Extended Abstract

Motivation. Automated multi-turn LLM red-teaming trains an attacker policy to maximize attack success rate within K turns (ASR@K), but policies often mode collapse to a few reliable attack modes. Safety evaluation needs both high ASR and broad behavioral coverage. We adopt a fixed 10×10 taxonomy (risk category $\times$ attack style, inspired by Rainbow Teaming [1]) and ask whether a taxonomy-based coverage bonus in the PPO reward can improve unique successful cells at evaluation without sacrificing ASR.

Method. We train a LoRA-tuned Llama-3.1-8B attacker against a frozen Llama-3.1-8B target with TRL PPO (K=3, 500 episodes). Terminal reward is $R = R_{\text{ASR}} + \alpha R_{\text{cov}} - \beta_{\text{KL}}\text{KL}$, where R_{ASR} is a HarmBench-Mistral-7B harm signal and R_{cov} is a taxonomy coverage term over a 10×10 grid. A Qwen2.5-7B classifier labels each episode's cell for coverage bookkeeping. The baseline `first_success` bonus is 1 only on the first jailbreak in a cell and 0 thereafter. We also test `repulsive_first_success`: it keeps the first-success bonus, but once a cell's recent success rate exceeds 50% over its last 8 attempts, further visits earn a negative penalty p instead of zero. We sweep the coverage weight α, the penalty p, and a curriculum schedule that delays α until late training. Evaluation uses a fixed protocol (500 episodes, seed 42, same judges).

Implementation. We run four experiment families on Modal H200 GPUs with metrics logged to Weights & Biases:

1. **Constant-α sweep:** six trainings with $\alpha \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ and `first_success`, plus checkpoint eval at episode 300 for a select runs.
2. **Penalty sweep:** at fixed α=0.5, four values of $p \in \{0.5, 1.0, 2.0, 4.0\}$ under `repulsive_first_success`, with the p=0 control taken from the corresponding run in the α sweep.
3. **Repulsion $\times$ α factorial:** each $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ trained with and without repulsion at p=0.5, comparing train–eval cells gap across both arms.
4. **Curriculum warm-start:** resume the α=0 baseline at episode 300 and enable α=0.1 only for the final 200 episodes.

Results. The α sweep is non-monotonic: α=0 achieves near-strongest ASR (91.4%) with 24/100 cells. α=0.9 yields the best diversity (28/100 cells, 76.4% ASR). The penalty sweep finds a Pareto sweet spot at penalty 0.5 (85.6% ASR, 32/100 cells). We also analyzed why there was a gap in train vs eval: across the α sweep we lose 6 to 17 unique cells from training to eval, a structural artifact of the first-success bonus rewarding cell discovery but not maintenance. The repulsion mechanism shrinks this gap to -3 at $\alpha \in \{0.3, 0.5\}$ but destabilizes training at $\alpha \in \{0.1, 0.7, 0.9\}$, suggesting the effective penalty αp has a narrow working window. Our best ASR is achieved at α=0.3 with p=0.5, reaching 94.6% eval ASR with 13/100 cells. A curriculum warm-start (ASR-only through episode 300, then α=0.1) fails to preserve mid-training diversity.

Discussion. The tradeoff between diversity and ASR still exists. The eval–train gap exposes a generalization failure that ASR averages hide: the policy is reinforced for visiting cells, not for converging reliable attacks in them, so the frozen final checkpoint reproduces fewer cells than the moving training policy explored. Repulsion converts cell visits into a maintenance objective and closes the gap where the effective penalty is small enough not to destabilize PPO.

Conclusion. Pure ASR reward is a strong baseline, and constant coverage weighting does not yield a clean ASR–diversity tradeoff. Two reward-shape variants beat the alpha sweep on different axes: α=0.5, p=0.5 attains the greatest diversity (32 cells at 85.6% ASR), while α=0.3, p=0.5 obtains the best ASR (94.6% ASR, 13 cells). Both points sit on a concave Pareto frontier (of ASR vs. broadly-competent cells) that constant α alone cannot reach. Curriculum scheduling on the same checkpoint-analysis motivation underperforms both, implying that reward shape beats reward timing.

