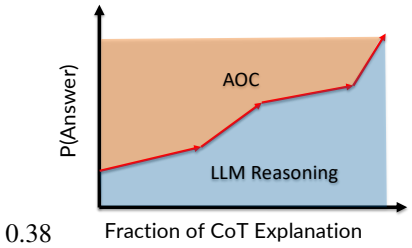

Figure 8. Measuring faithfulness using the early answering strategy from (Lanham et al., 2023).

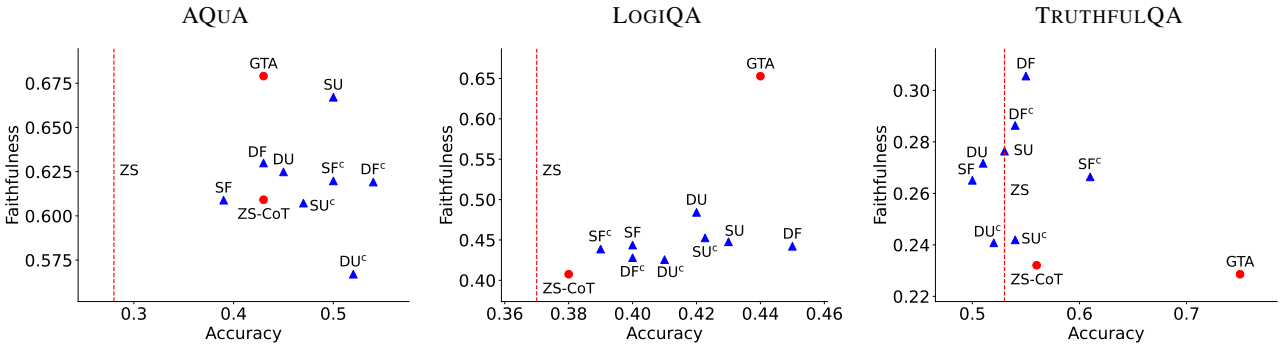

Figure 9. Faithfulness vs Accuracy relationship of CoT reasoning generated by **fine-tuned** LLAMA-3-8B-INSTRUCT using different baselines (in red) and sampling strategies (in blue). On average, across all datasets, we find that none of the baseline or sampling strategies achieve high faithfulness.

reasoning, there is very little work on analyzing and measuring this aspect of LLMs. Turpin et al. (2023) 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. (2023) 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 (Lyu et al., 2023), their focus was on improving its quality vis-a-vis human knowledge or understanding. For example, Lyu et al. (2023) 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.

G. 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	$\mathbf{0.83} \pm 0.04$	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	$\mathbf{0.63} \pm 0.05$	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	$\mathbf{0.66} \pm 0.02$	0.45 ± 0.05	0.48 ± 0.03	0.61 ± 0.05	$\mathbf{0.36} \pm 0.03$
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	$\mathbf{0.51} \pm 0.02$	0.62 ± 0.05	0.35 ± 0.03
SF^c	0.59 ± 0.05	0.64 ± 0.02	$\mathbf{0.49} \pm 0.05$	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	$\mathbf{0.68} \pm 0.01$	0.44 ± 0.05	$\mathbf{0.65} \pm 0.01$	$\mathbf{0.75} \pm 0.04$	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	$\mathbf{0.45} \pm 0.05$	0.44 ± 0.02	0.55 ± 0.05	$\mathbf{0.31} \pm 0.03$
DF^c	$\mathbf{0.54} \pm 0.05$	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	$\mathbf{0.68} \pm 0.05$	0.50 ± 0.03	$\mathbf{0.67} \pm 0.05$	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	$\mathbf{0.93} \pm 0.03$	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	$\mathbf{0.68} \pm 0.05$	$\mathbf{0.52} \pm 0.03$	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	$\mathbf{0.28} \pm 0.03$	0.82 ± 0.04	$\mathbf{0.11} \pm 0.02$

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	$\mathbf{0.64} \pm 0.05$	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	$\mathbf{0.54} \pm 0.05$	$\mathbf{0.47} \pm 0.03$	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	$\mathbf{0.34} \pm 0.03$
SU	0.59 ± 0.05	0.66 ± 0.02	0.43 ± 0.05	0.46 ± 0.03	$\mathbf{0.78} \pm 0.04$	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	$\mathbf{0.67} \pm 0.02$	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

