Enhancing Reasoning Alignment in LLMs via GRPO Fine-Tuning: A Comparative Evaluation

Haiwei Du, Linwei Wu, Yiran Wang

Objective

Our project aims to evaluate the effectiveness of **Group Relative Policy Optimization (GRPO)** in fine-tuning **Large Language Models (LLMs)** for structured reasoning alignment. Specifically, we analyze how it improves **answer accuracy** and adherence to **reasoning formats** across several benchmarks. To enable efficient training, we adopt the **Unsloth** fine-tuning framework. Furthermore, we investigate how variations in **fine-tuning datasets**, **prompt designs**, and **reward function formulations** influence the model's ability to generate coherent step-by-step reasoning and generate correct answers.

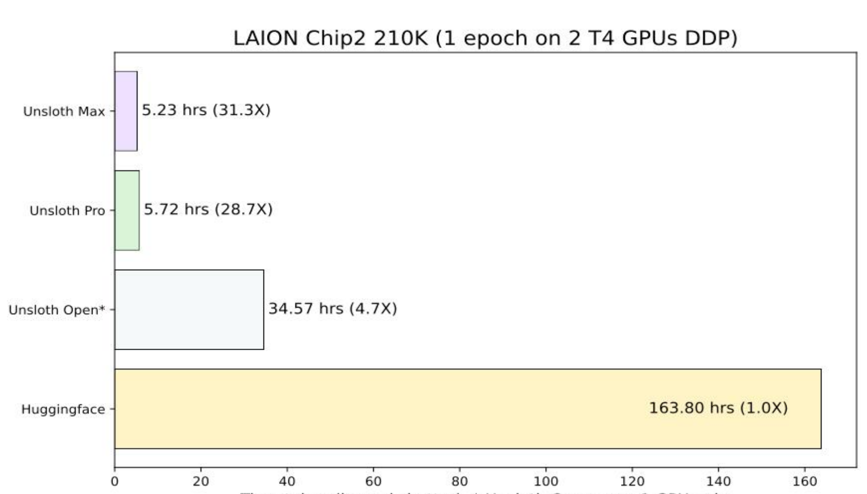
Overall Framework

Basic Components

1.Qwen2.5-3B: It offers a balanced trade-off between performance and efficiency, making it a suitable foundation for exploring alignment and reasoning strategies.

2.LoRA (Low-Rank Adaptation): LoRA is used to efficiently fine-tune Qwen2.5-3B by reducing memory and compute needs, enabling GRPO training with limited resources.

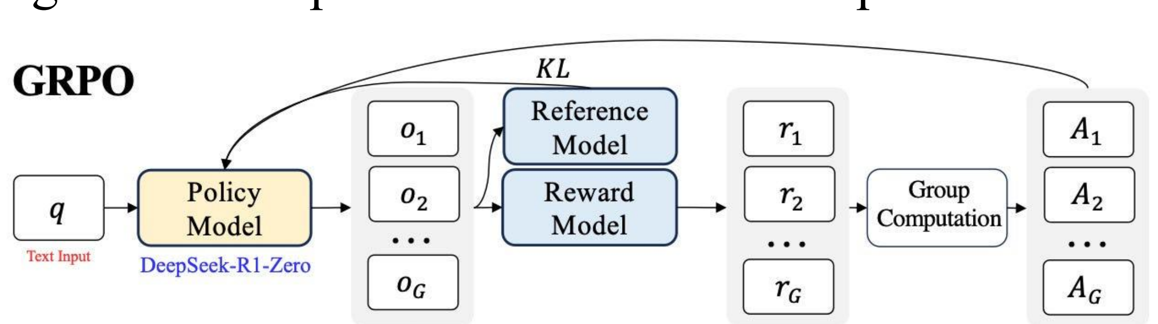
3.Unsloth: Unsloth is an open-sourced framework library designed for efficient, low memory, and high-speed fine-tuning of LLMs.



