EXPLORATION-PRESERVING POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards improves reasoning, while the allocation of learning signal shapes which solutions remain accessible under repeated sampling. Group-relative objectives assign equal advantages to equally rewarded responses, making aggregate credit proportional to sampled mode frequency. We introduce Exploration-Preserving Policy Optimization (ExPPO), a lightweight advantage-shaping rule that redistributes credit using prompt-relative, length-normalized response surprisal and prompt pass rate. ExPPO combines bounded shaping with shared normalization to preserve verifier polarity and approximately maintain each prompt group's total absolute sequence-advantage mass. Our analysis characterizes response-level credit allocation alongside sampled mode updates, deriving local conditions for gains in entropy and correct-mode discovery. Experiments show improved in-domain and out-of-domain reasoning coverage, higher aggregate response accuracy, and strong coverage at large sampling budgets. A controlled multi-answer evaluation further demonstrates increased correct-mode yield and gains in diversity among verified-correct responses.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) post-trains language models from automatically checked outcomes and has produced strong gains in mathematical reasoning (DeepSeek-AI et al., 2025). At each update, the policy samples several responses for each prompt and learns from their sequence-level verifier rewards. Reasoning systems often use repeated sampling, so their utility depends on both single-sample accuracy and the probability of finding a correct solution across K samples. Recent studies report that RLVR can improve pass@1 while reducing pass@K or answer diversity (Yue et al., 2025; Wu et al., 2025).

Group Relative Policy Optimization (GRPO) is widely used for RLVR because group-normalized rewards replace PPO's learned critic (Schulman et al., 2017; Shao et al., 2024). Equal-reward responses receive equal sequence advantages, so aggregate coefficient mass follows sampled mode frequency. Proposition 5 characterizes the corresponding relative-odds change in a categorical mode model, and finite groups give rare correct modes fewer updates (Plyusov et al., 2026). GRPO variants share this credit structure (Yu et al., 2025; Zheng et al., 2025), and recent analyses connect it to support concentration and entropy decline (Lochab et al., 2026; Gai et al., 2025). Cui et al. (2025) connect entropy change to log-probability–advantage covariance and propose Clip-Cov and KL-Cov to restrict high-covariance token updates. These controls target entropy; useful exploration also depends on allocating credit across distinct correct and incorrect responses.

Recent work addresses exploration through entropy control, collection-level objectives, and rarity-aware credit. Direct entropy bonuses have task-dependent benefits in MuJoCo (Andrychowicz et al., 2021; Choe and Kim, 2024) and show strong coefficient sensitivity in reasoning LMs (Cui et al., 2025). For GRPO-H, the bonus promotes token uncertainty independently of verifier credit (Appendix D.2). Entropy methods regulate policy uncertainty; REPO-R uses token-level surprisal to rescale positive and negative advantages toward an entropy target (Petrenko et al., 2026; Zhang et al., 2025; 2026; Cheng et al., 2026). Pass@K and set-level methods optimize the collective utility of sampled responses (Walder and Karkhanis, 2025; Chen et al., 2025; Orney et al., 2026). Rewarding the Unlikely emphasizes low-probability correct responses, F-GRPO scales prompt groups by their success rates, and correct-set methods redistribute credit across successful solutions (He et al., 2025; Plyusov et al., 2026; Cao et al., 2026). These approaches motivate prompt-relative control over how each group's learning signal is allocated across responses.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce **Exploration-Preserving Policy Optimization (ExPPO)**, a prompt-relative, mass-normalized advantage-shaping rule for group-relative RLVR. ExPPO combines group-standardized length-normalized rollout surprisal, a bounded transformation, pass-rate gating, and shared absolute-advantage-mass normalization. The rule preserves verifier polarity and assigns higher signed credit to higher-surprisal responses within each equal-advantage class. Using existing rollout quantities, it integrates into GRPO and DAPO.

We derive sampled mode-logit updates for GRPO and ExPPO in a common geometry and characterize response-level allocation with a KL-local model. The analysis establishes polarity preservation and surprisal ordering, together with conditional local results for entropy and correct-mode discovery. Experiments with Qwen2.5-7B, Qwen3-0.6B, and Phi-3.5-mini measure response accuracy, large-K coverage, entropy dynamics, and diversity among verifier-checked answer modes.

Our contributions are threefold:

- We characterize sampled mode-level credit allocation in group-relative RLVR and its dependence on response multiplicity.
- Our method, ExPPO, combines bounded response surprisal, pass-rate gating, and absolute-advantage-mass normalization in a prompt-relative redistribution rule.
- We evaluate accuracy, coverage, and correct-mode diversity, including closest-baseline comparisons and independently replicated Phi-3.5-mini training.

2 Preliminaries

Reinforcement learning with verifiable rewards. For a prompt q from training distribution $\mathcal{D}$, a frozen rollout policy π_{old} independently samples G responses. We use Y_i for a random response and $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ for its realized value, with length T_i and token context $x_{i,t} = (q, y_{i,<t})$. A verifier assigns reward $r_i = R(q, y_i)$. The current policy π_θ learns from these sequence rewards.

Group-relative policy optimization. GRPO assigns the detached sequence advantage

$$A_i = \frac{r_i - \mu_{r,q}}{\sigma_{r,q} + \varepsilon}, \qquad \mu_{r,q} = \frac{1}{G}\sum_{j=1}^{G} r_j, \qquad \sigma^2_{r,q} = \frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_{r,q})^2, \tag{1}$$

where $\varepsilon > 0$ is a numerical stabilizer (Shao et al., 2024). The same A_i applies to every token in y_i. Writing $\mathcal{B}$ for prompt/group sampling and $\langle\cdot\rangle$ for token averaging, GRPO and its entropy-regularized baseline, GRPO-H, maximize

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\mathcal{B}}[\langle \ell_{\text{clip}}(w_{i,t}(\theta), A_i) - \beta_{\text{KL}} d_{i,t}(\theta)\rangle], \tag{2}$$

$$\mathcal{J}_{\text{GRPO-H}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) + \beta_{\text{ent}}\mathbb{E}_{\mathcal{B}}[\langle H(\pi_\theta(\cdot \mid x_{i,t}))\rangle]. \tag{3}$$

Here $w_{i,t}$ is the token importance ratio, ℓ_{clip} the clipped surrogate, $d_{i,t}$ the reference-policy KL, and H Shannon entropy (Schulman et al., 2017; Mnih et al., 2016). Details appear in Appendix D.

3 From Reward Credit to Support Allocation

This section examines sampled mode-level coefficients and motivates the allocation rule in Section 4.

Equal reward credit gives multiplicity-weighted coefficients. Under binary rewards, every correct response in a mixed group receives the same positive GRPO advantage A^+, and every incorrect response receives the same negative value A^-; the closed form is given in Appendix C. Suppose a correct response mode m appears n_m times. Its aggregate positive coefficient is

$$C_m = \sum_{i:y_i \in m} A_i = n_m A^+. \tag{4}$$

Equal per-response credit therefore produces aggregate mode-level credit proportional to sample multiplicity. The categorical mode update in Proposition 5 converts a count difference into the relative log-odds shift $\eta A^+(n_m - n_{m'})$ between two correct modes m and m'. Recurring failure modes receive the same exposure effect on the negative side. Finite-group analyses show that rare-correct trajectories also enter fewer updates (Plyusov et al., 2026).

Global entropy combines correct and failure diversity. PPO/TRPO-style entropy bonuses reward uncertainty at sampled states; MuJoCo studies report task-dependent gains and benefits from trajectory-level entropy credit (Andrychowicz et al., 2021; Choe and Kim, 2024). For GRPO-H, the bonus favors unlikely tokens regardless of correctness, while equal-reward responses retain equal GRPO advantages. Entropy can therefore grow through failed continuations (Appendix D.2). To connect this signal to correct support, let $Y \sim \pi(\cdot \mid q)$ be a random response and $C = R(q, Y) \in \{0, 1\}$ its verifier-defined success indicator. Writing $p_q = \mathbb{P}(C = 1 \mid q)$, the entropy chain rule gives

$$H(Y \mid q) = H_b(p_q) + p_q H(Y \mid C = 1, q) + (1 - p_q) H(Y \mid C = 0, q). \tag{5}$$

where H_b is binary entropy. The three terms measure uncertainty over success, diversity among correct responses, and diversity among failures.

The decomposition shows that global entropy combines these three sources of policy spread. Consider two four-mode policies with success probability 0.5. One assigns correct masses $\{0.49, 0.01\}$ and failure masses $\{0.25, 0.25\}$; the other swaps these pairs. The policies share global entropy and total correct mass; the second distributes correct probability evenly across both valid modes. Correct-mode diversity therefore complements global entropy.

4 Exploration-Preserving Policy Optimization

ExPPO adds a detached correction to the verifier advantage: compare response surprisal within each prompt, use pass rate to allocate the correction, and normalize the shaped advantages. Table 1 summarizes the resulting credit allocation.

Prompt-relative response surprisal. For response $y_i = (y_{i,1}, \ldots, y_{i,T_i})$, we use its length-normalized surprisal under the rollout policy,

$$s_i = -T_i^{-1} \log \pi_{\text{old}}(y_i \mid q), \qquad u_i = \tanh\left(\frac{s_i - \mu_{s,q}}{\max(\sigma_{s,q}, \varepsilon)} \right). \tag{6}$$

where $\mu_{s,q}$ and $\sigma_{s,q}$ are computed across the G responses for prompt q. The score $u_i \in (-1, 1)$ indicates whether length-normalized surprisal is above or below the group mean. We use *response rarity* as shorthand for this policy-relative statistic; the controlled evaluation defines semantic modes through exact answer identity. Length normalization controls length effects; group standardization puts responses on a prompt-relative scale; $\tanh$ bounds extreme scores. Rollout-policy quantities are detached.

Pass-rate-gated advantage shaping. Sampled prompt pass rate determines the strength assigned to the surprisal signal. Let $c_i = \mathbf{1}[r_i > 0]$ denote verifier-defined success and $\rho_q = G^{-1} \sum_{j=1}^{G} c_j$ the group pass rate. Graded rewards retain the original standardized advantage A_i and use a predeclared success threshold. The gate g_i and coefficient $\lambda \in [0, 1]$ define

$$g_i = (1 - \rho_q)\mathbf{1}[u_i > 0] + \rho_q \mathbf{1}[u_i < 0]. \tag{7}$$

$$\widehat{A}_i = A_i + \lambda |A_i| g_i u_i, \qquad 0 \le \lambda \le 1. \tag{8}$$

The factor $|A_i|$ ties correction strength to verifier credit. Together, $|u_i| < 1$ and $\lambda \le 1$ preserve each advantage's sign. Low pass rates emphasize rare alternatives; high pass rates reduce repeated reinforcement and strengthen penalties on common errors (Table 1).

Table 1: Response-level credit allocation before shared normalization. Rarity is prompt-relative. The gate emphasizes rare responses in low-success groups and common responses in high-success groups; bounded shaping preserves the verifier sign.

Response	$A_i > 0$: reinforcement	$A_i < 0$: penalty	Gate g_i
Rare ($u_i > 0$)	Stronger	Softer	$1 - \rho_q$
Common ($u_i < 0$)	Weaker	Stronger	ρ_q

Shared normalization keeps the prompt group's total credit comparable to the base rule. We use the absolute mass so reinforcement and penalties both count toward this budget:

$$\alpha_q = \frac{\sum_{j=1}^{G} |A_j|}{\sum_{j=1}^{G} |\widehat{A}_j| + \varepsilon}, \qquad \widetilde{A}_i = \alpha_q \widehat{A}_i. \tag{9}$$

with $\widetilde{A}_i = 0$ when $\sum_{j=1}^{G} |A_j| = 0$. The shared positive factor preserves response ordering and approximately matches the base absolute sequence-advantage mass. Shaping reallocates credit both within and between reward classes; token aggregation determines the corresponding loss weights.

Integration with group-relative RLVR. ExPPO substitutes $\widetilde{A}_i$ for A_i in Eq. (2). Its DAPO instantiation retains asymmetric clipping, dynamic sampling, token-level aggregation, and overlong-response treatment (Yu et al., 2025). The method reuses rollout log-probabilities, preserves the model set and rollout count, and adds $O(G)$ prompt-group reductions. See Appendices E.1 and F.

5 Analysis

We first establish polarity and ordering of the shaped coefficients, then connect their allocation to entropy and correct-mode discovery in a KL-local model. Throughout, Y denotes a random response and y a realized value; indexed coefficients such as A_i and B_i are evaluated on y_i. A notation key appears in Appendix G, with proofs in Appendix H.

