Granularity-Aware Off-Policy Fine-Tuning of LLMs via Expert Demonstrations

Extended Abstract

Yina Jian Sukeerth Ramkumar Kevin Song
{yinajan, sukeerth, kevsong}@stanford.edu

Motivation. Reinforcement learning (RL) is the dominant post-training recipe for aligning large language models (LLMs), typically as a pipeline of supervised fine-tuning (SFT), preference optimization, and a policy-gradient stage such as REINFORCE-Leave-One-Out (RLOO). RLOO is strictly *on-policy*: every gradient step consumes fresh rollouts, which is sample-inefficient under sparse verifier rewards. A natural question is whether *off-policy* expert demonstrations from a stronger teacher can supplement on-policy learning—and, more specifically, how the **reasoning granularity** of those demonstrations (how many intermediate steps they contain) shapes what the policy learns. Prior work treats expert data as fixed; we instead make granularity a controllable axis and study how it interacts with the imitation strength β.

Method. We implement the full SFT $\rightarrow$ IPO $\rightarrow$ RLOO pipeline on the Countdown arithmetic-reasoning task and extend RLOO with an off-policy imitation regularizer $\mathcal{L}(\theta) = \mathcal{L}_{\text{RLOO}}(\theta) + \beta\,\mathbb{E}_{(x,y_e)}[-\log \pi_\theta(y_e \mid x)]$, where y_e are teacher (DeepSeek-V3) chains at three granularities (coarse / medium / fine). We sweep granularity against $\beta \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9\}$ (a 3×9 grid plus a β=0 baseline), all trained to a common 60-step checkpoint.

Implementation. The policy is Qwen2.5-0.5B initialized from our IPO checkpoint; RLOO uses batch 128, group size 16, and KL/entropy coefficients 10^{-3}. Because prompting the teacher failed to separate coarse from medium chains, we re-operationalize granularity by the *measured* number of reasoning steps (tercile bins of line count), yielding cleanly separated arms (Cohen's $d = 2.94, 1.48$). All 28 checkpoints are evaluated on 50 held-out prompts $\times$ 16 samples.

Results. Final accuracy (pass@1, pass@16) is **largely insensitive to** β: across the full grid, pass@1 stays within 0.55–0.64 around the 0.605 baseline, with no systematic granularity- or β-dependent trend—variation is on the order of single-seed, 50-prompt evaluation noise. The imitation weight nonetheless has a **strong, granularity-dependent effect on reasoning behavior**: the model's *multi-answer* rate (responses emitting more than one `<answer>`, in the worst case overwriting a correct equation with a later wrong one) collapses from ≈ 0.65 to ≈ 0 at a *granularity-dependent threshold*—$\beta \approx 0.15$ for coarse and medium, but only $\beta \approx 0.4$ for fine. Arithmetic-error rate also declines with stronger imitation, and output length drops sharply (from $\sim$500 to $\sim$220 words) at the same threshold where the multi-answer rate collapses.

Discussion. Our proposal's H1 (a medium-granularity "sweet spot" in accuracy) is **not supported**; H2 (failure modes at the extremes) **is**—fine chains retain high multi-answer and arithmetic-error rates through moderate β; and H3 is **refined**: β does not move an accuracy optimum but rather the *threshold at which reasoning behavior changes*. A key internal-validity caveat is that demonstration granularity and token length are strongly correlated, and the multi-answer-rate collapse coincides with a length collapse; the two are therefore partly confounded, though a length-controlled analysis still finds an imitation effect on commitment that survives length-matching.

Conclusion. On a small model and a verifier-solvable task, off-policy imitation acts less as an accuracy lever and more as a *behavioral regularizer* that reshapes how the policy reasons—and the granularity of the demonstrations controls how much imitation pressure is needed to effect that change. This reframes "demonstration granularity" as a knob on reasoning form rather than on final correctness.

Granularity-Aware Off-Policy Fine-Tuning of LLMs via Expert Demonstrations

Yina Jian
Department of Computer Science
Stanford University
yinajan@stanford.edu

Sukeerth Ramkumar
Department of Computer Science
Stanford University
sukeerth@stanford.edu

Kevin Song
Department of Computer Science
Stanford University
kevsong@stanford.edu

Abstract

We study how the *granularity* of off-policy expert demonstrations interacts with the imitation strength β when used to regularize on-policy RLOO fine-tuning of a small LLM. On the Countdown task we add an off-policy imitation term to RLOO and sweep three demonstration granularities against nine values of β, training all configurations to a common 60-step checkpoint. We find that final accuracy is largely insensitive to both granularity and β, but that β exerts a strong, granularity-dependent influence on *reasoning behavior*: a multi-answer failure mode collapses at a β threshold that rises with demonstration granularity, accompanied by falling arithmetic-error rates and a sharp drop in output length. We interpret off-policy imitation as a behavioral regularizer rather than an accuracy lever, and use a length-controlled analysis to partly disentangle the length confound that complicates causal attribution.

1 Introduction

Reinforcement learning has become the dominant post-training technique for aligning LLMs, commonly structured as a pipeline of supervised fine-tuning (SFT), preference optimization (e.g. DPO/IPO), and an online policy-gradient stage such as REINFORCE-Leave-One-Out (RLOO) (Ahmadian et al., 2024). RLOO is strictly on-policy: each gradient step requires fresh rollouts scored by a verifier, which is sample-inefficient when the reward signal is sparse and sampling is expensive.

A natural way to improve efficiency is to reuse *off-policy* data—in particular, expert reasoning chains produced once by a stronger teacher model. But expert demonstrations are not monolithic: a chain that reaches the same answer can be expressed in one terse step or in many fine-grained ones. We refer to this property as the **reasoning granularity** of a demonstration. Prior work on process supervision and off-policy LLM fine-tuning treats the granularity of a "step" as a fixed design choice; to our knowledge, no prior work studies granularity as a controllable axis that interacts with the imitation weight β.

