NAACL 2025

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

Elita Lobo, Chirag Agarwal, Himabindu Lakkaraju, Luis Chiruzzo, Alan Ritter, Lu Wang

University of Massachusetts Amherst , University of Virginia, Harvard University

2025.09.22
HyorinJung

Index

Introduction

Fine tuning

Adapting LLMs to specialized tasks through small domain-specific datasets to improve performance in unseen domains

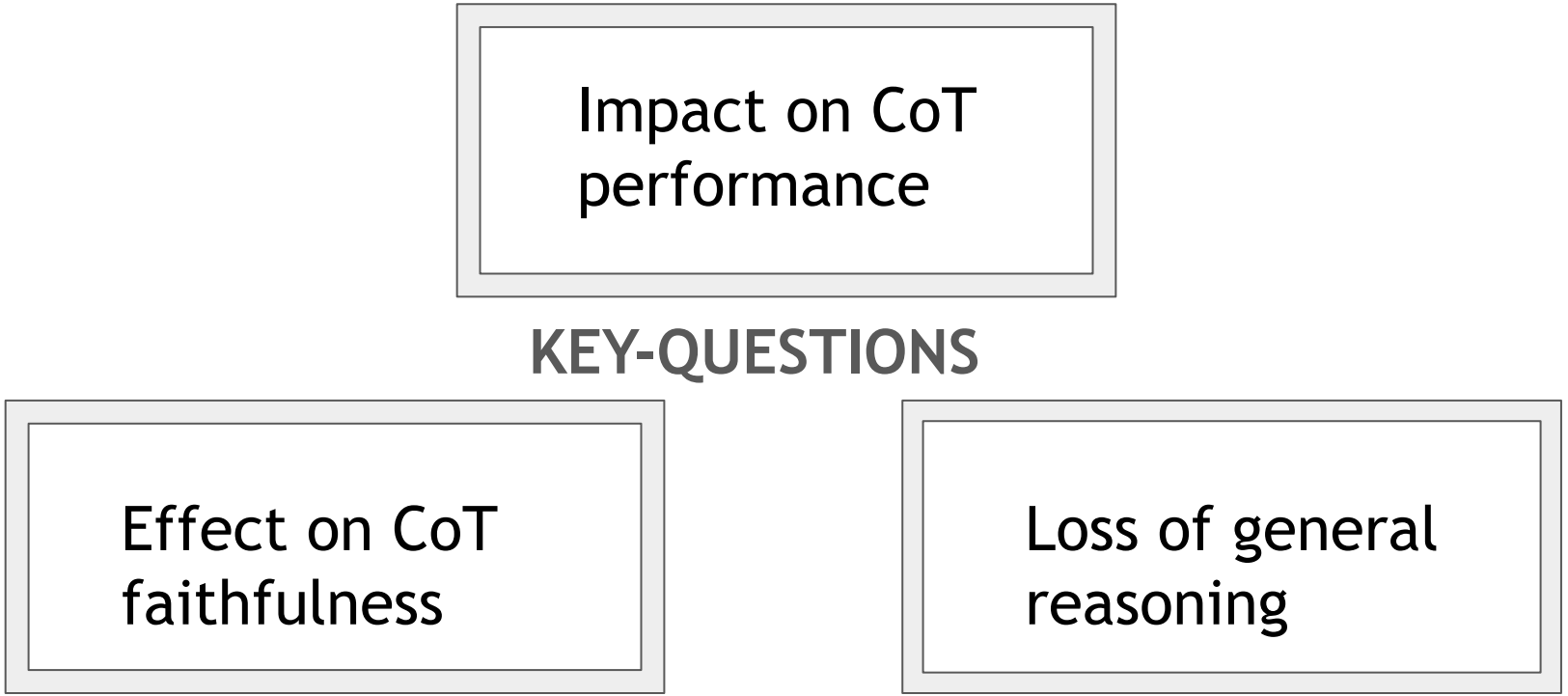
However, fine-tuning also brings several challenges and risks

- **Catastrophic Forgetting**: Performance loss on other tasks
- **Safety Risks**: Deactivation of toxicity/safety filters
- **Privacy Risks**: Higher chance of data leakage

👉 Little research on impact to reasoning abilities

Introduction

Chain-of-Thought : generate step-by-step reasoning paths



Approach & Benefits

- **Techniques:** LoRA, QLoRA → improve efficiency via low-rank gradient decomposition
- **This Work:** Combines QLoRA + supervised fine-tuning (SFT)
- **Benefits:** Enhances task-specific performance

Background

LoRA

- Fine-tuning method that updates only a small subset of weight parameters.
- The majority of pre-trained weights are **kept frozen**.
- Efficient: reduces computation and memory cost compared to full fine-tuning.

