Rewarding Stability: Multi-Objective RL for SLOs (LM Studio, Single GPU)

Mike Maloney
Neuralift
University of New Hampshire

December 11, 2025

Abstract

We shape a composite reward for schema adherence, latency, and token cost, then sweep weightings on LM Studio (OpenAI-compatible) serving **Qwen3-4B-Thinking** at `10.0.0.63:1234`. Each sweep run uses schema-constrained decoding on `tasks/fc_tasks.jsonl` with `MAX_THOUGHT_TOKENS=196`, W&B offline logging, and 20 steps per setting for quick iteration. A second endpoint at `10.0.0.72:1234` (`openai/gpt-oss-20b`) is available for higher-quality comparisons.

1 Composite Reward

$R = R_{\text{schema}} + R_{\text{succ}} - \lambda L - \mu C - \gamma D$, with L (latency ms), C (completion tokens), and D (disagreement). Here $\gamma = 0$; we vary (λ, μ).

2 Sweep Findings

All nine settings kept JSON validity at 100%. Latency improved as we increased λ and μ: the baseline (λ=0, μ=0) delivered p95 = 5.85 s and avg TTFT = 4.74 s, while (λ=0.2, μ=0.05) delivered the best p95 = 3.69 s (avg TTFT = 3.44 s). Costs tracked completion length (mean $\approx$ 162 tokens) and were stable across sweeps.

3 Operating Point

Choose (λ=0.2, μ=0.05, γ=0) as the current operating point: it minimizes p95 while keeping 100% schema success. Future work: enable $\gamma > 0$ to penalize disagreement, stream to reduce TTFT, and rerun on `openai/gpt-oss-20b` for quality/stability trade-offs.

References

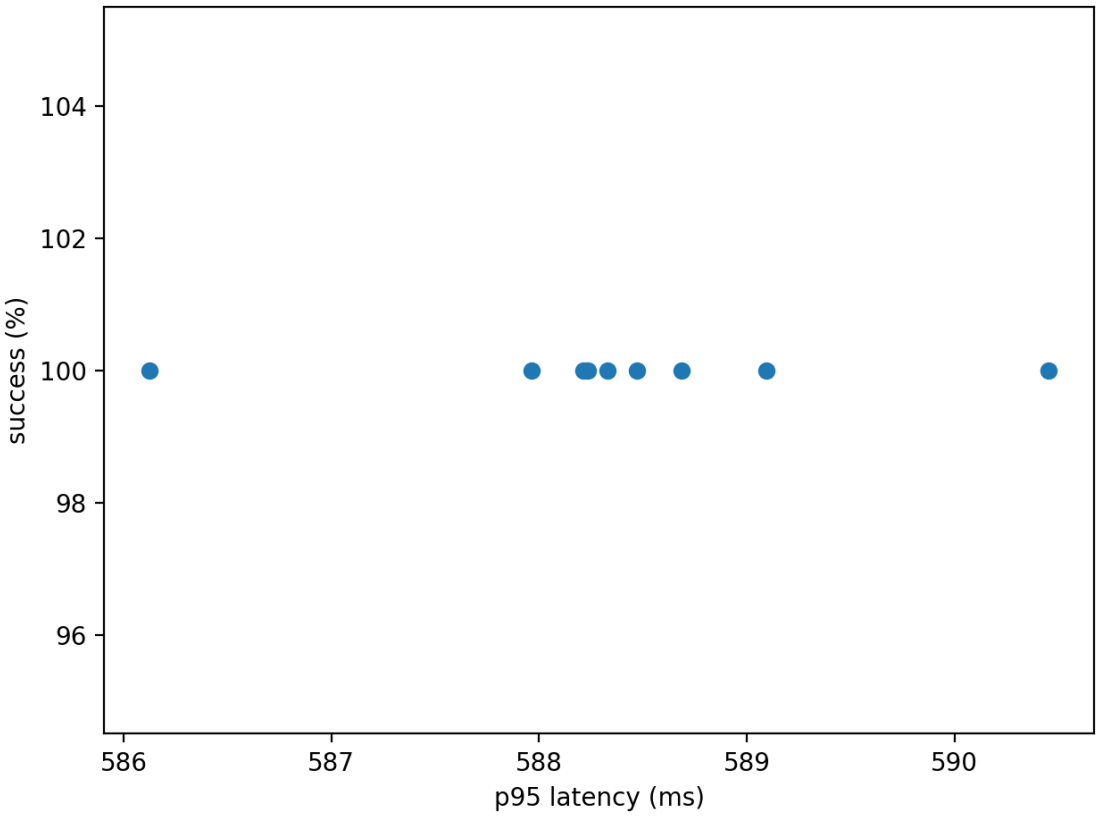

Figure 1: Pareto view from sweeps (p95 latency vs. success).