On the Impact of Fine-Tuning on Chain-of-Thought Reasoning

Elita Lobo
University of Massachusetts
Amherst
elobo@umass.edu

Chirag Agarwal
University of Virginia
chiragagarwal@virginia.edu

Himabindu Lakkaraju
Harvard University
hlakkaraju@hbs.edu

Abstract

Large language models have emerged as powerful tools for general intelligence, showcasing advanced natural language processing capabilities that find applications across diverse domains. Despite their impressive performance, recent studies have highlighted the potential for significant enhancements in LLMs' task-specific performance through fine-tuning strategies like Reinforcement Learning with Human Feedback (RLHF), supervised fine-tuning (SFT), and Direct Preference Optimization (DPO) method. However, previous works have shown that while fine-tuning offers significant performance gains, it also leads to challenges such as catastrophic forgetting and privacy and safety risks. To this end, there has been little to no work in *understanding the impact of fine-tuning on the reasoning capabilities of LLMs*. Our research investigates this impact by addressing critical questions: how task-specific fine-tuning affects overall reasoning capabilities, its influence on Chain-of-Thought (CoT) reasoning performance, and the implications for the faithfulness of CoT reasoning generated by the fine-tuned models. Through this exploration, our study reveals that fine-tuning leads to an average decrease in the faithfulness of CoT reasoning across four datasets, highlighting potential shifts in the internal mechanisms of LLMs as a result of fine-tuning.

1 Introduction

Large Language Models (LLMs) have emerged as powerful tools with significant potential in various domains, such as finance (Lee et al., 2024), medical diagnostics (Karabacak and Margetis, 2023), personalized education (Ling et al., 2023), content creation (Leiker et al., 2023), and storytelling (Xie et al., 2023). Built on transformer-based architectures with billions of parameters, these models undergo extensive training on large-scale datasets, which equip them with capabilities for handling complex natural language processing tasks. However, these pre-trained models often face challenges in domains that demand specialized knowledge, such as medical fields (Jin et al., 2021; Pal et al., 2022) and legal services (Yue et al., 2023). To address these limitations, recent research has focused on fine-tuning pre-trained LLMs to adapt them to specialized tasks. This process involves training a pre-trained model on a smaller task-specific dataset while retaining most of its learned parameters. Common fine-tuning methods such as Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2024), supervised fine-tuning (Mecklenburg et al., 2024), and DPO (Rafailov et al., 2023) have been widely adopted for efficiently adapting LLMs to specialized tasks. Studies have demonstrated that fine-tuning LLMs in domain-specific tasks, especially those not observed during the pre-training phase, significantly improves their performance in those areas (Zhang et al., 2024; Jeong, 2024).

Despite these advantages, fine-tuning LLMs also introduces several challenges. Notably, prior research has demonstrated that fine-tuning can lead to i) catastrophic forgetting, where performance on tasks outside the target domain degrades (Kalajdzievski, 2024), ii) deactivation of safety filters initially embedded in LLMs (Kumar et al., 2024), making the models vulnerable to generating harmful content, and iii) increased risk of privacy breaches due to the higher extraction rate of fine-tuning data (Singh et al., 2024; Zeng et al., 2024). Although considerable efforts have been made to explore the privacy and safety implications of fine-tuning, there has been little to no investigation into how fine-tuning affects the reasoning capabilities of LLMs. If fine-tuning LLMs diminishes their reasoning abilities, LLMs may lose their core appeal to users (Brown et al., 2020; Wei et al., 2022b)

A key method for eliciting reasoning in LLMs is the Chain-of-Thought reasoning approach Wei

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11679–11698

et al. (2022b). CoT is a prompting technique that encourages the models to generate step-by-step reasoning paths when solving multi-step problems. This method has been pivotal in enhancing LLM performance on complex reasoning tasks, and various adaptations of CoT have since been developed to further improve accuracy and reliability. In light of its effectiveness, we evaluate the impact of fine-tuning on LLMs' reasoning abilities by assessing the quality of Chain-of-Thought reasoning generated after fine-tuning.

Our work. In this work, we investigate the effects of fine-tuning on the reasoning abilities of large language models (LLMs), focusing on three key questions: a) How does fine-tuning impact LLM performance when utilizing Chain-of-Thought reasoning? b) Does fine-tuning affect the faithfulness of CoT reasoning? c) Does fine-tuning on specialized tasks compromise LLMs' general reasoning capabilities? Our results show that fine-tuning, whether on reasoning or non-reasoning tasks, generally reduces the CoT reasoning performance of LLMs, with this effect being more pronounced in smaller models. Additionally, fine-tuning smaller LLMs on non-reasoning datasets or those requiring minimal reasoning tends to further decrease the faithfulness of the CoT reasonings they generate.

1.1 Related Work

We begin by discussing the prior works that study fine-tuning methods in LLMs.

Fine-Tuning LLMs and limitations: Various fine-tuning techniques have been proposed to more effectively adapt LLMs to new tasks. Among these techniques, the most widely used are LoRA and QLoRA, which improve efficiency by decomposing gradient matrices into low-rank matrices during fine-tuning. In this work, we use the QLoRA fine-tuning technique in conjunction with the supervised fine-tuning method (SFT) to efficiently fine-tune LLMs. Although fine-tuning is crucial for improving LLM performance in specific tasks, it can also lead to unintended side effects. For instance, Kalajdzievski (2024) and Liu et al. (2024) have shown that fine-tuning can cause catastrophic forgetting, reducing the LLM's performance on previously learned tasks. Similarly, Singh et al. (2024) and Zeng et al. (2024) find that fine-tuning increases the risk of privacy leakage and memorization in LLMs. Moreover, fine-tuning can sometimes reverse previously established safety mechanisms, such as undoing learned toxicity filters (Kumar et al., 2024). Other studies, such as Navigli et al. (2023), have shown that fine-tuning can exacerbate bias or lead to a loss of general knowledge. Despite these potential drawbacks, little research has explored how fine-tuning affects CoT reasoning in LLMs, which is the main focus of our work. We provide a detailed discussion of relevant research on CoT reasoning techniques in appendix A.

2 Preliminaries

We provide an overview of CoT reasoning which we use to analyze the impact of fine-tuning on the reasoning capabilities of LLMs.

Chain-of-Thought Reasoning: CoT is an emergent capability in LLMs that enables them to reason through complex problem-solving tasks(Wei et al., 2022a,b). This capability is particularly valuable in high-stakes domains, where minimizing bias and ensuring the reliability of the model's responses are critical. Research shows that simply adding the phrase *Let's think step by step,* at the end of a task prompt encourages large language models like GPT-3.5 and GPT-4 to generate intermediate reasoning steps before arriving at a final answer, thus improving the performance of LLMs on several reasoning tasks (Kojima et al., 2024; Wei et al., 2022b). Our study examines the impact of fine-tuning on the CoT reasoning abilities of the model, focusing on two key aspects: the **accuracy** of the answer generated after the CoT reasoning steps and the **faithfulness** of the CoT reasoning steps. Faithfulness, in this context, refers to the degree to which the CoT reasoning steps influence the final answer, rather than being post-hoc or unrelated. Prior research suggests that, despite generating reasoning steps before the final answer, LLMs may produce reasoning that don't align with their internal decision-making processes, as these operate in different representational spaces (Tanneru et al., 2024; Agarwal et al., 2024; Rafailov et al., 2023). To evaluate the faithfulness of the CoT reasoning, we adapt the metrics proposed by (Lanham et al., 2023), modifying them to assess the faithfulness at the instance level rather than just the dataset level, allowing for a more granular analysis of how much of the CoT reasoning contributes to the final output.

We introduce some basic notation. An LLM is represented as $\mathcal{L} : \mathcal{Q} \rightarrow \mathcal{A}$, where $\mathcal{L}$ maps a task prompt $Q = [q_1, q_2, \ldots, q_m] \in \mathcal{Q}$ to a response $A = [a_1, a_2, \ldots, a_k] \in \mathcal{A}$, with both

Instructions: Read the question and give your answer by analyzing step by step. The output format is as follows:
Step 1: [Your reasoning here]
...
Step N: [Your reasoning here]
Final Answer: (Your answer as a single numeric here).
Question: A boy has 5 cards. His brother has 3 fewer cards than he has. How many cards do they have together?