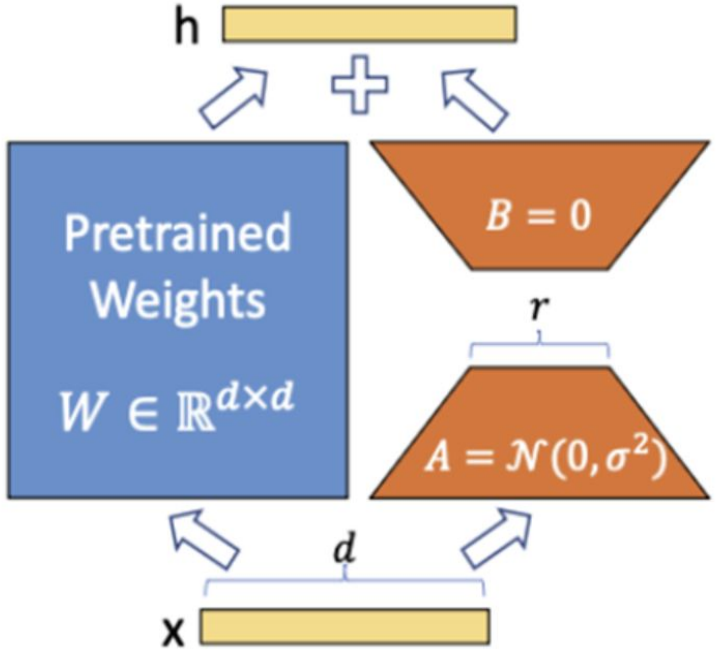


Figure 1: Our reparametrization. We only train A and B.

Background

QLoRA = LoRA + 4-bit quantization

Advantages:

- Enables fine-tuning of large LLMs with limited GPU resources
- Improves training speed and memory efficiency

Background

supervised fine-tuning method (SFT)

- **Target:** Pretrained large language models (LLMs)
- **Method:** Further training (fine-tuning) using supervised data
- **Goal:** Preserve pretrained knowledge while adapting to specific tasks

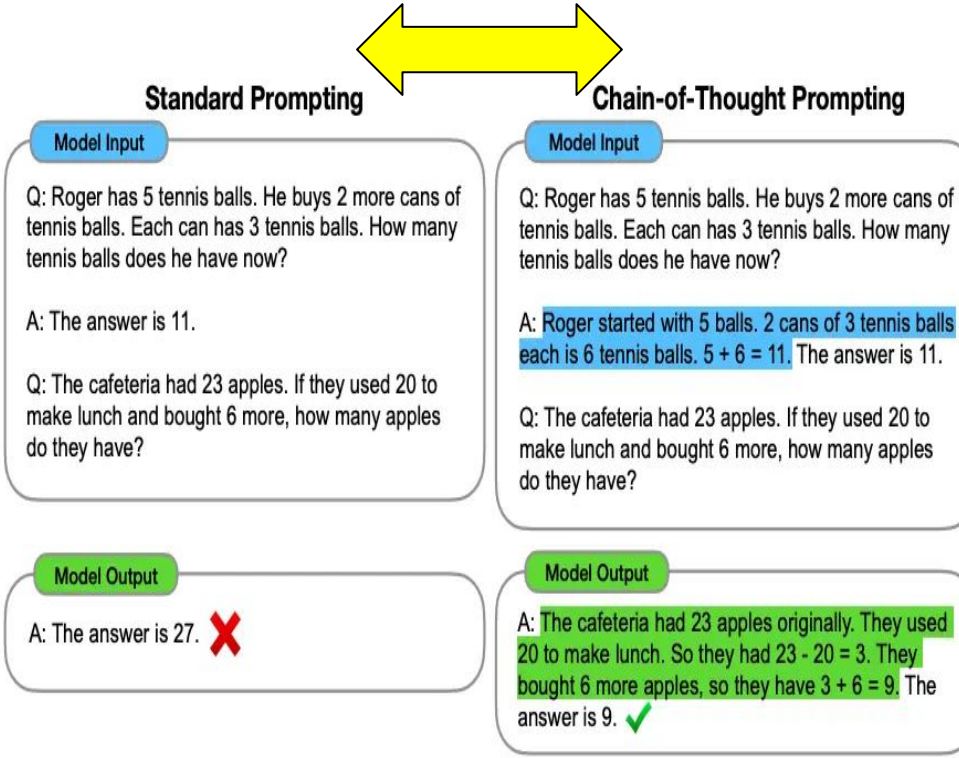
Methods

Experiment 1

- Fine-tune on reasoning & non-reasoning QA
- No intermediate reasoning steps (except GSM8K)
- Keep performance consistent

Experiment 2

- Compare task performance before vs. after fine-tuning



Faithfulness of CoT Reasoning

Faithfulness Tests:

1. **Early Termination**
    - Truncate the reasoning chain step by step.
    - Measure the fraction of steps that truly influence the final answer.
2. **Paraphrasing**
    - Reword last i steps of the reasoning.
    - If the answer stays the same, the reasoning is driven by logic, not phrasing.
3. **Filler Substitution**
    - Replace steps after each point with placeholders like “(...)”.
    - If the answer doesn’t change, later steps are not faithful.