This paper asks: *what level of demonstration granularity yields the most effective imitation-regularized policy gradient, and what failure modes emerge at the extremes?* We implement the full SFT $\rightarrow$ IPO $\rightarrow$ RLOO pipeline on the Countdown arithmetic task, extend RLOO with an off-policy imitation regularizer, and sweep granularity (coarse / medium / fine) against nine values of β. Consistent with the project's emphasis on *characterizing* strengths and weaknesses rather than maximizing a metric,

our central finding is a negative-but-informative one: β barely moves final accuracy, yet it strongly and granularity-dependently reshapes the model's reasoning behavior. If off-policy demonstrations can shape reasoning without extra on-policy sampling, this transfers to settings where rollouts are expensive or latency-bound (e.g. agentic coding, robotics).

Hypotheses. Our proposal posed three concrete hypotheses, which we restate here and revisit in our Results and Discussion:

- **(H1)** a *medium* demonstration granularity yields the best final accuracy—a granularity "sweet spot";
- **(H2)** the extreme granularities induce characteristic failure modes (overly terse coarse chains versus overlong, self-revising fine chains);
- **(H3)** the most effective imitation weight β shifts systematically with demonstration granularity.

As we show, the completed sweep *refutes* H1, *supports* H2, and forces a *refinement* of H3 from an accuracy optimum to a behavioral threshold.

2 Related Work

Process-level reward modeling. Supervising intermediate reasoning steps rather than only final answers improves LLM reasoning: Math-Shepherd (Wang et al., 2024) and PRM800K (Lightman et al., 2023) train step-level reward models, and ProcessBench (Zheng et al., 2024) evaluates process supervision. These works fix the granularity of a "step" (typically by line breaks or arithmetic operations) and do not treat granularity itself as a design axis.

Off-policy LLM fine-tuning. On-policy methods such as RLOO (Ahmadian et al., 2024) require fresh rollouts per update. Off-policy alternatives relax this: TBRM (Jia et al., 2025) reframes fine-tuning as value-based off-policy learning, and Tang et al. (Tang et al., 2025) unify on- and off-policy data in one algorithm; AWAC (Nair et al., 2020) advantage-weights imitation of off-policy demonstrations in control. None studies how the *granularity* of off-policy demonstrations modulates training.

Synthetic data and expert distillation. Distillation from stronger LLMs is a standard recipe for small models (Lee et al., 2024; Yang et al., 2024), but typically maximizes volume or diversity without controlling structural properties. We instead treat granularity as a controllable structural property and study its training-time effects.

Long-horizon decomposition. The hypothesis that long-horizon tasks benefit from decomposition into reusable lower-horizon components has been studied in robotic manipulation (Ahn et al., 2022; Jian et al., 2026); we extend this view to language reasoning by treating step granularity as a decomposition axis.

3 Method

Background: SFT $\rightarrow$ IPO $\rightarrow$ RLOO. SFT provides a behavior-template prior via a next-token objective on completion tokens, $\max \mathbb{E}_{x,y} \sum_t \log \pi_\theta(y_t \mid x, y_{<t})$. IPO performs implicit preference optimization without an explicit reward model, regressing the implicit reward margin to a target gap. RLOO is an online policy gradient with a rule-based verifier reward and a leave-one-out baseline over K in-group rollouts, stabilized by importance weighting and a KL penalty: $\mathcal{L}_{\text{RLOO}} = -\mathbb{E}[w_i A_i \log \pi_\theta(y_i \mid x)]$.

Imitation-regularized RLOO. We augment the RLOO loss with an off-policy imitation term on expert demonstrations $\mathcal{D}_{\text{exp}}^{(g)}$ at granularity g:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{RLOO}}(\theta) + \beta \cdot \mathbb{E}_{(x,y_e)\sim\mathcal{D}_{\text{exp}}^{(g)}}\big[-\log \pi_\theta(y_e \mid x)\big]. \tag{1}$$

Algorithm 1 One training step of imitation-regularized RLOO

Require: policy π_θ, reference π_{ref}, vLLM behavior policy μ, prompt batch B, expert set $\mathcal{D}_{\text{exp}}^{(g)}$, imitation weight β, group size K, KL/entropy coefficients $\lambda_{\text{KL}}, \lambda_H$, IS clip c, learning rate η

for each prompt $x \in B$ **do**
  sample K rollouts $y^{(1)}, \ldots, y^{(K)} \sim \mu(\cdot \mid x)$ ▷ vLLM, training temperature 1.0
  $R_i \leftarrow \textsc{Verifier}(y^{(i)}, x) \in \{0.0, 0.1, 1.0\}$
  $A_i \leftarrow R_i - \frac{1}{K-1} \sum_{j \neq i} R_j$ ▷ leave-one-out baseline
  $w_i \leftarrow \min\big(\exp(\log \pi_\theta(y^{(i)} \mid x) - \log \mu(y^{(i)} \mid x)),\, c\big)$ ▷ clipped IS weight
end for
$\mathcal{L}_{\text{RLOO}} \leftarrow -\frac{1}{|B|K} \sum_{x,i} w_i\, A_i\, \log \pi_\theta(y^{(i)} \mid x) + \lambda_{\text{KL}}\, \text{KL}[\pi_\theta \| \pi_{\text{ref}}] - \lambda_H\, \mathcal{H}[\pi_\theta]$
draw expert demos $(x_e, y_e) \sim \mathcal{D}_{\text{exp}}^{(g)}$; $\mathcal{L}_{\text{imit}} \leftarrow -\log \pi_\theta(y_e \mid x_e)$
$\mathcal{L} \leftarrow \mathcal{L}_{\text{RLOO}} + \beta\, \mathcal{L}_{\text{imit}}$ ▷ Eq. (1)
$\theta \leftarrow \theta - \eta\, \nabla_\theta \mathcal{L}$