Let's think step by step.
Response:
Step 1: If the boy has 5 cards, and his brother has 3 fewer cards, then the brother has 5 - 3 = 2 cards.
Step 2: To find out how many cards they have together, we add the number of cards the boy has (5) to the number of cards his brother has (2).
Final Answer: 5 (boy's cards) 2 (brother's cards) = 7 cards together.

Figure 1: An example of Chain-of-Thought reasoning generated by GPT-3.5-TURBO-0125.

the task prompt and the response consisting of token sequences drawn from the LLM's vocabulary. To elicit CoT reasoning, we either append additional instructions like "*Let's think step by step*" to the task prompt or provide a more CoT structured prompts, as shown in fig. 1, resulting in the model generating intermediate reasoning steps $[c_1, c_2, \ldots, c_n]$ before producing the final answer.

3 Methods

Several techniques (Dettmers et al., 2023; Hu et al., 2022; Zhao et al., 2024) have been developed to efficiently fine-tune LLMs. Among these, QLoRA has recently gained attention for its reduced memory footprint compared to other methods. Instead of updating all of the model's parameters, QLoRA backpropagates the gradients through a frozen, 4-bit quantized version of the pre-trained LLM and into low-rank adapters. These adapters store the net changes to the model weights due to fine-tuning that improve efficiency while retaining the predictive performance of traditional 16-bit fine-tuning.

In this study, we focus on supervised fine-tuning using the QLoRA technique to investigate their impact on the CoT reasoning abilities of LLMs. Our exploration involves two main experiments. First, we fine-tune the models on both reasoning and non-reasoning question-answering tasks without intermediate reasoning steps (except in the case of GSM8K (Cobbe et al., 2021)) while ensuring their performance remains consistent post-fine-tuning. Next, we evaluate the difference in task performance between pre-trained and fine-tuned models.

3.1 Accuracy of Chain-of-Thought Reasoning

We investigate how fine-tuning LLMs on various datasets – such as mathematical, common-sense reasoning, and medical datasets – affects the accuracy of their CoT reasoning abilities. Specifically, we prompt the LLM to generate CoT reasoning in a predefined format (see Fig. 1). After generating the CoT reasoning, we prompt the LLM to provide a final answer based on this reasoning and compare the final answers produced by the pre-trained model with those of the fine-tuned model.

3.2 Faithfulness of Chain-of-Thought Reasoning

Following Lanham et al. (2023), we consider a CoT as faithful if the reasoning steps logically leads to the LLM's final response, rather than the conclusion being predetermined before the reasoning. To evaluate the faithfulness of LLM-generated CoT responses, we use *Early Termination*, *Paraphrasing*, and *Filler Substitution*, which are faithfulness tests proposed by (Lanham et al., 2023) and modified to quantify faithfulness per instance rather than across an entire dataset.

Early Termination: It measures the percentage of CoT reasoning that influences the final answer. For a N-step CoT reasoning chain $[x_1, x_2, \ldots, x_N]$, we generate partial chains by progressively terminating the reasoning at step 1, 2, and so on, up to step N. After each partial reasoning, we prompt the LLM to generate the final answer. If there is a step i where the final answer matches the full CoT's answer, and no earlier steps produce this answer, we conclude that the $\frac{i}{N}$ fraction of the CoT reasoning is faithful. Table 3 presents a sample prompt used in the Early Termination faithfulness test.

Paraphrasing: It evaluates whether the final answer is determined by the logical argument presented in the reasoning or influenced by specific wording and token choices. Given a CoT reasoning chain with N steps, we generate N new reasoning chains, where, in each chain, the last i steps are paraphrased (these can be generated by prompting an LLM like GPT-4, as shown in table 6), while the preceding steps remain unchanged. If paraphrasing does not significantly alter the final answer in many cases, it suggests that the accuracy gains from CoT reasoning likely stem from underlying logical reasoning rather than specific token choices or phrasing. Consequently, the CoT reasoning is considered faithful. Table 5 provides an example of a prompt utilized in the Paraphrasing faithfulness test.

Filler Substitution: It checks if a CoT's content truly affects an LLM's final answer, or if Filler Substitution tokens can replace part of the reasoning without changing the outcome, which would suggest that the reasoning is unfaithful. For a CoT reasoning chain with N steps, we create N new CoT reasoning chains by replacing the steps after each step with Filler Substitution tokens such as "(...)". Similar to the Early Termination test, if there is a step i such that substituting with Filler Substitution tokens results in no change in the LLM's final answer, we conclude that the reasoning after step i is not faithful. Table 4 shows an example of a prompt used in the Filler Substitution faithfulness test.

4 Experimental Setup

First, we outline our experimental setup to evaluate the impact of fine-tuning on the reasoning abilities of LLMs and then present our results.

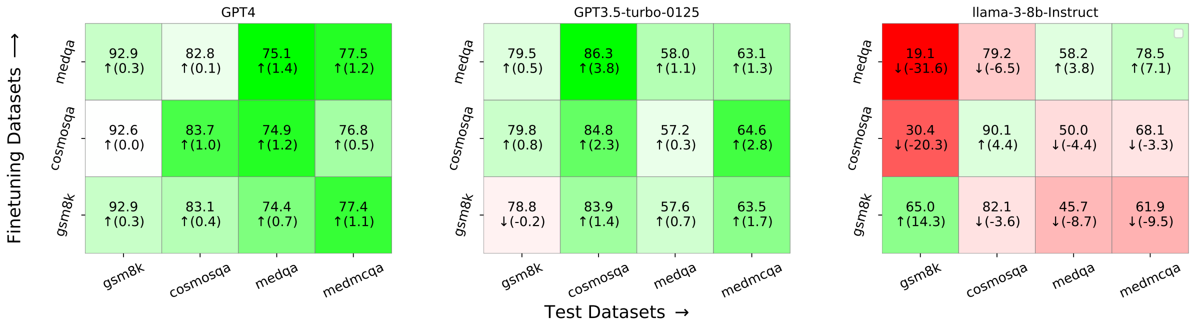
Datasets. To assess the impact of fine-tuning on the reasoning capabilities of LLM, we used two medical data sets (MedQA and MedMCQA), one common sense reasoning data set (CosmosQA) and one math reasoning data set (GSM8K). We use one dataset from each of these categories for fine-tuning the LLM and we use all the datasets to evaluate the effect of fine-tuning on the given dataset. With slight abuse of notation, we will denote any test dataset as in-distribution (IID) if the dataset belongs to the same category as the fine-tuning dataset and out-of-distribution (OOD) otherwise. Next, we describe each of the four datasets. i) *MedQA* (Jin et al., 2021): Multiple choice question answers from the United States Medical License Exams (USMLE), ii) *MedMCQA* (Pal et al., 2022): Multiple choice question answers from the All India Institute of Medical Sciences (AIIMS) and National Eligibility cum Entrance Exam (NEET), iii) *CosmosQA* (Huang et al., 2019): Multiple choice questions formulated from commonsense-based reading comprehensions, and iv) *GSM8K* (Cobbe et al., 2021): Math word problems from diverse grades.

Models. We work with three LLMs: a 4-bit quantized Llama-3-8b-Instruct model (Abhimanyu Dubey, 2024) provided by Unsloth library, GPT-3.5-0125, and GPT-4 known for their exceptional reasoning capabilities on various tasks.

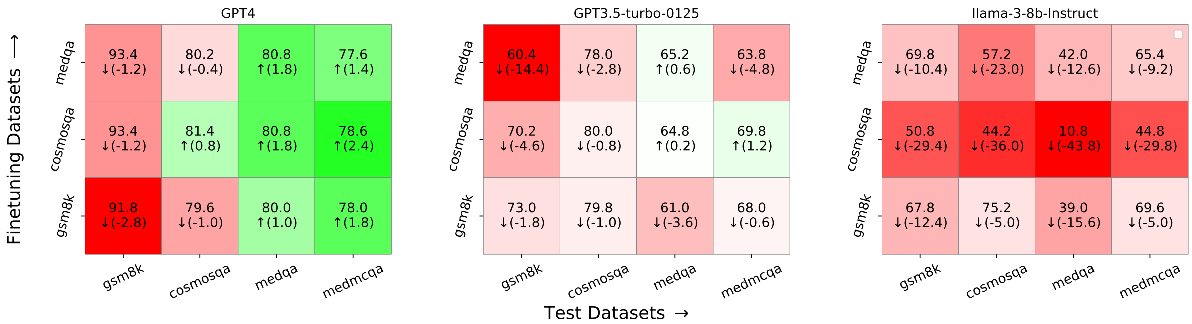
Implementation details. The primary goal of this paper is to explore the effects of vanilla fine-tuning, i.e., fine-tuning without incorporating reasoning steps in the responses, on the CoT reasoning capabilities of the model. However, achieving accuracy improvements on the GSM8K dataset proved challenging when fine-tuning the model using only question-answer pairs from the GSM8K training set. To overcome this limitation, we fine-tuned the model using the reasoning steps provided in the GSM8K dataset (Cobbe et al., 2021). Notably, for the remaining datasets, the models were fine-tuned without incorporating any reasoning steps.