Local response-space update. Fix a prompt and its rollout-group statistics, and extend Eqs. (6), (8) and (9) to responses in the rollout support using these frozen statistics. The score $h = \widetilde{A}$ describes response-level advantage allocation. A parameter update direction d induces the score $h_d(y) = \nabla_\theta \log \pi_\theta(y \mid q)^\top d$, evaluated at $\theta = \theta_{\text{old}}$, whose exponential tilt matches the parameter update's first-order density change. See Appendix C.

Lemma 1 (KL-regularized local update)**.** *For step size $\eta > 0$, the solution of*

$$\max_{\pi} \left\{ \mathbb{E}_{Y\sim\pi}[h(q,Y)] - \eta^{-1} D_{\text{KL}}(\pi \| \pi_{\text{old}}) \right\}. \tag{10}$$

is the exponential tilt

$$\pi_{\eta,h}(y \mid q) = \frac{\pi_{\text{old}}(y \mid q) e^{\eta h(q,y)}}{\mathbb{E}_{Y'\sim\pi_{\text{old}}}[e^{\eta h(q,Y')}]}. \tag{11}$$

and $\frac{d}{d\eta}\mathbb{E}_{\pi_{\eta,h}}[V]\Big|_{\eta=0} = \text{Cov}_{\pi_{\text{old}}}(V,h)$ *for any fixed response utility V.*

This exponential-tilt model follows the KL geometry of natural-gradient and trust-region policy optimization and makes relative probability changes explicit (Kakade, 2001; Schulman et al., 2015).

Bounded polarity-preserving redistribution. Write $B_i = |A_i| g_i u_i$, so that $\widehat{A}_i = A_i + \lambda B_i$. Equivalently, $\widehat{A}_i = A_i M_i$ with $M_i = 1 + \lambda g_i \operatorname{sign}(A_i) u_i$.

Proposition 2 (Bounded reward-polarity preservation)**.** *For $0 \le \lambda \le 1$, $g_i \in [0,1]$, and $u_i \in [-1,1]$,*

$$1 - \lambda \le M_i \le 1 + \lambda, \qquad A_i \widehat{A}_i \ge 0, \qquad |\widehat{A}_i - A_i| \le \lambda |A_i|. \tag{12}$$

The normalized advantage $\widetilde{A}_i$ has the same sign as $\widehat{A}_i$. For $\lambda < 1$, every nonzero base advantage strictly preserves its sign.

Every nonzero shaped advantage retains the sign of its base verifier advantage. Relative surprisal changes the response's share of the group coefficients. The intervention ratio $\kappa(\lambda) = (1+\lambda)/(1-\lambda)$ bounds relative multiplier spread for $\lambda < 1$ (Appendix H).

Surprisal ordering and response-space odds. The ordering result concerns responses that receive equal credit under the base objective. For a generic sampled pass rate $\rho \in (0,1)$, define

$$\psi_\rho(u) = u\big((1-\rho)\mathbf{1}[u > 0] + \rho \mathbf{1}[u < 0]\big). \tag{13}$$

We write $\pi_{\eta,\widetilde{A}}$ for the exponential tilt in Eq. (11) with response score $h(q, y_i) = \widetilde{A}_i$ on the group.

Theorem 3 (Surprisal-ordered response preference)**.** *Fix a mixed group with $0 < \rho_q < 1$, a nonzero base advantage A, and $\lambda > 0$. The score $h_\lambda(u) = A + \lambda |A| \psi_{\rho_q}(u)$ is strictly increasing in u. For any response policy π, define $\Delta_{ij}(\pi) = \log \pi(y_i \mid q) - \log \pi(y_j \mid q)$. Hence, two sampled responses with $A_i = A_j$ and $u_i > u_j$ satisfy $\widetilde{A}_i > \widetilde{A}_j$ and*

$$\Delta_{ij}(\pi_{\eta,\widetilde{A}}) - \Delta_{ij}(\pi_{\text{old}}) = \eta(\widetilde{A}_i - \widetilde{A}_j) > 0. \tag{14}$$

Within this KL-local family, the base score A preserves equal-credit odds, and ExPPO shifts those odds toward the higher-surprisal response. The sampled logit counterpart appears in Appendix C.

Surprisal alignment, entropy, and matched-KL utility. The pairwise theorem describes one equal-credit class. At the group level, the correction is aligned with the implemented surprisal statistic. We write $\widehat{\mathrm{Cov}}_q$ for covariance under the uniform distribution over the realized group.

Proposition 4 (Empirical surprisal alignment)**.** *Under the uniform distribution over a prompt group,*

$$\widehat{\mathrm{Cov}}_q(s, B) = G^{-1} \sum_{i=1}^{G} (s_i - \mu_{s,q}) B_i \geq 0. \tag{15}$$

To connect this direction to Shannon entropy, freeze the group statistics and draw $Y \sim \pi_{\mathrm{old}}(\cdot \mid q)$. Write $A = A(Y)$, $B = B(Y)$, and $S = -\log \pi_{\mathrm{old}}(Y \mid q)$ for the resulting random variables. Unmarked covariances and variances use this response distribution. For $h = \alpha_q(A + \lambda B)$ with fixed $\alpha_q > 0$, Lemma 1 gives

$$\left.\frac{dH(\pi_{\eta,h}(\cdot \mid q))}{d\eta}\right|_0 = \alpha_q \big[\mathrm{Cov}(S, A) + \lambda \mathrm{Cov}(S, B)\big]. \tag{16}$$

The first term is the entropy direction induced by the reward advantage; the second gives ExPPO's additional contribution. Proposition 4 establishes empirical alignment with length-normalized surprisal s, and Eq. (16) characterizes response entropy through full-response surprisal S. Logged next-token entropy in Section 6.3 provides a complementary training diagnostic.

Correct-mode discovery gives a concrete exploration utility. For correct modes $m \in \mathcal{M}_q^+$ with masses $p_m(\pi) = \mathbb{P}_{Y\sim\pi}(Y \in m \mid q)$, the expected number observed in K independent responses is $F_K(\pi) = \sum_m [1 - (1 - p_m(\pi))^K]$. Its local response utility is $V_K(y) = K(1 - p_{m(y)}(\pi_{\mathrm{old}}))^{K-1}$ for correct y and zero for failures. For $K \geq 2$, this utility gives greater weight to low-mass correct modes. See Appendix H. For a fixed utility V, write $J_V(\pi) = \mathbb{E}_{Y\sim\pi}[V(Y)]$. For a small KL budget $\delta > 0$, let $\eta_h = \eta_h(\delta)$ satisfy $D_{\mathrm{KL}}(\pi_{\eta_h,h} \| \pi_{\mathrm{old}}) = \delta$ and define

$$\Gamma_V(h) = \frac{\mathrm{Cov}(V, h)}{\sqrt{\mathrm{Var}(h)}}. \tag{17}$$

At a common KL budget, the local gain is $J_V(\pi_{\eta_h,h}) - J_V(\pi_{\mathrm{old}}) = \sqrt{2\delta}\,\Gamma_V(h) + O(\delta)$, where $O(\delta)$ collects higher-order terms. Assuming $\mathrm{Var}(A) > 0$, define the component of B orthogonal to A as

$$B_\perp = B - \frac{\mathrm{Cov}(A, B)}{\mathrm{Var}(A)} A, \qquad \mathrm{Cov}(A, B_\perp) = 0. \tag{18}$$

Its alignment with V determines the gain:

$$\left.\frac{d}{d\lambda}\Gamma_V(A + \lambda B)\right|_{\lambda=0} = \frac{\mathrm{Cov}(V, B_\perp)}{\sqrt{\mathrm{Var}(A)}}. \tag{19}$$

For $V = V_K$, positive residual alignment increases the leading matched-KL gain in expected correct-mode yield for sufficiently small positive λ. Under binary rewards, A is affine in correctness: $V_1(Y) = C$ gives $\mathrm{Cov}(V_1, B_\perp) = 0$, while $K \geq 2$ introduces variation across correct modes. Verifier coverage depends on total correct mass through $1 - (1 - p_q)^K$; mode yield and correct-pair diversity characterize its allocation across correct answers. See Appendix I.

6 Experiments

We evaluate whether response-level rarity shaping preserves repeated-sample coverage, sustains response accuracy, and discovers distinct correct modes.

6.1 Experimental Setup

Eight benchmarks including four In-Domain (ID) and four Out-of-Domain (OOD) for evaluation. Training uses the DAPO-Math-17k RL prompt set and its rule-based verifier (Yu et al., 2025).

The ID suite comprises AIME24, AIME25, Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). The OOD suite comprises GPQA (Rein et al., 2023), MMLU-Pro (Wang et al., 2024), a safety set, and a four-choice TruthfulQA-MC construction (Lin et al., 2022), adapting the ID/OOD evaluation setup of Jin et al. (2025; 2026). We call pass@k, the probability of at least one verified success in k responses, *coverage*. High-k curves estimate it from the first k generated responses per prompt. We use $k = 16$ for AIME24/25 (30 prompts each), $k = 8$ for Minerva (272 prompts), and $k = 4$ for the other five benchmarks. Response accuracy is the mean verified-response share. Both metrics use unweighted suite averages; see Appendix E.

Three backbones span scale and checkpoint type. We study **Qwen2.5-7B**, **Qwen3-0.6B**, and **Phi-3.5-mini** (Qwen et al., 2024; Yang et al., 2025; Abdin et al., 2024). Core methods are **GRPO** (Shao et al., 2024), entropy-regularized **GRPO-H** (Eq. (3)), and **ExPPO**. Focused comparisons add Rewarding the Unlikely (**GRPO-UR**), **REPO-R**, and **F-GRPO** (He et al., 2025; Petrenko et al., 2026; Plyusov et al., 2026). See Appendix E.

Matched budgets support controlled comparisons. The main comparisons share data, optimizer, clipping, token budget, and a 900-step schedule with $G = 16$ rollouts per prompt; evaluation uses step 900. The Phi-3.5-mini comparison of F-GRPO, GRPO-UR, REPO-R, and ExPPO uses 300 matched steps and three independent training runs per method. Standard evaluation uses temperature 0.7 and top-$p = 0.95$.

6.2 Performance During Training

ExPPO leads both ID and OOD across all backbones. ExPPO finishes highest across all six model–suite trajectories (Fig. 1). It is the only core method to exceed base-model OOD coverage on every backbone during training; GRPO and GRPO-H show OOD forgetting on Qwen2.5-7B and Phi-3.5-mini. This is consistent with exploration supporting knowledge retention during adaptation. OOD representation probes appear in Appendix J.9.

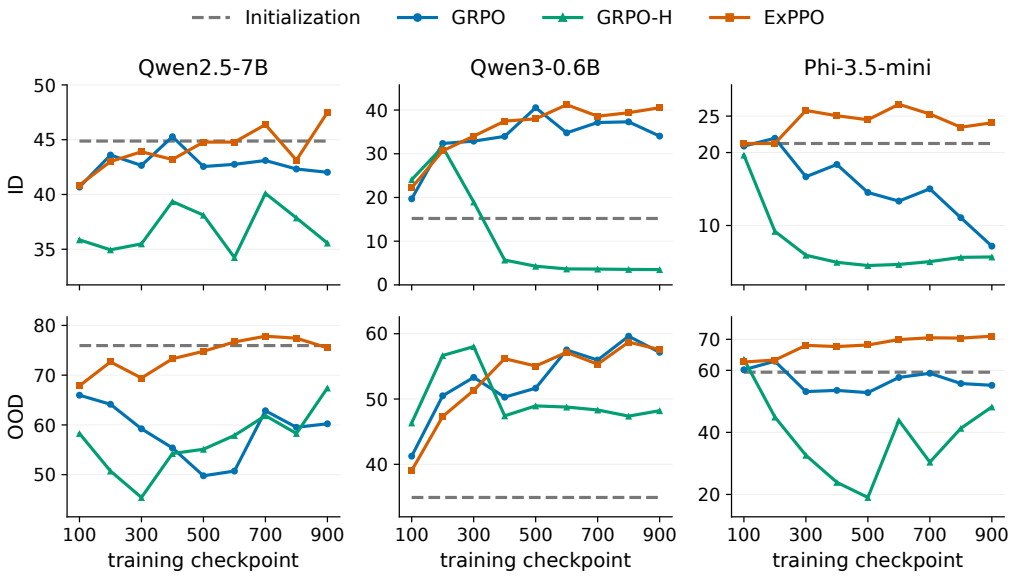

Figure 1: Checkpoint-wise any-success coverage. Columns: Qwen2.5-7B, Qwen3-0.6B, Phi-3.5-mini; rows: ID and OOD four-task means (percentage points). Budgets: $k = 16$ for AIME24/25, 8 for Minerva, 4 otherwise. Dashed lines: initialization. Per-task trajectories: Appendix J.5.

