SFT Overtraining Predicts Rank Inversion via Entropy Collapse Under RLVR

Siddharth Aphale, Kelly Liu

Stanford University · DL4C Workshop @ ICML 2026, Seoul

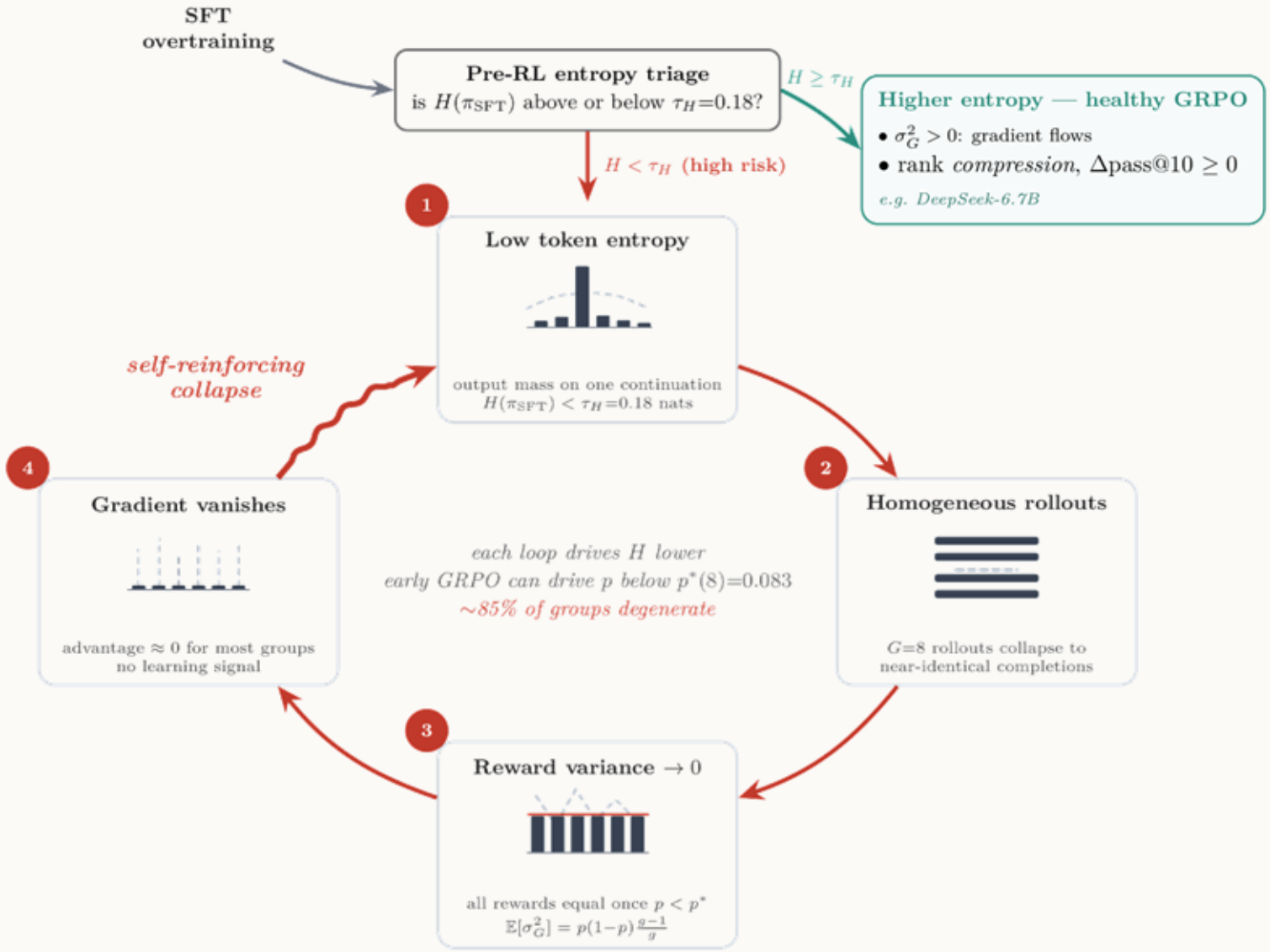
Motivation

- **Standard recipe:** RLVR via GRPO [2, 4] runs after SFT and keeps the highest-scoring SFT checkpoint.
- **The trap:** this can pick the worst checkpoint for RL. As SFT deepens, pass@1 rises but the GRPO pass@10 it reaches *falls* [1], so RL compute is spent on a doomed checkpoint.