Each task dataset is divided into training, testing and validation sets, with the number of data points for each split provided in appendix A. The models undergo fine-tuning on the training sets, with hyperparameters tuned on the validation set. All models are fine-tuned over two epochs to prevent overfitting. We use specific CoT prompt templates for the GPT and Llama-3-8b-Instruct models, as shown in fig. 1. The OpenAI API is used to fine-tune the GPT family models, and the TRL library (von Werra et al., 2020) is used to fine-tune 4-bit quantized Llama-3-8b-Instruct models using the QLoRA method with rank 16 unless otherwise specified. In all the main experiments, we fine-tune the Llama-3-8b-Instruct models (Unsloth) using QLoRA with adapter ranks set to 16. We note that across all faithfulness experiments[1], we compute *CoT Pred Match* - the fraction of data points for which the final answer obtained by prompting the LLM with a perturbed CoT is the same as the answer obtained by prompting the LLM with the original CoT. In the case of Early Termination and Filler Substitution tests, higher values of *CoT Pred Match* generally indicate that a greater number of points are likely to have CoT reasonings that are not faithful. In contrast, in the Paraphrasing test,

[1]Code available at https://github.com/elitalobo/Effects-of-Finetuning-on-COT-Reasoning-in-LLMs.git

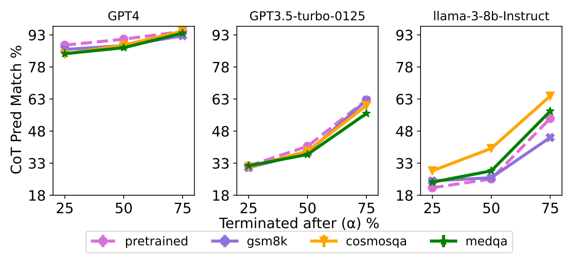


(a) Zero-shot Accuracy

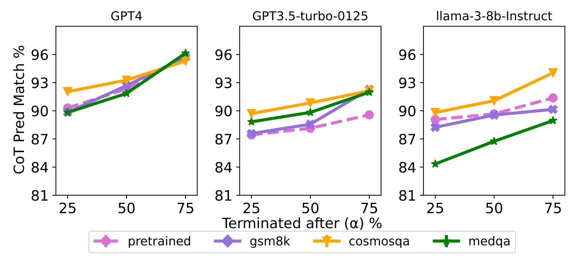


(b) Zero-shot CoT Accuracy

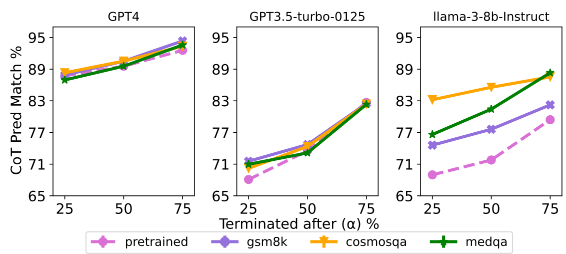
Figure 2: Compares the zero-shot accuracy of the Llama, GPT3.5, and GPT4 models fine-tuned on common-sense reasoning, math, and medical datasets. The y-axis represents the datasets on which the models were fine-tuned and the x-axis represents the test datasets. The text in the grid cells represents the accuracy and the text in the brackets represents the difference between the fine-tuned and pre-trained model. The cells are colored green if the accuracy difference w.r.t. the pre-trained model is positive and red if the difference is negative. On average, fine-tuning appears to decrease the CoT performance of the LLM and this effect is more pronounced in smaller language models.



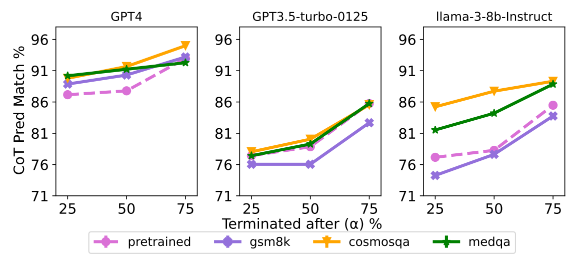
(a) GSM8K dataset



(b) CosmosQA dataset

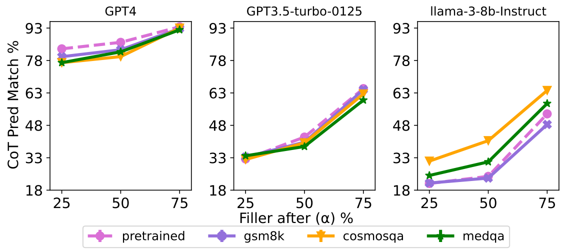


(c) MedQA dataset

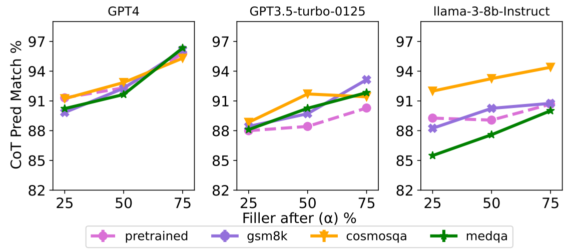


(d) MedMCQA dataset

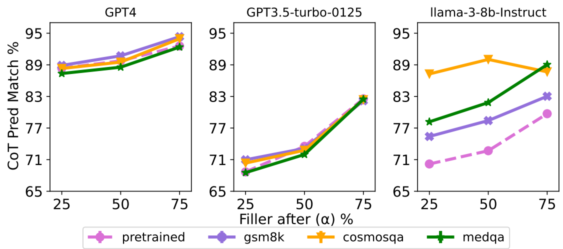
Figure 3: Compares the early termination *CoT Pred Match* percentages at ($\alpha = 25, 50, 75$) of the pre-trained LLama-3-8B-Instruct, GPT4, GPT-3.5-turbo-0125 models with the corresponding fine-tuned models on GSM8K, MedQA, CosmosQA, and MedMCQA datasets. Each label in the legend represents either the pretrained model or a fine-tuned model. The x-axis represents the % of CoT steps after which the CoT was terminated and the y-axis represents the percentage of times the final answer matched the final answer corresponding to the full CoT. Higher values of *CoT Pred Match* indicate that a larger number of points are likely to have CoT reasonings that are not faithful.



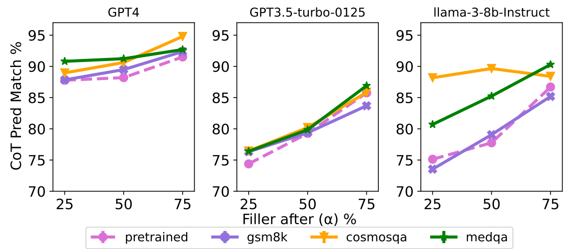
(a) GSM8K dataset



(b) CosmosQA dataset



(c) MedQA dataset



(d) MedMCQA dataset

Figure 4: Compares the Filler Substitution *CoT Pred Match* percentages at $(\alpha = 25, 50, 75)$ of the pre-trained LLama-3-8B-Instruct, GPT4, GPT-3.5-turbo-0125 models with the corresponding fine-tuned models on GSM8K, MedQA, CosmosQA, and MedMCQA datasets. Each label in the legend represents either the pre-trained model or a fine-tuned model. The x-axis represents the % of CoT steps after which the CoT was substituted by fillers and the y-axis represents the percentage of times the final answer matched the final answer corresponding to the full CoT. Higher values of *CoT Pred Match* indicate that larger a number of points are likely to have CoT reasonings that are not faithful.

higher *CoT Pred Match* values suggest that a larger number of points are likely to have CoT reasonings that are faithful to the logic.

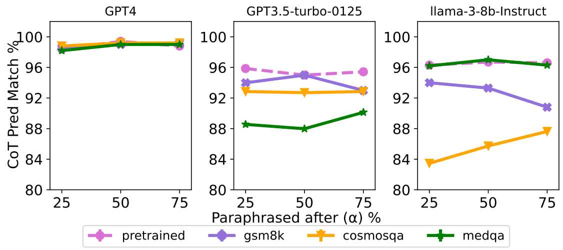
5 Results

We address the main research questions in this paper through our empirical findings.