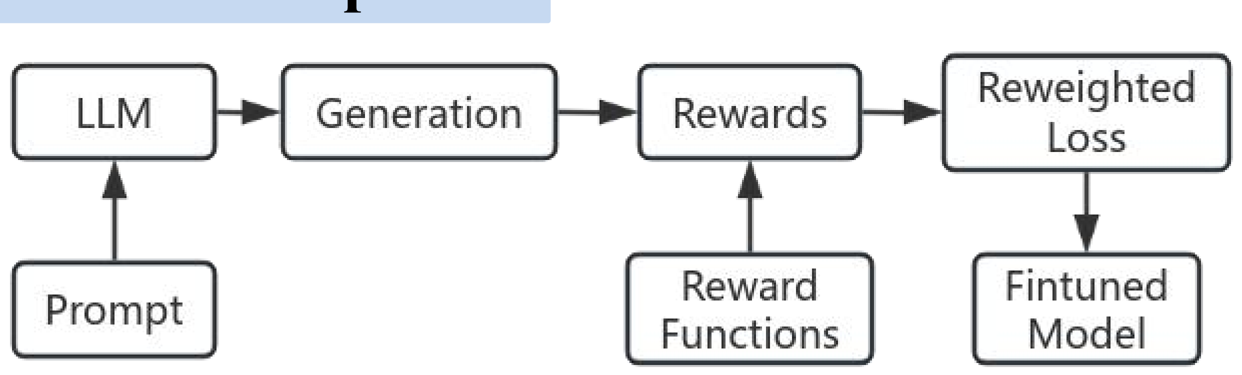
4.FastLanguageModel: It's Unsloth's core module for generation, integrating LoRA and quantization into causal LLMs like LLaMA and Qwen, and streamlining both training and deployment with high speed and less computing resources.

Fine-tuning Method

GRPO (Group Relative Policy Optimization) :
It's a reward-guided fine-tuned algorithm designed to align large language model outputs with human-defined preferences.



Reinforcement process



Control Dimensions

1. Fine-tuning Datasets

We fine-tuned the base model separately on the **GSM8K (Grade School Math 8K)** and **SQuAD (Stanford Question Answering Dataset)** datasets, and the results are as follows:

Dataset	Base Qwen 2.5	GRPO+Qwen2.5	Improvement
GSM8K	**69.52%**	**71.11%**	**+2.29%**
SQuAD	**14.1%**	**13.2%**	**-6.38%**

(a)**Baseline Performance**: It achieves notably higher accuracy on GSM8K than on SQuAD, indicating that Qwen is more naturally suited for **text generation and reasoning tasks** rather than **extractive question answering**.

(b)**GRPO fine-tuning Performance**: Fine-tuning on GSM8K leads to an improvement in accuracy, indicating the effectiveness of GRPO. Fine-tuning on SQuAD results in a performance decline, which suggests that the method may face challenges in adapting to extractive Q-A tasks.

2. Reward Functions

(1) **Improvement** on the original functions:
We enhance the original reward function from the Unsloth implementation by extending the reasoning evaluation from a **single line** to **multiple lines**, aiming to improve training robustness and better capture multi-step reasoning quality.

(2) **Comparison** with GSM8K's functions:

Functions	SQuAD	GSM8K
Answer Structure	**Lists of strings**	**Unique numbers**
correctness_ reward_func	**F1 score + exact match + empty answer penalty**	**Strictly match whether the answer is equal to ground truth**

In addition, for SQuAD, additional reward functions are used:
<u>Penalty mechanism</u>: Deduct points for generated empty words.
<u>Combination rewards mechanism</u>: Combined use of correctness, formatting, and penalty mechanisms.

3. Fine-tuned System Prompt

(1) **Basic System Prompt (Partial)**:
"<reasoning>...</reasoning> <answer>...</answer>"
Using this system prompt enhances evaluation robustness by promoting more consistent model behavior during testing.

(2) **More Detailed System Prompt**:
For SQuAD, the system prompt can add **"You are a helpful assistant answering questions based on the given context."**.