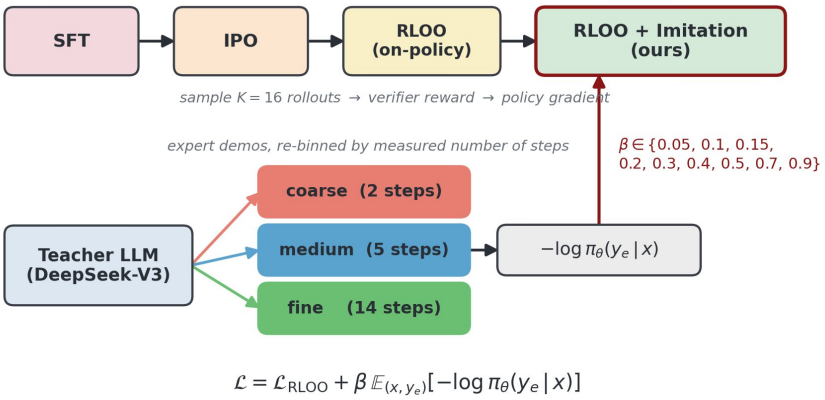

Figure 1: Method overview. A teacher (DeepSeek-V3) produces expert chains at three reasoning granularities; the imitation NLL of these chains is added to the RLOO loss with weight β. The same valid `<answer>` is shared across granularity arms; only the depth of the `<think>` trace varies.

At each step we (1) sample K=16 responses with vLLM, (2) score them with the verifier, and (3) take an RLOO step plus the imitation gradient. The imitation term anchors the policy toward the expert distribution at granularity g, while RLOO supplies on-policy verifier-grounded improvement. Holding everything else fixed and varying (g, β) isolates the effect of demonstration structure.

Operationalizing granularity. We first controlled granularity *by prompting* the teacher for coarse / medium / fine chains, but a manipulation check showed prompting alone failed to separate coarse from medium (Cohen's $d \approx 0.46$ on reasoning steps). We therefore re-operationalize granularity by the *measured* number of reasoning steps—tercile bins of line count—yielding cleanly separated arms (Cohen's $d = 2.94$ and 1.48; no range overlap), faithful to the proposal's definition of granularity as average steps per chain. Crucially, the same valid `<answer>` is shared across all three arms, so granularity is isolated from solution correctness. Figure 2 shows one real demonstration from each arm.

4 Experimental Setup

Task. Countdown: given a set of numbers and a target, output an arithmetic expression $(+, -, \times, \div)$ evaluating to the target. The verifier scores 1.0 (correct format and arithmetic), 0.1 (parseable `<answer>` with wrong arithmetic), or 0.0 (unparseable). **Demonstrations.** 393 verifier-checked DeepSeek-V3 chains ($\sim$131 per granularity arm), re-binned by measured step count. **Policy.** Qwen2.5-0.5B initialized from the IPO checkpoint. **RLOO config.** batch 128, group size 16, KL and entropy

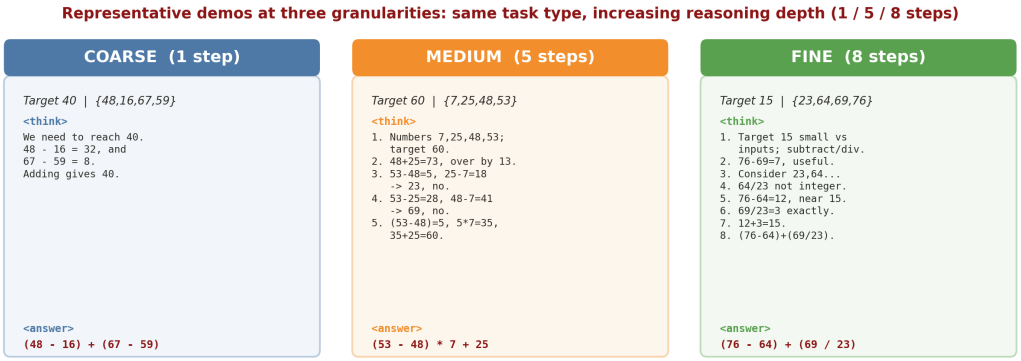

Figure 2: Representative expert demonstrations, one per granularity arm. All three solve a Countdown instance and terminate in a valid `<answer>`; only the depth of the `<think>` trace varies—from a single combined step (coarse) to an eight-step search (fine). These individual examples have $1/5/8$ reasoning steps; the arms' tercile-bin *mean* step counts are $\approx 2/5/14$. This is the structural axis our imitation term varies while holding the target answer fixed.

coefficients 10^{-3}, learning rate 10^{-5}, 60 training steps (run to a common checkpoint across all arms). **Baselines.** Our primary baseline is the β=0 run (pure RLOO at the same 60-step checkpoint), which isolates the effect of the imitation term. The upstream pipeline stages provide context for where this sits: on the same eval, pass@1 rises 0.28 (SFT) $\rightarrow$ 0.35 (IPO) $\rightarrow$ 0.55 (RLOO), confirming each stage is correctly implemented and clearing the project's expected thresholds before any imitation is added. **Sweep.** granularity $\{$coarse, medium, fine$\}$ $\times$ $\beta \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9\}$, plus a β=0 baseline (28 configurations). **Evaluation.** 50 held-out prompts $\times$ 16 samples, temperature 0.6; pass@k uses the unbiased estimator. We additionally measure arithmetic-error rate, the multi-answer rate (responses emitting more than one `<answer>`—a failure to commit to a single final answer), and mean output length. **Decoding.** Following the project's constrained-decoding specification, training rollouts are sampled at temperature 1.0 with top-k, top-p, and min-p disabled; evaluation uses temperature 0.6, top-k=20, top-p=0.95, min-p=0.

5 Results

Figure 3 gives a single-glance overview of the full sweep: final accuracy (top-left) is flat across β, whereas the three behavioral metrics (multi-answer rate, arithmetic error, output length) change sharply and granularity-dependently. We unpack each below.