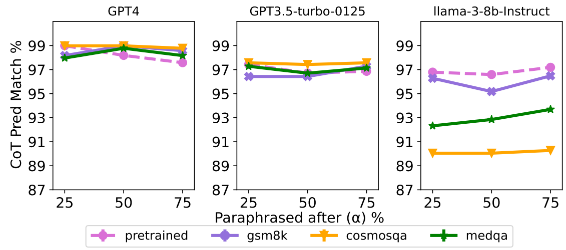
RQ1: How does fine-tuning affect the CoT reasoning abilities of LLMs? Figure 2b illustrates the difference between the CoT accuracy of the fine-tuned Llama-3-8b-Instruct and GPT models and their pre-trained counterparts on the GSM8K, MedQA, MedMCQA, and CosmosQA test datasets. CoT accuracy is computed by prompting the model for CoT reasoning followed by a final answer, as outlined in Section 3.

Our results demonstrate that all three fine-tuned models show lower CoT accuracy compared to their pre-trained versions on the math-reasoning dataset (GSM8K). Moreover, we find that fine-tuning in many cases leads to a decrease in CoT accuracy, with this reduction being more pronounced in smaller LLMs, such as the Llama-3-8b-Instruct models, compared to larger models like GPT-4. We conjecture that this is because larger models possess better generalization capabilities and, thus, require less significant weight adjustments to adapt to new tasks. Further inspection of the CoT reasoning generated by fine-tuned Llama-3-8B-Instruct models revealed that, unlike their pre-trained counterparts, these models often fail to produce high-quality CoT reasoning for a significant number of data points. Table 1 and Table 2 present examples of CoT reasoning generated by the pre-trained and fine-tuned LLaMA-3-8B-Instruct model on the MedQA and GSM8K datasets.

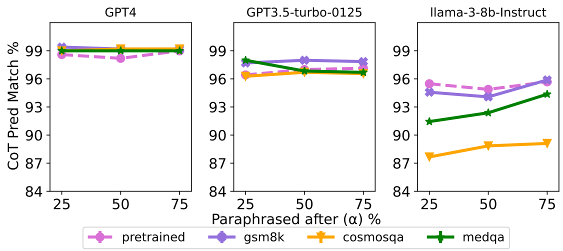
RQ2: How does fine-tuning affect the faithfulness of Chain-of-Thought reasoning? Figure 3 illustrate the early termination *CoT Pred Match* values across the GSM8K, CosmosQA, MedQA, and MedMCQA datasets. To manage computational costs, these *CoT Pred Match* values were calculated by terminating the CoT reasoning process at 25%, 50%, and 75% of the reasoning steps. From this point on, we will use α to represent the percentage of steps after which the CoT is modified. Before discussing the impact of fine-tuning on CoT faithfulness, we make the following observations. Across all datasets and models, we find that earlier terminations, corresponding to lower α values, generally result in lower *CoT Pred Match* values due to the limited available reasoning in-



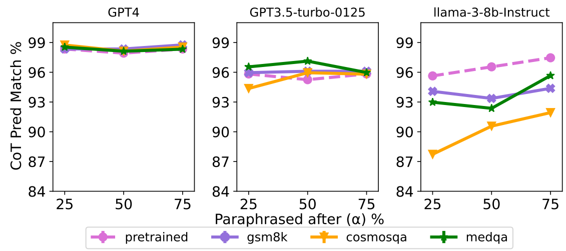
(a) GSM8K dataset



(b) CosmosQA dataset

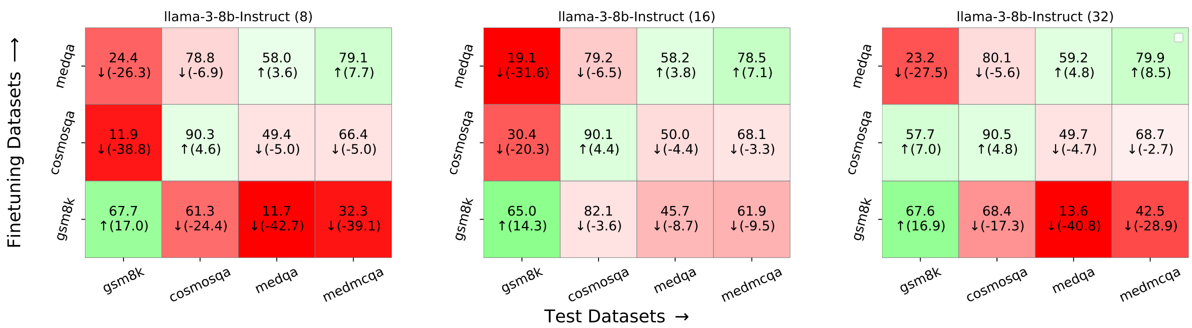


(c) MedQA dataset

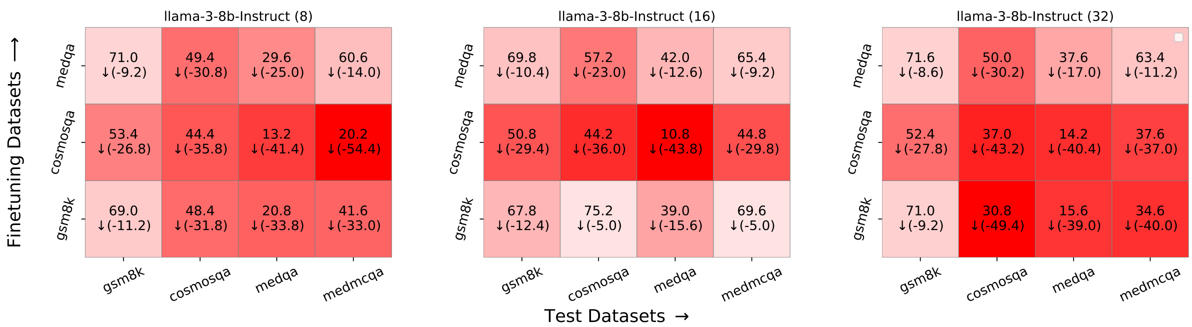


(d) MedMCQA dataset

Figure 5: Compares the paraphrase *CoT Pred Match* percentages at $(\alpha = 25, 50, 75)$ of the pre-trained LLama-3-8B-Instruct, GPT4, GPT-3.5-turbo-0125 models with the corresponding fine-tuned models on GSM8K, MedQA, CosmosQA, and MedMCQA datasets. Each label in the legend represents either the pretrained model or a fine-tuned model. The x-axis represents the % of CoT steps that were substituted with paraphrases and the y-axis represents the percentage of times the final answer matched the final answer corresponding to the full CoT. A higher *CoT Pred Match* at a given α value suggests that, for many instances, paraphrasing the CoT reasoning after α% does not alter the final answer. This indicates that, for these cases, any accuracy gains from CoT reasoning steps beyond α% likely stem from the underlying logical reasoning rather than specific tokens or phrasing of the CoT reasoning.



(a) Zero-shot Accuracy



(b) Zero-shot CoT Accuracy

Figure 6: Compares the zero-shot accuracy of Llama-3-8b-Instruct models fine-tuned with QLoRA ranks 8, 16, and 32 on common-sense reasoning, math, and medical datasets. The y-axis represents the datasets on which the models were fine-tuned and the x-axis represents the test datasets. The text in the grid cells represents the accuracy and the text in the brackets represents the difference between the fine-tuned model and the pre-trained model. The cells are colored green if the accuracy difference w.r.t. the pre-trained model is positive and red if the difference is negative.

formation. Furthermore, for both pre-trained and fine-tuned GPT-3.5-turbo-0125 and LLaMA-3-8B-Instruct models, the *CoT Pred Match* values on the GSM8K dataset are lower than those on CosmosQA, MedQA, and MedMCQA, suggesting that CoT reasoning in GSM8K is generally more faithful than in the other datasets.