6.3 Entropy Dynamics and Large-Sample Coverage

Entropy dynamics accompany coverage gains. ExPPO retains higher entropy than GRPO on three models (Fig. 2) with stronger coverage, while GRPO-H combines high entropy with low ID coverage. Qwen2.5-7B initialization remains competitive at the largest budgets (Fig. 3).

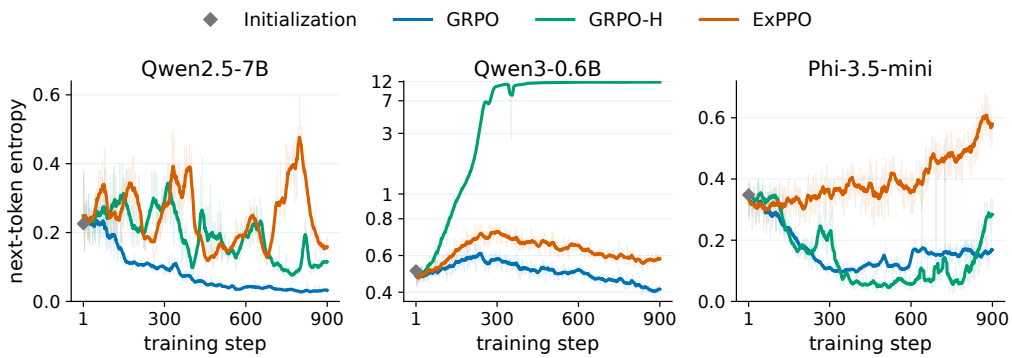

Figure 2: Next-token entropy through training. Faint/solid traces: per-step observations/step means; diamonds: initial states. Qwen3-0.6B uses a linear scale through 0.8 and continuous compression above 0.8, with ticks in entropy units. The other panels use model-specific linear ranges.

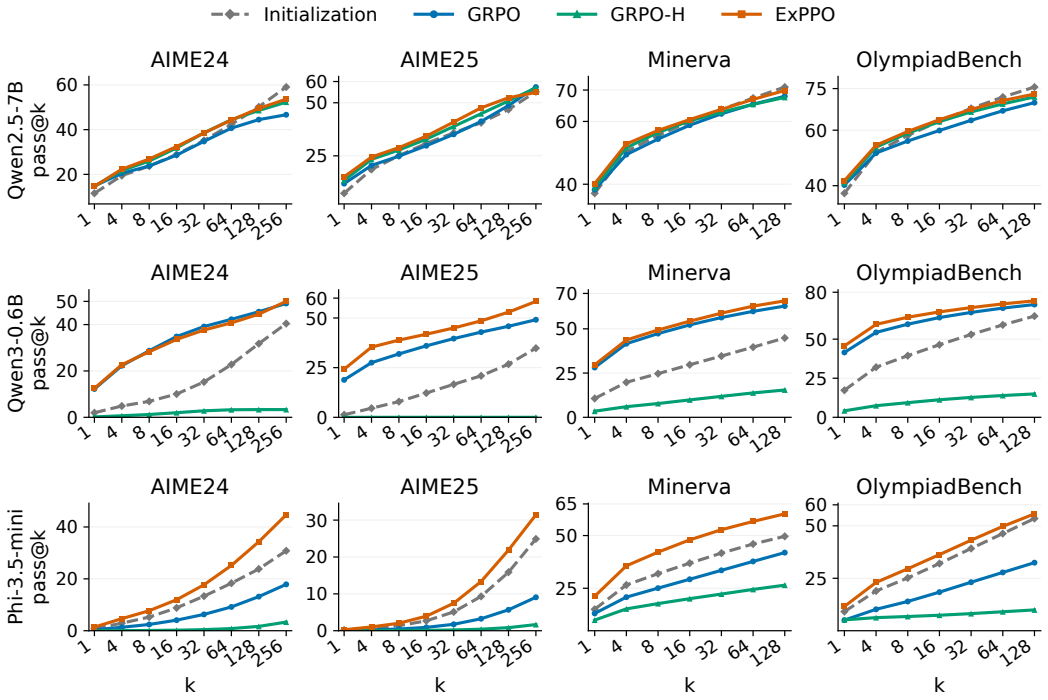

Figure 3: Large-K coverage at step 900, using the first k responses per prompt. Rows: the three backbones; columns: four ID tasks. Curves include initialization and the core training rules. At the largest budget, ExPPO matches or exceeds the other trained core methods on both smaller backbones and three Qwen2.5-7B tasks.

6.4 Controlled Multi-Answer Evaluation

We use 64 synthetic coin-count prompts, each asking for four bounded integer counts whose total coin value matches a target. Each distinct verified count vector defines a correct mode, so measured diversity reflects different valid answers.

ExPPO increases fixed-budget mode yield and conditional diversity. ExPPO increases observed and effective correct-mode yields over GRPO and GRPO-H across backbones (Table 2). On Qwen2.5-7B, it also broadens conditional diversity relative to both core baselines and F-GRPO, despite F-GRPO's higher accuracy. REPO-R has a similar diversity point estimate and lower yields. Qwen3-0.6B gains diversity at near-saturated accuracy. Full comparisons appear in Appendix J.7.

Table 2: Last-checkpoint controlled evaluation ($K = 128$; 64 prompts). A: Accuracy/Solved: correct-response share (0–100)/successful prompts. Observed/effective yields: mean distinct correct modes/exponentiated mode entropy (unsolved: zero). Correct/prompt: successes/64. B: Shared: prompts with ≥ 2 successes per method; D: correct-pair mode disagreement averaged on Shared (0–100); ΔD: ExPPO minus baseline; CI: 95% paired prompt-bootstrap interval. Qwen2.5-7B: updated ExPPO for A and F-GRPO/REPO-R; earlier for GRPO/GRPO-H (Appendix J.7).

A. Fixed-budget correctness and correct-mode yield

Model	Method	Accuracy	Solved / 64	Observed yield	Effective yield	Correct / prompt
Qwen2.5-7B	GRPO	6.45	57	1.750	1.546	8.25
	GRPO-H	9.68	59	1.906	1.610	12.39
	F-GRPO	**35.51**	**64**	3.234	2.389	**45.45**
	REPO-R	15.42	**64**	3.578	2.820	19.73
	ExPPO	33.50	**64**	**4.172**	**3.013**	42.88
Qwen3-0.6B	GRPO	**98.43**	**64**	5.422	3.013	**125.98**
	GRPO-H	0.02	1	0.016	0.016	0.03
	ExPPO	97.74	**64**	**6.016**	**3.481**	125.11
Phi-3.5-mini	GRPO	0.90	22	0.438	0.407	1.16
	GRPO-H	0.50	9	0.219	0.206	0.64
	ExPPO	**2.26**	**33**	**0.703**	**0.634**	**2.89**

B. Conditional correct-pair diversity

Model	Baseline	Shared / 64	Baseline D	ExPPO D	ΔD
Qwen2.5-7B	GRPO	51	33.98	**50.57**	+16.59
	GRPO-H	55	35.14	**48.77**	+13.63
	F-GRPO	63	46.53	**56.28**	+9.74
	REPO-R	63	56.12	**56.28**	+0.16
Qwen3-0.6B	GRPO	64	51.46	**58.44**	+6.98

6.5 Comparison with the Closest Baselines

Table 3: Qwen2.5-7B last-checkpoint coverage (percentage points): mean ± sample standard deviation across three decoding seeds at one fixed step-900 checkpoint per method. Initialization: one evaluation. Budgets: Section 6.1; Avg.: within-run four-task mean. Bold: best trained mean.

Domain	Method	AIME24	AIME25	Minerva	Olympiad	Avg.
ID	Initialization	36.70	36.70	55.10	51.00	44.88
	GRPO	23.33±3.33	24.44±9.62	53.55±0.77	50.96±0.90	38.07±2.74
	GRPO-H	26.67±10.00	27.78±10.18	56.50±1.39	53.28±1.13	41.06±4.76
	GRPO-UR	25.56±1.92	21.11±1.92	56.74±1.49	47.95±0.87	37.84±0.88
	REPO-R	26.67±3.33	33.33±3.33	57.60±1.29	**56.20±1.05**	43.45±0.90
	F-GRPO	27.78±3.85	**36.67±3.33**	56.37±0.93	53.83±0.34	43.66±0.67
	ExPPO	**34.44±5.09**	**36.67±3.33**	**58.46±1.33**	55.01±0.34	**46.14±1.94**
Domain	Method	GPQA	MMLU-Pro	Safety	TruthfulQA	Avg.
OOD	Initialization	59.80	81.40	84.10	78.50	75.95
	GRPO	57.66±3.29	77.62±0.82	32.30±0.38	70.57±0.61	59.54±0.88
	GRPO-H	56.99±1.23	81.43±1.43	52.03±0.47	**80.17±1.08**	67.66±0.57
	GRPO-UR	58.11±2.17	80.95±2.18	81.27±0.55	74.27±0.67	73.65±1.30
	REPO-R	55.73±1.58	80.95±0.82	34.17±0.49	77.83±0.55	62.17±0.18
	F-GRPO	57.29±0.52	77.62±0.82	83.00±0.20	79.78±0.72	74.42±0.42
	ExPPO	**61.38±0.89**	**83.33±3.30**	**83.70±0.50**	79.43±0.95	**76.96±0.54**

ExPPO leads both ID and OOD coverage. ExPPO leads both coverage means and six task means, including a tie with F-GRPO on AIME25 (Table 3). After training, ExPPO's OOD coverage exceeds initialization, while other methods generally show OOD forgetting.

Phi-3.5-mini. Across three seeds per method, ExPPO leads ID coverage and accuracy; REPO-R closely matches OOD coverage with less variation. GRPO-UR leads OOD accuracy but has lower, more variable coverage. F-GRPO has the lowest suite means and greatest coverage variation (Table 4).

Table 4: Phi-3.5-mini at 300 steps: mean ± sample standard deviation across three independent training seeds per method. Each score is an unweighted four-task suite mean (%). Coverage is the share of prompts with a verified success; response accuracy is the verified-response share. Bold: highest mean per column. Task and seed details: Appendix J.2.

	Coverage		Response accuracy	
Method	ID	OOD	ID	OOD
F-GRPO	13.88±5.76	42.91±25.37	6.42±0.81	36.57±22.19
GRPO-UR	21.66±0.75	62.76±5.48	9.79±1.11	**52.85±3.62**
REPO-R	21.94±1.89	**66.35±0.38**	9.64±0.19	52.16±0.15
ExPPO	**25.67±1.31**	66.33±1.59	**10.59±0.27**	52.26±1.17

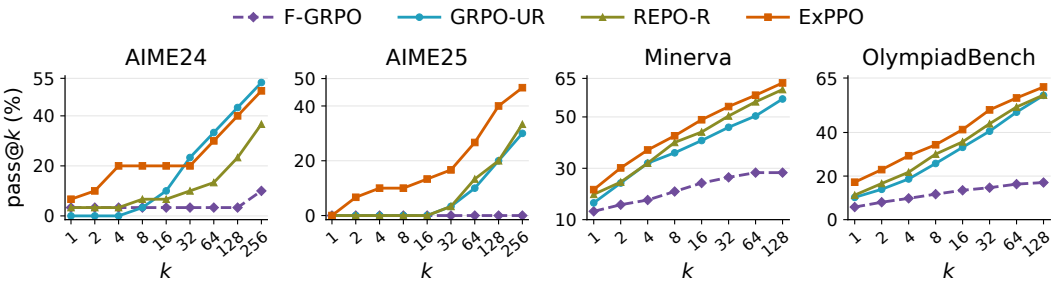

Figure 4: Phi-3.5-mini coverage (%) from the first k responses per prompt, for the methods in Table 4. ExPPO leads on three of four tasks at their largest evaluated budgets ($k = 256$ for AIME24/25; $k = 128$ otherwise), while GRPO-UR leads on AIME24. At the shared budget $k = 128$, ExPPO's four-task macro coverage of 51.03 exceeds GRPO-UR's 44.34, REPO-R's 40.30, and F-GRPO's 12.17.

6.6 Beyond Mathematics: Code Reasoning

Training Phi-3.5-mini on CodeReasoner-derived Python execution reasoning (Tang et al., 2025) tests a second adaptation domain. ExPPO leads both code metrics and OOD means (Table 5); its OOD response accuracy is 8.92 points above GRPO-H. This extends the accuracy–coverage benefit beyond mathematical training. See Appendix J.4.

