No Probllama: Fine-Tuning Llama Using GRPO

Brandt Buchda, Devi Mahajan, Peyton Nash, Kyler Rosen

3/12/2025

Agenda

- Algorithm
- Environment and Agent
- Training
- Results

Agenda

- **Algorithm**
- Environment and Agent
- Training
- Results

DeepSeek

- DeepSeek released its R1 model in January 2025
  - Produced results comparable to OpenAI and Google Gemini on reasoning tasks
  - Introduced innovations in reinforcement learning
- R1 builds on Proximal Policy Optimization with Group Relative Policy Optimization (GRPO)
  - Reduces memory and compute usage
  - Produces more stable advantage estimations

Proximal Policy Optimization

Objective function:

$$L(\theta) = \min\left(\frac{\pi_\theta(o_t|q)}{\pi_{\theta_{old}}(o_t|q)} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q)}{\pi_{\theta_{old}}(o_t|q)}, 1-\epsilon, 1+\epsilon\right) A_t\right)$$

where:

- π_θ is the policy model and $\pi_{\theta_{old}}$ is the old policy model
- A_t is the advantage – the difference between the reward and the value function

$$A_t = Q(s_t, a_t) - V(s_t)$$

Advantage vs Generalised

Advantage Function

- Measures how much better an action is compared to the baseline value function
- Can be high variance if using one-step TD
- Is noisy and has high variance during long episodes, making learning unstable
- No explicit method to adjust bias and variance

$$A_t = Q(s_t, a_t) - V(s_t)$$

Generalised Advantage Estimation

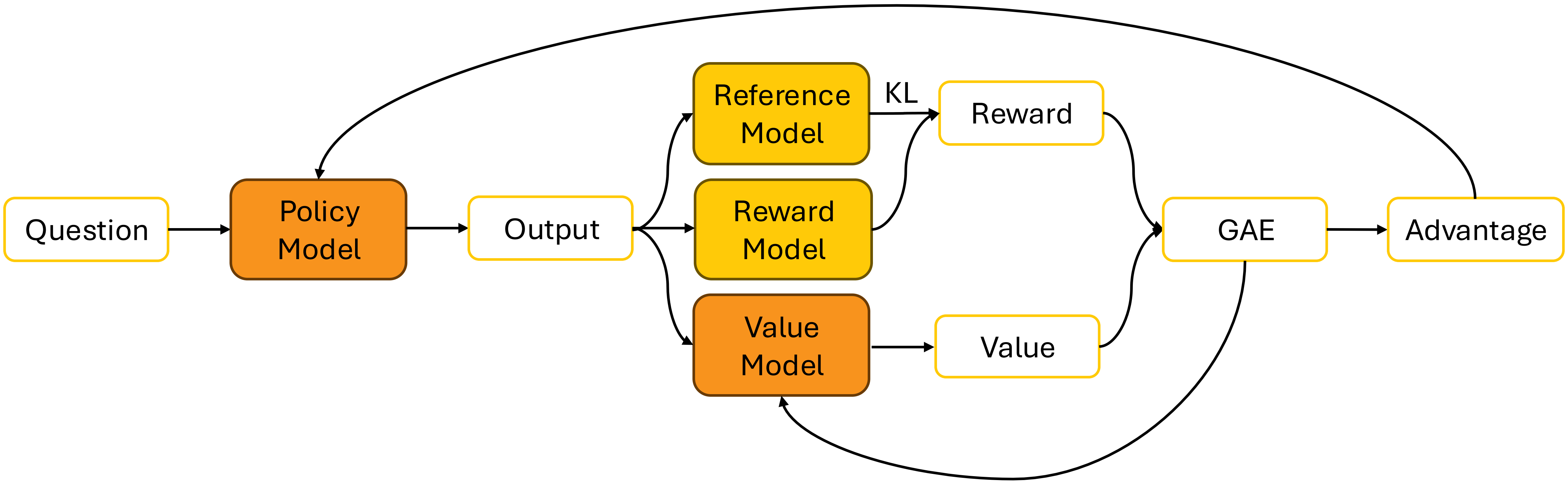
- A weighted sum of TD errors, smoothing the estimate
- Allows tuning via λ to balance bias and variance
- Ultimately smoother, more tunable, and more generalised than simulating the advantage function

$$A_t^{GAE} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$$

PPO is a powerful algorithm, but is computationally and memory intensive

- PPO uses the Generalized Advantage Estimation: $A_t = r_t - V(s_t)$
  - The difference between the reward at t and the predicted value of that reward
- In PPO the value model, $V(s_t)$, is trained alongside the policy model
  - Typically a model of comparable size to the policy model
  - Computationally and memory intensive to train