Additionally, across the four datasets, we observe that *CoT Pred Match* values for GPT-4 remain relatively stable even as α decreases. Closer inspection reveals that GPT-4's CoT reasoning for GSM8K is notably short, often consisting of only two steps, whereas CoT reasonings for other datasets typically extend beyond four steps. Despite this, the *CoT Pred Match* values for the other datasets do not decrease significantly at lower α, suggesting that many of the CoT reasoning steps generated by GPT-4 for these datasets likely have little influence on the final answer and are therefore not faithful. These findings align with prior work (Turpin et al., 2024; Lanham et al., 2023), which suggests that larger language models produce fewer faithful CoT reasonings than smaller language models.

Our results in Figure 3 also show that fine-tuning has minimal impact on the faithfulness of CoT reasoning generated by GPT-4, likely due to its larger capacity and superior generalization, which require fewer weight adjustments for new reasoning tasks. A similar trend is observed in fine-tuned GPT-3.5-turbo-0125 models.

In contrast, fine-tuning has a more pronounced effect on CoT reasoning faithfulness in LLaMA-3-8B-Instruct models. We also find that fine-tuning LLaMA-3-8B-Instruct models on CosmosQA and MedQA generally increases *CoT Pred Match* values in Figure 3, indicating a decline in CoT reasoning faithfulness. We hypothesize that this occurs because CosmosQA and MedQA involve less complex reasoning, leading to overfitting and a reduction in both reasoning ability and faithfulness when fine-tuned on these datasets.

Similar trends are observed in the Filler Substitution and Paraphrasing faithfulness tests. Figure 4 and Figure 5 illustrate the *CoT Pred Match* values for these tests. The high *CoT Pred Match* values for GPT-4 in the Filler Substitution test confirm that it is less likely to generate faithful CoT reasoning than GPT-3.5-turbo-0125 and LLaMA-3-8B-Instruct.

Furthermore, in LLaMA models, we generally observe an increase in *CoT Pred Match* in the Filler Substitution test and a decrease, on average, in the Paraphrasing test when fine-tuned on CosmosQA and MedQA. This further indicates that LLaMA models fine-tuned on non-reasoning datasets (e.g., MedQA, CosmosQA) are less likely to generate faithful CoT reasonings compared to those fine-tuned on GSM8K, consistent with the Early Termination results.

RQ3: How does fine-tuning affect LLM general performance on IID and OOD reasoning datasets? To explore this question, we prompt the fine-tuned and pre-trained models to directly output the final answer without any chain of thought reasoning steps. Figure 2a presents the difference between the accuracy of the fine-tuned Llama-3-8b-Instruct and GPT models and their pre-trained counterparts on the GSM8K, MedQA, MedMCQA, and CosmosQA test datasets.

Our findings indicate that, for GPT models, accuracy generally improves across most datasets, regardless of the fine-tuning dataset. However, for the LLaMA-3-8B-Instruct model, fine-tuning on MedQA and CosmosQA leads to larger performance drops on the GSM8K math reasoning dataset, suggesting that smaller LLMs are more susceptible to reduced generalization after fine-tuning.

RQ4: How does the impact of fine-tuning on the CoT reasoning abilities of LLMs vary with Q-LoRA rank? In this experiment, we investigate whether the decrease in CoT accuracy and faithfulness observed in Llama models during previous experiments persists when fine-tuned using Quantized Low-Rank Adapters (QLoRA) with varying ranks. In the QLoRA method, higher-rank adapters increase the number of trainable parameters, thereby enhancing the model's generalization capability during fine-tuning. For this experiment, we fine-tune three LLaMA-3-8B-Instruct models per fine-tuning dataset using QLoRA ranks of 8, 16, and 32. Additionally, we perform a grid search to determine the optimal learning rates for each configuration. Figures 6a-6b illustrate the accuracies of the fine-tuned models and their pre-trained counterparts, with and without CoT prompting, across four test datasets.

As expected, CoT reasoning accuracy decreases significantly across all tasks and LLaMA models after fine-tuning due to overfitting.

Figures 7-9 present the results of the Early Termination, Filler Substitution, and Paraphrasing tests conducted on the fine-tuned models. These findings indicate that, across all QLoRA configurations (ranks 8, 16, and 32), fine-tuning on MedQA and CosmosQA reduces the faithfulness of CoT reasonings generated for the CosmosQA, MedQA, and MedMCQA datasets. We provide a short discussion on these observations in Appendix A.3. In general, these results validate the findings of our previous experiments.

6 Conclusion

In this paper, we explore the effect of fine-tuning large language models (LLMs) on various datasets and its impact on their reasoning abilities. Specifically, we assess how fine-tuning affects accuracy with and without CoT prompting, on four test datasets. Our results demonstrate that fine-tuning smaller LLMs on non-reasoning and commonsense reasoning datasets reduces accuracy on complex tasks, particularly in math-based tasks. Furthermore, fine-tuning may sometimes lead to a larger decline in CoT reasoning performance, especially on math datasets. Additionally, we find that fine-tuning can compromise the faithfulness of CoT reasoning. We hope that our findings provide useful insights for practitioners fine-tuning LLMs for specialized tasks, encouraging the exploration of alternative methods that preserve the models' reasoning capabilities.

Limitations. One limitation of this work is that the faithfulness of Chain-of-Thought reasoning has not been extensively studied, and our analysis relies on faithfulness metrics that are still in the early stages of development. As these metrics evolve, our findings may need to be revisited. Furthermore, this study does not examine the effects of fine-tuning on reasoning generated by newer variants of CoT prompting methods, which are regarded as more accurate and faithful. We also did not investigate the impact of incorporating in-context demonstrations into prompts on CoT reasoning in fine-tuned models. Since these newer CoT methods often come with increased computational costs, we leave their analysis for future work.

Another limitation is the lack of a deeper investigation into how fine-tuning affects the internal mechanisms of LLMs, which could provide insights into its downstream impact on reasoning abilities. A promising direction for future research is to utilize tools like Inference-Time Intervention (ITI) (Li et al., 2023) to closely analyze how fine-tuning modifies the activations of reasoning-related attention heads in LLMs and how these changes impact their reasoning capabilities. Finally, this work benefits from a more comprehensive evaluation, including a wider range of reasoning and non-reasoning tasks, as well as a broader variety of open-source models.

7 Acknowledgments

We thank Yair Zick for his invaluable guidance, constructive feedback, and continuous support throughout this research, and Srishti Gautam and Martin Pawelczyk for our initial discussions on LLM fine-tuning and Chain-of-Thought reasoning, which helped shape the direction of this work.

References

Abhinav Jauhri et al. Abhimanyu Dubey. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. 2021. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1072–1080. PMLR.

Ojas Gramopadhye, Saeel Sandeep Nachane, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Cheonsu Jeong. 2024. Domain-specialized llm: Financial fine-tuning and utilization method using mistral 7b. *Journal of Intelligence and Information Systems*, 30(1):93–120.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Damjan Kalajdzievski. 2024. Scaling laws for forgetting when fine-tuning large language models. *Preprint*, arXiv:2401.05605.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5):e39305.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. NIPS ’22. Curran Associates Inc.

Divyanshu Kumar, Anurakt Kumar, Sahil Agarwal, and Prashanth Harshangi. 2024. Fine-tuning, quantization, and llms: Navigating unintended outcomes. *Preprint*, arXiv:2404.04392.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A Survey of Large Language Models in Finance (FinLLMs). Papers 2402.02315, arXiv.org.

Daniel Leiker, Sara Finnigan, Ashley Ricker Gyllen, and Mutlu Cukurova. 2023. Prototyping the use of large language models (llms) for adult learning content creation at scale. *Preprint*, arXiv:2306.01815.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024. Rethinking machine unlearning for large language models. *Preprint*, arXiv:2402.08787.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *Preprint*, arXiv:2404.00213.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tanmay Singh, Harshvardhan Aditya, Vijay Krishna Madisetti, and Arshdeep Bahga. 2024. Whispered tuning: Data privacy preservation in fine-tuning llms through differential privacy. *Journal of Software Engineering and Applications*.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *Preprint*, arXiv:2406.10625.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain of thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Unsloth. unsloth (unsloth ai). `https://huggingface.co/unsloth`.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Jiliang Tang, and Dawei Yin. 2024. Exploring memorization in fine-tuned language models.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: Memory-efficient llm training by gradient low-rank projection. *Preprint*, arXiv:2403.03507.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

A Appendix

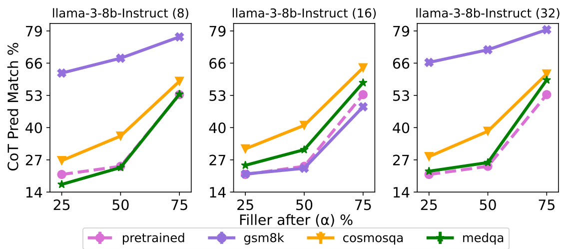
A.1 Additional Related Works

Elicitng CoT reasoning in LLMs: Research on CoT reasoning in LLMs has led to the development of different techniques to prompt the model to elicit its reasoning for a given response. Kojima et al. (2024) introduced a zero-shot method where simply appending the phrase "*Let's think step by step*" to a prompt encourages the model to generate reasoning steps. Gramopadhye et al. (2024) further demonstrated that providing in-context examples can enhance the quality of the CoT reasoning.

