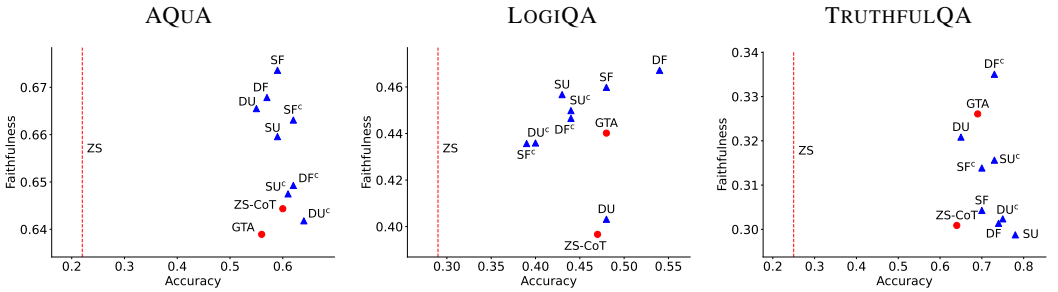

Figure 5: Faithfulness vs Accuracy relationship of CoT reasoning generated by GPT-3.5-TURBO using different baseline (in red) and **ICL** strategies (in blue). On average, across all three datasets, we find that deterministic faithful (DF) sampling strategy achieve better accuracy-faithful trade-off.

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 [14] 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 [16] 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 [13] 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. 1) 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 CoT reasoning. Our findings indicate that current techniques do not conclusively improve the 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.

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. 10,5,6), 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

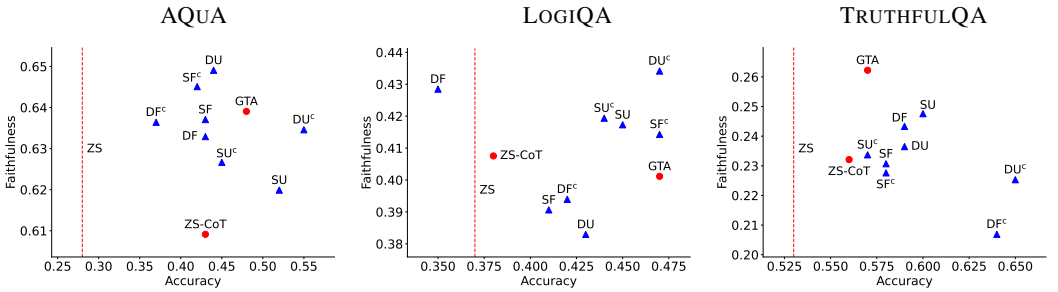

Figure 6: Faithfulness vs Accuracy relationship of CoT reasoning generated by LLAMA-3-8B-INSTRUCT using different baseline (in red) and **ICL** strategies (in blue). Results show that none of the baseline or sampling strategy consistently achieve high accuracy and faithfulness.

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. 10,5,6. 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. One exception is LLAMA-3-8B-INSTRUCT on TRUTHFULQA dataset. We suspect this is due to TRUTHFULQA being a dataset of human falsehoods relies less on reasoning to arrive at an answer. While ICL improves faithfulness, this often comes with a drop in accuracy as shown in Figs. 10,5,6.

Certain sampling strategies provide better trade-offs. Using top-K faithful samples (DF), on average, improves the faithfulness of the CoT reasoning but takes a hit on the accuracy, whereas the stochastic uniform sampling (SU) obtains better accuracy without improving faithfulness. Stochastic faithful sampling (SF) provides a middle ground. Moreover, we find better accuracy-faithfulness trade-offs when we perform ICL prompting using only the CoT reasoning from correctly predicted question-answer pairs by the LLM.

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[2] using different baselines (Sec. 4.1) and sampling techniques (Sec. 3.2).

Fine-tuned LLMs show contrasting faithfulness performance. Our results in Figs. 7 and 8 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. 8), it does not follow the same trend for fine-tuned GPT-3.5-TURBO (Fig. 7). 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 ($\sim$20% drop in accuracy) (see TRUTHFULQA; Fig. 8).

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 faithfulness 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. 7; DF^c in AQUA and SF^c in LOGIQA).