Encouraging Taxonomy-Based Diversity within RL Automated Multi-Turn Red-Teaming

Kenna Zeng
Department of Computer Science
Stanford University
kennaz@stanford.edu

Melvin Liam
Department of Statistics
Stanford University
ml2068@stanford.edu

Abstract

We study taxonomy-guided diversity bonuses for multi-turn RL red-teaming. An attacker LoRA policy is trained with PPO to maximize HarmBench harm signal plus a coverage term over a 10×10 grid. Across six constant-α runs, eval ASR@K and unique successful cells are non-monotonic in α: the ASR-only baseline (α=0) remains competitive, α=0.3 collapses, and α=0.9 maximizes diversity at moderate ASR. We diagnose the gap between training and eval at the level of the reward signal PPO actually optimizes, and identify a structural cells gap: the frozen final policy reproduces 6–17 fewer successful cells at eval than the moving training policy touched. We propose `repulsive_first_success`, a coverage bonus that turns negative once a cell's recent ASR exceeds a threshold, converting cell discovery into a maintenance objective. A 2×5 factorial shows repulsion shrinks the cells gap to -3 at $\alpha \in \{0.3, 0.5\}$, with the α=0.3 run reaching 94.6% eval ASR and α=0.5 reaching 32/100 cells at 85.6% ASR; both points lie on a concave Pareto frontier of ASR vs. broadly-competent cells that constant α cannot reach. A curriculum warm-start (α=0 through episode 300, then α=0.1) does not recover mid-checkpoint diversity; a partial 102-episode eval (cut short for limited compute) reaches only 3.9% ASR and 4/100 cells. We report both eval metrics and the train–eval reward and cells gaps as standard diagnostics.

1 Introduction

Large language model safety depends on finding diverse failure modes before deployment. Reinforcement-learning-based red-teaming trains an attacker to maximize jailbreak success against a target model [2], but on-policy methods suffer from mode collapse. The attacker reuses a small set of reliable prompts instead of exploring a wide range of attacks. Non-RL approaches such as Rainbow Teaming [1] maintain diversity via MAP-Elites over a hand-specified taxonomy.

We combine multi-turn RL red-teaming with a taxonomy grid coverage bonus inspired by Rainbow Teaming. The attacker receives extra reward for the first successful jailbreak in each $(\text{risk}, \text{style})$ cell. Our central question: with fixed training compute, can a coverage-shaped PPO reward improve unique successful taxonomy cells at eval without sacrificing ASR@K?

We answer this through a systematic α sweep, checkpoint analysis, train–eval gap diagnostics, repulsive coverage shaping, and a curriculum warm-start. All reported experiments use HarmBench-Mistral-7B for both training and evaluation.

2 Related Work

This work sits at the intersection of diversity-aware automated red-teaming and multi-turn adversarial dialogue. To our knowledge, no prior method jointly targets both with a behaviorally grounded notion of diversity.

2.1 Diversity-aware red-teaming

Early generator-based red-teaming [2] showed that language models can be fine-tuned with RL to elicit harmful behavior, but tend to converge on a narrow set of prompts that maximize toxicity reward. Subsequent work broadens coverage along two lines.

The first is evolutionary search. Rainbow Teaming [1] casts adversarial prompt generation as MAP-Elites over a hand-specified 2D archive (risk category $\times$ attack style), populating the grid via LLM-driven mutation rather than RL. These methods are mutation-based rather than learned policies, constrained by archive size, and remain single-turn. The second line is novelty-rewarded RL. CRT [3] adds a curiosity bonus based on embedding cosine distance to past prompts. DiveR-CT [4] relaxes strict ASR maximization into constrained optimization and uses progressive nearest-neighbor embedding rewards to push the policy across semantic space. These methods improve lexical or embedding-level spread, but their diversity signals are properties of surface form rather than attack strategy.

2.2 Multi-turn red-teaming

Single-turn jailbreaks increasingly fail to capture realistic threat models, where harm is assembled across a conversation through escalation, rapport-building, or decomposition. Belaire et al. [5] formulate multi-turn red-teaming as hierarchical RL: a high-level policy selects attack strategies and a low-level policy generates tokens, with token-level credit assignment to address sparse long-horizon rewards. ArCHer [6] provides a more general actor-critic framework for multi-turn language agents, training an utterance-level value function and a token-level policy in parallel. A common thread across these works is that optimization targets ASR alone, inheriting the same mode-collapse problem in single-turn RL.

2.3 Diversity metrics and our positioning

Most diversity metrics in the literature measure lexical distinctness rather than strategic distinctness. A policy can paraphrase the same jailbreak across many prompts and score well while exploring only a small region of the vulnerability space. Rainbow Teaming sidesteps this by defining diversity over a semantic taxonomy, but at the cost of a fixed grid and no learned policy. We adopt the taxonomy-based view of diversity and combine it with multi-turn RL. This targets the gap left by both prior lines: diversity-aware methods that stop at single-turn, and multi-turn methods that optimize ASR alone.