Training Using PPO



Group Relative Policy Optimization

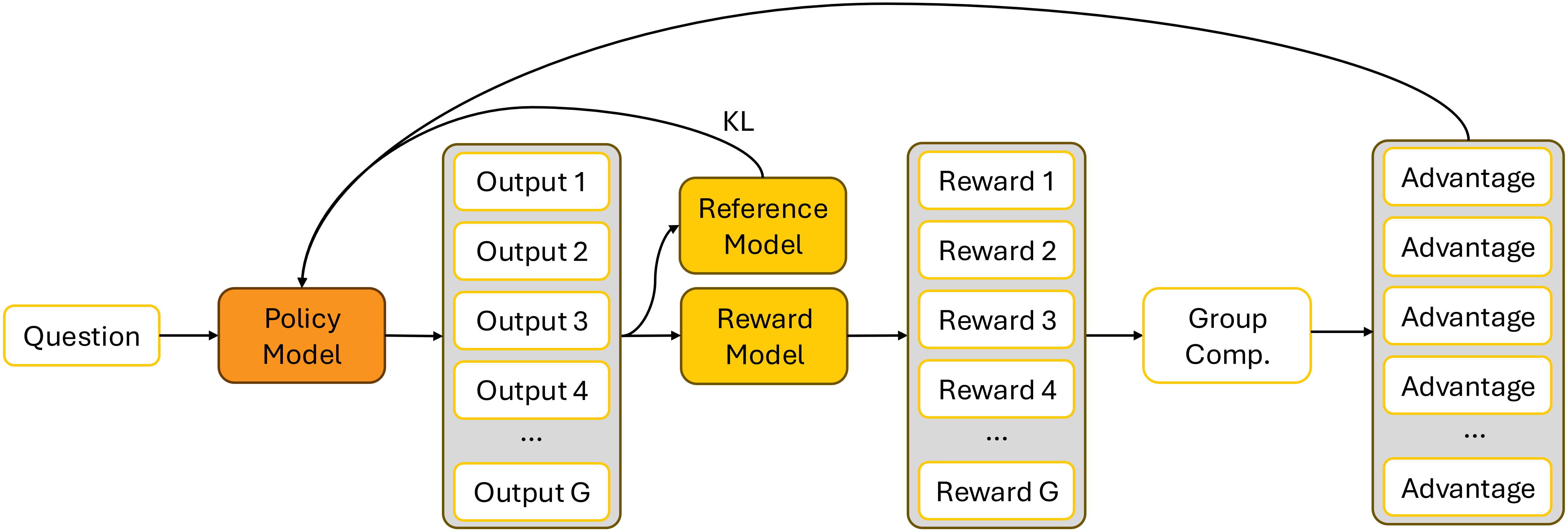
- The advantage function for GRPO does not require training a value model
- Uses the average reward of multiple sampled outputs produced in response to the same question using the old policy, $\pi_{\theta_{old}}$
- Collects a sample of outputs and computes advantage as:

$$\widehat{A_t} = \frac{r_t - \text{mean}(\boldsymbol{r})}{\text{std}(\boldsymbol{r})}$$

KL Penalty in GRPO

- The KL (Kullback-Leibler) penalty quantifies the difference between the updated policy and the reference policy to ensure controlled deviation
- It is useful in that it:
  - Prevents over-exploration
  - Encourages stable learning
  - Has adjustable penalty strength as a hyperparameter
- It is an additional term in the optimization objective, thereby influencing gradient updates (keeping the policy within a trust region) while continuing to improve expected rewards

GRPO


KL
Question
Policy Model
Output 1
Output 2
Output 3
Output 4
...
Output G
Reference Model
Reward Model
Reward 1
Reward 2
Reward 3
Reward 4
...
Reward G
Group Comp.
Advantage
Advantage
Advantage
Advantage
...
Advantage