Table 5: Code adaptation and OOD generalization at the last checkpoint. Scores (%): mean ± sample standard deviation across three independent Phi-3.5-mini training runs. Code: CodeReasoner; OOD: unweighted mean over GPQA, MMLU-Pro, Safety, and TruthfulQA-MC ($k = 4$). Initialization: fixed base model. Bold: best trained mean.

	Coverage		Response accuracy	
Method	Code	OOD	Code	OOD
Initialization	32.11	59.40	6.92	41.93
GRPO	85.94±0.00	52.24±0.44	75.22±0.04	36.42±1.18
GRPO-H	86.46±0.08	60.51±1.20	74.62±0.02	42.10±0.34
ExPPO	**86.69±0.33**	**65.74±0.71**	**75.46±0.05**	**51.02±0.87**

7 Conclusion

ExPPO combines bounded prompt-relative surprisal, pass-rate gating, and shared advantage-mass normalization to support exploration through response-level credit allocation. Our analysis identifies local conditions for gains in entropy and correct-mode discovery and characterizes how the rule preserves verifier polarity. Experiments across three backbones show improved in-distribution coverage with competitive out-of-distribution generalization, while sustaining policy entropy and increasing diversity among verified-correct responses. Large-sample evaluations and code-reasoning results further support the benefits across sampling budgets and training domains. ExPPO integrates into GRPO and DAPO using existing rollout statistics with negligible computational overhead.

Reproducibility Statement

Appendices E, F and J.7 document model variants, training/evaluation settings, the final-answer parser, verifier, and metrics. The generation script and archived evidence cover checkpoint metrics, entropy histories, maximum-budget pools, the controlled $K = 128$ rollout bank, paired-bootstrap summaries, Qwen decoding logs, sweeps, independent Phi-3.5-mini training runs, and derived summaries. A code package accompanies this submission as supplementary material.

AI Use Statement

An AI assistant supported editing, reported-data arithmetic checks, and LaTeX formatting.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 151–160, 2019.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? A large-scale study. In *International Conference on Learning Representations*, 2021.

Zhe Cao, Miaowen Wen, and Fangjiong Chen. When correct solutions repeat: Rarity-aware credit redistribution for GRPO. *arXiv preprint arXiv:2608.03467*, 2026.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

Guangran Cheng, Chengqi Lyu, Songyang Gao, Wenwei Zhang, and Kai Chen. Group entropy-controlled policy optimization. *arXiv preprint arXiv:2607.16850*, 2026.

Jean Seong Bjorn Choe and Jong-Kook Kim. Maximum entropy on-policy actor-critic via entropy advantage estimation. *arXiv preprint arXiv:2407.18143*, 2024.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Jingchu Gai, Guanning Zeng, Huaqing Zhang, and Aditi Raghunathan. Differential smoothing mitigates sharpening and improves LLM reasoning. *arXiv preprint arXiv:2511.19942*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1861–1870, 2018.

Chaoqun He, Renjie Luo, Yuzhuo Bai, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. *arXiv preprint arXiv:2506.02355*, 2025.

Hangzhan Jin, Sitao Luan, Tianwei Ni, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. RL fine-tuning heals OOD forgetting in SFT. *arXiv preprint arXiv:2509.12235*, 2025.

Hangzhan Jin, Tianwei Ni, Lu Li, Pierre-Luc Bacon, Mohammad Hamdaqa, and Doina Precup. Rotation-preserving supervised fine-tuning. *arXiv preprint arXiv:2605.10973*, 2026.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems 14*, 2001.

Aitor Lewkowycz, Anders Andreassen, David Dohan, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252, 2022.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking RLVR's indifference to diversity. *arXiv preprint arXiv:2605.00365*, 2026.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1928–1937, 2016.

Ifdita Hasan Orney, Jubayer Ibn Hamid, Shreya S. Ramanujam, Shirley Wu, Hengyuan Hu, Noah Goodman, Dorsa Sadigh, and Chelsea Finn. Poly-EPO: Training exploratory reasoning models. *arXiv preprint arXiv:2604.17654*, 2026.

Aleksei Petrenko, Ben Lipkin, Kevin Chen, Erik Wijmans, Marco Cusumano-Towner, Raja Giryes, and Philipp Krähenbühl. Entropy-preserving reinforcement learning. *arXiv preprint arXiv:2603.11682*, 2026.

Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, Daria Korotyshova, and Daniil Gavrilov. F-GRPO: Don't let your policy learn the obvious and forget the rare. *arXiv preprint arXiv:2602.06717*, 2026.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, et al. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Ruslan Salakhutdinov, Haiwen Feng, and Andrea Zanette. Maximum likelihood reinforcement learning. *arXiv preprint arXiv:2602.02710*, 2026.

Lingxiao Tang, He Ye, Zhongxin Liu, Xiaoxue Ren, and Lingfeng Bao. CodeReasoner: Enhancing the code reasoning ability with reinforcement learning. *arXiv preprint arXiv:2507.17548*, 2025.

Christian Walder and Deep Karkhanis. Pass@K policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why RLVR may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Xiaoyun Zhang, Xiaojian Yuan, Di Huang, Wang You, Chen Hu, Jingqing Ruan, Ai Jian, Kejiang Chen, and Xing Hu. Revisiting entropy regularization: Adaptive coefficient unlocks its potential for LLM reinforcement learning. *arXiv preprint arXiv:2510.10959*, 2025.

Jiazheng Zhang, Ziche Fu, Junrui Shen, et al. Entropy polarity in reinforcement fine-tuning: Direction, asymmetry, and control. *arXiv preprint arXiv:2605.11775*, 2026.

APPENDIX CONTENTS

A Related Work **14**

B Scope and Limitations **14**

C Binary-Reward GRPO and Mode-Level Credit **14**

D GRPO Background and Entropy Bonuses **15**

D.1 GRPO and GRPO-H Objective Details 15

D.2 Entropy Bonuses in MuJoCo and Their Connection to RLVR 16

E Experimental Details **17**

E.1 Measured Training Cost 18

F Optimization Implementation **18**

G Notation and Probability Conventions **19**

H Proofs **21**

I Finite-λ Matched-KL Certificate **23**

J More Experimental Results **24**

J.1 Coefficient Sensitivity and Component Ablations 24

J.2 Closest Baselines at the Phi-3.5-mini Ablation Budget 25

J.3 Independent-Run Stability on Phi-3.5-mini 27

J.4 Code-Reasoning Adaptation and OOD Generalization 27

J.5 Per-Benchmark Checkpoint Trajectories 28

J.6 DAPO-Based ExPPO 29

J.7 Controlled Multi-Answer Task 30

J.8 Common-Checkpoint Continuation and Normalized Credit Allocation 32

J.9 Matched-Input OOD Representations 34

A Related Work

RLVR objectives and support concentration. GRPO introduced critic-free group-relative optimization for mathematical reasoning, DAPO developed a stronger large-scale training recipe, and GSPO aligned importance weighting and clipping with sequence-level rewards (Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025). Subsequent studies report higher pass@1 together with lower large-K coverage, narrower answer support, and stronger dependence on behaviors already present in the base policy (Yue et al., 2025; Wu et al., 2025). MaxRL reweights low-success prompts through objectives that approach maximum likelihood as rollout compute grows (Tajwar et al., 2026). ExPPO examines the within-prompt allocation of credit across sampled responses.

Entropy control and set-level exploration. Entropy-focused work characterizes policy-entropy dynamics and develops explicit control through covariance clipping, adaptive regularization, advantage shaping, and asymmetric clipping (Cui et al., 2025; Zhang et al., 2025; Petrenko et al., 2026; Zhang et al., 2026; Cheng et al., 2026). Set-level methods optimize the collective utility of several responses, including Pass@K Policy Optimization, analytical pass@k training, and Poly-EPO (Walder and Karkhanis, 2025; Chen et al., 2025; Orney et al., 2026). Differential Smoothing modifies correct-trajectory rewards to counteract distribution sharpening (Gai et al., 2025). These approaches motivate exploration measurements beyond pass@1. ExPPO operates at response granularity and approximately preserves prompt-group absolute sequence-advantage mass while reallocating it across samples.

Rarity-aware and correct-support credit. The closest methods directly alter credit for underrepresented responses. Rewarding the Unlikely increases pressure on low-probability correct trajectories, and F-GRPO downweights high-success prompts to protect rare correct modes under finite groups (He et al., 2025; Plyusov et al., 2026). UCPO encourages a more uniform distribution within the correct set, and Cue-GRPO redistributes positive advantages using correct-strategy partitions (Lochab et al., 2026; Cao et al., 2026). ExPPO specifies a prompt-relative allocation through group-standardized response surprisal, a bounded transformation, pass-rate gating, and absolute-advantage-mass normalization. The controlled task measures correct-mode yield and conditional correct-pair diversity, and Table 3 compares ExPPO with F-GRPO, GRPO-UR, and REPO-R. REPO-R controls token-level rescaling through entropy feedback. ExPPO uses prompt-relative response standardization, pass-rate gating, bounded corrections, and a shared absolute-mass budget.

B Scope and Limitations

Evaluation scope. The evaluation covers three backbones, matched 900-step reasoning runs, and a four-method Phi-3.5-mini comparison with three training runs per method at 300 steps under limited computational resources. Larger models, broader benchmark suites, and longer schedules would further characterize scaling with model capacity and optimization time.

C Binary-Reward GRPO and Mode-Level Credit

This derivation gives the closed-form exposure effect used in Section 3. Consider a mixed group with binary rewards $r_i \in \{0, 1\}$ and sampled pass rate $\rho_q \in (0, 1)$. Ignoring the numerical constant in the denominator, the group reward mean and standard deviation are $\mu_{r,q} = \rho_q$ and $\sigma_{r,q} = \sqrt{\rho_q(1-\rho_q)}$. Therefore,

$$A_i = \begin{cases} \sqrt{(1-\rho_q)/\rho_q}, & r_i = 1, \\ -\sqrt{\rho_q/(1-\rho_q)}, & r_i = 0. \end{cases} \tag{20}$$

All successes receive the same positive coefficient and all failures receive the same negative coefficient. If a correct mode m appears n_m times in the group, its aggregate positive coefficient is $C_m = n_m\sqrt{(1-\rho_q)/\rho_q}$. This identity is the mode-level multiplicity effect that motivates the controlled experiment.

Proposition 5 (Multiplicity changes relative mode odds)**.** *Let $\mathcal{M}_q$ be a finite set of response modes for prompt q, let $p_z(m \mid q) = e^{z_m} / \sum_{a\in\mathcal{M}_q} e^{z_a}$ be a categorical mode policy with logit z_m, and let m_i be the mode containing response y_i. Define $C_m = \sum_{i:m_i=m} A_i$, the empirical policy-gradient*

surrogate $\mathcal{L}_q(z) = \sum_{i=1}^{G} A_i \log p_z(m_i \mid q)$*, and the relative log-odds* $\Delta_{mm'}(p) = \log p(m \mid q) - \log p(m' \mid q)$*. Because group-relative advantages satisfy* $\sum_i A_i = 0$*, one gradient-ascent step* $z_m^+ = z_m + \eta\, \partial \mathcal{L}_q / \partial z_m$ *with step size* $\eta > 0$ *gives, for any two correct modes* m *and* m'*,*

$$\Delta_{mm'}(p_{z^+}) - \Delta_{mm'}(p_z) = \eta(C_m - C_{m'}) = \eta A^+(n_m - n_{m'}). \tag{21}$$

A count advantage $n_m > n_{m'}$ *therefore gives mode* m *a larger relative log-odds update.*

Proof. For the softmax mode policy, $\partial \mathcal{L}_q / \partial z_m = C_m - p_z(m \mid q) \sum_i A_i = C_m$. The identity $\Delta_{mm'}(p_z) = z_m - z_{m'}$ then gives Eq. (21). This mode-level derivation follows the score-function policy-gradient update (Williams, 1992).