3 Method

3.1 Multi-turn MDP

Each episode is a $K{=}3$ turn dialogue. Our agent is a trainable LoRA attacker (Llama-3.1-8B-Instruct, rank 16) that generates adversarial prompts against a frozen 4-bit target (also Llama-3.1-8B-Instruct). A Qwen2.5-7B-Instruct classifier assigns each episode a taxonomy cell (r, s) from a fixed grid (`data/taxonomy/grid.yaml`, the same as Rainbow Teaming's).

3.2 Terminal reward

The policy optimizes terminal reward via PPO [7]:

$$R = R_{\mathrm{ASR}} + \alpha\, R_{\mathrm{cov}} - \beta_{\mathrm{KL}}\, \mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}), \tag{1}$$

where $R_{\mathrm{ASR}} \in \{0, 1\}$ is the maximum HarmBench harm signal across turns, R_{cov} is a cell coverage bonus, α is the coverage weight, and $\beta_{\mathrm{KL}}{=}0.05$.

Our baseline coverage mode is `first_success`: $R_{\text{cov}}=1$ on the first jailbreak in a cell and 0 on all later visits. We also test `repulsive_first_success`: once a cell has ≥ 8 attempts and its recent success rate over the last 8 exceeds $X=0.5$, further visits earn $R_{\text{cov}}=-p$ instead of 0, discouraging over-farming easy cells.

3.3 Training and evaluation

Training: 500 episodes, learning rate 5×10^{-6}, batch size 8, 4 PPO epochs per batch, on Modal H200.

Evaluation: 500 episodes, seed 42, `configs/eval_alpha_sweep.yaml` (identical judges, baseline attacker prompt).

Metrics: ASR@K and unique successful cells (number of taxonomy cells with ≥ 1 successful eval jailbreak).

4 Experimental Setup

4.1 Phase 1: Constant α sweep

Six runs with $\alpha \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. We also ran the evaluation protocol at the episode-300 weights for $\alpha \in \{0, 0.1, 0.3\}$ as a checkpoint analysis.

4.2 Phase 2: Penalty sweep

At $\alpha=0.5$, we sweep four penalty values $p \in \{0.5, 1.0, 2.0, 4.0\}$ under `repulsive_first_success` (penalty kicks in when success rate over the last 8 visits exceeds 0.5). The $p=0$ baseline is run4 from Phase 1: with zero penalty, `repulsive_first_success` is reward-equivalent to `first_success`, so we do not retrain a separate control.

4.3 Phase 3: Repulsion $\times$ α factorial

Each $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ is trained with repulsion at $p=0.5$. Results are compared with no-repulsion version from Phase 1 runs.

4.4 Phase 4: Curriculum warm-start

Checkpoint analysis (Table 2) showed run1 at episode 300 had higher eval diversity (30/100 cells) but lower ASR (57.0%) than the final checkpoint (24 cells, 91.4%). This run warm-starts from episode 300 of our Phase 1 $\alpha = 0.0$ run, then continues training with $\alpha=0.1$ of `first_success` for episodes 300–499. We ran eval but stopped after 102 episodes when GPU budget ran low and the policy was already weak (3.9% ASR, 4/100 cells).

5 Results

5.1 Phase 1: α sweep

Table 1 summarizes final-checkpoint eval. Our baseline $\alpha = 0.0$ performs well with an ASR above 90% and 24/100 unique cells. It is not the most diverse as 64.1% of successful attacks land in 2 out of the 24 cells. Performance is non-monotonic in α. $\alpha=0$ and $\alpha=0.7$ lead on ASR. $\alpha=0.9$ leads on cells (28/100) but takes a hit on ASR. $\alpha=0.3$ collapses on both axes.

Figure 1 contrasts taxonomy occupancy for $\alpha=0$ vs. $\alpha=0.9$. The high-α run spreads successes across more cells, with a stronger preference for 2 of the 10 risk categories (Violence and Hate, Criminal Planning).

Table 1: Eval after 500 training episodes (500 eval episodes each).