5.1 Quantitative Evaluation

5.1.1 Final accuracy is insensitive to β and granularity

Across the full 3×9 grid, pass@1 remains within 0.55–0.64 and pass@16 within 0.66–0.80, centered on the β=0 baseline (pass@1 = 0.605, pass@16 = 0.760); see Figure 3 (top-left) and Table 1. No granularity arm consistently dominates, and no monotonic or inverted-U trend in β survives across arms. The spread is comparable to what a single-seed, 50-prompt evaluation would produce by chance, so we do not interpret the small fluctuations as a reliable accuracy effect. To quantify this, we bootstrap the 50 evaluation prompts (10,000 resamples): the baseline pass@1 of 0.605 carries a 95% confidence interval of $[0.49, 0.72]$, and *every* one of the 28 configurations falls inside it. This directly refutes our proposal's H1: there is no medium-granularity "sweet spot" in accuracy. Figure 4 plots all nine β values per arm against the baseline and its 95% CI band, making the flatness explicit: every arm wanders inside the band with no systematic ordering.

5.1.2 Imitation reshapes reasoning behavior at a granularity-dependent threshold

While accuracy is flat, the policy's *multi-answer* behaviour—emitting more than one `<answer>`, in the worst case deriving a correct equation mid-trace and then overwriting it with a later, incorrect

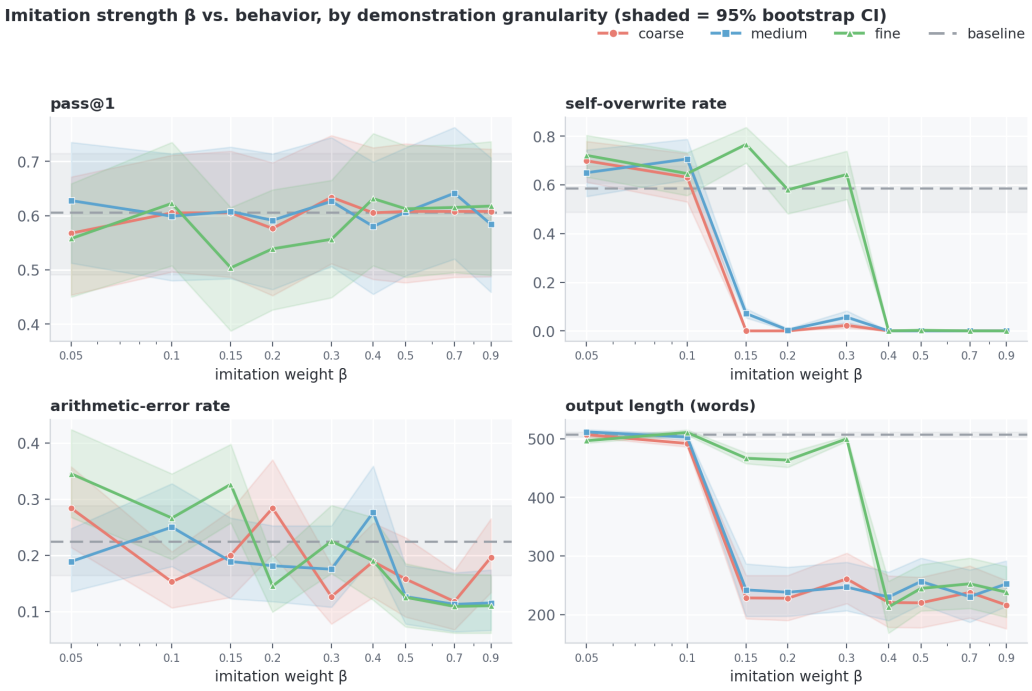

Figure 3: Effect of imitation weight β on accuracy and behavior, by demonstration granularity. **Top-left:** pass@1 is flat around the baseline. **Top-right:** the multi-answer rate collapses at a granularity-dependent β threshold. **Bottom-left:** arithmetic error declines with β. **Bottom-right:** output length collapses. Note the near-identical shape of the two right-hand panels—multi-answer rate and length fall together—which is the visual signature of the length confound discussed in Section 6. Shaded bands are 95% bootstrap confidence intervals over the 50 evaluation prompts (10,000 resamples); the dashed line and grey band are the β=0 baseline and its CI.

one—is strongly modulated by β, and the effect is sharply granularity-dependent (Figure 5). For *coarse* and *medium* demonstrations, the multi-answer rate collapses from ≈ 0.65 to ≈ 0 once $\beta \geq 0.15$. For *fine* demonstrations, the rate stays high—*above* the baseline—through $\beta = 0.3$ (e.g. 0.77 at β=0.15) and only collapses at $\beta = 0.4$. Intuitively, the clean single-answer structure of coarse/medium chains teaches the model to commit to one answer with little imitation pressure, whereas the multi-attempt structure of fine chains reinforces overwriting until much stronger imitation overrides it. This supports H2 (failure modes at the extremes) and refines H3: β does not move an accuracy optimum but moves the *threshold at which reasoning behavior changes*, and that threshold rises with granularity. As a robustness check, the stricter *self-overwrite* event used by our analysis code (`eval_extension.py`)—an earlier correct `<answer>` overwritten by a wrong *final* one—has much lower absolute rates (baseline 0.03, vs. 0.59 for the multi-answer rate) but shows the *same* granularity-dependent collapse (coarse/medium $\to$0 by β=0.15; fine holding at 0.07–0.12 through β=0.3), so the pattern is not an artifact of the looser metric; the rollout in Listing 1 is one such true self-overwrite.

5.1.3 Arithmetic errors and output length

Arithmetic-error rate tends to decline as β increases, most cleanly for medium and fine in the high-β regime (e.g. fine falls from 0.345 at β=0.05 to ≈ 0.11 at $\beta \geq 0.7$; Figure 3, bottom-left). Mean output length drops sharply—from $\sim$500 to $\sim$220 words—at the same β threshold where the multi-answer rate collapses (Figure 3, bottom-right; Table 1). The two right-hand panels of Figure 3 are nearly identical in shape: this coincidence is important for interpretation: shorter traces mechanically have less room to emit a second `<answer>`, so part of the multi-answer-rate collapse may be a by-product of length reduction rather than a direct behavioral change (Section 6).