GRPO and ExPPO in the same logit geometry. For arbitrary detached sample scores h_i, set $p_m = p_z(m \mid q)$, $H_m = \sum_{i:m_i=m} h_i$, and $H_\Sigma = \sum_i h_i$. Differentiating $\mathcal{L}_q^h = \sum_i h_i \log p_z(m_i \mid q)$ gives

$$\frac{\partial \mathcal{L}_q^h}{\partial z_m} = H_m - p_m H_\Sigma.$$

Let $C_m^B = \sum_{i:m_i=m} B_i$ and $B_\Sigma = \sum_i B_i$. Applying the same gradient-ascent step size η to $h_i = A_i$ and $h_i = \widetilde{A}_i$ yields the respective relative log-odds changes

$$\begin{aligned}
\delta_{mm'}^A &= \eta(C_m - C_{m'}), \\
\delta_{mm'}^{\mathrm{E}} &= \alpha_q \delta_{mm'}^A + \eta \alpha_q \lambda \big[C_m^B - C_{m'}^B - (p_m - p_{m'}) B_\Sigma\big].
\end{aligned}$$

The first term records shared rescaling; the bracket gives the additional allocation contribution in the same geometry as Proposition 5. For a rare/common pair, a positive bracket shifts relative odds toward the rare mode compared with the shared-scale base update. The B_Σ term accounts for the group-sum change introduced by shaping.

The effective score $h_{\mathrm{eff}}(m) = H_m - p_m H_\Sigma$ also gives the exact updated distribution,

$$p_{z^+}(m \mid q) = \frac{p_z(m \mid q) e^{\eta h_{\mathrm{eff}}(m)}}{\sum_a p_z(a \mid q) e^{\eta h_{\mathrm{eff}}(a)}}.$$

Thus the sampled logit step is an exponential tilt with its gradient-induced score. Setting $h = \widetilde{A}$ in Theorem 3 evaluates the response-level advantage allocation within the same exponential-tilt family.

Neural gradients and induced response scores. At θ_{old}, the token importance ratios equal one and the clipping interval contains them in its interior. For the sampled policy-loss term, let $\omega_{i,t}$ denote the fixed token aggregation weights and evaluate the following derivatives at $\theta = \theta_{\mathrm{old}}$:

$$f_i = \sum_t \omega_{i,t} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x_{i,t}), \qquad \mathcal{G}_q(h) = \sum_i h_i f_i.$$

The prompt's gradient contribution satisfies $\mathcal{G}_q(\widetilde{A}) = \alpha_q \mathcal{G}_q(A) + \alpha_q \lambda \mathcal{G}_q(B)$. Response averaging uses $\omega_{i,t} = 1/(GT_i)$; token-mean aggregation uses the reciprocal of the total generated-token count in the loss aggregation batch. These expressions retain parameter sharing through the vectors f_i. For a fixed parameter direction d and differentiable policy on a finite response space, the score $h_d(y) = \nabla_\theta \log \pi_\theta(y \mid q)^\top d$, evaluated at $\theta = \theta_{\mathrm{old}}$, has $\mathbb{E}_{\pi_{\mathrm{old}}}[h_d] = 0$ and satisfies

$$\left.\frac{d\pi_{\theta_{\mathrm{old}}+\eta d}(y \mid q)}{d\eta}\right|_0 = \pi_{\mathrm{old}}(y \mid q) h_d(y) = \left.\frac{d\pi_{\eta,h_d}(y \mid q)}{d\eta}\right|_0.$$

This identity connects a chosen optimizer direction to the local density model in Lemma 1.

D GRPO Background and Entropy Bonuses

D.1 GRPO and GRPO-H Objective Details

Equations (2) and (3) give the objectives used in the main text. Here $\mathcal{B}$ denotes joint sampling of a prompt from $\mathcal{D}$ and a group of G responses from $\pi_{\mathrm{old}}(\cdot \mid q)$. For realized response y_i, the sequence

probability factors as $\pi_{\text{old}}(y_i \mid q) = \prod_{t=1}^{T_i} \pi_{\text{old}}(y_{i,t} \mid x_{i,t})$. GRPO uses the detached advantage in Eq. (1) and the token importance ratio

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x_{i,t})}{\pi_{\text{old}}(y_{i,t} \mid x_{i,t})}.$$

For clipping radius $\epsilon_{\text{clip}} > 0$, the clipped surrogate is

$$\ell_{\text{clip}}(w, A) = \min\{wA, \text{clip}(w, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}})A\}.$$

The reference-policy penalty is $d_{i,t}(\theta) = D_{\text{KL}}(\pi_\theta(\cdot \mid x_{i,t}) \| \pi_{\text{ref}}(\cdot \mid x_{i,t}))$, where π_{ref} is fixed and $\beta_{\text{KL}} \geq 0$ controls its strength. Canonical response averaging uses $\langle z_{i,t} \rangle = G^{-1} \sum_{i=1}^{G} T_i^{-1} \sum_{t=1}^{T_i} z_{i,t}$ (Shao et al., 2024). Our implementation uses token-mean aggregation, as specified in Appendix E; the response-level shaping applies with either convention.

The GRPO-H objective adds Shannon next-token entropy, $H(\pi_\theta(\cdot \mid x)) = -\sum_a \pi_\theta(a \mid x) \log \pi_\theta(a \mid x)$, with coefficient $\beta_{\text{ent}} \geq 0$ (Williams and Peng, 1991; Mnih et al., 2016; Schulman et al., 2017). This term averages uncertainty over sampled token contexts using the same aggregation as the policy loss.

D.2 Entropy Bonuses in MuJoCo and Their Connection to RLVR

Evidence from continuous control. In a large PPO study on five MuJoCo tasks, Andrychowicz et al. (2021) find similar performance across most tested regularizers, with entropy constraints benefiting HalfCheetah. They suggest that careful initialization already supplies useful exploration and that PPO clipping controls policy changes. Ahmed et al. (2019) show that higher entropy can smooth policy optimization in some environments. Together, these results tie the value of an entropy bonus to the initial policy and task dynamics.

Action spread and task direction. For a diagonal Gaussian policy at state x, the differential entropy is

$$H\big(\mathcal{N}(\mu_\theta(x), \operatorname{diag} \sigma_\theta^2(x))\big) = \frac{d_a}{2} \log(2\pi e) + \sum_{j=1}^{d_a} \log \sigma_{\theta,j}(x), \tag{22}$$

where d_a is the action dimension. With separate mean and scale parameters, the direct entropy gradient changes action spread and has zero derivative with respect to the mean; the reward advantage supplies the task direction. Increasing spread can therefore spend samples on low-return actions. Its balance with reward depends on the bonus coefficient, reward scale, and initialization. This formula describes the Gaussian before action clipping or squashing.

Local entropy and long-horizon credit. A direct entropy term at fixed rollout states promotes uncertainty at those states. A maximum-entropy return also credits earlier actions for reaching states with high future entropy (Haarnoja et al., 2018). Choe and Kim (2024) separate reward and entropy advantage estimates and match or improve entropy-bonus PPO/TRPO on their MuJoCo comparisons, whose main experiments use discretized actions. These results show how exploration benefits from assigning credit across a trajectory.

GRPO-H and verifier credit. The bonus in Eq. (3) supplies a reward-independent gradient at each sampled prefix. For a softmax token policy $p_a = \exp(z_a)/\sum_b \exp(z_b)$,

$$\frac{\partial H(p)}{\partial z_a} = p_a\big[-\log p_a - H(p)\big]. \tag{23}$$

Thus the bonus favors tokens with above-average surprisal, including continuations that fail the verifier, while the GRPO term retains equal sequence advantages for equal rewards. Its scale also matters: Cui et al. (2025) report weak entropy control at small coefficients, rapid entropy growth at a large coefficient, and performance at or below the compared baselines at a coefficient that stabilizes entropy. These mechanisms help explain the high entropy and low ID coverage of GRPO-H in our mathematical runs (Figs. 1 and 2). The code experiment shows gains from GRPO-H in a different task regime (Table 5). ExPPO ties exploration credit to the verifier advantage and prompt pass rate, with bounded shaping and shared normalization; Eq. (5) explains why diversity among correct responses is the relevant target.

E EXPERIMENTAL DETAILS

Two instruction checkpoints and one base checkpoint span model regimes. The short names in the main paper correspond to `Qwen/Qwen2.5-7B-Instruct` (**Qwen2.5-7B**), `Qwen/Qwen3-0.6B` (**Qwen3-0.6B**), and `microsoft/Phi-3.5-mini-instruct` (**Phi-3.5-mini**). Qwen2.5 and Phi use their instruction templates, while Qwen3 uses the base checkpoint and tokenizer. Every method for a model starts from the same checkpoint and tokenizer.

Shared optimization isolates advantage shaping. Training uses the fixed 17,398-prompt DAPO-Math-17k set, with left truncation and removal of prompts longer than 2,048 tokens. The maximum response length is 10,240 tokens. Each optimization step samples $G = 16$ responses per prompt with temperature 1.0, top-$p = 1.0$, and top-k disabled. The prompt batch size is 64, the policy mini-batch size is 32, and the rollout-generation batch size is 32. We optimize with AdamW at learning rate 10^{-6}, 10 warmup steps, weight decay 0.1, and gradient clipping at 1.0. The policy loss uses token-mean aggregation, symmetric PPO clipping $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.2)$, zero reward KL coefficient, and zero actor KL coefficient. The main reasoning-benchmark runs use 900 steps; the Phi-3.5-mini coefficient, component, and four-method baseline comparisons use checkpoint 300. Core methods save every 100 steps; GRPO-UR, REPO-R, and F-GRPO save every 200 steps and at the final step.

GRPO uses the group-standardized verifier advantage and a zero auxiliary exploration coefficient. ExPPO uses Eq. (8) with $\lambda = 0.2$. GRPO-H uses Eq. (3) with $\beta_{\text{ent}} = 0.001$. In the closest-method experiment, GRPO-UR uses unlikeliness coefficient 0.25, REPO-R uses $(\zeta_{\text{init}}, \zeta_{\min}, \zeta_{\max}) = (10^{-3}, 10^{-4}, 0.05)$, and F-GRPO uses focal exponent $\gamma = 0.5$. These modifiers are mutually exclusive and otherwise retain the GRPO configuration.

Closest baselines target rarity and entropy through distinct signals. F-GRPO multiplies every normalized advantage in prompt group q by $(1-\rho_q)^\gamma$, with $\gamma = 0.5$ (Plyusov et al., 2026). Our GRPO-UR reproduction follows Rewarding the Unlikely (He et al., 2025). Let $\bar{\ell}_i = T_i^{-1} \sum_t \log \pi_{\text{old}}(y_{i,t} \mid x_{i,t})$ and let $\text{rank}_i = \text{rank}_\uparrow(\bar{\ell}_i) \in \{0, \dots, G-1\}$, where rank zero is least likely. Before group normalization, nonzero-signal samples use

$$\widehat{r}_i^{\text{UR}} = r_i \left(1 - \beta_{\text{UR}} \frac{\text{rank}_i}{G}\right), \qquad \beta_{\text{UR}} = 0.25. \tag{24}$$

All-pass and all-fail groups retain their original zero advantage. Our REPO-R reproduction (Petrenko et al., 2026) applies the detached current-policy token log probability $\ell_{i,t} = \text{sg}[\log \pi_\theta(y_{i,t} \mid x_{i,t})]$ to the GRPO advantage:

$$\widetilde{A}_{i,t}^{\text{R}} = \begin{cases} \max\{A_i(1 - \zeta\ell_{i,t}), 0\}, & A_i > 0, \\ \min\{A_i(1 + \zeta\ell_{i,t}), 0\}, & A_i < 0, \\ 0, & A_i = 0. \end{cases} \tag{25}$$

Its target is the token-mean entropy observed on the first step. The signed controller doubles or halves $|\zeta|$ according to the target-entropy error and bounds its nonzero magnitude to $[10^{-4}, 0.05]$, starting from 10^{-3}.

Fixed budgets separate coverage from response accuracy. Standard evaluations sample with temperature 0.7, top-$p = 0.95$, and at most 10,240 new tokens. Generation ends at the tokenizer end-of-sequence token or the length cap. We render each model's serialized reasoning prompt with its tokenizer and request a boxed final answer; multiple-choice records additionally request one option letter. The benchmark-specific rule-based extractor scores the final answer.

AIME24 and AIME25 use $k = 16$, Minerva uses $k = 8$, and OlympiadBench, GPQA, MMLU-Pro, Safety, and TruthfulQA-MC use $k = 4$. For prompt q, avg@k is the fraction of its k responses that pass the verifier; reported benchmark scores average this quantity over prompts and call it *response accuracy*. The main-paper coverage measure follows Section 6.1. Checkpoint trends use one fixed decoding configuration. The Qwen2.5-7B closest-method comparison evaluates one fixed step-900 checkpoint per method with three decoding seeds. Its sample standard deviations quantify decoding variability conditional on the trained checkpoint; suite averages are computed within each decoding seed before aggregation.

Table 6: Reasoning-evaluation suites and prompt counts. Benchmark: evaluation set; Suite: In-Domain (ID) adaptation or Out-of-Domain (OOD) generalization; Prompts: evaluated questions. Each suite contains four tasks, and suite averages are unweighted.