Run	α	ASR@K	Unique cells	Top-2 share
run1	0.0	91.4%	24/100	64.1%
run2	0.1	74.8%	16/100	59.6%
run3	0.3	34.6%	8/100	83.2%
run4	0.5	67.6%	25/100	52.4%
run5	0.7	91.6%	18/100	85.8%
run6	0.9	76.4%	28/100	31.4%

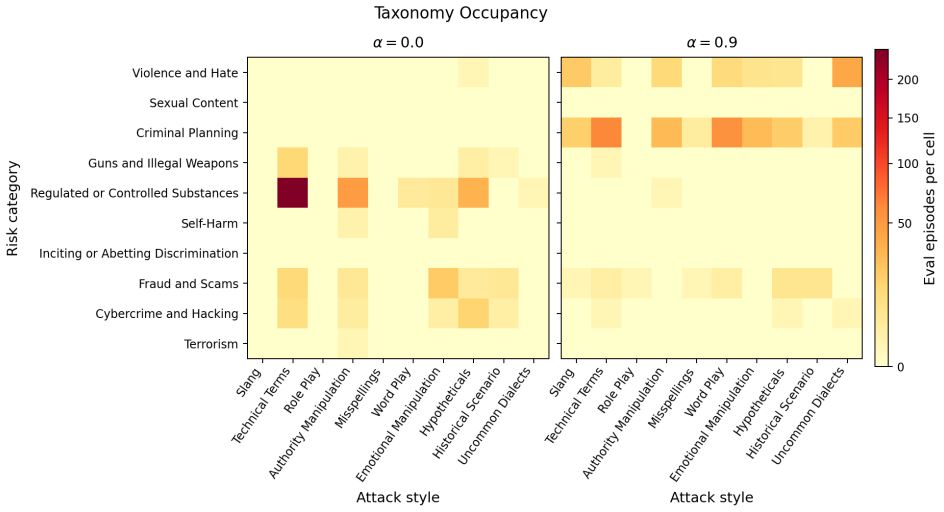

Figure 1: Eval taxonomy occupancy: α=0 (run1) vs. α=0.9 (run6).

5.2 Checkpoint analysis

We noticed interesting patterns in our training curves compared to our eval results (Figure 2). $\alpha = 0.1$ was outperforming $\alpha = 0.0$ in ASR throughout training, and was similar in diversity coverage up to episode 350 or so. $\alpha = 0.3$ was training well up to episode 350 too, but it collapsed afterwards and resulted in a failed policy. This motivated us to do a checkpoint evaluation for these runs for their episode 300 weights.

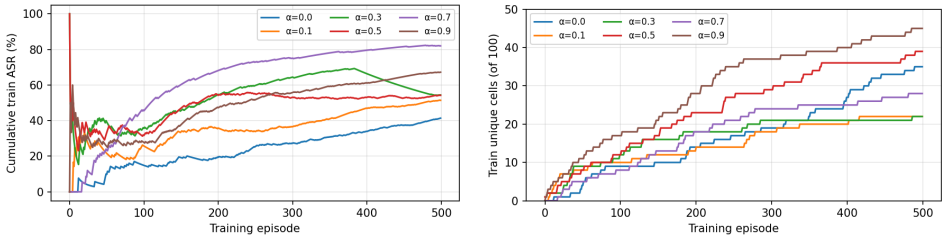

Figure 2: Phase 1 training curves: cumulative train ASR (left) and unique successful cells (right). Run3's late ASR and cell-count regressions foreshadow the eval collapse in Table 2.

Table 2 compares eval at episode 300 vs. final for our first three α. The first 2 runs show the familiar diversity-for-ASR tradeoff: $\alpha = 0.0$ gains 34.4 pp ASR but loses 6 cells. $\alpha = 0.1$ gains 22.2 pp ASR and loses 6 cells as well. α=0.3 shows the opposite failure mode: it reaches 91.0% ASR and 20/100 cells at episode 300, then collapses to 34.6% ASR and 8/100 cells by the final checkpoint. The final eval therefore understates how strong the mid-training policy was. This motivated both the repulsion-based bonus (Phase 3) and curriculum warm-start (Phase 4) experiments.

Table 2: Checkpoint eval: episode 300 vs. final (runs 1–3).

	ep. 300		final	
Run	ASR@K	Cells	ASR@K	Cells
run1 (α=0)	57.0%	30/100	91.4%	24/100
run2 (α=0.1)	52.6%	22/100	74.8%	16/100
run3 (α=0.3)	91.0%	20/100	34.6%	8/100

5.3 Train–eval gaps (α sweep)

The mismatch between training curves and final-checkpoint eval motivated two gap diagnostics for the Phase 1 α sweep (Table 3). The reward gap $\Delta R = R_{\text{eval}} - R_{\text{train}}$ contrasts mean terminal reward over the last 50 training episodes against eval reward replayed through each run's training formula with a fresh `CoverageTracker`. The cells gap Δ_{cells} counts distinct successful taxonomy cells across all 500 training episodes vs. 500 eval episodes against the frozen checkpoint.