[2]Due to OpenAI API errors at the time of experimentation [20], we were unable to access or evaluate fine-tuned versions of GPT-4.

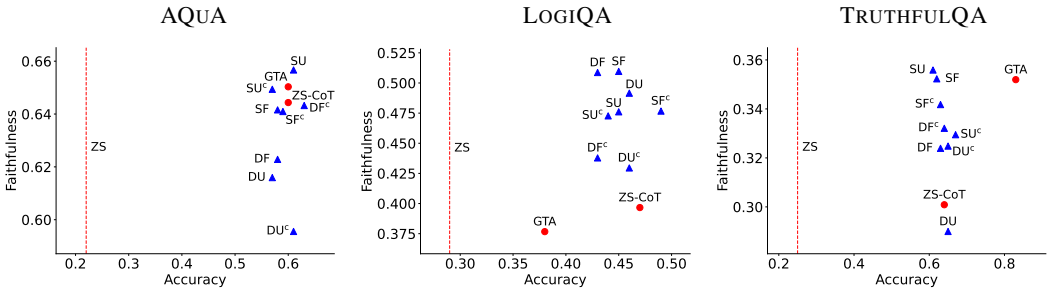

Figure 7: Faithfulness vs Accuracy relationship of CoT reasoning generated by **fine-tuned** GPT-3.5-Turbo using different baselines (in red) and sampling strategies (in blue). Results show that while the baseline *GTA* achieves good accuracy-faithfulness trade-off (top-right corner) for AQuA and TruthfulQA dataset, it achieves the worst trade-off (bottom-left corner) for LogiQA dataset.

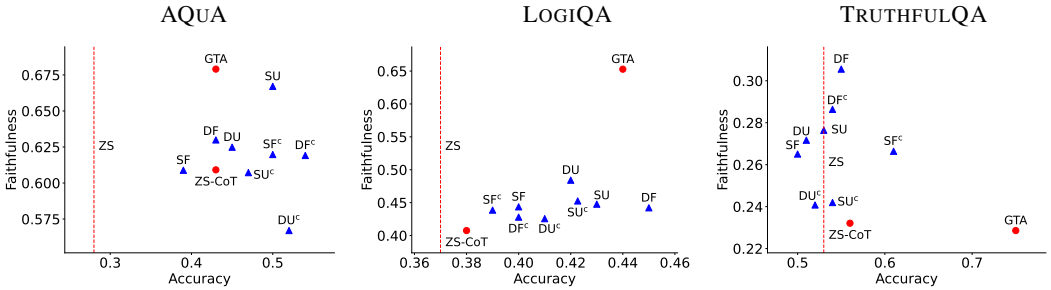

Figure 8: Faithfulness vs Accuracy relationship of CoT reasoning generated by **fine-tuned** Llama-3-8b-Instruct using different baselines (in red) and sampling strategies (in blue). On average, across all datasets, we find that none of the baseline or sampling strategies achieve high faithfulness.

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 4 has a hyperparameter α indicating the intervention strength α. Furthermore, we intervene only on the top-K faithful heads (as identified in Fig. 3) in order to be minimally invasive. The results in Fig. 9 show faithfulness and accuracy on TruthfulQA and AQuA upon intervening on different number of attention heads of Llama-3-8b-Instruct, *i.e.,* $K = \{2, 4, 8\}$, and intervention strengths, *i.e.,* $\alpha = \{0.25, 0.50, 1.0\}$.

Intervening on attention heads leads to a drop in accuracy with a marginal gain in faithfulness. The results in Fig. 9 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. 9; columns (a),(c)) and faithfulness (Fig. 9; 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. In addition, our analysis demonstrates contrasting behaviors in Llama-3-8b-Instruct, where activation editing works for improving truthfulness but shows mixed results for faithfulness, underscoring the challenge of eliciting faithful CoT reasoning from LLMs. Finally, our results also highlight the dichotomy between accuracy and faithfulness, where the values of $\{\alpha, K\}$ for the most faithful attention head (dark green in Fig. 9) are not always equivalent to the most accurate one (dark blue in Fig. 9).

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