Benchmark	Suite	Prompts
AIME24	ID	30
AIME25	ID	30
Minerva	ID	272
OlympiadBench	ID	675
GPQA	OOD	448
MMLU-Pro	OOD	70
Safety	OOD	1,000
TruthfulQA-MC	OOD	684

The ID evaluation uses the four benchmark files summarized in Table 6. The OOD files use GPQA's `gpqa_main`/train split, MMLU-Pro validation, the train split of the pairwise safety set[1], and TruthfulQA MC1. TruthfulQA-MC retains questions with at least four choices and deterministically samples four choices when more are available. This ID/OOD setup adapts the checkpoint-wise evaluation of Jin et al. (2025; 2026).

At each displayed k, large-sample coverage is the percentage of prompts with at least one verified success among their first k generated responses. This ordered-prefix estimate uses the responses available at each displayed budget. The maximum displayed budget is 256 for AIME24/25 and 128 for Minerva/OlympiadBench. The entropy curves read `actor/entropy` against `training/global_step`; duplicate resumed-log rows are deduplicated by step, and all plots stop at the final checkpoint.

E.1 Measured Training Cost

Table 7: Qwen2.5-7B-Instruct training cost on one eight-H200 node. Although the configured training prompt batch is 64, the synchronous data loader uses a generation batch of 32 prompts with 16 responses each, yielding 512 responses per logged step; Samples/s therefore equals 512 times Steps/s. Each profile runs for 100 steps, and throughput and per-GPU peak memory use steps 11–100. Estimated GPU-hours scale the full eight-GPU allocation wall time of the corresponding 100-step job by nine; † marks the separately profiled F-GRPO job. GRPO-H uses the unchunked entropy-loss profile with activation checkpointing disabled. Peak alloc./reserved are GiB per GPU; GRPO-UR abbreviates Rewarding the Unlikely.

Method	GPU-hours	Samples/s	Steps/s	Peak alloc.	Peak reserved
GRPO	151.70	6.997	0.01367	88.9 GiB	101.2 GiB
GRPO-H	157.76	6.710	0.01311	88.5 GiB	100.8 GiB
GRPO-UR	136.18	8.107	0.01583	86.5 GiB	98.6 GiB
REPO-R	151.66	7.055	0.01378	87.8 GiB	100.0 GiB
F-GRPO	146.08	7.289	0.01424	87.3 GiB	99.6 GiB
ExPPO	150.12	7.230	0.01412	87.7 GiB	99.9 GiB

ExPPO matches GRPO's compute envelope. ExPPO and GRPO occupy the same end-to-end compute band under the common schedule. GRPO-UR and F-GRPO are faster, GRPO-H takes more time, and all six methods remain closely matched in peak memory.

F Optimization Implementation

ExPPO transfers across GRPO and DAPO objectives. The GRPO implementation evaluates Eq. (2) with the substitution $A_i \leftarrow \widetilde{A}_i$. The DAPO implementation applies the same sequence-level

[1] https://huggingface.co/datasets/ThWu/safety_benchmark

substitution inside its configured asymmetric clipped objective and retains its dynamic sampling, token-level aggregation, and overlong-response shaping. This construction keeps the rollout and optimizer pipeline fixed while changing the allocation of detached response advantages.

GRPO-H supplies a policy-wide exploration contrast. Equation (3) defines the GRPO-H objective used in the main comparison. The implementation computes token-distribution entropy over the same generated-token mask and applies the same token aggregation convention as the policy loss. This baseline supplies a policy-wide exploration signal for comparison with the response-level allocation implemented by ExPPO.

ExPPO adds one bounded detached correction per response. Algorithm 1 summarizes the response-level shaping for one prompt group.

Algorithm 1 Exploration-Preserving Policy Optimization for one prompt group

Require: Prompt q, rollout policy π_{old}, group size G, coefficient $\lambda \in [0, 1]$

1: Sample responses $\{y_i\}_{i=1}^{G}$ and compute rewards $\{r_i\}_{i=1}^{G}$
2: Compute detached base advantages A_i and verifier-defined success indicators c_i
3: Compute s_i and u_i using Eq. (6); set $\rho_q \leftarrow G^{-1} \sum_i c_i$
4: Compute gates g_i and raw shaped advantages $\widehat{A}_i$ using Eq. (8)
5: Normalize $\widehat{A}_i$ into $\widetilde{A}_i$ using Eq. (9)
6: Broadcast $\widetilde{A}_i$ across response tokens and optimize the selected base policy loss

G Notation and Probability Conventions

Random responses and realized values. For a fixed prompt q, $Y_i \sim \pi_{\text{old}}(\cdot \mid q)$ is a random rollout and y_i its realized value. After observing $y_{1:G}$, indexed quantities such as T_i, A_i, and B_i, and group statistics ρ_q and α_q, are fixed. Uppercase A_i follows standard advantage notation. A fresh response $Y \sim \pi_{\text{old}}(\cdot \mid q)$ induces the random variables $A = A(Y)$, $B = B(Y)$, $S = S(Y)$, and $V = V(Y)$; $A(y)$, $B(y)$, $S(y)$, and $V(y)$ are their values at response y. The verifier $R(q, \cdot)$ is a deterministic function; under binary rewards, $C = R(q, Y)$ is the random success indicator. The prompt success probability p_q is a population quantity, and ρ_q is the observed group pass rate.

Frozen statistics and probability measures. Local analysis fixes q, $y_{1:G}$, and all group statistics before varying the response distribution. It uses the finite response space induced by the generation length limit. For any response in the rollout support, set $A(y) = (R(q, y) - \mu_{r,q})/(\sigma_{r,q} + \varepsilon)$ and $s(y) = -T(y)^{-1} \log \pi_{\text{old}}(y \mid q)$; compute $u(y)$, $g(y)$, $B(y)$, and $\widetilde{A}(y)$ using the same frozen group statistics as in the method. At $y = y_i$, these functions equal their indexed values. Unless a measure is shown explicitly, expectations, variances, and covariances in the proofs use $Y \sim \pi_0 = \pi_{\text{old}}(\cdot \mid q)$. Empirical group covariance uses the uniform measure on the realized sample indices:

$$\widehat{\text{Cov}}_q(v, z) = \frac{1}{G} \sum_{i=1}^{G} (v_i - \bar{v}_q)(z_i - \bar{z}_q), \qquad \bar{v}_q = \frac{1}{G} \sum_{i=1}^{G} v_i, \quad \bar{z}_q = \frac{1}{G} \sum_{i=1}^{G} z_i.$$

Thus Proposition 4 concerns observed, length-normalized surprisal, and Eq. (16) concerns population full-response surprisal. The scalar α_q remains fixed as the tilt step η varies. It may vary with λ; positive shared scaling cancels from Γ_V.

Table 8: Notation for the method and analysis. Indexed quantities are evaluated on the realized rollout group.

Symbol	Meaning
Rollouts, objectives, and shaping	
$q, \mathcal{D}, G$	Prompt, training prompt distribution, and responses per prompt group.
$Y_i, y_i;\ Y, y$	Random rollout and realized rollout; fresh random response and a value in its support.
$T_i, y_{i,t}, x_{i,t}$	Realized response length, token t, and context $(q, y_{i,<t})$.

Table 8 (continued)

Symbol	Meaning
$\theta, \pi_\theta, \pi_{\text{old}}, \pi_{\text{ref}}$	Parameters; current, frozen rollout, and fixed reference policies.
$R(q,y), r_i;\ C, c_i$	Verifier and realized reward; random and realized success indicators ($c_i = r_i$ for binary rewards).
$\mu_{r,q}, \sigma_{r,q}, \varepsilon$	Group reward mean, reward standard deviation, and numerical stabilizer.
$A_i, A(y), A = A(Y)$	Realized advantage, its frozen-statistic response function, and the induced random variable.
$\mathcal{B}, \langle\cdot\rangle$	Joint prompt/group sampling and canonical response-averaged token aggregation.
$w_{i,t}, \ell_{\text{clip}}, \epsilon_{\text{clip}}$	Token importance ratio, clipped surrogate, and clipping radius.
$d_{i,t}, \beta_{\text{KL}}, \beta_{\text{ent}}$	Reference-policy KL divergence and the reference-KL and entropy coefficients.
$d_a, \mu_\theta(x), \sigma_\theta(x)$	Action dimension, Gaussian policy mean, and componentwise standard deviation in Eq. (22).
z_a, p_a	Token logit and softmax probability at a fixed prefix in Eq. (23).
$s_i, \mu_{s,q}, \sigma_{s,q}$	Realized length-normalized response surprisal and its group mean and standard deviation.
$u_i, \rho_q, g_i, \lambda$	Bounded relative surprisal, observed pass rate, pass-rate gate, and shaping strength.
$B_i, \widehat{A}_i, \widetilde{A}_i$	Correction $\lvert A_i\rvert g_i u_i$, raw advantage $A_i + \lambda B_i$, and normalized advantage.
α_q, M_i	Shared absolute-mass scale and raw multiplier $1 + \lambda g_i \operatorname{sign}(A_i) u_i$.
$\psi_\rho(u)$	Gated score $u((1-\rho)\mathbf{1}[u>0] + \rho\mathbf{1}[u<0])$.
$\bar{M}_A, \xi_q, \kappa(\lambda)$	Absolute-advantage-weighted mean multiplier, stabilizer divided by base absolute mass, and relative multiplier-spread bound.
Local response updates and discovery	
$\pi_0, \pi_{\eta,h}, \eta$	Baseline response law, exponential tilt with score h, and tilt step size.
$h(q,y), h_d(y)$	Frozen response score and score induced by parameter direction d.
$S(y), S = S(Y)$	Full-response surprisal $-\log \pi_0(y)$ and its random-variable form; $S(y_i) = T_i s_i$.
H, H_b, D_{KL}	Shannon entropy, binary entropy, and KL divergence.
$\widehat{\text{Cov}}_q, \text{Cov}, \text{Var}$	Uniform empirical group covariance, population covariance, and population variance.
$V(y), J_V(\pi)$	Frozen response utility and its expectation under π.
$\delta, \eta_h(\delta), \Gamma_V(h)$	KL budget, score-specific step attaining it, and $\text{Cov}(V,h)/\sqrt{\text{Var}(h)}$.
$B_\perp$	Correction after removing its covariance-aligned component along A.
$\mathcal{M}_q, \mathcal{M}_q^+, m(y)$	Response-mode partition, its correct modes, and the mode of response y.
$p_m(\pi), p_m, p_q, p_+$	Mode mass under π; baseline mode mass; success mass; the proof notation $p_+ = \sum_{m\in\mathcal{M}_q^+} p_m = p_q$.
K, F_K, V_K	Evaluation sample budget, expected number of distinct correct modes, and its frozen local utility.
$w_m, Q, D_{\text{pair}}, V_D$	Correct-conditional mode mass, concentration $Q = \sum_m w_m^2$, pair disagreement $1-Q$, and its frozen local utility.
$\pi_\tau, \dot{p}_m, \dot{p}_+$	Differentiable policy path and mode/success-mass derivatives at $\tau = 0$.
$Z_h(\eta), K_h(\eta), \nu$	Tilt normalizer, its logarithm, and the normalization Lagrange multiplier.
Mode logits, gradients, and finite-λ conditions	
$z_m, p_z, m_i, n_m, A^+, A^-$	Mode logit, softmax mode policy, sampled mode, its count, and positive/negative binary-GRPO advantages.
C_m, C_m^B, B_Σ	Mode-summed base credit, mode-summed correction, and total group correction.
$\mathcal{L}_q^h, H_m, H_\Sigma, h_{\text{eff}}$	Sampled logit surrogate, mode/group score sums, and effective logit-step score.
$\Delta_{ij}, \Delta_{mm'}, \delta^A_{mm'}, \delta^{\text{E}}_{mm'}$	Response/mode relative log-odds and their base/ExPPO logit-step changes.
$\omega_{i,t}, f_i, \mathcal{G}_q(h), d$	Token aggregation weight, response gradient feature, prompt gradient, and parameter update direction.
a, b, v, c, d	Scalar proof abbreviations: $\text{Cov}(V,A)$, $\text{Cov}(V,B)$, $\text{Var}(A)$, $\text{Cov}(A,B)$, and $\text{Var}(B)$. Here d is a scalar variance.
a_R, b_R, b_D	Binary-advantage coefficients in $A = a_R C + b_R$ and diversity alignment $b_D = \text{Cov}(V_D, B)$.
$P_0, P_1, \lambda_{\text{num}}, \lambda_{\text{perf}}$	Coefficients of the finite-gain inequality and its numerator/utility thresholds.