- **Why:** entropy collapse from SFT overtraining puts almost all mass on one continuation, so GRPO rollouts go homogeneous and the reward variance vanishes.
- **Mechanism:** inside a GRPO group the advantage variance vanishes once pass@1 drops below p*(g) (p*(8)=0.083), leaving most groups with no gradient:

$$\mathbf{E}\left[\sigma_G^2\right] = \mathbf{pass@1}(1 - \mathbf{pass@1}) \cdot \frac{\mathbf{g} - \mathbf{1}}{\boldsymbol{g}}$$

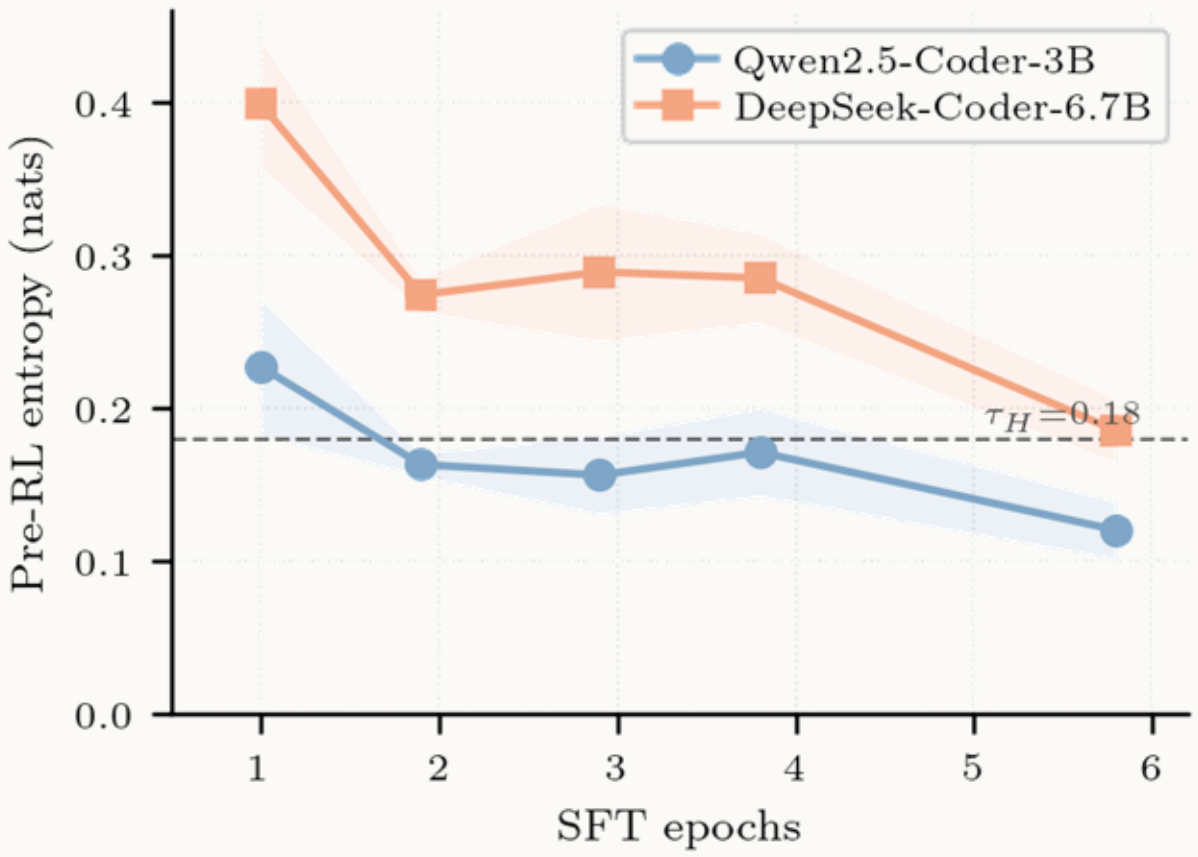
- **Our fix:** an entropy screen that rejects doomed checkpoints at zero RL cost, validated on 3B and 6.7B code ladders.