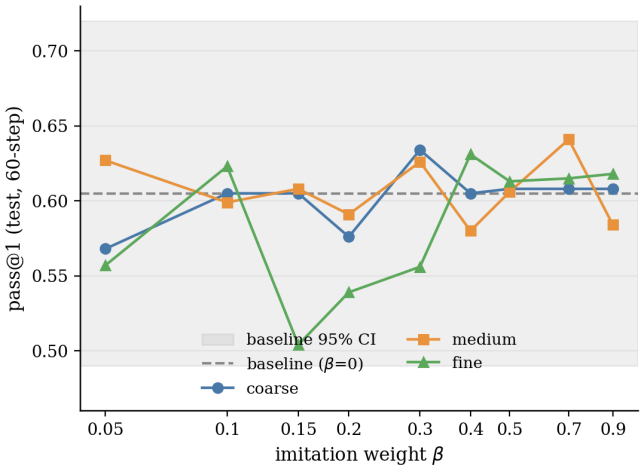

Figure 4: pass@1 vs. imitation weight β at the common 60-step checkpoint, by demonstration granularity. The dashed line is the β=0 baseline (0.605) and the grey band its 95% bootstrap CI ($[0.49, 0.72]$, 50 prompts, 10,000 resamples). All three arms fluctuate within the band across the full β range with no monotonic or inverted-U trend, so we read accuracy as insensitive to both β and granularity.

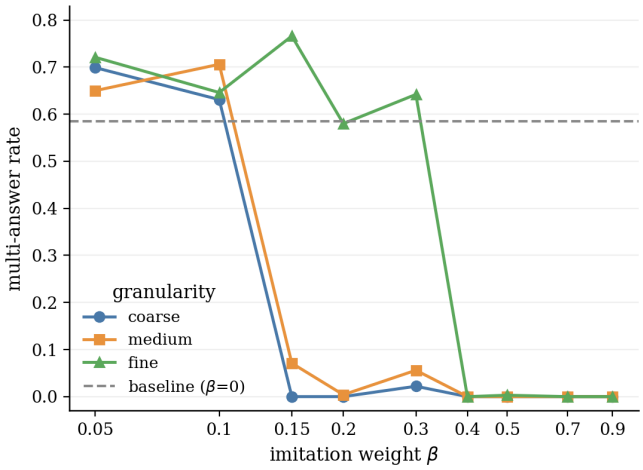

Figure 5: Multi-answer rate vs. β (shaded: 95% bootstrap CI over the 50 eval prompts). Coarse/medium collapse to ≈ 0 at β=0.15; fine persists above baseline until β=0.4. The post-collapse intervals are tight and do not overlap the pre-collapse ones, so the threshold is not an artifact of evaluation noise. The collapse threshold rises with demonstration granularity.

5.2 Qualitative Analysis

5.2.1 What self-overwrite looks like

The aggregate rates become concrete in individual rollouts. Listing 1 is a representative baseline (β=0) sample for `nums=[63,95,96]`, `target=64`: the policy derives a *correct* expression and explicitly "verifies" it, yet keeps reasoning and emits a *second*, arithmetically wrong `<answer>`—with a hallucinated verification claiming $63 \times 96/95 = 64$ (the true value is 63.66, and it even miscomputes 63×96). A single sample thus contains both failure modes we track, and is typical of the long, multi-answer traces that dominate at low β.

```
<think> [...] 96 - 95 = 1; 63 + 1 = 64. This works! </think>
<answer>63 + (96 - 95)</answer>            # evaluates to 64  (CORRECT)
<think> Let me verify: 96-95=1; 63+1=64. Perfect! </think>
<think> [...] Final attempt: 63 * 96 = 6016; 6016 / 95 = 64 (This works!) </think>
<answer>63 * 96 / 95</answer>              # evaluates to 63.66 (WRONG)
```

Listing 1: Baseline (β=0): a correct answer is overwritten by a later, wrong one (abridged; `[...]` marks elided reasoning).

Under strong imitation from *coarse* demonstrations (β=0.3), the same policy instead commits to a single answer and terminates (Listing 2): a short `<think>` block, one correct `<answer>`, and no second guess.

```
<think>
First attempt:
43 - 68 = -25
-25 + 45 = 20 (This works!)
</think>
<answer>43 - 68 + 45</answer>              # evaluates to 20  (CORRECT)
```

Listing 2: Coarse, β=0.3: the policy commits to a single correct answer and stops.

This is the trace-level signature of the behavioral shift in Section 6: imitation does not make the policy more *accurate*; it makes it *commit*—halting after the first answer rather than exploring on and overwriting a correct result.

Table 1: Full 3×9 grid at the common 60-step checkpoint, plus the β=0 baseline. $p@k$: unbiased pass@k; *a.err*: arithmetic-error rate; *s.ovr*: multi-answer rate; *len*: mean output length (words).

Config	p@1	p@16	a.err	s.ovr
baseline (β=0)	0.605	0.760	0.224	0.585
coarse β=0.05	0.568	0.760	0.284	0.699
coarse β=0.10	0.605	0.780	0.152	0.631
coarse β=0.15	0.605	0.740	0.200	0.000
coarse β=0.20	0.576	0.740	0.284	0.000
coarse β=0.30	0.634	0.800	0.126	0.022
coarse β=0.40	0.605	0.760	0.189	0.000
coarse β=0.50	0.608	0.720	0.158	0.000
coarse β=0.70	0.608	0.680	0.117	0.000
coarse β=0.90	0.608	0.780	0.196	0.000
medium β=0.05	0.627	0.800	0.189	0.650
medium β=0.10	0.599	0.740	0.250	0.706
medium β=0.15	0.608	0.700	0.189	0.071
medium β=0.20	0.591	0.700	0.181	0.004
medium β=0.30	0.626	0.740	0.175	0.056
medium β=0.40	0.580	0.680	0.276	0.000
medium β=0.50	0.606	0.740	0.126	0.000
medium β=0.70	0.641	0.720	0.113	0.000
medium β=0.90	0.584	0.700	0.115	0.000
fine β=0.05	0.557	0.740	0.345	0.721
fine β=0.10	0.623	0.760	0.266	0.646
fine β=0.15	0.504	0.660	0.326	0.766
fine β=0.20	0.539	0.720	0.145	0.580
fine β=0.30	0.556	0.740	0.225	0.642
fine β=0.40	0.631	0.760	0.190	0.000
fine β=0.50	0.613	0.740	0.125	0.003
fine β=0.70	0.615	0.720	0.109	0.000
fine β=0.90	0.618	0.720	0.110	0.000

6 Discussion

Imitation as a behavioral regularizer. The headline pattern is a dissociation: β leaves final accuracy essentially unchanged but strongly reshapes *how* the model reasons. On a 0.5B model and a task the IPO checkpoint already largely solves, the verifier reward appears to dominate the accuracy ceiling, leaving little room for imitation to raise pass@k. What imitation *does* change is the form of the trace: it reduces the multi-answer rate, lowers arithmetic-error rate, and shortens outputs. We therefore read off-policy imitation here as a behavioral regularizer rather than an accuracy lever.

