Advancing Math Reasoning Capabilities in Small Language Models

Georgia Institute of Technology

Sai Paresh Karyekar, Sharika Menon Rajeev, Alyan Khan

Georgia Institute of Technology, School of Electrical and Computer Engineering

Georgia Tech College of Engineering
School of Electrical and Computer Engineering

Motivation

LLMs excel in various NLP tasks but struggle with **mathematical reasoning** and are very resource intensive. Small Language Models offer a lightweight alternative but also perform poorly on reasoning tasks.

This project explores methods to enhance reasoning capabilities in SLMs, specifically focusing on the T5-small Transformer model.

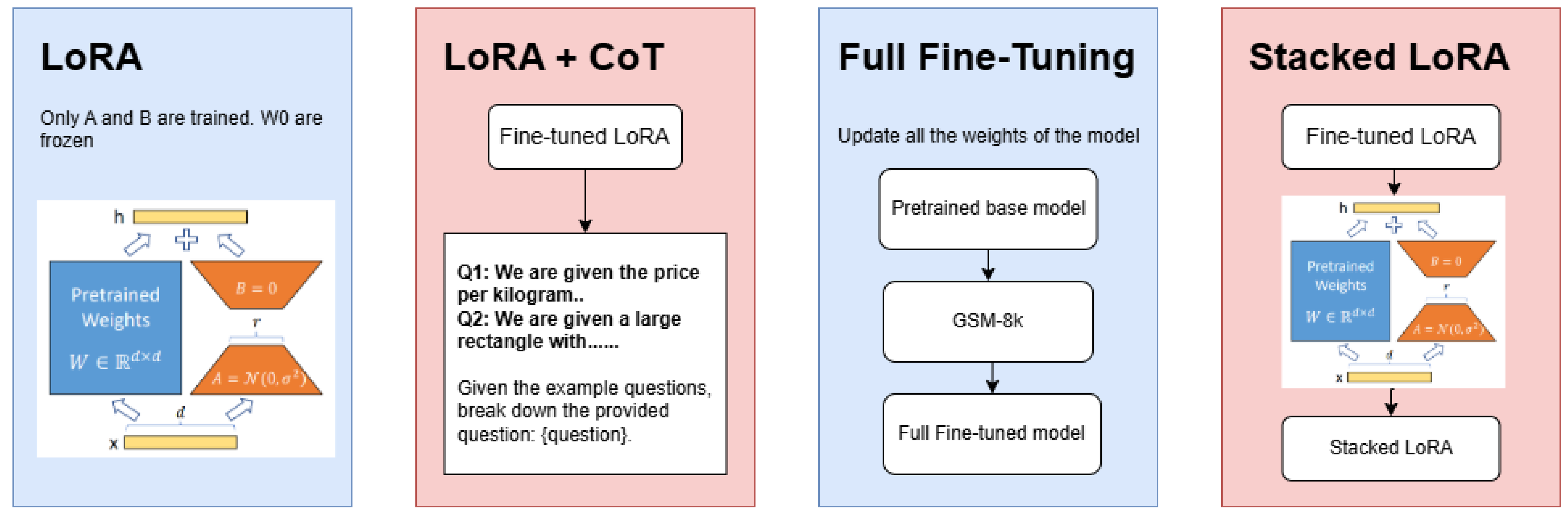


Figure 1: Overview of Fine-Tuning strategies

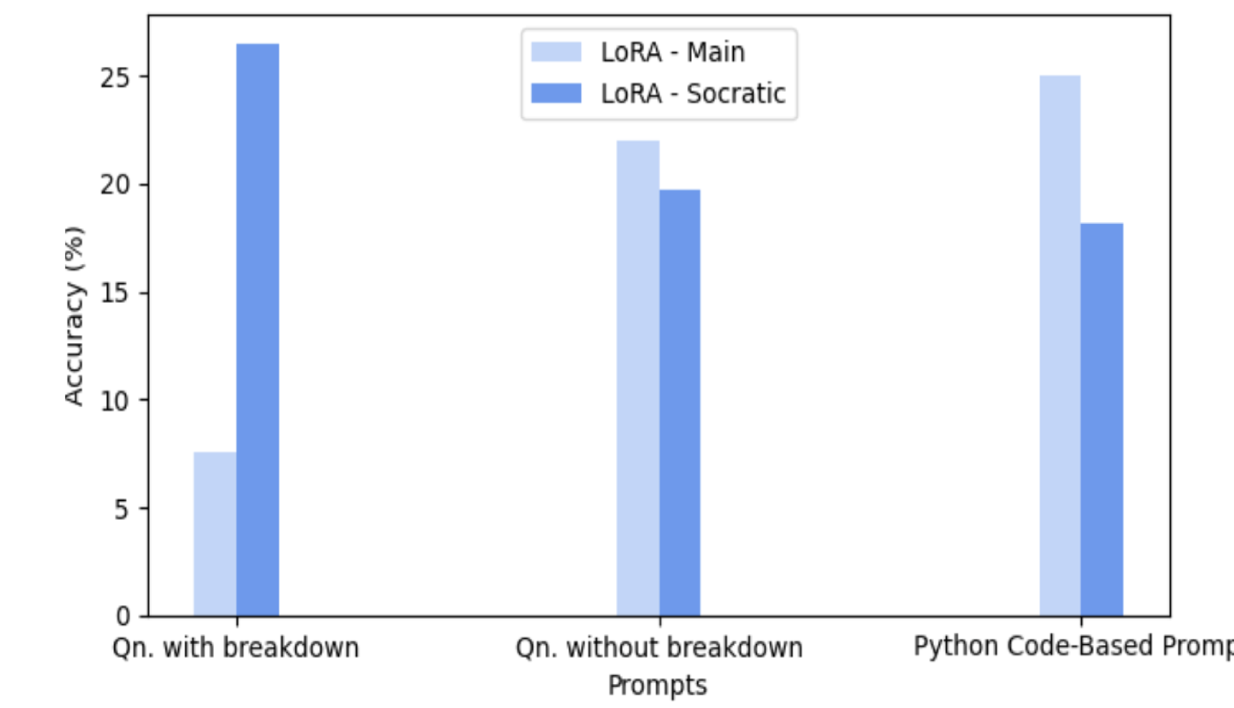


Figure 3: Chain-Of-Thought for LoRA

Methodology

Dataset: GSM8k (Grade School Math 8K) a dataset for mathematical reasoning tasks consisting of 8000 problems.

Base Model: T5 Small Transformer architecture with 60M parameters. Chosen for its efficiency and lightweight architecture.

Fine-Tuning Approaches:

- **Low-Rank Adaptation (LoRA)**
  - Re-train only low-rank matrices A and B, keeping pretrained model weights frozen, reducing memory requirements and computation costs. (Hu et al., 2021)
- **LoRA with Chain-of-Thought (CoT) Prompting**
  - The model is guided to reason through problems step-by-step using structured prompts. Fine-tuned LoRA model is used to generate intermediate reasoning steps.
- **Full Fine-Tuning**
  - All model parameters are updated during training resulting in a fully adapted version. Highest task performance but computationally expensive.
- **Quantized LoRA**
  - An efficient variant of LoRA applied to T5-Base (220M parameters) using 8-bit quantization.

Stacked LoRA (Our approach)