Experiments : Setting

Datasets

- **MedQA:** USMLE multiple-choice questions(Medical)
- **MedMCQA:** AIIMS & NEET multiple-choice questions(Medical)
- **CosmosQA:** Commonsense reading comprehension(Commonsense reading)
- **GSM8K:** Math word problems(Math)
- **IID vs OOD:** Test dataset is *in-distribution* if same category as fine-tuning dataset, else *out-of-distribution*

Experiments : Setting

Models

- 4-bit quantized **Llama-3-8b-Instruct** (QLoRA, rank 16)
- **GPT-3.5-0125**
- **GPT-4**

Implementation Details

- Vanilla fine-tuning (QA pairs only), except GSM8K uses reasoning steps
- Datasets split into train/validation/test, models trained for 2 epochs
- GPT models: OpenAI API; Llama: TRL library with QLoRA
- CoT prompt templates used for reasoning evaluation

Experiments : Setting

Faithfulness Evaluation (CoT Pred Match)

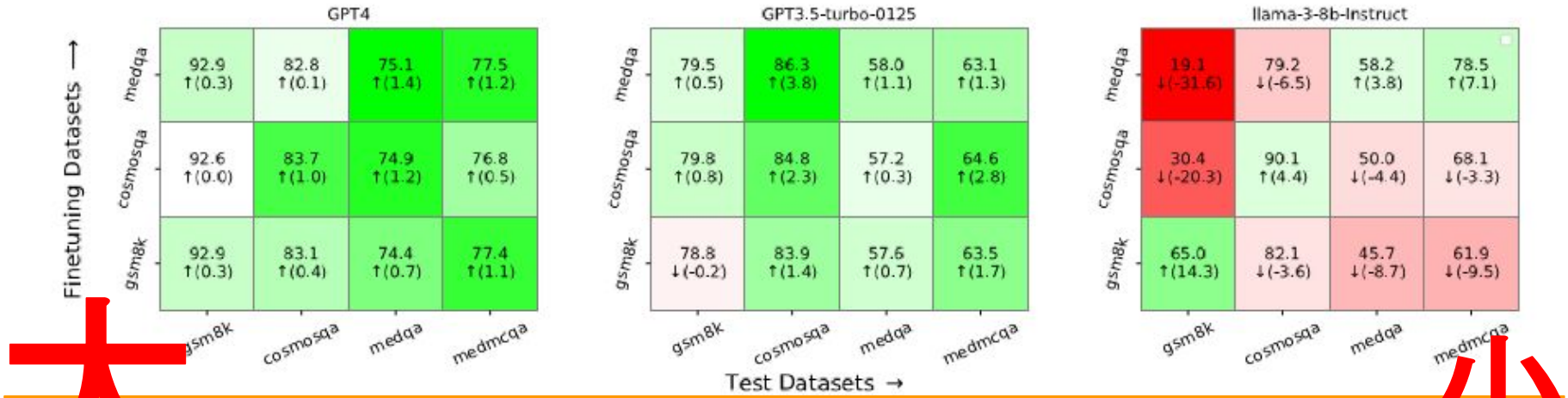
- Fraction of data points where perturbed CoT gives same answer as original
- **Early Termination & Filler Substitution:** Higher values → more unfaithful reasoning
- **Paraphrasing:** Checks if reasoning is robust to wording changes

  > Higher values → more faithful reasoning

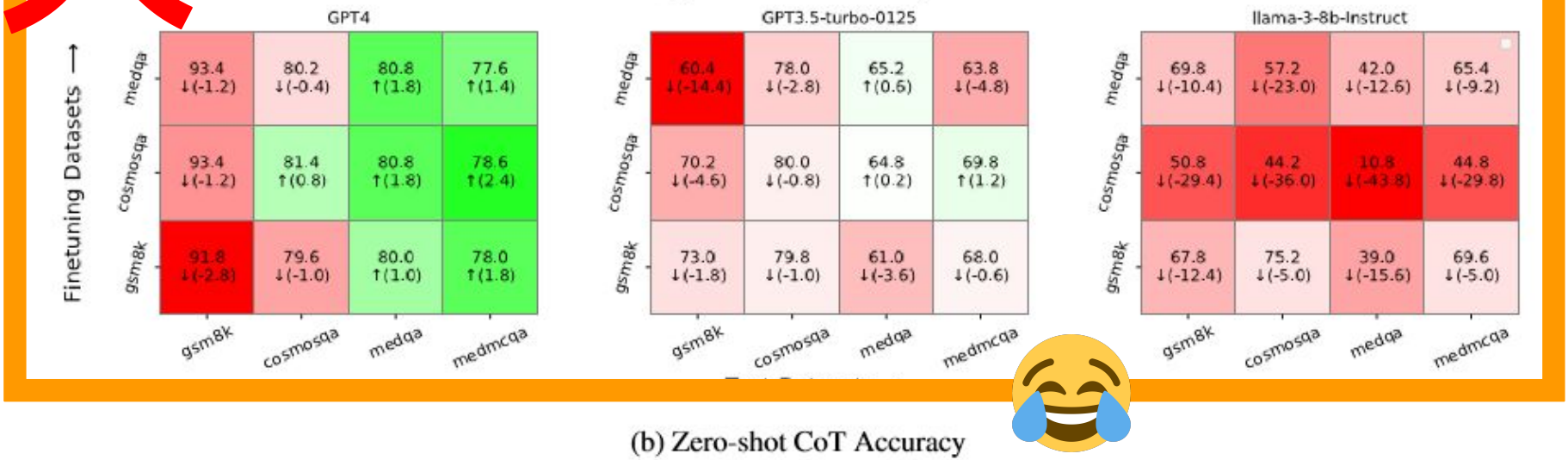
Experiments : Results

RQ1: How does fine-tuning affect CoT reasoning?

- Fine-tuned LLaMA-3-8b-Instruct and GPT models show **lower CoT accuracy on GSM8K (math reasoning)**.
- Accuracy drop is **more pronounced in smaller models (LLaMA)** than larger models (GPT-4).
- Fine-tuned LLaMA often fails to produce **high-quality CoT reasoning**.