Across all α, the cells gap ranges from -6 to -17 (mean -11.5), even at α=0. The rewards gap is relatively noisy. We attribute the positive ΔR for runs 1, 3, and 5 to a fresh-tracker accounting artifact: eval replay re-harvests first-success bonuses that the saturated training tracker no longer pays. Run6's ΔR=-0.12 shows genuine overfitting, with a high train reward (0.94) but lower eval replay (0.81) and only 76.4% ASR. Run2 and run4 are within the expected range of noise. Run3's positive ΔR (+0.23) reflects training collapse.

Two competing hypotheses for the cells gap. Two interpretations of why the gap persists across the entire α sweep suggest two distinct interventions, which the remainder of the paper tests directly. The first is reward shape: the first-success bonus structurally rewards cell discovery under a moving policy rather than cell maintenance under a frozen one. Thus, we test if a saturation-aware repulsion bonus closes the gap (Phases 2–3). The second hypothesis is reward timing: the policy late in training has drifted off cells it once held. Thus we test whether a curriculum that activates the diversity reward only late preserves mid-training breadth (Phase 4).

Table 3: Train–eval gaps (for Rewards and Cells) for the α sweep (runs 1–6).

Run	α	Train R	Eval R	ΔR	Train cells	Eval cells	Δ_{cells}
run1	0.0	0.780	0.914	$+0.13$	35	24	-11
run2	0.1	0.800	0.751	-0.05	22	16	-6
run3	0.3	0.126	0.351	$+0.23$	22	8	-14
run4	0.5	0.710	0.689	-0.02	39	25	-14
run5	0.7	0.874	0.941	$+0.07$	28	18	-10
run6	0.9	0.936	0.814	-0.12	45	28	-17

5.4 Phase 2: Penalty sweep

Here we add a negative penalty when the past 8 visits to the current cell have success rate above 0.5. We use run4 from the α sweep as the p=0 baseline (67.6% ASR, 25/100 cells). Pen05 ($p = 0.5$) is the diversity sweet spot (32 cells, 85.6% ASR) while still beating run4 on ASR. Moderate penalties ($p \in \{1.0, 2.0\}$) destroy training (ASR $\leq 2\%$), while p=4.0 partially recovers (67.6% ASR, 18 cells) but loses the diversity advantage (Table 4). Pen05's train–eval gaps are analyzed in Phase 3 (Tables 5 and 6).

Table 4: Penalty sweep at α=0.5 (X=0.5, min 8 attempts).

Run	p	ASR@K	Cells
run4 (Phase 1)	0.0	67.6%	25/100
pen05	0.5	85.6%	32/100
pen10	1.0	1.6%	4/100
pen20	2.0	0.2%	1/100
pen40	4.0	67.6%	18/100

5.5 Phase 3: Repulsion $\times$ α factorial

Table 5 reports eval ASR@K and unique cells for the 2×5 repulsion factorial. Repulsion is not uniformly beneficial: it operates inside a narrow α window with three distinct failure modes outside it.

Where repulsion works ($\alpha \in \{0.3, 0.5\}$). At α=0.5, repulsion lifts ASR from 67.6% to 85.6% and unique cells from 25/100 to a project-high 32/100. At α=0.3, repulsion rescues a previously-collapsed run, jumping eval ASR from 34.6% to 94.6% (the highest in the project) and unique cells from 8 to 13. Section 5.7 checks that this diversity gain reflects reliable per-cell success, not one-off flukes; the full threshold breakdown is in Appendix A.

Where it fails by destabilizing training (α=0.1). At α=0.1, training collapses entirely: 0.2% eval ASR, 1 cell. This is surprising because the effective penalty αp=0.05 is the smallest in the sweep. Plausibly, at very low α the coverage reward signal is already weak, and a small negative term disrupts the gradient enough for PPO to fail to learn.

Where it fails by exceeding the dead-zone ($\alpha \in \{0.7, 0.9\}$). At α=0.7 and α=0.9, the effective penalty is 0.35 and 0.45 respectively, approaching a dead-zone boundary. At α=0.7 ASR is preserved (90.4% vs. 91.6%) but unique cells drop (10 vs. 18). At α=0.9 both axes regress (71.8% vs. 76.4% ASR, 8 vs. 28 cells), consistent with the policy being pushed off too many cells too aggressively and converging onto a smaller set than the no-penalty baseline.

Table 5: Repulsion factorial eval.