Self Reinforcing Collapse

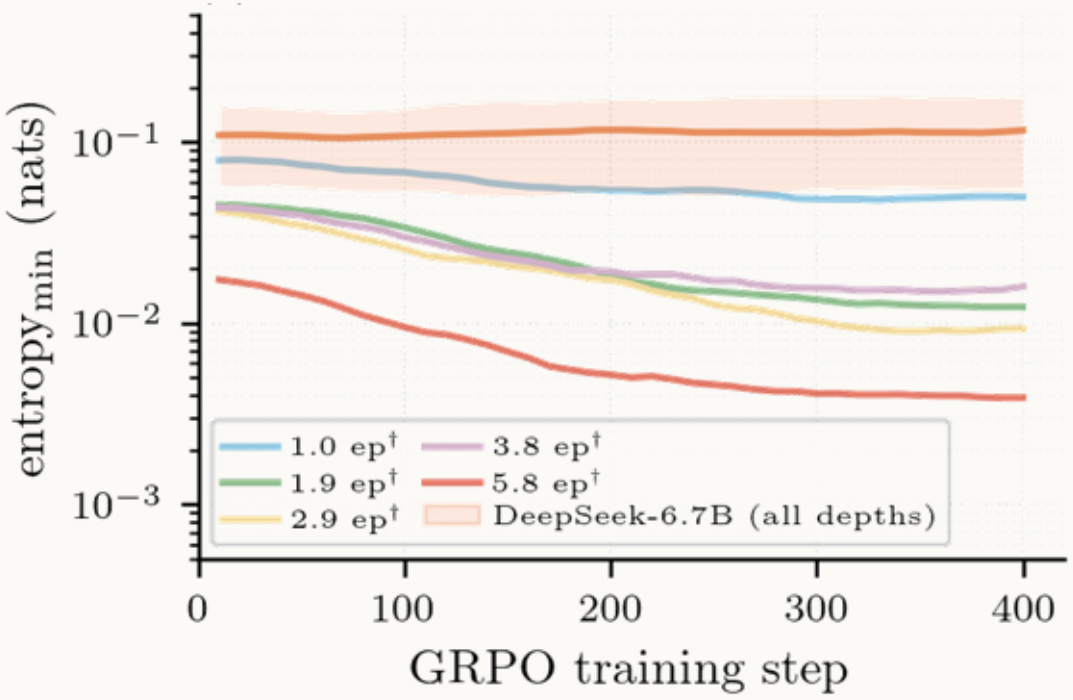


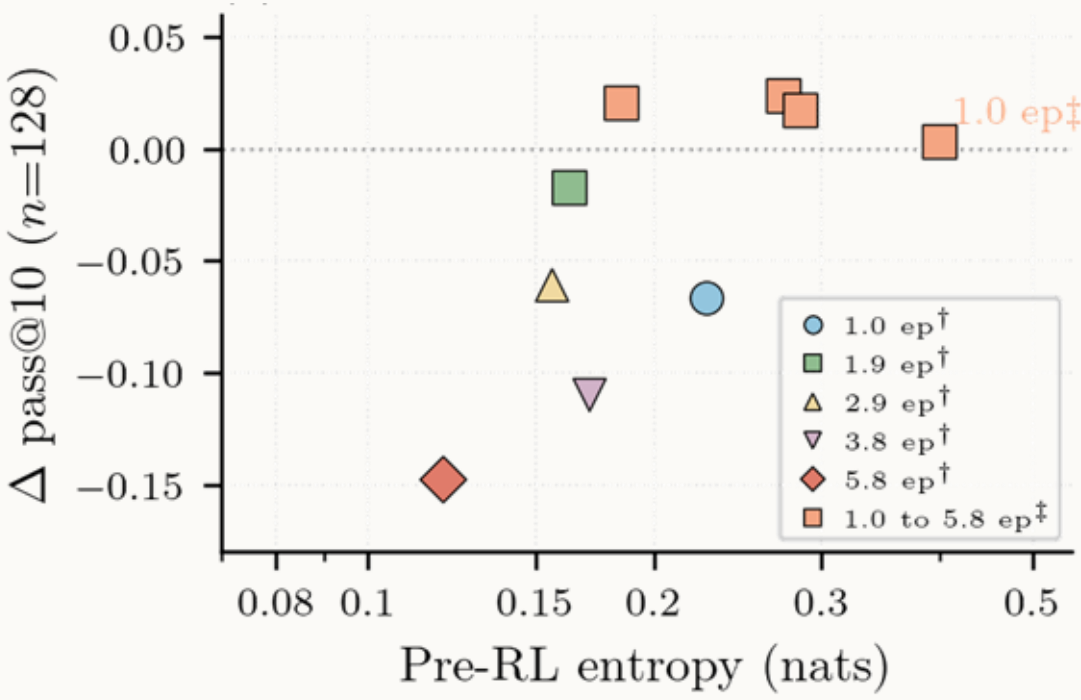
SFT Checkpoints

- **Setup**
  - Two SFT ladders, identical recipe:
    - Qwen2.5-Coder-3B-Base
    - DeepSeek-Coder-6.7B-Base
  - LoRA (r=128, α=128)
  - Sweep only SFT depth (1.0–5.8 epochs), holding data, hyperparameters, and eval fixed
- **Data**
  - 5,000 KodCode-V1 problems with Gemini 2.5 Flash CoT traces
  - decontaminated against HumanEval+/MBPP
- **Pre-RL Screen:** Qwen's pass@1 climbs with SFT depth while entropy falls, the highest scorer is the worst checkpoint to pick.
- **Verdict:** Qwen's entropy bottoms at 0.12 nats (below τ_H); DeepSeek stays ≥ 0.185, confirming collapse for Qwen and stability for DeepSeek.



GRPO





- **Setup:** 400-step GRPO (DAPO, G=8, β=0) on ~1,100 held-out KodCode-V1 problems.
- **Eval:** 40 frozen HumanEval+ problems, pass@10 every 50 steps; entropy = mean next-token entropy (nats).

- **Lower pre-RL entropy (Qwen):** produces inversion, Δpass@10 < 0.
- **Higher pre-RL entropy (DeepSeek):** produces compression, Δpass@10 ≥ 0.

Pre RL Diagnostic

- **Stage 1: pre-RL entropy screen** (no RL compute). Reject a checkpoint if

$$H(\pi_{\mathrm{SFT}}) < \tau_H = 0.18 \text{ nats}$$

  rejecting a failed checkpoint saves 100% of its RL budget.