(a) Zero-shot Accuracy



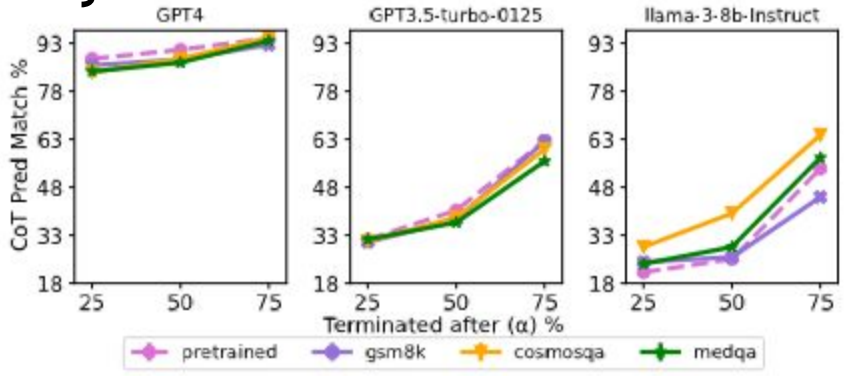
(b) Zero-shot CoT Accuracy

Experiments : Results

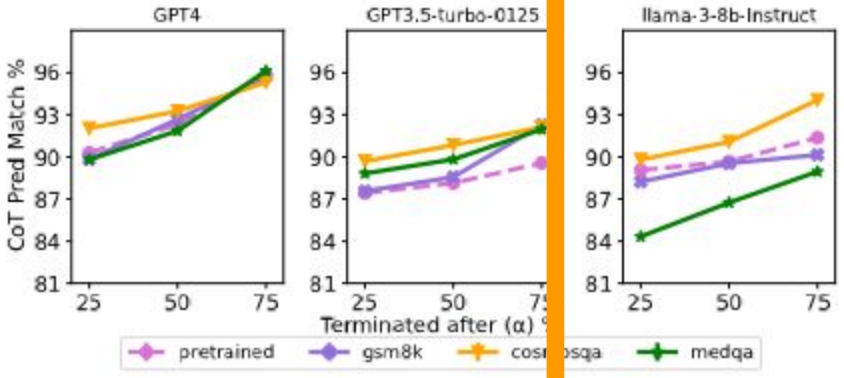
RQ2: How does fine-tuning affect CoT faithfulness?

- **Early Termination, Filler Substitution, Paraphrasing** tests conducted.
- GSM8K CoT reasoning generally **more faithful** than CosmosQA/MedQA/MedMCQA.
- GPT-4: **stable** CoT Pred Match → many reasoning steps have **little influence** → some steps unfaithful
- LLaMA-3-8b-Instruct fine-tuned on CosmosQA/MedQA → **increased CoT Pred Match** → reduced faithfulness.

