One-Shot RLVR

Collapsing Reinforcement Learning to a Single Gradient Step

Research Proposal

April 2026

The Observation

RLVR training is **surprisingly linear**:

- Weights evolve with $R^2 > 0.7$ against training step (Wang et al., 2026)
- Log-probabilities follow the same linear trend
- Validated across GRPO, Reinforce++, GSPO; 1.5B–8B models

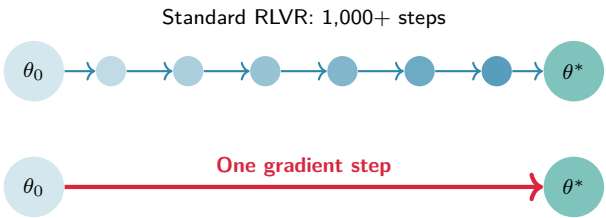
And the destination is **predictably close**:

- Pretrained weights are surrounded by dense task-expert solutions (Gan & Isola, 2026)
- Solution density scales with model size
- Experts are specialists, not generalists

The Implication

If the direction doesn't change and the expert is close:

Why take 1,000 steps when one suffices?



Method

One-Shot RLVR in four steps:

1. **Sample** a massive rollout batch from π_{base} ($N \sim$ 10K–50K trajectories, high τ)
2. **Score** with binary verifier
3. **Compute** one policy gradient step (GRPO-style advantage normalization)

$$g = \nabla_\theta \sum_{i=1}^{N} A_i \log \pi_\theta(y_i|x_i)\Big|_{\theta=\theta_0}$$

4. **Line-search** over step size α on held-out validation set

The key shift: allocate compute to rollout breadth (large batch, one step) rather than rollout depth (small batch, many steps).

Preventing Overshoot

The expert lives at some unknown distance. Too large α exits the thicket.

Line search with the verifier:

- Evaluate $\theta_0 + \alpha \cdot g$ for $\alpha \in \{0.5, 1, 2, 5, 10\} \times \alpha_{\text{base}}$
- Pick α^* maximizing held-out accuracy
- Cost: 5 forward-pass evaluations (trivial vs. RL training)

Entropy monitoring (free):

- Track $H(\pi_{\theta_0+\alpha g})$ across step sizes
- Entropy collapse $\Rightarrow$ overshoot into degenerate mode

Expected landscape: inverted-U performance vs. step size (consistent with Wang et al. Figure 6).

Validating the Direction

How do we know the one-shot gradient is correct?

Cosine similarity diagnostic:

- Run k actual RL steps, compute $\Delta\theta_{\text{RL}} = \theta_k - \theta_0$
- Measure $\cos(\Delta\theta_{\text{RL}},\ g_{\text{one-shot}})$
- High similarity $\Rightarrow$ direction matches RL trajectory

Cross-subset consistency:

- Split rollout batch into M subsets, compute gradient on each
- High pairwise cosine similarity $\Rightarrow$ batch is large enough

Correct/incorrect decomposition:

- Gradient from correct vs. incorrect rollouts should be anti-parallel

Experimental Plan

Axis	Settings
Models	Qwen3-4B, DeepSeek-R1-Distill-7B, OLMo3-7B
Benchmarks	AIME 24/25, MATH-500, LiveCodeBench v6, MBPP+
Baselines	Standard GRPO, RL-Extra, RandOpt, STaR, DPO
Key sweep	Pareto frontier: (steps $\times$ batch) at fixed rollout budget
Diagnostics	Cosine sim, entropy curves, α landscape

Central experiment: 1 step $\times$ 50K batch vs. 1,000 steps $\times$ 50 batch.
Same total rollout budget. Compare final accuracy.

Scale sweep: Repeat at 1.5B, 4B, 7B, 14B. Thicket paper predicts this works better at larger scale.

No-RL Alternatives

Methods that skip RL entirely:

- **Rejection sampling + SFT (STaR):** Sample, keep correct, finetune. One round.
- **DPO on base-model rollouts:** Pair correct/incorrect, one epoch.
- **SVD probe:** Mean(h_L^{correct}) − Mean($h_L^{\text{incorrect}}$), backprop to $\Delta\theta$.
- **RandOpt:** Zero training. Random perturbations + ensemble.

All require ≥ 1 correct sample. At pass@$N = 0$: the Invisible Leash binds.

Fallback for sparse reward:

- Partial-credit verifiers (execution traces, intermediate step validity)
- Relative ranking among all-incorrect rollouts (less wrong $\Rightarrow$ positive advantage)

Failure Modes

When this breaks:

- Base model pass rate too low $\Rightarrow$ insufficient positive signal in one batch
- Code tasks (sparser reward) may need more samples than math
- One-shot gradient may overfit to batch distribution
- Very small models (< 1B): thicket is sparse, direction is noisy

Fallback:

- 5–10 steps with large batches (still massive speedup over 1,000 steps)
- The interesting result is the *Pareto frontier shape*, not the $k = 1$ point

The Story

"RLVR's compute is almost entirely wasted."

- **Thickets** (Gan & Isola): the expert is close
- **Linearity** (Wang et al.): the direction doesn't change
- **This work**: just take the step

A single policy gradient step with a sufficiently large batch achieves the same effect as thousands of RL steps.

The remaining steps are not exploring—they are walking a straight line the first step already found.