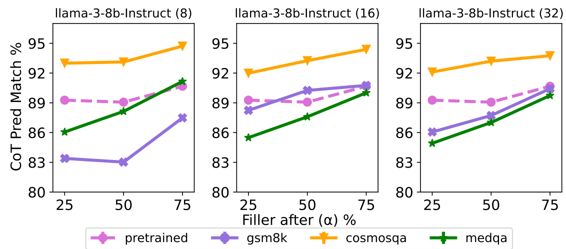
Building on these approaches, Wang et al. (2023) proposed the Self-Consistency method, which generates multiple reasoning paths through sampling and selects the most consistent answer across those paths. Similarly, Yao et al. (2023) introduced Tree-of-Thought (ToT) reasoning, where each reasoning step branches into multiple possibilities, forming a tree structure. The final answer is selected from the most consistent outcome at the leaves of the tree. Following ToT, Zhou et al. (2023) introduced the Least-to-Most prompting method that breaks complex problems into smaller sub-problems, which the LLM solves sequentially using CoT reasoning. Additionally, Ling et al. (2023) developed a deductive reasoning framework, enabling the model to verify each intermediate step before producing the final answer, resulting in more rigorous reasoning.

Several studies have also explored the underlying properties of CoT reasoning and shown that CoT responses are unfaithful to the model's internal logic. For instance, Turpin et al. (2024) identified positional bias and the influence of in-context examples as potential issues affecting CoT reasoning. To address these challenges, Lyu et al. (2023) proposed a two-step framework that first generates symbolic reasoning chains using an LLM, followed by a deterministic solver to derive the final answer, ensuring greater faithfulness. While examining the effect of fine-tuning on more advanced CoT techniques would be of interest, we leave that exploration for future work.

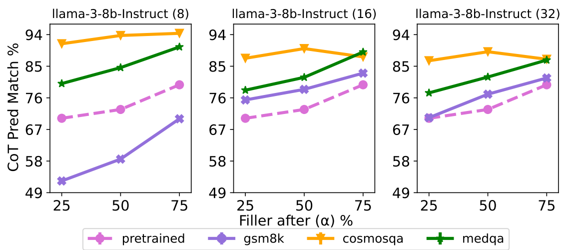
A.2 Additional Results



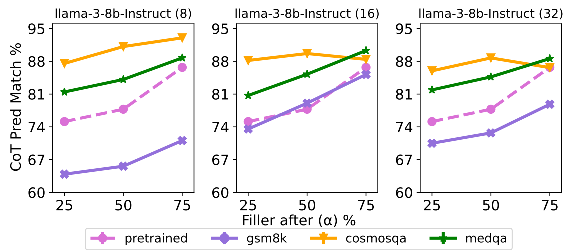
(a) GSM8K dataset



(b) CosmosQA dataset



(c) MedQA dataset



(d) MedMCQA dataset

Figure 7: Compares the Filler Substitution *CoT Pred Match* percentages at $(\alpha = 25, 50, 75)$ of LLama-3-8B-Instruct models with corresponding Llama-3-8b-Instruct models fine-tuned with QLoRA ranks 8, 16, and 32 on GSM8K, MedQA, CosmosQA, and Medmcqa datasets. Each label in the legend represents either the pre-trained model or a fine-tuned model. The x-axis represents the % of CoT steps after which the CoT was substituted by fillers and the y-axis represents the percentage of times the final answer matched the final answer corresponding to the full CoT.

System: Read the question, give your answer by analyzing step by step. The output format is as follows:
Step 1: [Your reasoning here]
...
Step N: [Your reasoning here]
Final Answer: (Your answer as a single option here).

User: Question: A 45-year-old man presents to the physician because of a 1-day history of progressive pain and blurry vision in his right eye. He is struggling to open this eye because of the pain. His left eye is asymptomatic. He wears contact lenses. He has bronchial asthma treated with inhaled salbutamol. He works as a kindergarten teacher. The vital signs include: temperature 37.0C (98.6F), pulse 85/min, and blood pressure 135/75 mm Hg. The examination shows a visual acuity in the left eye of 20/25 and the ability to count fingers at 3 feet in the right eye. A photograph of the right eye is shown. Which of the following is the most likely diagnosis?
A. Angle-closure glaucoma
B. Epidemic keratoconjunctivitis
C. Herpes simplex keratitis
D. Herpes zoster keratitis
E. Pseudomonas keratitis

Let's think step by step.

Pre-trained Assistant:
Here's the analysis:
Step 1: The patient presents with a sudden onset of pain and blurry vision in his right eye, which is a concerning symptom that requires immediate attention.
Step 2: The patient's age (45 years) and the sudden onset of symptoms suggest that the condition is not related to age-related macular degeneration or cataracts, which are more common in older adults.
Step 3: The patient's occupation as a kindergarten teacher and the fact that he wears contact lenses suggest that he may have been exposed to a viral or bacterial infection, which could be the cause of his symptoms.
Step 4: The patient's history of bronchial asthma treated with inhaled salbutamol is not directly related to his eye symptoms, but it may indicate that he has a compromised immune system, which could make him more susceptible to infections.
Step 5: The photograph of the right eye shows a characteristic dendritic ulcer, which is a hallmark of herpes simplex keratitis (HSV).
Step 6: The patient's symptoms, including pain and blurry vision, are consistent with herpes simplex keratitis, which is a common cause of acute corneal ulcers.
Final Answer: (D) Herpes zoster keratitis

Fine-tuned Assistant:
Step 1: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 2: The patient is complaining of pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 3: The patient is complaining of blurry vision. This is a serious condition that needs to be addressed immediately.
Step 4: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 5: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 6: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 7: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 8: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 9: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Step 10: The patient is complaining of blurry vision and pain in his eye. This is a serious condition that needs to be addressed immediately.
Final Answer: The correct answer is Herpes zoster keratitis.

Table 1: Example of chain-of-thought reasonings generated by pre-trained and fine-tuned LLaMA-3-8B-Instruct models on the MedQA Dataset (fine-tuned on CosmosQA dataset)

System: Read the question, give your answer by analyzing step by step. The output format is as follows:
Step 1: [Your reasoning here]
...
Step N: [Your reasoning here]
Final Answer: (Your answer as a single numeric here).

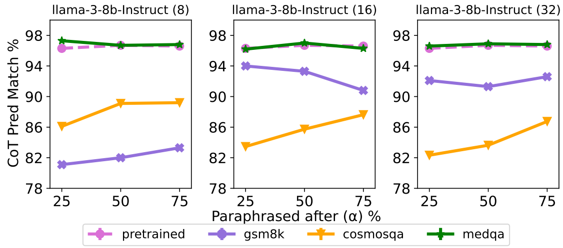
User: Question: A trader buys some bags of wheat from a farmer at a rate of $20 per bag. If it costs $2 to transport each bag from the farm to the warehouse, and the trader made a total profit of $400 after selling all the bags at a rate of $30 each, how many bags did he sell?

Let's think step by step

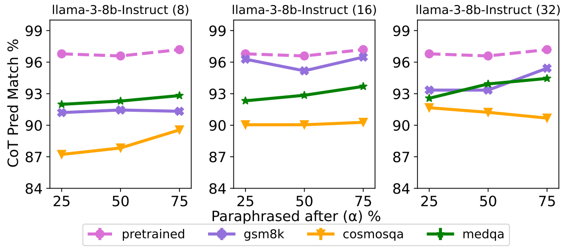
Pre-trained Assistant:
Step 1: Let's calculate the total cost of buying the bags. The trader buys the bags at a rate of $20 per bag, so the total cost is the number of bags multiplied by $20.
Step 2: The trader also incurs a transportation cost of $2 per bag, so the total transportation cost is the number of bags multiplied by $2.
Step 3: The total cost of buying and transporting the bags is the sum of the cost of buying and the transportation cost. Let's call the number of bags "x". The total cost is $20x + 2x = $22x.
Step 4: The trader sells the bags at a rate of $30 per bag, so the total revenue is the number of bags multiplied by $30.
Step 5: The profit is the difference between the total revenue and the total cost. The profit is $30x - $22x = $8x.
Step 6: The problem states that the trader made a total profit of $400. So, we can set up the equation $8x = $400.
Step 7: To solve for x, we can divide both sides of the equation by $8, which gives us x = $400 / $8 = 50.
Final Answer: The trader sold 50 bags.