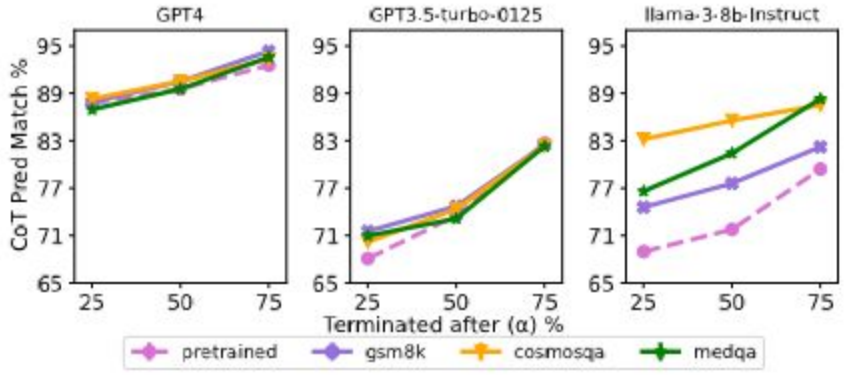
Early Termination



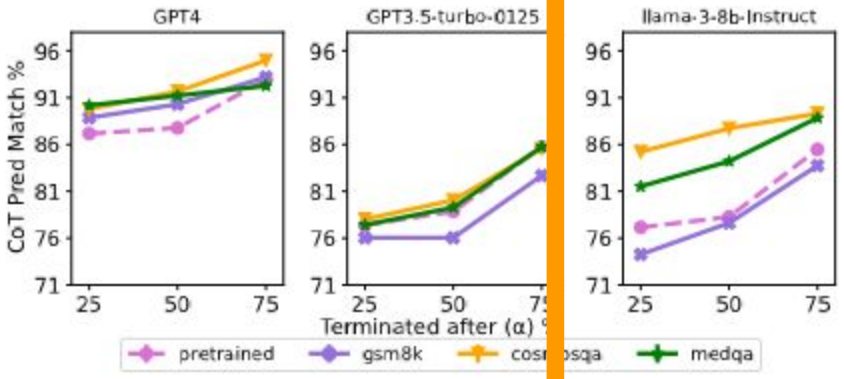
(a) GSM8K dataset



(b) CosmosQA dataset



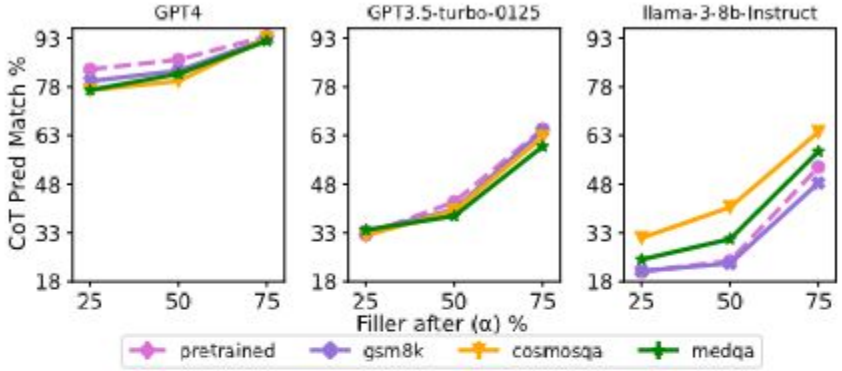
(c) MedQA dataset



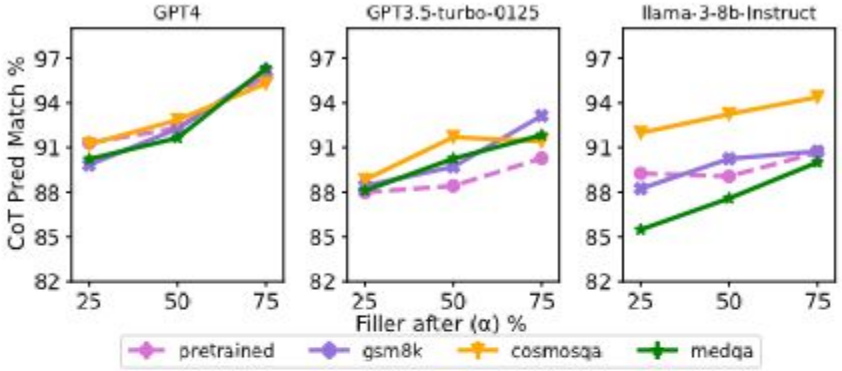
(d) MedMCQA dataset

If the CoT Pred Match value is high → a larger number of data points likely have unfaithful CoT reasoning