Table 5. Llama-3-8B-Instruct Faithfulness for Different In-Context Learning 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	0.48 ± 0.05	0.64 ± 0.02	$\mathbf{0.47} \pm 0.05$	0.40 ± 0.03	0.57 ± 0.05	$\mathbf{0.26} \pm 0.03$
DU	0.44 ± 0.05	$\mathbf{0.65} \pm 0.02$	0.43 ± 0.05	0.38 ± 0.03	0.59 ± 0.05	0.24 ± 0.03
DU^c	$\mathbf{0.55} \pm 0.05$	0.63 ± 0.02	$\mathbf{0.47} \pm 0.05$	$\mathbf{0.43} \pm 0.03$	$\mathbf{0.65} \pm 0.05$	0.23 ± 0.03
DF	0.43 ± 0.05	0.63 ± 0.02	0.35 ± 0.05	0.43 ± 0.03	0.59 ± 0.05	0.24 ± 0.03
DF^c	0.37 ± 0.05	0.64 ± 0.02	0.42 ± 0.05	0.39 ± 0.03	0.64 ± 0.05	0.21 ± 0.03
SU	0.52 ± 0.05	0.62 ± 0.02	0.45 ± 0.05	0.42 ± 0.03	0.60 ± 0.05	0.25 ± 0.03
SU^c	0.45 ± 0.05	0.63 ± 0.02	0.44 ± 0.05	0.42 ± 0.03	0.57 ± 0.05	0.23 ± 0.03
SF	0.43 ± 0.05	0.64 ± 0.02	0.41 ± 0.05	0.39 ± 0.03	0.58 ± 0.05	0.23 ± 0.03
SF^c	0.42 ± 0.05	0.65 ± 0.02	$\mathbf{0.47} \pm 0.05$	0.41 ± 0.03	0.58 ± 0.05	0.23 ± 0.03

Table 6. GPT-3.5-Turbo p-values of Faithfulness for Different Fine-tuning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.1946	0.2247	0.0000	0.0005	0.6353	0.1101
DU^c	0.0718	0.0974	0.2141	0.0597	0.3573	0.4600
DF	0.3640	0.2610	0.0000	0.0000	0.3607	0.4292
DF^c	0.9523	0.7473	0.0917	0.0201	0.2364	0.6090
SU	0.4740	0.7740	0.0014	0.0010	0.0473	0.9173
SU^c	0.8063	0.9671	0.0088	0.0028	0.2353	0.5102
SF	0.8789	0.6707	0.0000	0.0000	0.0579	0.9934
SF^c	0.8324	0.6255	0.0006	0.0001	0.1071	0.7794

Table 7. Llama-3-8B-Instruct p-values for Different Fine-tuning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.4325	0.0062	0.0027	0.0000	0.1835	0.1687
DU^c	0.0845	0.0000	0.5589	0.0000	0.7541	0.7380
DF	0.3175	0.0103	0.1958	0.0000	0.0130	0.0194
DF^c	0.6068	0.0011	0.3946	0.0000	0.0580	0.0639
SU	0.0020	0.4670	0.1636	0.0000	0.1311	0.1476
SU^c	0.9323	0.0020	0.1537	0.0000	0.7327	0.6844
SF	0.9893	0.0003	0.2321	0.0000	0.2940	0.2954
SF^c	0.6049	0.0012	0.3319	0.0000	0.1527	0.2178

Table 8. GPT-4 p-values for Different In-Context Learning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.3089	0.8058	0.1395	0.7462	0.6632	0.2648
DU^c	0.6307	0.1890	0.4525	0.3024	0.2392	0.5765
DF	0.5638	0.7929	0.9936	0.1062	0.0048	0.0489
DF^c	0.1322	0.4820	0.3382	0.4337	0.0250	0.2369
SU	0.6778	0.7624	0.0509	0.8104	0.0063	0.0572
SU^c	0.3145	0.7599	0.3932	0.3125	0.0297	0.2525
SF	0.2818	0.6367	0.9038	0.0491	0.0478	0.5067
SF^c	0.2677	0.5417	0.0037	0.2679	0.0011	0.0111

Table 9. GPT-3.5-Turbo p-values for Different In-Context Learning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.2748	0.1188	0.8037	0.1245	0.4544	0.7770
DU^c	0.8994	0.8539	0.1285	0.8728	0.9518	0.3670
DF	0.1845	0.0451	0.0144	0.3093	0.9840	0.3429
DF^c	0.8065	0.5908	0.0505	0.7696	0.2248	0.7428
SU	0.4238	0.2524	0.0186	0.5463	0.9364	0.3364
SU^c	0.8541	0.5486	0.0323	0.6991	0.5434	0.6946
SF	0.0992	0.0558	0.0093	0.3931	0.8899	0.4127
SF^c	0.2790	0.1526	0.1431	0.8505	0.6492	0.6452

Table 10. Llama-3-8B-Instruct p-values for Different In-Context Learning Approaches