- **Stage 2: step-150 monitor**. Stop the run if the relative entropy drop

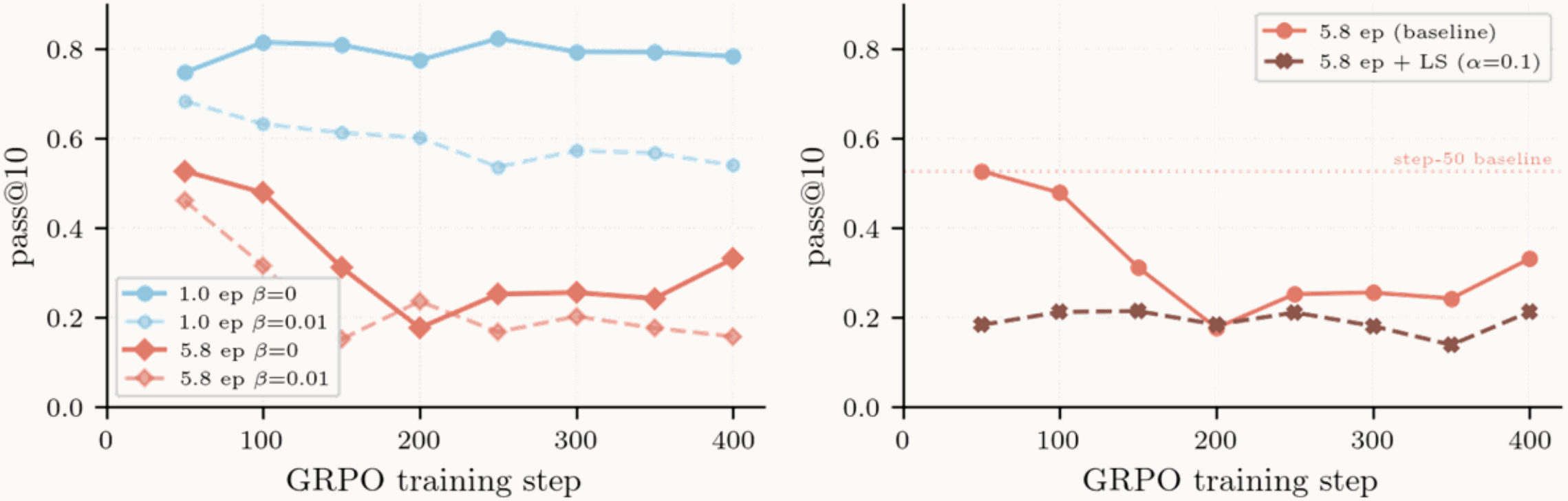
$$\frac{\Delta H(10 \rightarrow 150)}{H(10)} > \tau_2 = 0.50$$

  saving 62.5% of the run.

- **Selection gain:** the diagnostic picks the 1.9-epoch checkpoint over the standard "deepest" (5.8-epoch) rule, lifting deep-eval pass@10 0.553 → 0.643 (+0.090, n=128).

Key takeaway: two cheap checks (a pre-RL entropy threshold and a step-150 monitor) cut wasted RL compute by up to 100% and 62.5%.

Interventions Do Not Rescue Collapse



Key takeaway: once a checkpoint has collapsed (5.8 ep), neither entropy regularization (β) nor label smoothing during RL closes the pass@10 gap to a healthy (1.0 ep) checkpoint. The fix has to happen before RL, not during it.

Key Results and Takeaways

- **Rank inversion (Qwen):** peak GRPO pass@10 falls 0.806 → 0.481 with SFT depth, peaking earlier (step 250 → 50); also seen in math reasoning [3].
- **DeepSeek compresses, not inverts:** the 6.7B ladder bunches; post-RL rank matches pre-RL (ρ = +1.00).
- **Entropy beats pass@1:** Spearman ρ with peak pass@10 is +0.69 for pre-RL entropy vs. −0.75 for pass@1 (n=15, p<0.01); higher pass@1 predicts worse RL.
- **Threshold mechanism:** by step 200, Qwen's pass@1 = 0.020 < p*(8) = 0.083, so ~85% of groups give zero gradient.
- **More rollouts don't help:** doubling g only halves p*(g); a collapsed policy stays collapsed.

Practitioner takeaway: measure mean token entropy before spending RL compute. If pre-RL entropy < $\tau_H \approx 0.18$ nats (or pass@1 < p*(g)), expect zero-gradient collapse and ladder inversion. Optimizing pass@1 directly (e.g., label smoothing) can make RL *worse*: the restoration paradox.

Conclusions

- **Entropy collapse + p*(g):** together they predict whether GRPO inverts (Qwen) or compresses (DeepSeek) the SFT ladder; pass@k alone misses this.
- **Pre-RL entropy signal:** a better RL-readiness indicator than pass@k, and the diagnostic acts before or early in RL, not after compute is spent.
- **Future work:** test the p*(g) diagnostic on math/reasoning tasks and non-binary rewards, larger g, and generalisation-loss predictors.

Key takeaway: a single pre-RL entropy check predicts GRPO's fate better than pass@k: catch a doomed checkpoint before you spend the RL compute.

Thank You

Questions & discussion welcome

saphale@stanford.edu

Paper: https://arxiv.org/abs/2606.18487

Project page: https://siddharthaphale.github.io/blog/2026/sft-checkpoint-rl-collapse