Agenda

- Algorithm
- **Environment and Agent**
- Training
- Results

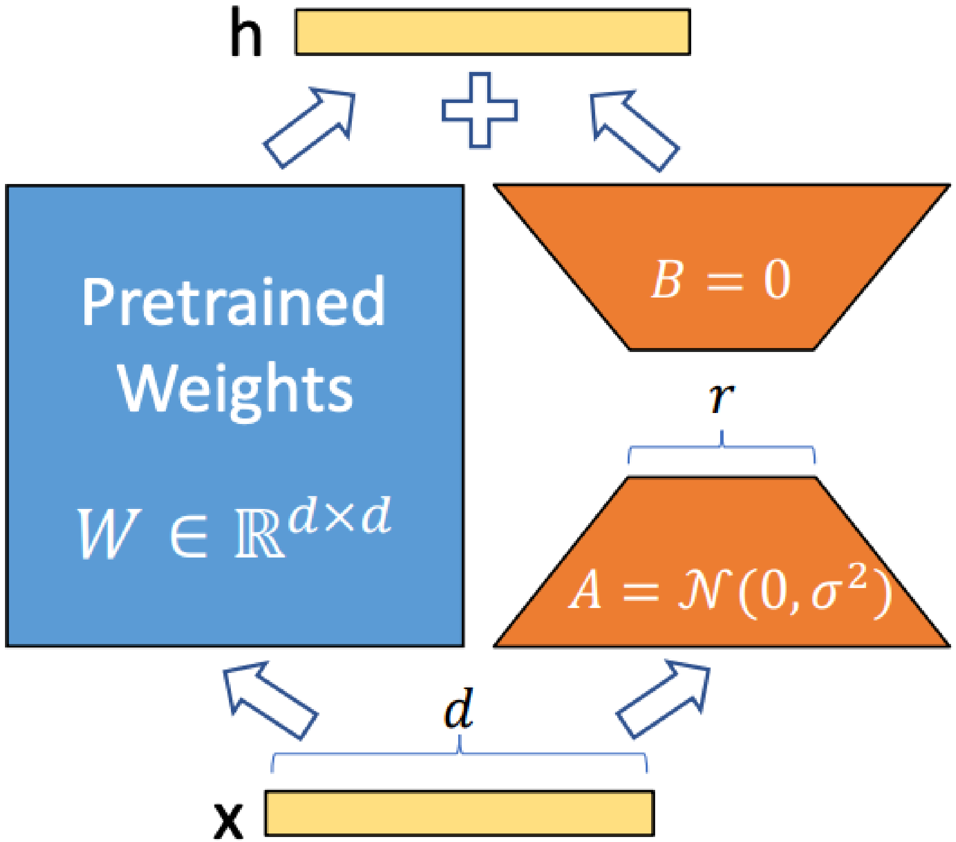
Environment

- Hugging Face 🤗
- Provides pre-trained models for NLP and LLMs via `transformers`
- Supports LLM fine-tuning, RLHF, and fast model deployment
- Contains libraries for optimized text tokenization for efficient model training and inference
- Has Model Hub for sharing and `Trainer` API for fine-tuning and RL workflows

Agent

- Llama 3.1 8B Instruct
- 4-bit quantized model
  - Model weights are truncated
- Compact version of the Llama 3 series designed for efficiency
- Allows for deployment across devices and servers with limited resources
  - Suitable for a wider range of applications

QLoRa (Quantized Low Rank Adaptation)



- **Full Fine-Tuning:** Updates all **d×d** model weights for each layer
- **LoRa:** Uses matrix decomposition to approximate weight updates with two **d×r** matrices
- **QLoRa:** Compresses 16-bit float weights into a custom 4-bit integer format optimized for a normal distribution which preserves more resolution near 0

Agenda

- Algorithm
- Environment and Agent
- **Training**
- Results

Training - Riddles

- Training on a dataset of 800, short answer riddles
- Targeting QKV, Output, Gate, Up and Down projection layers.

```
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "meta-llama/meta-Llama-3.1-8B-Instruct",
    max_seq_length = 512,
    max_lora_rank = 64,
    load_in_4bit = True,
    fast_inference = True,
    gpu_memory_utilization = 0.8)

model = FastLanguageModel.get_peft_model(
    model,
    r = 64,
    target_modules = [
        "q_proj", "k_proj", "v_proj", "o_proj",
        "gate_proj", "up_proj", "down_proj",
    ],
    lora_alpha = 64,
    use_gradient_checkpointing = "unsloth",
    random_state = 3407)
```

