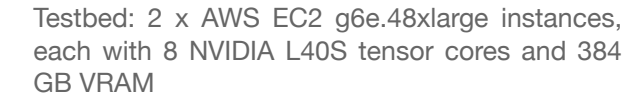



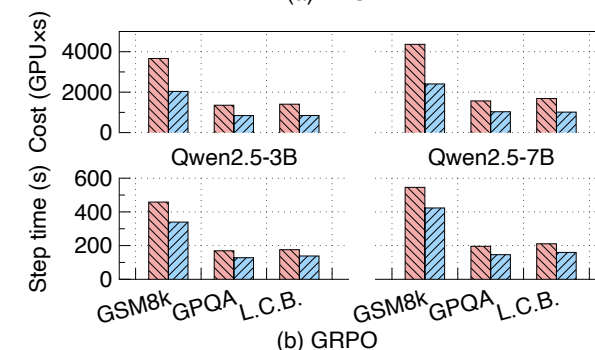
Stevens Institute of Technology¹, Northeastern University², Stony Brook University³, Missouri University of Science & Technology⁴



Evaluation



Up to
1.35x
Speedup 



Overall performance

Figure 1: RLHFless w/o DP+PA. (a) Diverse responses: A violin plot showing response lengths for Math, Science, and Coding datasets. (b) Changing lengths: A line plot showing mean response length vs. training step for ReMax+Qwen-7B, PPO+Qwen-3B, and PPO+DeepSeek-7B. The diagram illustrates the cost calculation: $\text{Cost} = \text{latency} \times \text{resources}$, where $\text{Resources} = \text{latency}$. A 'Sweet spot' is identified to balance cost and speed. The right side shows three stacked plots: # of actors, Time (s), and Cost (GPUxs) vs. Training step, comparing RLHFless w/o DP+PA (blue) and RLHFless w/o DP+AS+PA (red).

Figure 10 illustrates the RLHFless w/o DP+AS configuration. (a) Prompt assignment: A diagram showing the assignment of prompts to Decode1 and Decode2 blocks. Predicted length is indicated by a blue dashed line with a circle, and idle time is indicated by a red dashed line with a square. (b) Cost reduction: A bar chart showing GPU memory-time (GB*s) before and after optimization. Before: ~20,000 GB*s, After: ~10,000 GB*s. (c) Heatmaps: Four heatmaps showing slot index vs time (second) for Actor 1, Actor 2, Actor 3, and Migrate. The heatmaps show the distribution of slots over time for each actor.

Intra-node:

- PCIe
- NVLink

Inter-node:

- Ethernet
- InfiniBand

Latencies can be overlapped

Model sync. latency

KV transfer latency

Shared prefix

Response

This research was supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot allocation 240269.

NAIRR Pilot