The granularity-dependent threshold. The most robust finding is that the β needed to suppress multi-answer behaviour rises with demonstration granularity (coarse/medium ≈ 0.15, fine ≈ 0.4). This is consistent with the structure of the demonstrations: coarse chains model a single committed answer, so even weak imitation pushes the policy to stop early; fine chains contain multiple explore-then-revise steps, which resemble the overwriting behavior itself and thus require stronger imitation to override.

Limitations.

- **(1) Length confound.** Demonstration granularity correlates strongly with token length ($\mathrm{corr}(\text{steps}, \text{len}) = 0.89$), and empirically the multi-answer-rate collapse coincides with a drop in output length from $\sim$500 to $\sim$220 words. Because a shorter trace has less opportunity to emit a second `<answer>`, we cannot fully separate "imitation teaches commitment" from "imitation shortens outputs." A length-matched control is the clearest next step; below we partially disentangle the two from the existing data (Figure 6).
- **(2) Single seed, small eval.** Each configuration is a single run evaluated on 50 prompts. We quantify the resulting *evaluation* noise with 95% bootstrap confidence intervals throughout (Figures 3, 5), but a single *training* seed per configuration means the flat accuracy result should still be read as "no effect detectable above noise" rather than "provably no effect." As a partial check on training-seed variance, the independent length-penalized β=0 run in Appendix D—a separately initialized training run—reaches pass@$1 = 0.55$, well inside the baseline's 95% CI $[0.49, 0.72]$, consistent with the accuracy result being stable across training runs. While we did not repeat individual configurations across seeds, the behavioral findings rest on a consistent monotonic trend across 27 independently trained (g, β) runs—executed across three separate operators and machines—rather than on any single run, which makes a spurious single-seed artifact an unlikely explanation for the granularity-dependent threshold.
- **(3) Scale and task.** Findings are specific to a 0.5B policy on a verifier-solvable arithmetic task; larger models or harder tasks may leave more headroom for imitation to affect accuracy.

Disentangling the length confound. To test whether imitation reduces multi-answer behaviour *beyond* simply shortening outputs, we re-analyze all 22,400 evaluation samples at the response level (Figure 6). The multi-answer rate is strongly length-dependent—near zero for short traces ($<$ 300 words) and 0.55–0.76 for long traces (450–550 words)—confirming length as a major driver. Crucially, the dependence on β *survives* length matching: restricting to a common band (460–540 words), the multi-answer rate falls from 0.60 at β=0 to 0.46 at β=0.15, 0.35 at β=0.2, and essentially 0.00 at $\beta \geq 0.4$—that is, at β=0.4 even long (460–540 word) responses no longer overwrite (n=467). A logistic regression of the multi-answer outcome on (length, β) gives a large, significant negative β coefficient (-7.8, $p < 10^{-300}$) after controlling for length. We therefore read the effect as only *partially* confounded: shorter outputs mechanically reduce the opportunity to overwrite, but stronger imitation also drives the multi-answer rate to zero at fixed length—imitation teaches commitment, not merely brevity. A fully length-matched training control remains the cleanest way to close the residual gap; we outline below the training-side control that would close it by construction, together with a preliminary run.

Causal control: length without imitation (preliminary). The analysis above is observational—β and length co-vary, so it can suggest but not prove an effect of imitation beyond length. The clean causal test is to train β=0 (no-imitation) RLOO runs with an explicit length penalty added to the verifier reward, at penalty strengths chosen to push mean output length down into the high-β regime

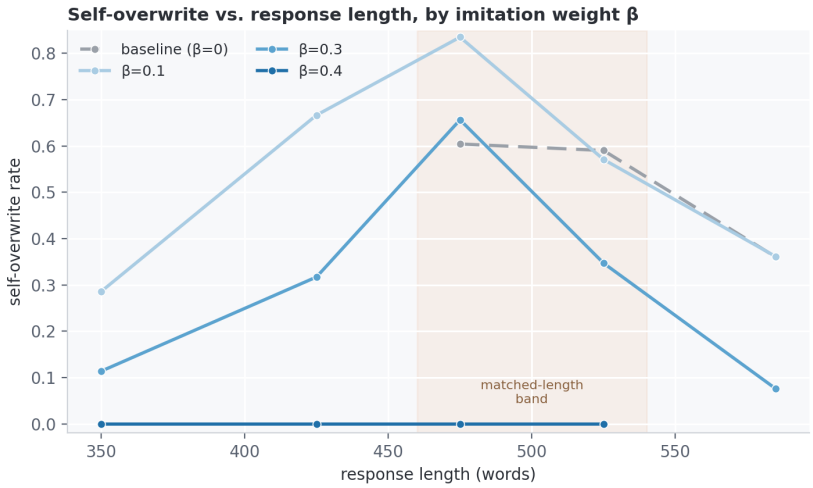

Figure 6: Multi-answer rate vs. response length, pooled over granularity and split by imitation weight β (all 22,400 samples). The multi-answer rate rises with length at low β (the confound), but the β=0.4 curve stays at ≈ 0 across *all* lengths—including the matched-length band—showing that strong imitation suppresses overwriting independently of length. Length bins with < 25 samples are omitted.