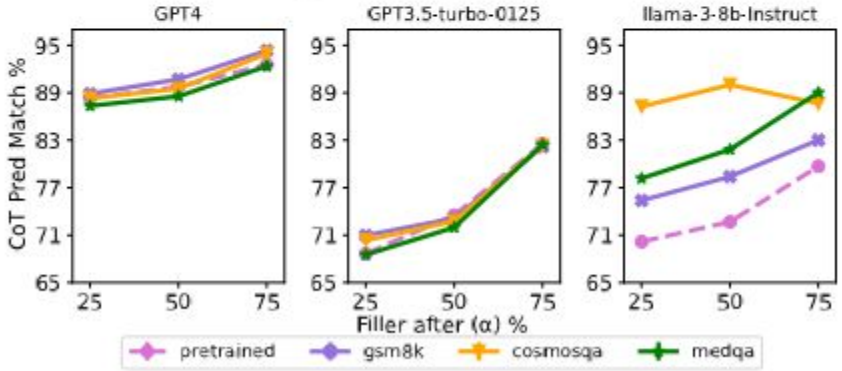
Filler Substitution



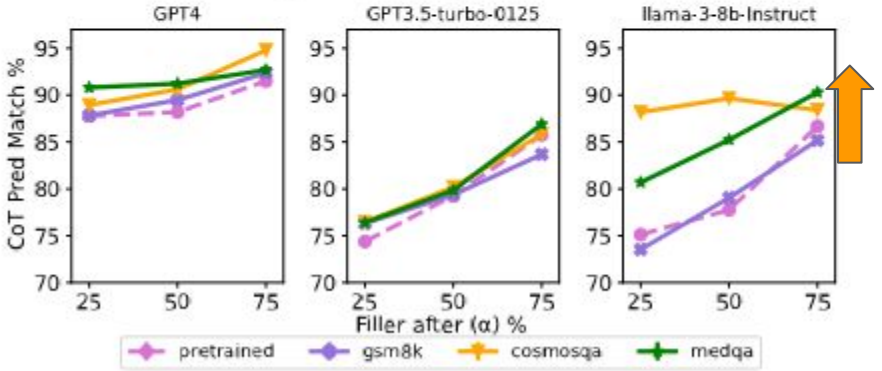
(a) GSM8K dataset



(b) CosmosQA dataset



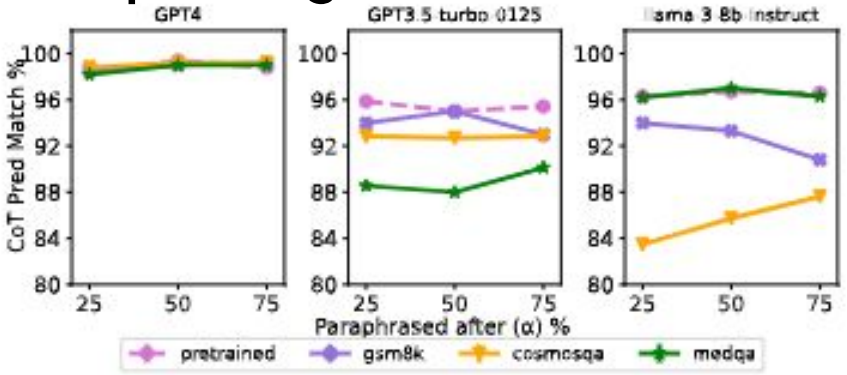
(c) MedQA dataset



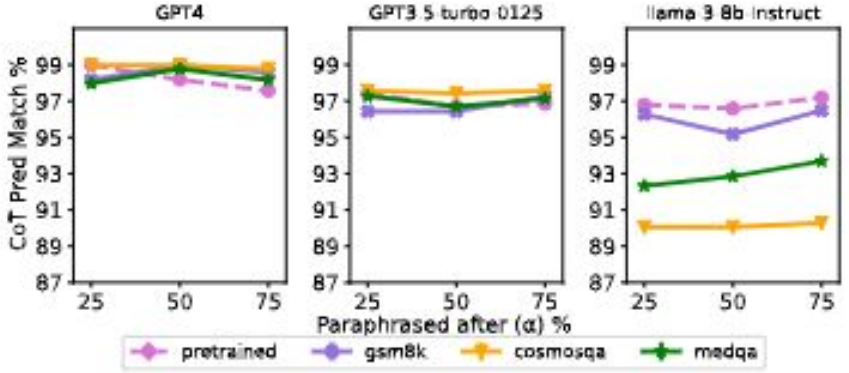
(d) MedMCQA dataset

If part of the reasoning is replaced with fillers, the final answer remains the same → some reasoning steps have no influence on the final answer → low faithfulness