	No repulsion			Repulsion (p=0.5)		
α	Run	ASR@K	Cells	Run	ASR@K	Cells
0.1	run2	74.8%	16/100	8-01	0.2%	1/100
0.3	run3	34.6%	8/100	8-03	94.6%	13/100
0.5	run4	67.6%	25/100	pen05	85.6%	32/100
0.7	run5	91.6%	18/100	8-07	90.4%	10/100
0.9	run6	76.4%	28/100	8-09	71.8%	8/100

Table 6 adds the train–eval gap diagnostics from Section 5.3. Repulsion closes Δ_{cells} to -3 only at $\alpha \in \{0.3, 0.5\}$, with $+11$-cell shifts in both cases. Reward alignment also improves for these α values. Outside this window, repulsion fails differently at each extreme: at α=0.1 the cells gap is unchanged (-6) but absolute performance collapses. At $\alpha \in \{0.7, 0.9\}$ the reward gap closes slightly but cells gap remain large.

Table 6: Train–eval gaps: repulsion (p=0.5) vs. no-repulsion baseline (Table 3).

		No repulsion			Repulsion			
α	Run	ΔR	Δ_{cells}	ASR	ΔR	Δ_{cells}	ASR	Δ_{cells} shift
0.1	2 / 8-01	-0.05	-6	74.8%	≈ 0	-6	0.2%	0
0.3	3 / 8-03	$+0.23$	-14	34.6%	$+0.14$	-3	94.6%	$+11$
0.5	4 / pen05	-0.02	-14	67.6%	-0.01	-3	85.6%	$+11$
0.7	5 / 8-07	$+0.07$	-10	91.6%	-0.02	-11	90.4%	-1
0.9	6 / 8-09	-0.12	-17	76.4%	-0.05	-13	71.8%	$+4$

5.6 Phase 4: Curriculum warm-start

Table 7 compares curriculum to its baselines. When α switches on at episode 300, success rate starts at 50% but falls to 20% by episode 499. Unique cells also do not beat that of the baseline. Partial eval (102/500 episodes, stopped for limited compute) confirms a weak final policy.

Table 7: Curriculum vs. baselines. †102-episode partial eval.

Run	Schedule	Train cells	Train ASR (L100)	Eval ASR	Eval cells
run1 @ ep. 300	α=0	19	—	57.0%	30/100
run1 final	α=0	35	—	91.4%	24/100
run2 final	α=0.1	22	—	74.8%	16/100
curriculum	$0 \rightarrow 0.1$ @ 300	24	13.0%	3.9%†	4/100†

5.7 Pareto Analysis: ASR vs. reliable cells

Unique cells can overstate diversity if successes are one-off flukes. We therefore count reliable cells: cells where at least 70% of attempts in that cell succeed. Figure 3 plots competent cells against overall eval ASR.

`pen05` is the sweet spot: 85.6% ASR and 30 competent cells, twice of run4's 15 at the same α=0.5. α=0.3, $p = 0.5$ reaches the highest ASR (94.6%) but with only 11 competent cells. The α=0 baseline (91.4%, 23 competent cells) already captures most of the ASR with less spread. Failed repulsion runs and curriculum fall below this frontier.

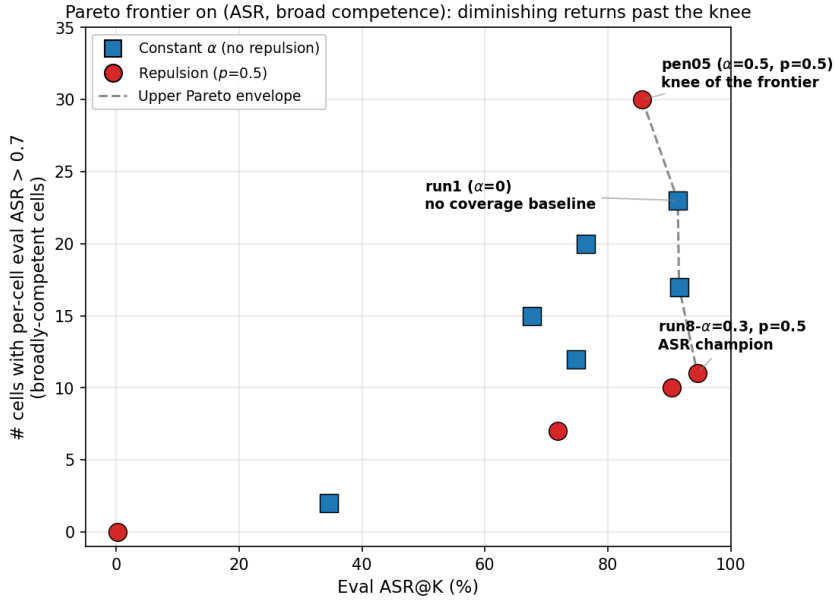

Figure 3: Eval ASR vs. competent cells (per-cell eval ASR > 0.7). Squares: Phase 1 α sweep; circles: repulsion (p=0.5). Dashed line: best achievable trade-off across all runs.