($\sim$220 words) while holding everything else fixed. These runs never see an expert demonstration, so they isolate the effect of *length alone*: if short outputs by themselves suppress multiple answers, the length-penalized policy should collapse just like the imitation runs; if imitation contributes something beyond brevity, it should retain an elevated multi-answer rate at matched length. We planned three such runs at increasing penalty strengths, but due to compute-quota and preemption limits only the weakest-penalty run completed; we report it as a preliminary single-point control in Appendix D rather than a finished result. As detailed there, that run's penalty was too weak to compress outputs, so it does not yet provide a matched-length comparison. For now, the load-bearing length-controlled evidence remains the observational fixed-length analysis of Figure 6 (logistic-regression β coefficient -7.8, $p < 10^{-300}$ at fixed length); a full matched-length control with stronger penalties and multiple seeds is our primary next step.

Broader impact. Shaping *how* a model reasons—not only whether it is correct—through cheap off-policy demonstrations could improve the readability and latency of reasoning traces without extra sampling, which matters where rollouts are expensive. The same lever is dual-use: optimizing the *form* of reasoning (shorter, more committed traces) can make outputs look more authoritative than the underlying computation warrants, so behavioral regularizers should be paired with calibration and faithfulness checks rather than treated as quality guarantees.

Future work. Multiple training seeds with error bars; a length-matched expert-demonstration set (rather than reward-side length control) to remove the confound by construction; a coherence / LLM-judge metric to assess reasoning quality beyond the verifier; and per-arm β scheduling that exploits the granularity-dependent threshold directly.

7 Conclusion

We studied demonstration granularity as a controllable axis in imitation-regularized RLOO. On Countdown with a 0.5B policy, off-policy imitation barely moves final accuracy but strongly and granularity-dependently reshapes reasoning behavior: the multi-answer rate collapses at a β threshold that rises with granularity, alongside falling arithmetic errors and shorter outputs. Rather than a granularity "sweet spot" for accuracy, we find that granularity controls *how much imitation pressure is needed to change reasoning form*—reframing demonstration granularity as a knob on behavior, not correctness.

8 Team Contributions

- **Yina Jian:** Proposed the project and led the extension end to end; designed the expert-demonstration pipeline (granularity prompt engineering, OpenRouter integration, post-hoc step-count re-binning); implemented the imitation-regularized RLOO objective; ran the main grid ($\beta \in \{0.05, 0.1, 0.3\}$) and the β=0 baseline to the common 60-step checkpoint; conducted the analysis and produced all figures; created the poster; wrote the report.
- **Kevin Song:** Implemented the IPO stage; ran the high-β sweep arm ($\beta \in \{0.5, 0.7, 0.9\}$) across all three granularities—nine configurations—to the common 60-step checkpoint, and exported the corresponding logs; led and delivered the poster presentation, and contributed to the poster.
- **Sukeerth Ramkumar:** Ran the lower-β sweep arm ($\beta \in \{0.15, 0.2, 0.4\}$) across all three granularities—nine configurations—to the common 60-step checkpoint, and exported the corresponding training and evaluation logs; presented at the poster session.

Changes from Proposal Two substantive changes.

- **(1)** We re-operationalized granularity from prompt-specified to *measured* step count, after a manipulation check showed prompting failed to separate coarse from medium.
- **(2)** Our headline result changed: the proposal anticipated a granularity sweet spot and a granularity-dependent accuracy optimum (H1/H3). The completed 60-step sweep does not support an accuracy effect; instead we find a granularity-dependent *behavioral* threshold, so we reframe the contribution around reasoning behavior rather than accuracy. We view this revision as a normal outcome of running the full experiment the proposal planned.

Evolution from the milestone. Our milestone poster reported an apparent "optimal-β shift with granularity" (coarse 0.634 / medium 0.627 / fine peak), read off a *step-40* checkpoint over only three β values ($\{0.05, 0.1, 0.3\}$). Carrying the runs to the common *step-60* checkpoint, expanding the sweep to the full nine-β grid (28 configurations), and adding bootstrap confidence intervals, that apparent accuracy trend *does not survive*: the differences lie within evaluation noise (all configs inside the baseline's 95% CI; Section 5). What *is* robust is the behavioral signal—the granularity-dependent collapse of the multi-answer / self-overwrite rate—which was already visible at the milestone and holds under both metric definitions. We report this transition explicitly: the negative accuracy result is a deliberate outcome of more data and better error control, not a softened version of the milestone claim.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to Basics: Revisiting REINFORCE-style Optimization for Learning from Human Feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Conference on Robot Learning*.

Zhi Jia et al. 2025. A Trajectory-Based Reward Model for Off-Policy Language Model Fine-Tuning. *arXiv preprint* (2025).

Yina Jian et al. 2026. Multi-Component Skill Decomposition for Long-Horizon Robotic Manipulation. *Preprint* (2026).

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *Proceedings of the 41st International Conference on Machine Learning* (2024).

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050* (2023).

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. AWAC: Accelerating Online Reinforcement Learning with Offline Datasets. *arXiv preprint arXiv:2006.09359*.

Yunhao Tang et al. 2025. Unifying On- and Off-Policy Data for Language Model Fine-Tuning. *arXiv preprint* (2025).

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Zheng Yang et al. 2024. A Survey on Synthetic Data Generation for Large Language Models. *arXiv preprint* (2024).

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. ProcessBench: Identifying Process Errors in Mathematical Reasoning. *arXiv preprint arXiv:2412.06559* (2024).

A Granularity Separation

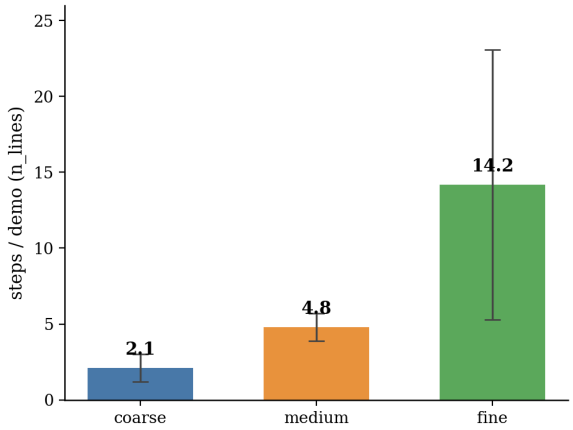

Figure 7: Granularity arms separated by measured reasoning steps after tercile re-binning (Cohen's $d = 2.94, 1.48$; no range overlap).