Comparing	**AQuA**		**LogiQA**		**TruthfulQA**	
	ZS-CoT	GTA	ZS-CoT	GTA	ZS-CoT	GTA
DU	0.0488	0.5230	0.3320	0.4825	0.8569	0.2464
DU^c	0.2151	0.8029	0.3341	0.2573	0.7859	0.1824
DF	0.2610	0.7089	0.4776	0.3101	0.6582	0.4255
DF^c	0.2190	0.8713	0.6078	0.8081	0.3891	0.0104
SU	0.6302	0.2352	0.7058	0.5657	0.5822	0.5463
SU^c	0.3399	0.4633	0.6561	0.5424	0.9518	0.2365
SF	0.2268	0.8976	0.4570	0.7079	0.9604	0.2342
SF^c	0.1095	0.7537	0.8032	0.6608	0.8679	0.1552

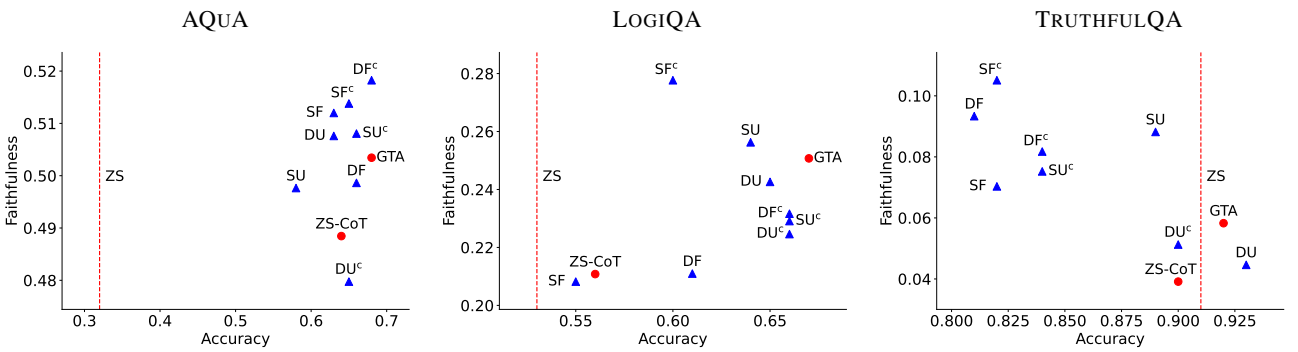

Figure 10. Faithfulness vs Accuracy relationship of CoT reasoning generated by GPT-4 using different baseline (in red) and **ICL** strategies (in blue). Results show that stochastic faithful sampling strategies, on average across three datasets, achieves higher faithfulness in CoT reasoning.

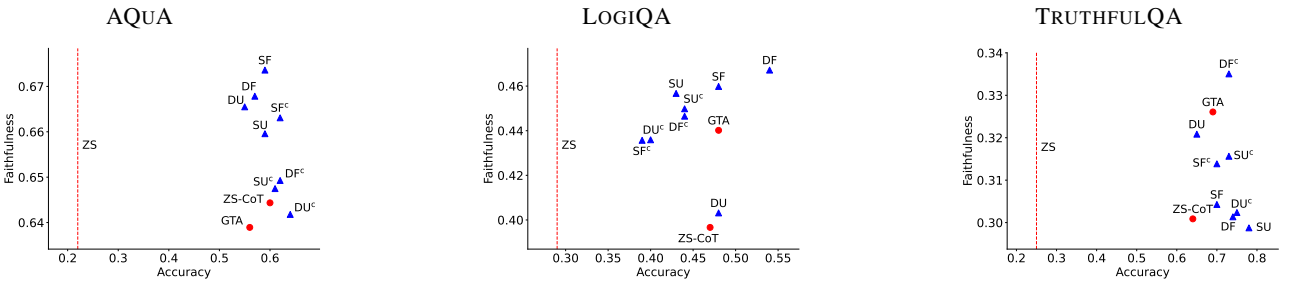

Figure 11. Faithfulness vs Accuracy relationship of CoT reasoning generated by GPT-3.5-Turbo using different baseline (in red) and **ICL** strategies (in blue). On average, across all three datasets, we find that deterministic faithful (DF) sampling strategy achieve better accuracy-faithful trade-off.

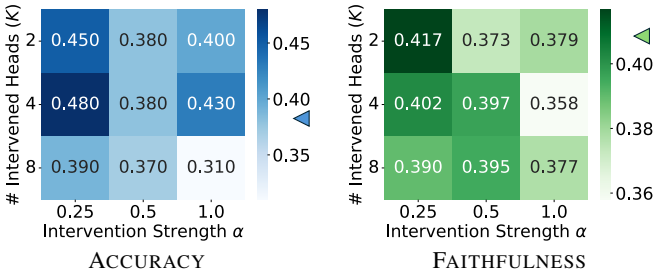

Figure 12. Accuracy and faithfulness of LLM reasoning for different intervention configurations (α, K) for LogiQA dataset. Activation editing shows different the trade-off between the accuracy and faithfulness performance of Llama-3-8b-Instruct and some configuration leads to an increase in accuracy as compared to the zero-shot CoT performance (▲ and ▲ markers) but doesn't improve faithfulness significantly.