Dataset	Basic Prompt	Detailed Prompt	Improvement
SQuAD	**11.9%**	**13.2%**	**+10.92%**

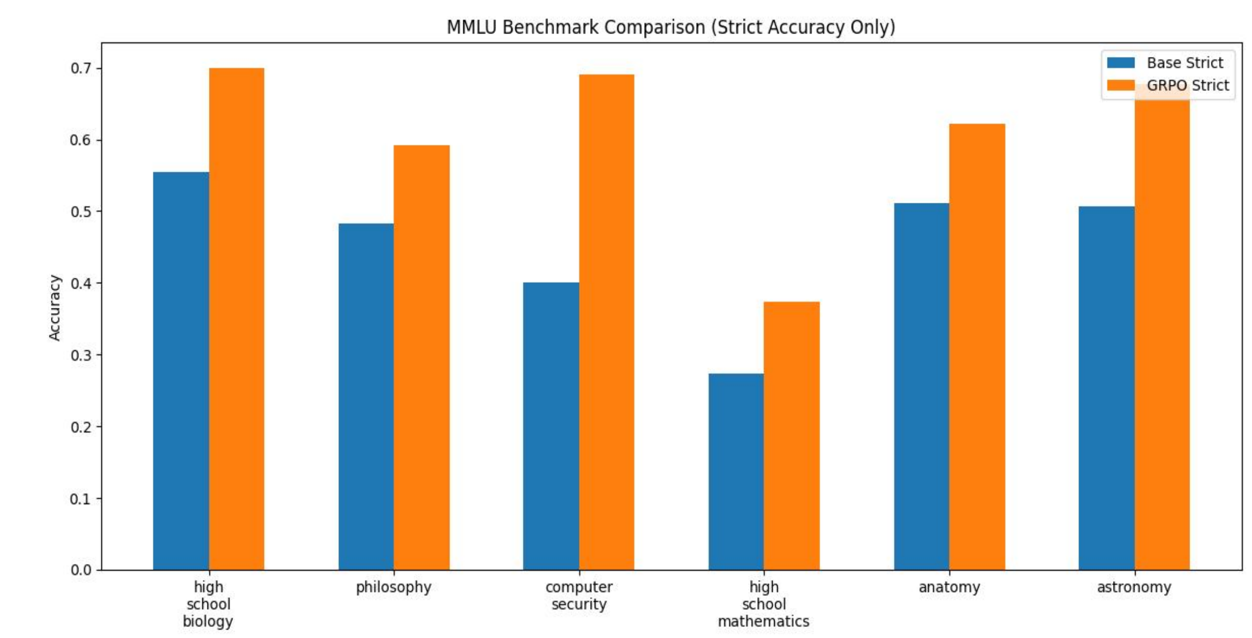
References

Unsloth Documentation <u>https://docs.unsloth.ai/</u>
DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models <u>https://arxiv.org/abs/2402.03300</u>
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning <u>https://arxiv.org/pdf/2501.12948</u>

Benchmark

We use the **Qwen2.5-3B** model to evaluate the performance of the GRPO fine-tuning strategy on the following benchmarks:

Dataset	Base without reasoning Accuracy	Base with reasoning Accuracy	GRPO with reasoning Accuracy
MMLU	**46.24%**	**28.68%**	**41.07%**
SWAMP	**50.00%**	**77.67%**	**83.67%**



On the **MMLU (Massive Multitask Language Understanding)** dataset, incorporating reasoning prompts degraded the performance of the base model, and GRPO fine-tuning did not surpass the base model without reasoning overall. However, GRPO still showed clear gains in reasoning-heavy subjects like math, biology, and philosophy. In contrast, the **SVAMP (Simple Variations on Arithmetic Math word Problems)** dataset—featuring open-ended numerical answers—better aligned with the CoT-style generation trained via GRPO. In this format, reasoning prompts substantially improved both base and GRPO-enhanced models, with the GRPO model achieving **83.67%** accuracy, a **7.72%** improvement over the reasoning base and a **47.34%** gain over the no-reasoning base, surpassing the MMOS-CODE-34B benchmark. These results highlight that while CoT-style prompting may cause overthinking and option confusion in multiple-choice tasks like MMLU, it greatly benefits structured generation tasks like SVAMP where step-by-step reasoning leads directly to the final answer.

Future Prospects

In future work, we will extend GRPO fine-tuning to larger models like **Qwen2.5-7B** and evaluate its generalizability on multilingual benchmarks. We plan to incorporate human preference data into the **reward function** for better alignment with human reasoning. Additionally, we will experiment with more expressive reasoning templates and test the model's robustness on out-of-distribution questions. We also aim to investigate the **output collapse issue** and quantify the impact of system prompts on fine-tuning stability. Another promising direction is to explore GRPO fine-tuning on more **diverse reasoning-intensive datasets** to further validate and enhance its generalization across complex cognitive tasks.