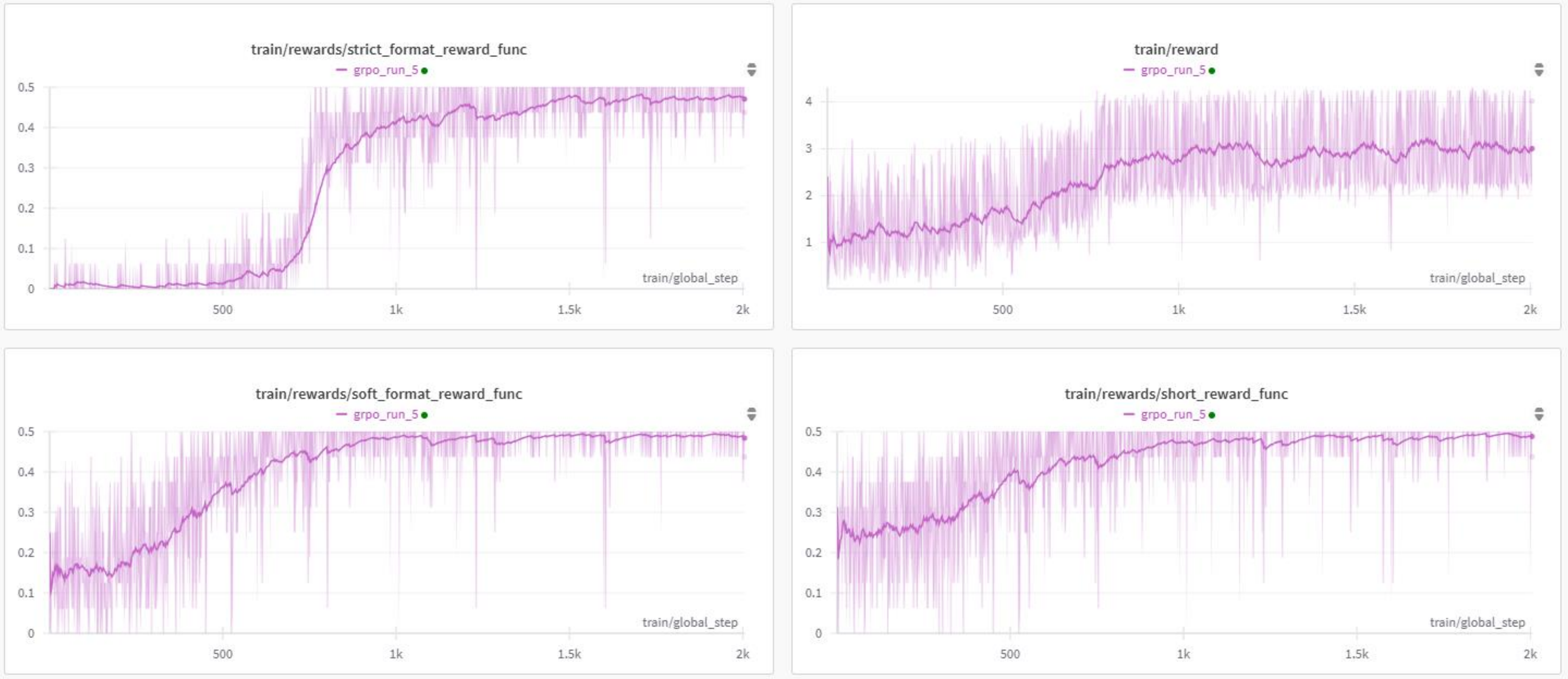
Riddle: I have streets, but no pavement. I have cities, but no buildings. I have forests, yet no trees. I have rivers, yet no water.

Answer: A map

Riddle: I have keys but can't open locks. People make me and they use forks. In some houses, I'm big, in others small.

Answer: A piano

Training - Riddle Progress



Training - Physical Interaction

- Dataset of 2,000+ questions and answers for testing physical world commonsense
- Trained for nearly 4,000 steps
- Targeted the QKV, Output, Gate, Up and Down projection layers

Question: How do you shake something?