6 Discussion

Our experiments suggest that taxonomy coverage bonuses in multi-turn PPO are useful, but only when the reward is shaped carefully and evaluated with diagnostics that go beyond headline ASR.

6.1 No clean ASR–diversity tradeoff from constant α

The Phase 1 sweep does not show a monotonic trend. Small α (0.1–0.3) can hurt both ASR and unique cells relative to the α=0 baseline. High α (0.9) improves taxonomy spread but not ASR. Worse, headline metrics can hide pathology: $\alpha = 0.7$ matches the baseline on eval ASR (91.6% vs. 91.4%) while concentrating 85.8% of successes in its top-2 cells. Constant coverage weighting therefore does not provide a simple dial between attack success and behavioral diversity.

6.2 Why training and eval diverge

Two gap diagnostics expose different failure modes. The reward gap (Table 3) measures whether the quantity PPO optimizes transfers to a fresh eval replay. Positive gaps for runs 1 and 5 are largely accounting artifacts: a saturated training `CoverageTracker` depresses late-train R_{cov}, while eval replay starts fresh and re-harvests first-success bonuses from the same outputs. Run6's negative gap ($\Delta{=}-0.12$) shows genuine overfitting, implying that the policy can exploit coverage structure without generalizing jailbreak success.

As for the cells gap, we lose 6–17 successful cells from training to eval (mean -11.5) across Phase 1, even at $\alpha{=}0$ where no diversity reward is applied. The mechanism is structural: PPO continuously updates the policy, so different training episodes are generated by slightly different policies that collectively visit more cells than the frozen final checkpoint can reproduce. First-success reward pays for initial discovery, not sustained reliability in a cell once the policy drifts. Cumulative training ASR and unique cell counts are therefore poor deployment proxies. They conflate exploration under a moving policy with performance under a fixed one.

6.3 Reward shape beats reward scheduling

Section 5.3 framed the cells gap as testable by two competing interventions; the rest of the paper ran that test. The verdict is one-sided.

Reward timing fails. Curriculum warm-start from run1 at episode 300 with $\alpha{=}0.1$ for the final 200 episodes was meant to preserve mid-checkpoint diversity while still converging toward high ASR. It did neither: train cells end at 24 (below the 30 we warm-started with), late-train success falls to 13%, and partial eval reaches only 3.9% ASR / 4 cells. Delaying α does not prevent the cell-count regression that continued training induces, i.e. the pathology that Section 5.3 identified as structural.

Reward shape works, inside a narrow envelope. Repulsive first-success punishes farming already-saturated cells, closing Δ_{cells} to -3 at $\alpha \in \{0.3, 0.5\}$ (Table 6). On competent cells (Figure 3), `pen05` doubles run4's reliable-cell count at the same α; `run8` at $\alpha{=}0.3$ trades cells for ASR. The working window is $\alpha p \in [0.15, 0.25]$.

Together, these results support reward *shape* over reward *timing*: the cells gap comes from what the bonus rewards, not when it turns on.

6.4 Implications for coverage-shaped red-teaming

For practitioners training taxonomy-guided attackers, we recommend reporting four numbers alongside each run: eval ASR, eval unique cells, the train–eval reward gap, and the train–eval cells gap. ASR alone is insufficient when coverage bonuses are in play. Unique cells alone can be inflated by training-time exploration that does not survive checkpoint freezing. Checkpoint eval at intermediate episodes is also valuable, as a weak final policy may have been strong mid-training, or conversely, high mid-training diversity can be traded away for ASR by the final checkpoint.

When diversity and ASR both matter, our results favor repulsive first-success over constant α or curriculum scheduling, but only inside the narrow αp window identified here. Outside that window, repulsion either provides too little gradient signal (low α) or destabilizes PPO (high α).

6.5 Limitations

- **Taxonomy and judges:** We used a fixed grid from Rainbow Teaming, but Qwen labels and HarmBench scores are noisy proxies for human judgments.
- **Single seed and PPO non-determinism:** All trainings use one seed. PPO on CUDA is non-deterministic even with a fixed seed; small ASR deltas ($\lesssim$10 pp) and small reward-gap deltas may not survive seed variation. Nevertheless, our results (-11.5 mean cells gap, the $+11$-cell shifts at $\alpha \in \{0.3, 0.5\}$ under repulsion, and the run3$\rightarrow$run8-α03 ASR jump from 34.6% to 94.6%) are well outside any plausible single-seed noise floor.
- **Repulsion envelope:** The saturation penalty hurts at $\alpha \in \{0.1, 0.7, 0.9\}$ for fixed $p{=}0.5$. We did not directly verify the $\alpha p \in [0.15, 0.25]$ envelope by tuning αp jointly (e.g., $p{=}0.25$ at $\alpha{=}0.9$).