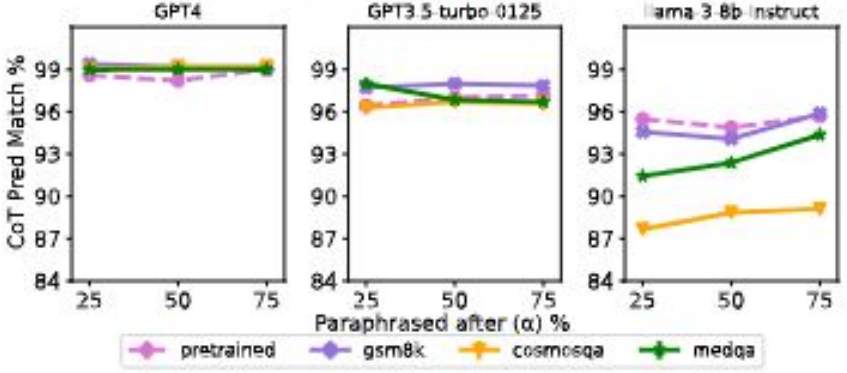
Paraphrasing



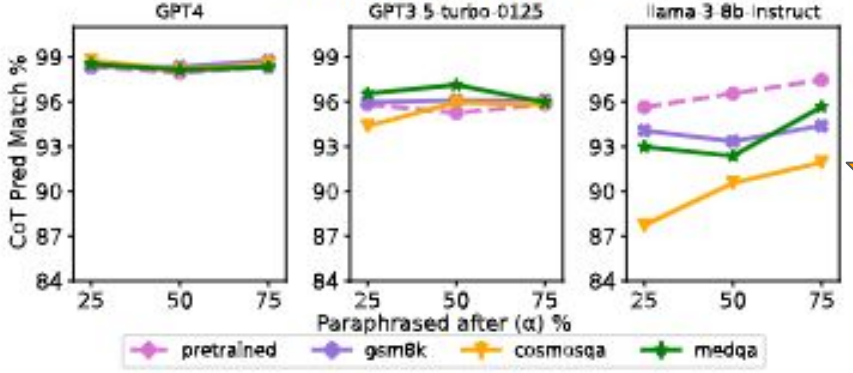
(a) GSM8K dataset



(b) CosmosQA dataset



(c) MedQA dataset

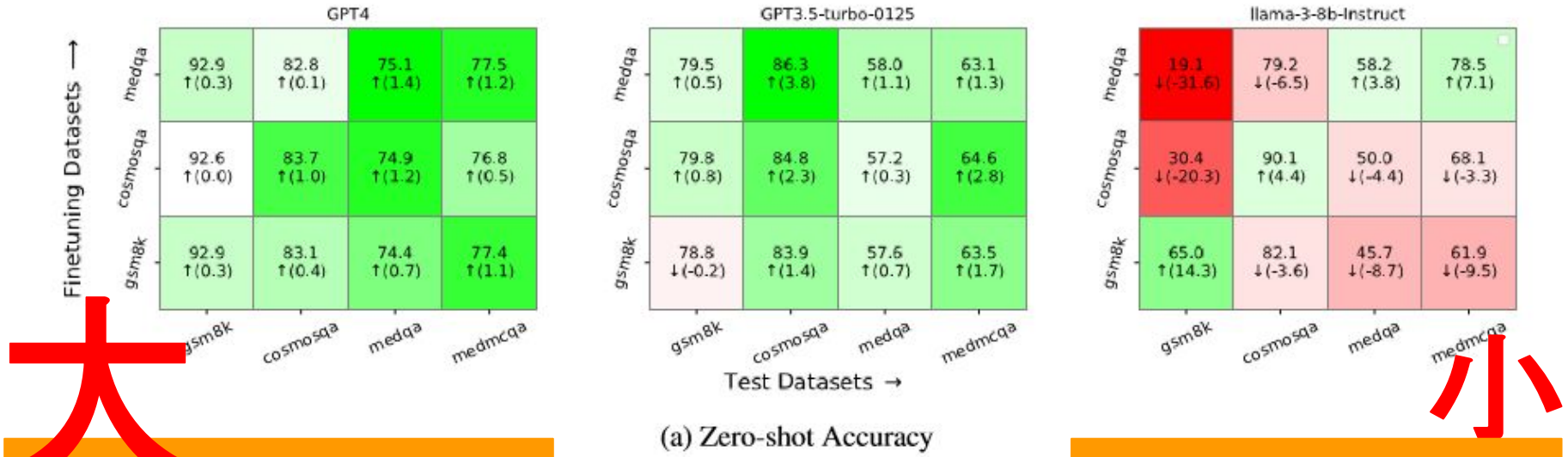


(d) MedMCQA dataset

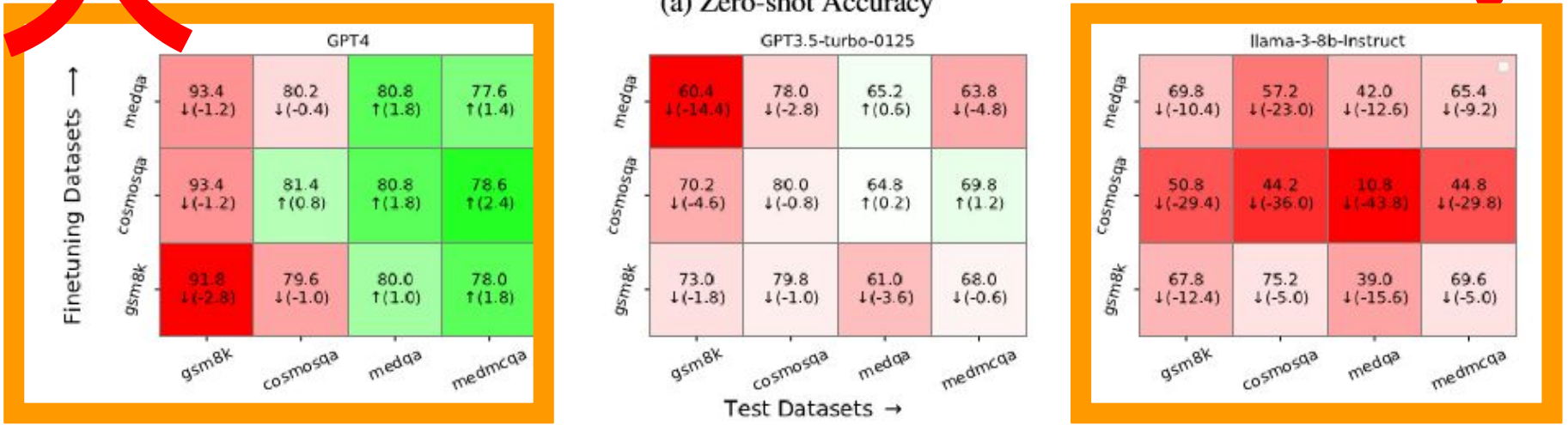
Experiments : Results

RQ3: How does fine-tuning affect general performance (IID/OOD)?

- GPT models: accuracy generally **improves** across datasets.
- LLaMA: fine-tuning on CosmosQA/MedQA → **performance drop on GSM8K**, smaller models are more sensitive to reduced generalization