C_B^H, λ_H	Entropy alignment $\mathrm{Cov}(S, B)$ and the nonnegative-entropy-slope threshold.

H PROOFS

The local-update lemma and local identities. For fixed q, abbreviate $\pi_0(y) = \pi_{\mathrm{old}}(y \mid q)$ and $h(y) = h(q, y)$. In expectations and covariances, h, V, and S mean $h(Y)$, $V(Y)$, and $S(Y)$ under the indicated response law. Adding a Lagrange multiplier ν for $\sum_y \pi(y) = 1$ to Eq. (10) gives

$$\mathcal{J}(\pi, \nu) = \sum_y \pi(y)h(y) - \frac{1}{\eta}\sum_y \pi(y) \log \frac{\pi(y)}{\pi_0(y)} + \nu\left(\sum_y \pi(y) - 1\right). \tag{26}$$

Stationarity with respect to $\pi(y)$ implies $\log(\pi(y)/\pi_0(y)) = \eta h(y) + \eta\nu - 1$. The last two terms are constant in y, and normalization yields Eq. (11).

Let $Z_h(\eta) = \mathbb{E}_{\pi_0}[e^{\eta h}]$. Differentiating the normalized tilt gives

$$\frac{d\pi_{\eta,h}(y)}{d\eta} = \pi_{\eta,h}(y)\left(h(y) - \mathbb{E}_{\pi_{\eta,h}}[h]\right). \tag{27}$$

At $\eta = 0$, $\pi_{\eta,h} = \pi_0$, so for any fixed V,

$$\frac{d}{d\eta}\mathbb{E}_{\pi_{\eta,h}}[V]\bigg|_0 = \mathbb{E}_{\pi_0}[V(h - \mathbb{E}[h])] = \mathrm{Cov}_{\pi_0}(V, h). \tag{28}$$

For response entropy $H(\pi_{\eta,h}) = -\sum_y \pi_{\eta,h}(y) \log \pi_{\eta,h}(y)$, the derivative of the normalization term sums to zero. Defining $S(y) = -\log \pi_0(y)$ and substituting Eq. (27) at zero yields

$$\frac{dH(\pi_{\eta,h})}{d\eta}\bigg|_0 = \mathrm{Cov}_{\pi_0}(S, h), \tag{29}$$

which proves Eq. (16) after setting $h = \alpha_q(A + \lambda B)$.

Finally, writing $K_h(\eta) = \log Z_h(\eta)$ gives

$$D_{\mathrm{KL}}(\pi_{\eta,h} \| \pi_0) = \eta K_h'(\eta) - K_h(\eta) = \frac{\eta^2}{2}\mathrm{Var}_{\pi_0}(h) + O(\eta^3). \tag{30}$$

Solving for the score-specific step size $\eta_h(\delta)$ at a small KL budget δ and substituting into the utility expansion gives

$$J_V(\pi_{\eta_h,h}) - J_V(\pi_0) = \sqrt{2\delta}\,\frac{\mathrm{Cov}(V, h)}{\sqrt{\mathrm{Var}(h)}} + O(\delta), \tag{31}$$

which justifies Eq. (17).

Bounded polarity preservation. Since $g_i \in [0, 1]$, $|u_i| \le 1$, and $|\mathrm{sign}(A_i)| \le 1$, we have $-1 \le g_i \,\mathrm{sign}(A_i) u_i \le 1$. Multiplying by λ and adding one gives $1 - \lambda \le M_i \le 1 + \lambda$. Because $\widehat{A}_i = A_i M_i$ and $M_i \ge 0$, $A_i \widehat{A}_i = A_i^2 M_i \ge 0$. Moreover,

$$|\widehat{A}_i - A_i| = \lambda |A_i| g_i |u_i| \le \lambda |A_i|. \tag{32}$$

The normalization in Eq. (9) is a common nonnegative factor and therefore preserves sign and ordering.

For the normalized distortion, define

$$\overline{M}_A = \frac{\sum_j |A_j| M_j}{\sum_j |A_j|}, \qquad \xi_q = \frac{\varepsilon}{\sum_j |A_j|}. \tag{33}$$

Whenever $A_i \ne 0$, $\widetilde{A}_i / A_i = M_i / (\overline{M}_A + \xi_q)$. Both M_i and its $|A|$-weighted average lie in $[1 - \lambda, 1 + \lambda]$, so for $\lambda < 1$,

$$\frac{1 - \lambda}{1 + \lambda + \xi_q} \le \frac{\widetilde{A}_i}{A_i} \le \frac{1 + \lambda}{1 - \lambda + \xi_q}. \tag{34}$$

Within a prompt group, the shared denominator cancels in the ratio of any two realized multipliers $\widetilde{A}_i / A_i$. Their largest-to-smallest ratio is therefore at most $\kappa(\lambda) = (1 + \lambda)/(1 - \lambda)$.

Surprisal-ordered response preference. For $u > 0$, $\partial h_\lambda / \partial u = \lambda |A|(1 - \rho_q) > 0$; for $u < 0$, $\partial h_\lambda / \partial u = \lambda |A| \rho_q > 0$. Both branches meet continuously at zero, so h_λ is strictly increasing. The normalization in Eq. (9) multiplies every score in the group by the same positive scalar and preserves the ordering. From Eq. (11),

$$\frac{\pi_{\eta,h}(y_i \mid q)}{\pi_{\eta,h}(y_j \mid q)} = \frac{\pi_0(y_i \mid q)}{\pi_0(y_j \mid q)} e^{\eta(h_i - h_j)}, \tag{35}$$

which gives Eq. (14) after taking logarithms.

Surprisal alignment and matched-KL derivative. Let $\overline{B}_q = G^{-1} \sum_{i=1}^{G} B_i$ be the group mean of the exploration correction. The empirical covariance expands as

$$\begin{aligned} \widehat{\mathrm{Cov}}_q(s, B) &= \frac{1}{G} \sum_i (s_i - \mu_{s,q})(B_i - \overline{B}_q) \\ &= \frac{1}{G} \sum_i (s_i - \mu_{s,q}) B_i, \end{aligned} \tag{36}$$

because centered surprisal sums to zero. Since u_i has the same sign as $s_i - \mu_{s,q}$ and $|A_i| g_i \geq 0$, each term $(s_i - \mu_{s,q}) B_i = |A_i| g_i (s_i - \mu_{s,q}) u_i$ is nonnegative.

For the matched-KL derivative, abbreviate

$$a = \mathrm{Cov}(V, A), \quad b = \mathrm{Cov}(V, B), \quad v = \mathrm{Var}(A), \quad c = \mathrm{Cov}(A, B), \quad d = \mathrm{Var}(B). \tag{37}$$

Then

$$\Gamma_V(A + \lambda B) = \frac{a + \lambda b}{\sqrt{v + 2\lambda c + \lambda^2 d}}. \tag{38}$$

Differentiating at zero gives $(bv - ac)/v^{3/2}$. By Eq. (18), $\mathrm{Cov}(V, B_\perp) = b - ac/v = (bv - ac)/v$, proving Eq. (19).

Utilities for correct-mode discovery and diversity. Fix a finite partition into correct modes $\mathcal{M}_q^+$ and write $p_m = \mathbb{P}_{Y \sim \pi_0}(Y \in m \mid q)$ and $p_+ = \sum_m p_m \in (0, 1)$. A mode with mass $p_m(\pi)$ appears at least once in K independent responses with probability $1 - (1 - p_m(\pi))^K$. Summing these probabilities gives $F_K(\pi)$ in Section 5. Define the frozen response utility

$$V_K(y) = \begin{cases} K(1 - p_{m(y)})^{K-1}, & R(q, y) = 1, \\ 0, & R(q, y) = 0. \end{cases}$$

For any differentiable policy path π_τ through π_0, differentiation gives

$$\left. \frac{dF_K(\pi_\tau)}{d\tau} \right|_0 = \sum_m K(1 - p_m)^{K-1} \left. \frac{dp_m(\pi_\tau)}{d\tau} \right|_0 = \left. \frac{d\mathbb{E}_{\pi_\tau}[V_K]}{d\tau} \right|_0.$$

Consequently, the exponential tilt has discovery slope $\mathrm{Cov}(V_K, h)$ and small-KL gain

$$F_K(\pi_{\eta_h,h}) - F_K(\pi_0) = \sqrt{2\delta}\, \Gamma_{V_K}(h) + O(\delta).$$

This is the population quantity estimated by fixed-budget observed correct-mode yield.

For conditional correct-pair diversity, let $w_m = p_m / p_+$, $Q = \sum_m w_m^2$, and $D_{\mathrm{pair}}(\pi_0) = 1 - Q$. Its frozen response utility is

$$V_D(y) = \begin{cases} \dfrac{2}{p_+}(Q - w_{m(y)}), & R(q, y) = 1, \\ 0, & R(q, y) = 0. \end{cases}$$

Indeed, writing $\dot{p}_m = dp_m(\pi_\tau)/d\tau|_0$ and $\dot{p}_+ = \sum_m \dot{p}_m$ gives

$$\left. \frac{dD_{\mathrm{pair}}(\pi_\tau)}{d\tau} \right|_0 = \frac{2}{p_+} \left(Q \dot{p}_+ - \sum_m w_m \dot{p}_m \right) = \left. \frac{d\mathbb{E}_{\pi_\tau}[V_D]}{d\tau} \right|_0.$$

Its tilt slope is therefore $\mathrm{Cov}(V_D, h)$, and its leading small-KL gain is $\sqrt{2\delta}\, \Gamma_{V_D}(h)$. This identifies the population counterpart of the correct-pair statistic in Appendix J.7.

With binary rewards and frozen normalization, $A = a_R C + b_R$ for $a_R > 0$. Hence $\mathrm{Cov}(C, B_\perp) = 0$. Also $\mathbb{E}[V_D] = \mathbb{E}[C V_D] = 0$, so $\mathrm{Cov}(V_D, A) = 0$ and $\mathrm{Cov}(V_D, B_\perp) = \mathrm{Cov}(V_D, B)$. The discovery utility V_K varies across correct modes for $K \geq 2$ whenever their masses differ, and the diversity utility directly measures redistribution within the correct set.

I FINITE-λ MATCHED-KL CERTIFICATE

The derivative in Eq. (19) certifies a neighborhood around $\lambda = 0$. The following result gives an observable finite interval. Use the abbreviations in Eq. (38) and assume $a > 0$ and $\mathrm{Cov}(V, B_\perp) > 0$. Define

$$P_0 = 2a(vb - ac), \qquad P_1 = vb^2 - a^2 d, \tag{39}$$

$$\lambda_{\mathrm{num}} = \begin{cases} +\infty, & b \geq 0, \\ a/(-b), & b < 0, \end{cases} \qquad \lambda_{\mathrm{perf}} = \begin{cases} \min\{\lambda_{\mathrm{num}}, P_0/(-P_1)\}, & P_1 < 0, \\ \lambda_{\mathrm{num}}, & P_1 \geq 0. \end{cases} \tag{40}$$

Then every $0 < \lambda < \min\{1, \lambda_{\mathrm{perf}}\}$ satisfying $a + \lambda b > 0$ obeys

$$\Gamma_V(A + \lambda B) > \Gamma_V(A). \tag{41}$$

To verify the result, both sides of the comparison are positive under $a > 0$ and $a + \lambda b > 0$. Squaring Eq. (41), cross-multiplying, cancelling $a^2 v$, and dividing by $\lambda > 0$ reduces the condition to $P_0 + \lambda P_1 > 0$. The residual-alignment assumption implies $vb - ac = v\mathrm{Cov}(V, B_\perp) > 0$, hence $P_0 > 0$; Eq. (40) is exactly the interval on which the remaining linear inequality and the positive-numerator condition hold.

For $V = V_K$ in the binary-reward setting above, $a = a_R(1 - p_+)\mathbb{E}[V_K] > 0$, so the discovery utility satisfies the positive-base-alignment requirement. For conditional diversity, $\mathrm{Cov}(V_D, A) = 0$. If $b_D = \mathrm{Cov}(V_D, B) > 0$, Eq. (38) gives

$$\Gamma_{V_D}(A + \lambda B) = \frac{\lambda b_D}{\sqrt{v + 2\lambda c + \lambda^2 d}} > \Gamma_{V_D}(A) = 0$$

for every $0 < \lambda \leq 1$ with $\mathrm{Var}(A + \lambda B) > 0$. These conditions compare the leading small-KL coefficients; for each fixed admissible λ, they yield a positive gain over the base update at sufficiently small KL budgets.