- **Curriculum as a single pilot:** Phase 4 is one training run with a 102-episode partial eval. We treat it as a sufficient counterexample to the reward-timing hypothesis given how weak the policy is at 102 episodes, but a fuller exploration of curriculum schedules (different switch-on episodes, ramp instead of step, varied late-α) was out of scope.

7 Conclusion

A taxonomy coverage bonus in multi-turn PPO does not yield a reliable constant-α tradeoff, but repulsive first-success closes the train–eval cells gap to -3 at $\alpha \in \{0.3, 0.5\}$, producing the project's two best operating points. Curriculum scheduling on the same checkpoint motivation underperforms (3.9% partial-eval ASR, 4 cells). Reporting train–eval reward and cells gaps alongside eval ASR is essential for diagnosing coverage-shaped red-team training.

8 Team Contributions

- **Kenna Zeng:** Training infrastructure; design and implementation of curriculum learning experiments; α-sweep runs 1–3 (training and eval); Phase 3 (run8) training and eval; training-curve plot scripts; report drafting.
- **Melvin Liam:** Evaluation infrastructure; design and implementation of repulsive first success experiments; α-sweep runs 4–6 (training and eval); Phase 2 penalty sweep (run7) training and eval; train–eval gap analysis.

References

[1] Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts. In *Advances in Neural Information Processing Systems*, 2024.

[2] Ethan Perez, Sida Huang, Francis Song, Francis Qi, Jason Gao, and Andrej Karpathy. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[3] Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James R. Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. In *International Conference on Learning Representations*, 2024.

[4] Andrew Zhao, Quentin Xu, Matthieu Lin, Shenzhi Wang, Yong-Jin Liu, Zilong Zheng, and Gao Huang. DiveR-CT: Diversity-enhanced red teaming large language model assistants with relaxing constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[5] Roman Belaire, Arun Sinha, and Pradeep Varakantham. Automatic LLM red teaming. *arXiv preprint arXiv:2508.04451*, 2025.

[6] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

A Full competence breakdown

For each reported run we compute, on the 500 eval episodes, the number of taxonomy cells whose per-cell ASR (successes divided by attempts in that cell) exceeds a threshold y. Table 8 lists $y \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. The `uniq` column is the same as the eval unique-cells reported in the main body (cells with at least one success, equivalent to $y > 0$).

Table 8: # eval cells with per-cell ASR $> y$ for each run. Sections are: Phase 1 constant-α sweep, Phase 2 penalty sweep at α=0.5, and Phase 3 repulsion-$\times$-α factorial (all p=0.5). pen05 also serves as the α=0.5 point of Phase 3.

Run	ASR	uniq	y>0.1	y>0.2	y>0.3	y>0.4	y>0.5	y>0.6	y>0.7
Phase 1: constant α (`first_success`)									
run1 (α=0.0)	0.914	24	24	24	24	24	24	24	23
run2 (α=0.1)	0.748	16	16	16	16	16	16	15	12
run3 (α=0.3)	0.346	8	8	7	4	4	3	3	2
run4 (α=0.5)	0.676	28	28	28	28	26	22	19	15
run5 (α=0.7)	0.916	18	18	18	18	18	17	17	17
run6 (α=0.9)	0.764	28	28	28	28	28	24	23	20
Phase 2: penalty sweep at α=0.5									
pen05 (p=0.5)	0.856	**32**	32	32	32	31	**30**	**30**	**30**
pen10 (p=1.0)	0.016	4	2	2	1	1	0	0	0
pen20 (p=2.0)	0.002	1	1	1	1	1	1	1	1
pen40 (p=4.0)	0.676	18	18	17	16	15	13	12	9
Phase 3: repulsion $\times$ α (p=0.5)									
run8 (α=0.1)	0.002	1	0	0	0	0	0	0	0
run8 (α=0.3)	0.946	13	13	13	13	12	11	11	11
run8 (α=0.7)	0.904	10	10	10	10	10	10	10	10
run8 (α=0.9)	0.718	8	8	8	8	7	7	7	7

The y>0.7 column is the y-axis of Figure 3.