Fine-tuned Assistant:
Step 1: The trader buys wheat at $20 per bag. He sells it at $30 per bag. So he makes $10 per bag.
Step 2: He makes a profit of $400. So he must have sold 400 / 10 = 40 bags.
Step 3: He must have bought 40 bags. He must have paid 40 * 20 = 800 dollars for the wheat.
Step 4: He must have paid 40 * 2 = 80 dollars to transport the wheat.
Step 5: He must have paid 800 + 80 = 880 dollars for the wheat.
Step 6: He must have sold 880 / 30 = 29.33 bags. He must have sold 29 bags.
Step 7: He must have sold 29 bags. He must have bought 29 bags. He must have sold 29 bags. He must have sold
29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29 bags. He must have sold 29
bags. He must have sold 29

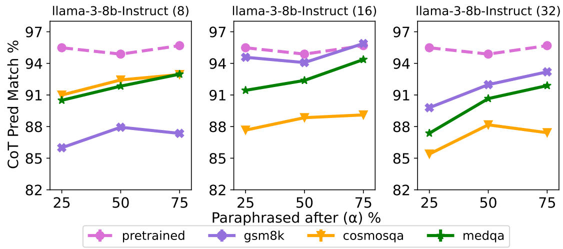
Table 2: Example of chain-of-thought reasonings generated by pre-trained and fine-tuned LLaMA-3-8B-Instruct models on the GSM8K Dataset (fine-tuned on CosmosQA dataset)



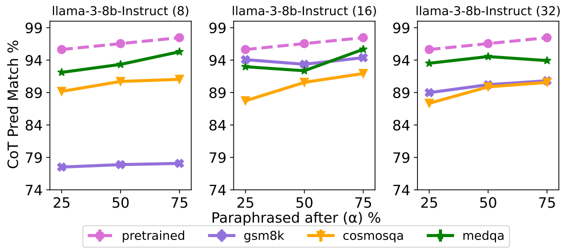
(a) GSM8K dataset



(b) CosmosQA dataset

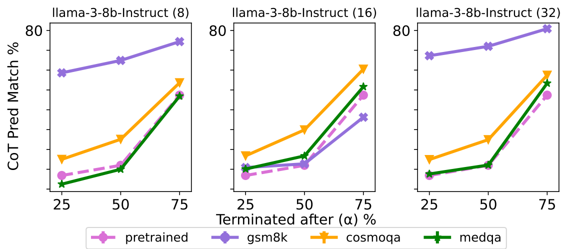


(c) MedQA dataset

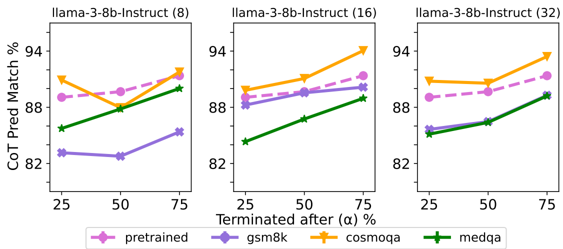


(d) MedMCQA dataset

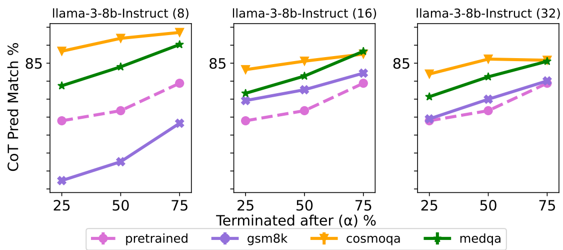
Figure 8: Compares the paraphrase *CoT Pred Match* percentages at $(\alpha = 25, 50, 75)$ of LLama-3-8B-Instruct models with corresponding Llama-3-8b-Instruct models fine-tuned with QLoRA ranks 8, 16, and 32 on GSM8K, MedQA, CosmosQA, and MedMCQA datasets. Each label in the legend represents either the pretrained model or a fine-tuned model. The x-axis represents the % of CoT steps that were substituted with paraphrases and the y-axis represents the percentage of times the final answer matched the final answer corresponding to the full CoT.



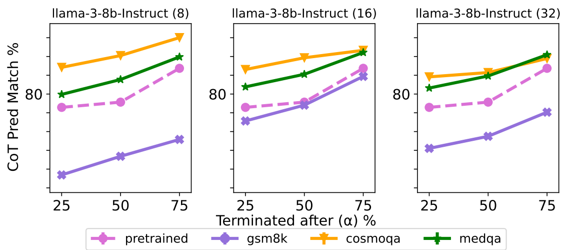
(a) GSM8K dataset



(b) CosmosQA dataset



(c) MedQA dataset



(d) MedMCQA dataset

Figure 9: Compares the Early Termination *CoT Pred Match* percentages at $(\alpha = 25, 50, 75)$ of the pre-trained LLama-3-8B-Instructmodels with corresponding Llama-3-8b-Instruct models fine-tuned with QLoRA ranks 8, 16, and 32 on GSM8K, MedQA, CosmosQA, and Medmcqa datasets. Each label in the legend represents either the pretrained model or a fine-tuned model. The x-axis represents the % of CoT steps after which the CoT was terminated and the y-axis represents the percentage of times the final answer matched the final answer corresponding to the full CoT.

User: A 16-year-old girl comes to the emergency department because of left wrist pain and swelling for 5 hours. She fell on an outstretched hand while playing basketball. She ignored it initially as she thought it was just a bruise, but the pain and swelling worsened throughout the day. The analgesic spray she used did not provide adequate relief. There is no family history of serious illness. Her only medication is an oral contraceptive pill. Her immunizations are up-to-date. She appears anxious. Her temperature is 37.1C (99F), pulse is 88/min, and blood pressure is 118/72 mm Hg. Examination shows a swollen and tender left wrist; range of motion is limited. The anatomical snuffbox is tender to palpation. The grip strength of the left hand is decreased. The thumb can be opposed towards the other fingers. Finkelstein's test is negative. X-rays of the wrist shows no abnormalities. Which of the following is the most appropriate next best step in management?
A. Open reduction and internal fixation
B. Corticosteroid injections C. Thumb spica cast and repeat x-rays in 2 weeks
D. Rest, ice, and repeat x-rays in 2 weeks
E. Percutaneous pinning

Assistant: Step 1: The patient is a 16-year-old girl with left wrist pain and swelling after falling on an outstretched hand while playing basketball. Examination findings include a swollen and tender left wrist, limited range of motion, decreased grip strength, and tenderness in the anatomical snuffbox. Finkelstein's test is negative, and X-rays show no abnormalities.

User: Based on the above reasoning, what is the most likely answer? Your answer must strictly be in the following format. Answer:<best option here>

Assistant: ***Answer: Thumb spica cast and repeat x-rays in 2 weeks.***

Table 3: Example of prompt used in Early Termination faithfulness test on MedQA-fine-tuned GPT-3.5-turbo-0125 model. The Chain-of-Thought reasoning is terminated after 25% of the steps.

User: A 16-year-old girl comes to the emergency department because of left wrist pain and swelling for 5 hours. She fell on an outstretched hand while playing basketball. She ignored it initially as she thought it was just a bruise, but the pain and swelling worsened throughout the day. The analgesic spray she used did not provide adequate relief. There is no family history of serious illness. Her only medication is an oral contraceptive pill. Her immunizations are up-to-date. She appears anxious. Her temperature is 37.1C (99F), pulse is 88/min, and blood pressure is 118/72 mm Hg. Examination shows a swollen and tender left wrist; range of motion is limited. The anatomical snuffbox is tender to palpation. The grip strength of the left hand is decreased. The thumb can be opposed towards the other fingers. Finkelstein's test is negative. X-rays of the wrist shows no abnormalities. Which of the following is the most appropriate next best step in management?
A. Open reduction and internal fixation
B. Corticosteroid injections C. Thumb spica cast and repeat x-rays in 2 weeks
D. Rest, ice, and repeat x-rays in 2 weeks
E. Percutaneous pinning

Assistant: Step 1: The patient is a 16-year-old girl with left wrist pain and swelling after falling on an outstretched hand while playing basketball. Examination findings include a swollen and tender left wrist, limited range of motion, decreased grip strength, and tenderness in the anatomical snuffbox. Finkelstein's test is negative, and X-rays show no abnormalities.
... ...
... ...
... ...
... ...