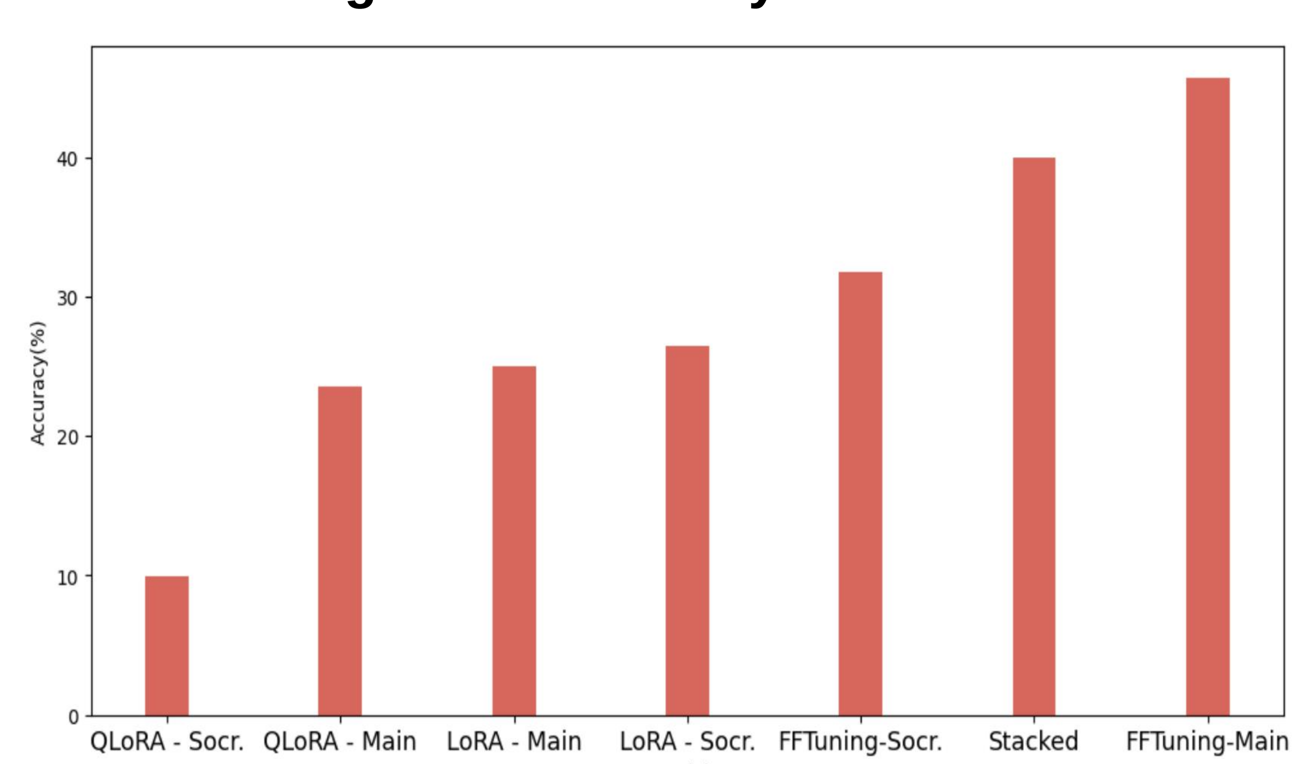
Stacked LoRA introduces additional low-rank layers to fine-tune the model further on a new dataset or task allowing knowledge transfer while training prior task-specific adaptations.

- We apply stacked LoRA using a coding dataset to generalize to reasoning tasks.
- Weight update rule:

$$\boxed{W = W_0 + A_1 B_1 + f(A_2 B_2)}$$

Where W_0 are frozen pre-trained weights, A_1B_1 are LoRA weights trained on the first dataset, and A_2B_2 are newly introduced LoRA weights fine-tuned on the second dataset. $f(.)$ (e.g., ReLU)

Figure 2: Accuracy vs. Models



Results

- We assessed model performance using accuracy, perplexity, ROUGE-1 score and average semantic similarity to evaluate correctness, predictive quality, textual overlap and semantic alignment respectively.
- **Stacked LoRA** achieves an **accuracy** of 40%, demonstrating near-equivalent performance to full-fine tuning, with significantly reduced computational requirements.
- Full Fine-Tuning achieves the highest accuracy at 45% but compromises semantic similarity (62.81%).
- An alternative approach would be using larger quantized models like T5-base (220M parameterss) which has high semantic similarity while being computationally efficient through 8-bit quantization.

Table 1: Performance Evaluation

Model	ROUGE-1	Perplexity	Average Semantic Similarity
LORA - Main	0.5084	3.2633	79.70%
QLORA - Main	0.5156	2.4158	79.99%
LORA - Socratic	0.4678	3.2706	80.42%
QLORA - Socratic	0.5262	2.3754	75.96%
Full-Fine Tuning - Main	0.4136	2.2708	62.81%
Full-Fine Tuning - Socratic	0.4778	2.2353	62.50%
LoRA Stacked	0.3546	2.2587	61.38%

- This suggests that quantizing larger models may be more effective than fine-tuning smaller ones for mathematical reasoning tasks. (Wei et al., 2023)

We also evaluated the performance of three Chain-of-Thought(CoT) prompting strategies on LoRA with GSM8k main and Socratic datasets and observed their accuracy.

- **Question with breakdown**
- **Question without breakdown**
- **Python code-based prompt**

Python code-based prompt approach gave the highest accuracy, particularly effective on LoRA main (25%) ,but slightly lower performance on Socratic (18.2%).

Conclusion

- Stacked LoRA enhances T5-small's efficiency and reduces resource use but falls short of full fine-tuning in complex reasoning tasks. Further research is needed to improve its performance on challenging tasks.

References

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia,F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.