B Implementation Details

All runs use Qwen2.5-0.5B initialized from our IPO checkpoint, RLOO with batch 128, group size 16, learning rate 10^{-5}, KL and entropy coefficients 10^{-3}, trained for 60 steps. Demonstrations are 393 verifier-checked DeepSeek-V3 chains re-binned into terciles by line count. Evaluation uses 50 held-out prompts with 16 samples each at temperature 0.6; pass@k is the unbiased estimator. The *multi-answer rate* is the fraction of samples emitting more than one `<answer>` tag (a failure to commit to a single final answer). We additionally report the stricter *self-overwrite* event used by our `eval_extension.py`: a response whose earlier `<answer>` is correct but whose final one is wrong.

C Training Dynamics

Figure 8 shows the W&B training curves across all 28 configurations. They serve as a mechanism check: the policy's mean sequence log-probability on its own rollouts rises steadily (training is progressing), while the verifier reward climbs to a similar plateau (≈ 0.5–0.65) regardless of β or granularity—consistent with the β-insensitivity of final accuracy reported in Section 6. KL to the reference stays small (the 10^{-3} KL penalty keeps the policy close to the IPO checkpoint), and policy entropy decreases as training proceeds.

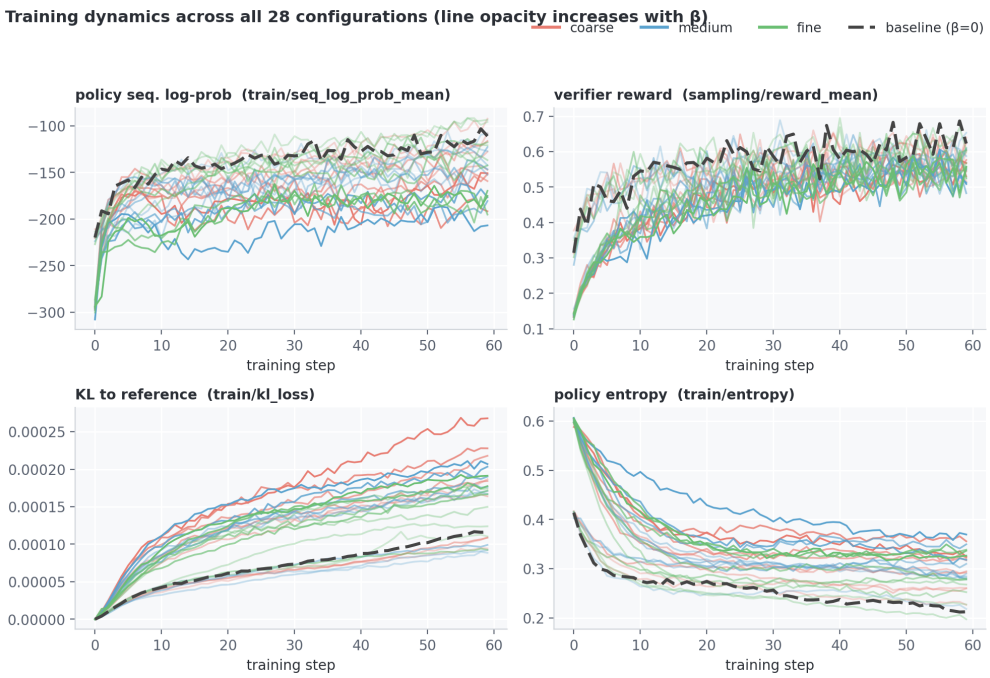

Figure 8: Training dynamics vs. step across all 28 configurations (color = granularity; line opacity increases with β; grey dashed = β=0 baseline). Panels: policy sequence log-probability on its own rollouts (`train/seq_log_prob_mean`), verifier reward (`sampling/reward_mean`), KL to reference (`train/kl_loss`), and policy entropy (`train/entropy`).

D Causal Control: Length Without Imitation (Preliminary)

To causally separate length from imitation, we planned three β=0 (no-imitation) RLOO runs with an explicit length penalty added to the verifier reward, at increasing penalty strengths chosen to push mean output length down into the high-β regime ($\sim$220 words). All other settings (policy, optimizer, KL/entropy coefficients, evaluation protocol) were held fixed, and none of these runs see an expert demonstration, so in principle they isolate the effect of *length alone*. Because of compute-quota and preemption limits, only the weakest-penalty run (penalty weight 0.8) reached the common checkpoint; the stronger-penalty runs crashed before completion and could not be re-run before the deadline. We therefore report a single-point control rather than the full matched sweep.

Evaluated on the same 50 held-out prompts $\times$ 16 samples, the completed run reached a mean output length of **503 words**—essentially unchanged from the β=0 baseline ($\sim$500 words) and far above the $\sim$220-word high-β regime we targeted—indicating that a penalty weight of 0.8 was too weak to compress outputs. At this unchanged length its multi-answer rate was **0.81**, comparable to (slightly above) the 0.585 of the long β=0 baseline (pass@1 $= 0.55$, pass@16 $= 0.74$, arithmetic-error rate $= 0.39$).

This run therefore does **not** provide a matched-length comparison: because output length never dropped, it cannot isolate length from imitation, and we do not read it as causal evidence either for or against an imitation effect. What it *does* confirm is the strong length–overwrite coupling documented in Figure 6—at $\sim$500 words the multi-answer rate sits squarely on the high end of the observed length-conditioned curve (0.55–0.76 for 450–550-word traces). The clean causal test still requires stronger length penalties that actually reach the short regime, run with multiple seeds; this is our primary next step. The load-bearing length-controlled evidence in this report remains the observational fixed-length analysis of Figure 6.