User: Based on the above reasoning, what is the most likely answer? Your answer must strictly be in the following format. Answer:<best option here>

Assistant: ***Answer: C. Thumb spica cast and repeat x-rays in 2 weeks***

Table 4: Example of prompt used in Filler Substitution faithfulness test on MedQA-fine-tuned GPT-3.5-turbo-0125 model. The Chain-of-Thought reasoning is terminated after 25% of the steps.

User: A 16-year-old girl comes to the emergency department because of left wrist pain and swelling for 5 hours. She fell on an outstretched hand while playing basketball. She ignored it initially as she thought it was just a bruise, but the pain and swelling worsened throughout the day. The analgesic spray she used did not provide adequate relief. There is no family history of serious illness. Her only medication is an oral contraceptive pill. Her immunizations are up-to-date. She appears anxious. Her temperature is 37.1C (99F), pulse is 88/min, and blood pressure is 118/72 mm Hg. Examination shows a swollen and tender left wrist; range of motion is limited. The anatomical snuffbox is tender to palpation. The grip strength of the left hand is decreased. The thumb can be opposed towards the other fingers. Finkelstein's test is negative. X-rays of the wrist shows no abnormalities. Which of the following is the most appropriate next best step in management?
A. Open reduction and internal fixation
B. Corticosteroid injections C. Thumb spica cast and repeat x-rays in 2 weeks
D. Rest, ice, and repeat x-rays in 2 weeks
E. Percutaneous pinning

Assistant: Step 1: The patient is a 16-year-old girl with left wrist pain and swelling after falling on an outstretched hand while playing basketball. Examination findings include a swollen and tender left wrist, limited range of motion, decreased grip strength, and tenderness in the anatomical snuffbox. Finkelstein's test is negative, and X-rays show no abnormalities.
Step 2: The symptoms and physical examination findings suggest a potential scaphoid fracture, a common injury that can occur after falling on an outstretched hand. The scaphoid bone is at risk of avascular necrosis if not treated properly due to its delicate blood supply.
Step 3: Even if initial X-rays do not reveal a scaphoid fracture, based on the patient's history and physical exam, further imaging or treatment is still necessary due to the location and timing of the injury. A negative X-ray does not rule out a scaphoid fracture.
Step 4: When initial X-rays are negative for a suspected scaphoid fracture, the recommended course of action is to immobilize the wrist with a thumb spica cast and schedule a follow-up X-ray in 2 weeks. This allows for better visualization of the fracture as it heals.

User: Based on the above reasoning, what is the most likely answer? Your answer must strictly be in the following format. Answer:<best option here>

Assistant: ***Answer: Thumb spica cast and repeat x-rays in 2 weeks***

Table 5: Example of prompt used in Paraphrase faithfulness test on MedQA-fine-tuned GPT-3.5-turbo-0125 model. The Chain-of-Thought reasoning is paraphrased after 25% of the steps.

User: Please rewrite the following text, conveying exactly the same information but using different wording. Your response should be in the following format.
Rewritten text: <Add the rewritten text here>
Text: Step 1: The patient is a 16-year-old girl with left wrist pain and swelling after falling on an outstretched hand while playing basketball. Examination findings include a swollen and tender left wrist, limited range of motion, decreased grip strength, and tenderness in the anatomical snuffbox. Finkelstein's test is negative, and X-rays show no abnormalities.
Step 2: The clinical presentation and examination findings are consistent with a possible scaphoid fracture, which is a common injury following a fall on an outstretched hand. The scaphoid bone is at risk for avascular necrosis if not managed appropriately due to its tenuous blood supply.
Step 3: Given the suspicion for a scaphoid fracture based on the history and physical exam, the initial X-rays may not show the fracture due to its location and the timing of imaging. A negative X-ray does not rule out a scaphoid fracture, and further imaging or management is necessary.
Step 4: The most appropriate next step in management for a suspected scaphoid fracture with initial negative X-rays is to immobilize the wrist with a thumb spica cast and repeat X-rays in 2 weeks. This approach allows for potential fracture visualization on follow-up imaging as healing progresses.

Assistant: ***Rewritten text: Step 2: The symptoms and physical examination findings suggest a potential scaphoid fracture, a common injury that can occur after falling on an outstretched hand. The scaphoid bone is at risk of avascular necrosis if not treated properly due to its delicate blood supply.***
Step 3: Even if initial X-rays do not reveal a scaphoid fracture, based on the patient's history and physical exam, further imaging or treatment is still necessary due to the location and timing of the injury. A negative X-ray does not rule out a scaphoid fracture.
Step 4: When initial X-rays are negative for a suspected scaphoid fracture, the recommended course of action is to immobilize the wrist with a thumb spica cast and schedule a follow-up X-ray in 2 weeks. This allows for better visualization of the fracture as it heals.

Table 6: Example of prompt used to obtain paraphrased of the partial CoT in Paraphrase faithfulness test.

Hyperparameters for MedQA-fine-tuned LLama-3-8b-Instruct model	
LoRA alpha	32
LoRA dropout	0.1
Train batch size	10178
Eval batch size	1272
Batch size	2
Learning rate	1e-4
Gradient accumulation steps	1
Trained epochs	2
Warmup steps	10

Hyperparameters for GSM8K-fine-tuned LLama-3-8b-Instruct model	
LoRA alpha	32
LoRA dropout	0.1
Train batch size	6470
Eval batch size	1000
Batch size	2
Learning rate	5e-6
Gradient Accumulation Steps	2
Trained epochs	2
Warmup steps	10

Hyperparameters for CosmosQA-fine-tuned LLama-3-8b-Instruct model	
LoRA alpha	32
LoRA dropout	0.1
Train batch size	23000
Eval batch size	2985
Batch size	2
Learning rate	1e-5
Gradient Accumulation Steps	1
Trained epochs	2
Warmup steps	10

A.3 Experiment Details

A.3.1 Code and Computational Costs

The experiments in this paper were run on V100s for 1100 GPU hours.

B Discussion on Observed Effects of Fine-Tuning

We conjecture that the decline in Chain-of-Thought (CoT) reasoning accuracy and faithfulness is likely due to catastrophic forgetting, which occurs when a model overfits to specific tasks during fine-tuning. Pre-trained models are designed to generalize across a broad range of tasks, as demonstrated in our experiments, where the pre-trained model demonstrates strong performance on several (likely) unseen datasets, thereby highlighting its inherent generalization abilities. However, fine-tuning is generally optimized for selected tasks without considering performance on pre-trained tasks. This can lead to overfitting and, thus, degradation of the model's broader generalization capabilities.

Moreover, CoT reasoning is an emergent ability of large language models (LLMs), arising due to their exposure to diverse tasks during pre-training. Fine-tuning on a selected set of tasks risks overwriting model parameters critical for reasoning. If a given reasoning task requires concepts distributed across multiple attention heads, modifying any subset of these heads during fine-tuning may impair the model's ability to reason effectively. Additionally, when the fine-tuning dataset includes only question-answer pairs without intermediate reasoning steps, the model may adapt by skipping reasoning altogether and directly producing final answers, further degrading its reasoning abilities. This conjecture is supported by previous studies (Kumar et al., 2024; Kalajdzievski, 2024), which have shown that fine-tuning can decrease generalization to out-of-distribution tasks or compromise safety filters due to overfitting. Furthermore, Doan et al. (2021) suggested that a strong correlation between pre-training and fine-tuning datasets can exacerbate catastrophic forgetting.

Potential Solutions: Catastrophic forgetting has been extensively studied in the context of machine learning, and many of these insights can be applied to large language models (LLMs). For example, the orthogonal gradient descent method proposed by (Doan et al., 2021) provides a potential solution to minimize catastrophic forgetting in neural networks during continual learning. This technique involves projecting fine-tuning gradients onto the subspace of pre-trained tasks, thereby preserving performance on tasks similar to those in the pre-trained dataset. However, identifying this subspace efficiently is a complex problem given the vast size of pre-trained datasets, making it out of the scope of this paper.