(a) Zero-shot Accuracy



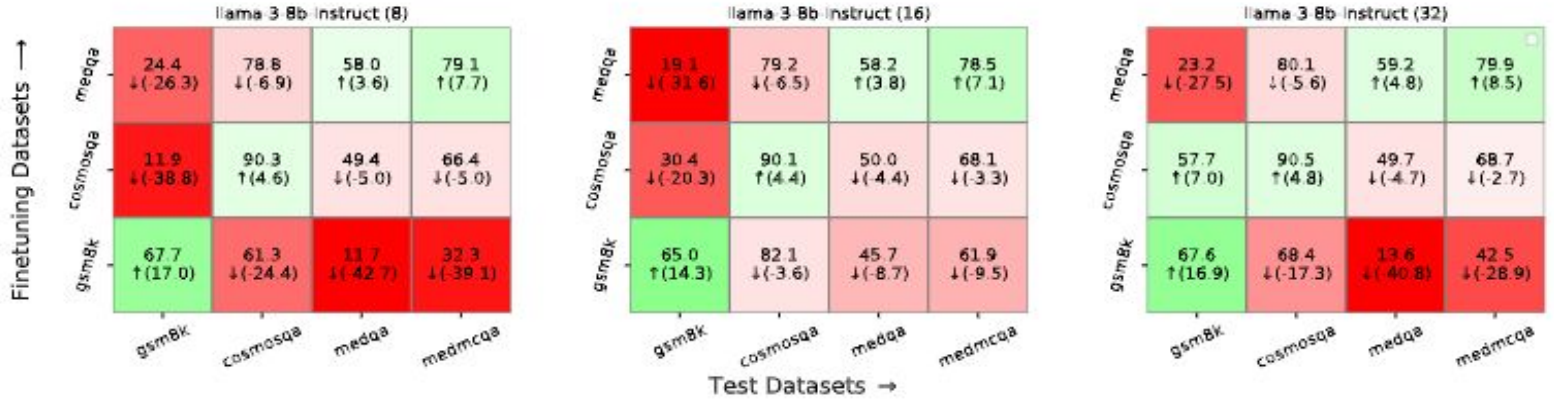
(b) Zero-shot CoT Accuracy

Experiments : Results

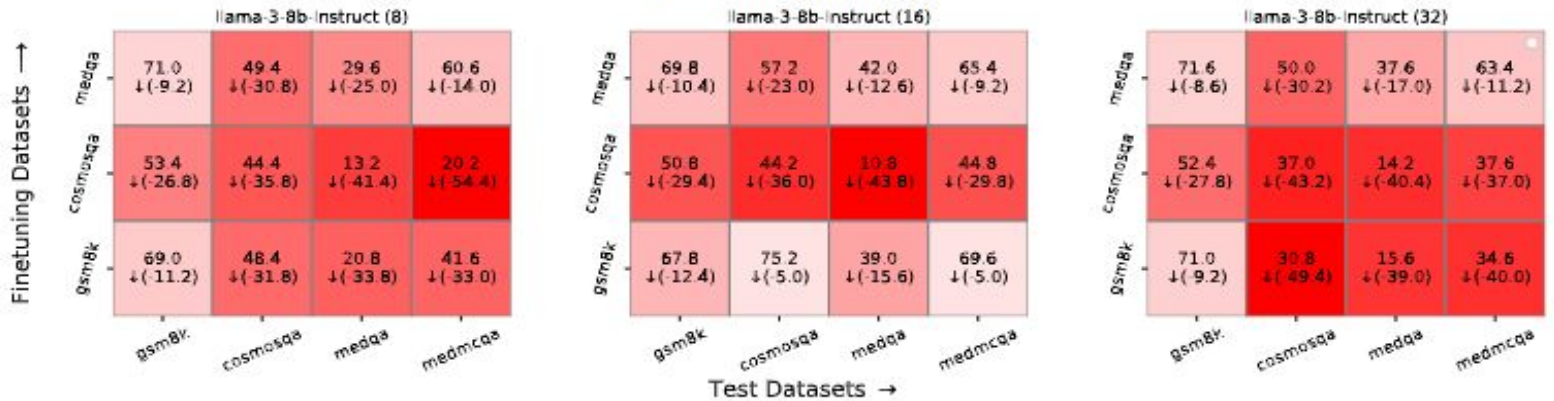
RQ4: Effect of QLoRA rank on fine-tuning

- LLaMA models fine-tuned with rank 8,16,32.
- Higher rank → more trainable parameters → **better generalization**, but CoT accuracy still drops due to overfitting.
- Fine-tuning on CosmosQA/MedQA → **reduced CoT faithfulness** across all ranks.

QLoRA rank on fine-tuning



(a) Zero-shot Accuracy



(b) Zero-shot CoT Accuracy

Increasing the rank -> not fully prevent CoT reasoning degradation

Experiments : Results

Model	Size	CoT Accuracy (Zero-shot)	CoT Faithfulness	Fine-tuning Impact
GPT-4	L	High	Moderate (Some steps unfaithful)	Low
GPT-3.5	M	Moderate	Moderate	Slight
LLaMA-3-8B-Instruct	S	Low	Lower	High

Limitation & Future Directions

- CoT faithfulness **metrics** are still **early-stage**; findings may need revisiting.
- Effects of **newer CoT prompting methods** and **in-context demonstrations** not explored.
- Internal mechanisms of LLMs under fine-tuning not deeply analyzed
- Broader evaluation across **more reasoning/non-reasoning tasks and models** needed.

Conclusion

skepticism about LLMs' "reasoning" abilities

- Fine-tuning **smaller LLMs** on non-reasoning or commonsense datasets **reduces accuracy on complex tasks**, especially math.
- Fine-tuning can also lead to a **larger decline in CoT reasoning performance**.
- CoT **faithfulness** may be compromised after fine-tuning.
- Highlights the need for **methods that preserve reasoning capabilities** when fine-tuning LLMs for specialized tasks

Proposal

- Enhancing CoT in Smaller LLMs
  - Since smaller models suffer the most degradation, explore synthetic CoT data augmentation, self-consistency decoding, or contrastive learning to improve faithfulness.
    - **“Child-friendly CoT-preserving fine-tuning”**
- Domain-Specific Analysis of CoT Degradation
  - Investigate why fine-tuning impacts math-heavy reasoning tasks more severely than commonsense or language-based reasoning tasks.
- Internal mechanisms of LLMs under fine-tuning
  - using ITI could provide insights

Appendix : Techniques for Eliciting CoT Reasoning in LLMs

1. **Zero-Shot CoT Prompting** : *"Let's think step by step"*
2. **In-Context Example Enhancement**: *Providing examples*
3. **Self-Consistency** : Generate multiple reasoning paths
4. **Tree-of-Thought (ToT) Reasoning** : Each step branches into multiple possibilities
5. **Least-to-Most Prompting** : Break complex problems into sub-problems
6. **Deductive Reasoning Framework** : Verify each intermediate step before final answer