Option A: Move it up and down and side to side quickly.

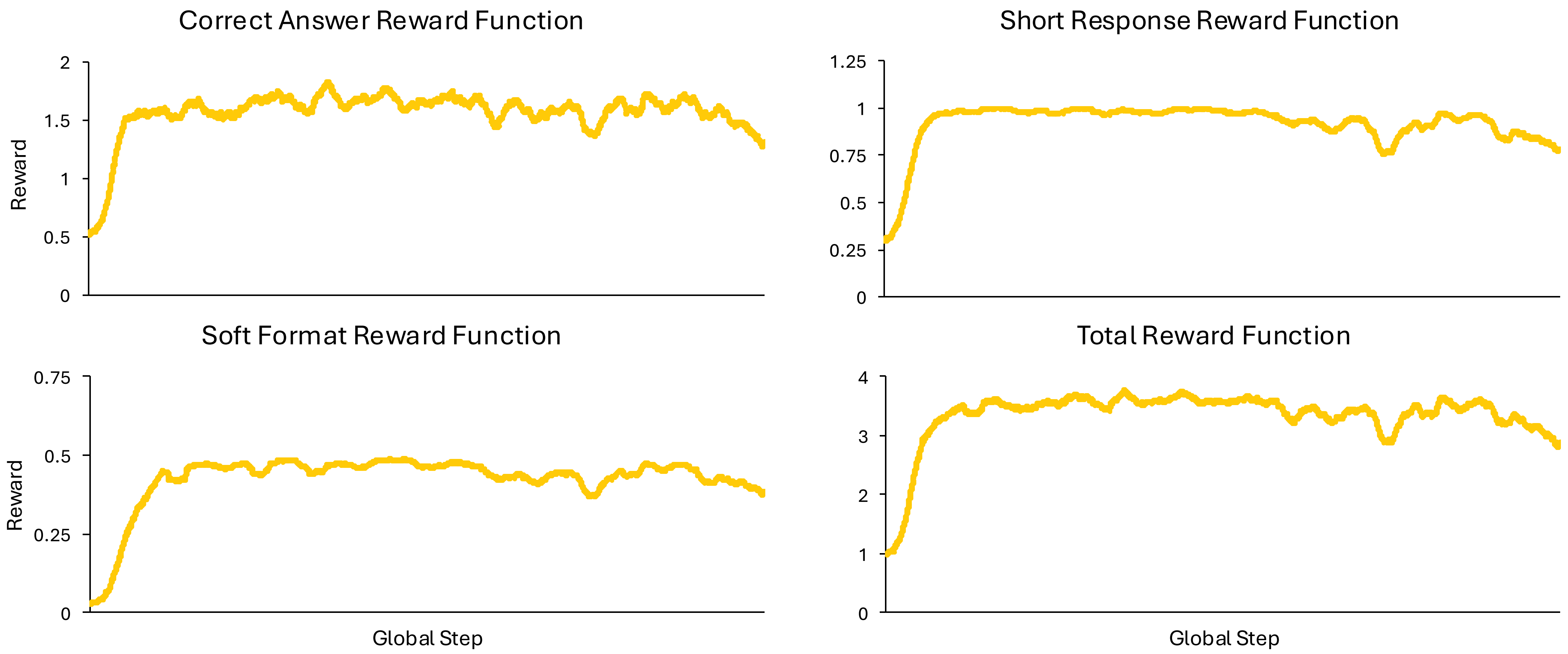
Option B: Stir it very quickly.

Question: How do you use a capped pen?

Option A: Replace cap, put tip of pen to paper and move across paper.

Option B: Remove cap, put tip of pen to paper and move across paper.

Training - Physical Interaction



Agenda

- Algorithm
- Environment and Agent
- Training
- **Results**

Model Evaluation Benchmark

- Evaluated using the AI2 Reasoning Challenge (ARC) dataset of grade school science questions
- Zero-shot learning to evaluate effectiveness of fine-tuning with a system prompt

Model	Without System Prompt	With System Prompt
Base Llama 3.1 8B	41.72%	66.04%
PIQA Tuned Llama 3.1 8B	-	50.46%
Riddler Tuned Llama 3.1 8B	42.24%	71.20%

Appendix