An analogous entropy threshold follows from Eq. (16). If $C_B^H = \mathrm{Cov}(S, B) > 0$, then the first-order entropy direction is nonnegative for

$$\lambda \geq \lambda_H := \max\left\{0, -\frac{\mathrm{Cov}(S, A)}{\mathrm{Cov}(S, B)}\right\}. \tag{42}$$

When the intervals overlap, the measured covariances identify values of λ that satisfy both the entropy-direction and matched-KL target-utility conditions.

J More Experimental Results

J.1 Coefficient Sensitivity and Component Ablations

Table 9: Phi-3.5-mini coefficient and component ablations at checkpoint 300. Coverage (%) at $k = 16$ for AIME24/25, $k = 8$ for Minerva, and $k = 4$ otherwise. Full ExPPO, component variants, and the shuffled control use $\lambda = 0.2$. Full ExPPO: three-seed mean; other settings: one run. Avg.: unweighted four-task mean; reported macros may differ by 0.01 from rounded task averages. Olympiad/TruthfulQA abbreviate OlympiadBench/TruthfulQA-MC. Bold: highest ExPPO domain mean.

Domain	Setting	AIME24	AIME25	Minerva	Olympiad	ID Avg.
ID	GRPO	3.33	3.33	40.81	19.26	16.68
	$\lambda = 0.10$	10.00	10.00	41.54	25.19	21.68
	$\lambda = 0.30$	13.33	3.33	41.91	28.74	21.83
	$\lambda = 0.50$	16.67	3.33	43.38	26.81	22.55
	$\lambda = 0.75$	16.67	3.33	43.38	27.85	22.81
	$\lambda = 1.00$	16.67	6.67	44.49	26.96	23.70
	No final normalization	10.00	10.00	40.81	27.26	22.02
	No pass-rate gate	16.67	3.33	43.75	27.56	22.83
	No $\tanh$ bound	10.00	10.00	43.01	27.85	22.72
	Positive only	13.33	6.67	44.49	27.85	23.09
	Negative only	10.00	6.67	42.28	26.96	21.48
	Within-class shuffled	10.00	6.67	42.65	28.30	21.90
	Full ExPPO	17.78	13.33	44.12	27.46	**25.67**
Domain	Setting	GPQA	MMLU-Pro	Safety	TruthfulQA	OOD Avg.
OOD	GRPO	43.08	57.14	67.50	44.88	53.15
	$\lambda = 0.10$	47.32	67.14	71.00	64.77	62.56
	$\lambda = 0.30$	54.24	72.86	77.20	76.61	70.23
	$\lambda = 0.50$	52.46	67.14	77.60	77.49	68.67
	$\lambda = 0.75$	57.14	75.71	77.10	76.17	**71.53**
	$\lambda = 1.00$	46.65	68.57	76.80	73.25	66.32
	No final normalization	50.67	67.14	75.40	70.18	65.85
	No pass-rate gate	46.65	65.71	76.20	75.00	65.89
	No $\tanh$ bound	50.67	65.71	75.80	69.59	65.44
	Positive only	42.86	60.00	72.40	63.74	59.75
	Negative only	46.65	72.86	74.80	69.44	65.94
	Within-class shuffled	46.65	61.43	74.90	70.91	63.47
	Full ExPPO	48.51	69.52	75.53	71.74	66.33

All swept coefficients improve coverage over GRPO. Every ExPPO coefficient exceeds the matched GRPO baseline on both domain macros. The default $\lambda = 0.2$ maximizes ID coverage, while OOD coverage peaks at $\lambda = 0.75$, showing a coefficient-dependent adaptation–generalization tradeoff.

Component removal reveals an adaptation–generalization tradeoff. Full ExPPO has the highest ID and OOD means among the component variants. Positive-only shaping favors ID coverage (23.09 vs. 21.48), whereas negative-only shaping favors OOD coverage (65.94 vs. 59.75).

Within-class shuffling tests response-specific credit assignment. To test credit assignment beyond class-level reweighting, we compute ExPPO advantages normally, then randomly permute the final sequence-level advantages within each prompt, separately among correct and incorrect responses, before broadcasting them to tokens. This preserves each class's coefficient distribution and total sequence-level mass while disrupting response assignment. Token-weighted update mass can change, so we log it alongside realized policy KL. The shuffled control reaches 21.90 ID, 63.47 OOD, and 42.69 overall coverage, below full ExPPO's three-seed means (25.67, 66.33, and 46.00). This single-run comparison is consistent with a benefit from response-specific assignment; token-weighted mass and realized KL provide additional factors for interpreting the result.

J.2 Closest Baselines at the Phi-3.5-mini Ablation Budget

Table 10: Phi-3.5-mini at checkpoint 300 under the ablation setting of Table 9. Entries are mean ± sample standard deviation across three training seeds per method. Avg.: unweighted four-task mean within each seed; Overall: eight-task mean. Task summaries use the reported seed scores; macro summaries retain author-reported precision, which can differ by 0.01 point from aggregation of rounded task entries. Individual seeds: Tables 11 and 12. Bold: highest mean across the four methods. ExPPO uses $\lambda = 0.2$.

A. Coverage (%)

Domain	Benchmark	k	F-GRPO	GRPO-UR	REPO-R	ExPPO
ID	AIME24	16	5.55±6.94	12.22±1.92	14.44±5.09	**17.78±3.85**
	AIME25	16	3.33±3.34	7.78±5.09	2.22±1.92	**13.33±0.00**
	Minerva	8	33.58±10.99	41.79±2.50	**46.20±1.12**	44.12±0.97
	Olympiad	4	13.04±3.71	24.84±1.77	24.89±0.40	**27.46±0.68**
	Avg.	—	13.88±5.76	21.66±0.75	21.94±1.89	**25.67±1.31**
OOD	GPQA	4	33.78±18.91	45.54±4.84	**50.60±0.25**	48.51±1.58
	MMLU-Pro	4	41.91±27.86	68.10±6.44	67.62±1.65	**69.52±2.18**
	Safety	4	55.73±24.30	70.53±5.23	75.30±0.26	**75.53±0.49**
	TruthfulQA	4	40.20±30.83	66.86±5.77	**71.88±0.75**	71.74±3.65
	Avg.	—	42.91±25.37	62.76±5.48	**66.35±0.38**	66.33±1.59
Overall	Avg.	—	28.39±15.22	42.21±2.65	44.14±1.00	**46.00±1.39**

B. Response accuracy (%)

Domain	Benchmark	k	F-GRPO	GRPO-UR	REPO-R	ExPPO
ID	AIME24	16	1.74±1.67	2.02±0.64	1.53±0.44	**2.85±0.43**
	AIME25	16	0.21±0.21	0.76±0.12	0.14±0.12	**1.11±0.32**
	Minerva	8	17.55±3.72	22.96±2.71	**24.25±0.57**	23.05±0.18
	Olympiad	4	6.18±0.61	13.42±2.33	12.65±0.47	**15.36±0.54**
	Avg.	—	6.42±0.81	9.79±1.11	9.64±0.19	**10.59±0.27**
OOD	GPQA	4	21.95±10.03	**27.71±1.27**	26.69±0.17	25.17±0.79
	MMLU-Pro	4	34.76±23.49	**52.50±2.86**	46.91±0.83	48.57±2.70
	Safety	4	51.58±25.65	68.12±5.62	**72.10±0.16**	71.98±0.20
	TruthfulQA	4	38.00±30.22	63.09±5.76	62.96±1.16	**63.32±1.85**
	Avg.	—	36.57±22.19	**52.85±3.62**	52.16±0.15	52.26±1.17
Overall	Avg.	—	21.50±11.43	31.32±2.37	30.90±0.08	**31.43±0.72**

Coverage and accuracy favor different allocations. ExPPO has the highest overall coverage (46.00), followed by REPO-R (44.14), GRPO-UR (42.21), and F-GRPO (28.39). ExPPO leads coverage on five benchmarks, while REPO-R leads on Minerva, GPQA, and TruthfulQA. GRPO-UR combines the highest OOD response accuracy with lower OOD coverage, illustrating the distinct roles of single-response success and prompt coverage at the reported sampling budgets. REPO-R has the smallest OOD variation in both metrics; F-GRPO has the largest. These are descriptive comparisons across three training seeds.

Table 11: Individual F-GRPO and ExPPO training seeds for Phi-3.5-mini at checkpoint 300. Each cell is coverage / response accuracy (percent), at $k = 16$ for AIME24/25, 8 for Minerva, and 4 otherwise. Task scores and macro rows retain author-reported precision; aggregation from rounded task entries can differ by 0.01 point. The other six runs appear in Table 12.

Method	Benchmark / suite	Seed 42	Seed 43	Seed 44
F-GRPO	AIME24	3.33 / 3.33	0.00 / 0.00	13.33 / 1.88
	AIME25	0.00 / 0.00	3.33 / 0.21	6.67 / 0.42
	OlympiadBench	9.19 / 6.19	13.33 / 5.56	16.59 / 6.78
	Minerva	21.69 / 13.28	35.66 / 19.30	43.38 / 20.08
	ID macro	**8.55 / 5.70**	**13.08 / 6.27**	**19.99 / 7.29**
	MMLU-Pro	10.00 / 8.21	54.29 / 43.21	61.43 / 52.86
	Safety	27.70 / 22.10	70.80 / 68.75	68.70 / 63.90
	TruthfulQA-MC	4.82 / 3.36	54.53 / 51.68	61.26 / 58.95
	GPQA	12.50 / 10.38	40.18 / 27.29	48.66 / 28.18
	OOD macro	**13.76 / 11.01**	**54.95 / 47.73**	**60.01 / 50.97**
	Overall macro	**11.15 / 8.36**	**34.02 / 27.00**	**40.00 / 29.13**
ExPPO	AIME24	20.00 / 3.33	13.33 / 2.71	20.00 / 2.50
	AIME25	13.33 / 1.46	13.33 / 0.83	13.33 / 1.04
	OlympiadBench	27.85 / 15.96	26.67 / 15.19	27.85 / 14.93
	Minerva	45.22 / 22.89	43.38 / 23.02	43.75 / 23.25
	ID macro	**26.60 / 10.91**	**24.18 / 10.44**	**26.23 / 10.43**
	MMLU-Pro	70.00 / 51.43	67.14 / 46.07	71.43 / 48.21
	Safety	76.10 / 71.93	75.30 / 72.20	75.20 / 71.80
	TruthfulQA-MC	75.88 / 65.17	69.01 / 63.30	70.32 / 61.48
	GPQA	50.22 / 25.89	48.21 / 25.28	47.10 / 24.33
	OOD macro	**68.05 / 53.60**	**64.92 / 51.71**	**66.01 / 51.46**
	Overall macro	**47.33 / 32.26**	**44.55 / 31.08**	**46.12 / 30.94**

Table 12: Individual GRPO-UR and REPO-R training seeds for Phi-3.5-mini at checkpoint 300. Each cell is coverage / response accuracy (percent), with budgets and rounding conventions as in Table 11. Together, the two seed tables report all twelve runs.

Method	Benchmark / suite	Seed 42	Seed 43	Seed 44
GRPO-UR	AIME24	13.33 / 1.88	10.00 / 1.46	13.33 / 2.71
	AIME25	6.67 / 0.83	3.33 / 0.63	13.33 / 0.83
	OlympiadBench	25.63 / 15.00	26.07 / 14.52	22.81 / 10.74
	Minerva	42.65 / 25.18	43.75 / 23.76	38.97 / 19.94
	ID macro	**22.07 / 10.72**	**20.79 / 10.09**	**22.11 / 8.56**
	MMLU-Pro	74.29 / 55.36	68.57 / 52.50	61.43 / 49.64
	Safety	73.80 / 71.55	73.30 / 71.18	64.50 / 61.63
	TruthfulQA-MC	72.08 / 67.65	67.84 / 64.99	60.67 / 56.62
	GPQA	48.66 / 29.07	47.99 / 26.56	39.96 / 27.51
	OOD macro	**67.21 / 55.91**	**64.42 / 53.81**	**56.64 / 48.85**
	Overall macro	**44.64 / 33.32**	**42.61 / 31.95**	**39.38 / 28.70**
REPO-R	AIME24	13.33 / 1.88	10.00 / 1.04	20.00 / 1.67
	AIME25	3.33 / 0.21	0.00 / 0.00	3.33 / 0.21
	OlympiadBench	24.44 / 12.22	25.04 / 12.59	25.19 / 13.15
	Minerva	45.22 / 23.62	45.96 / 24.72	47.43 / 24.40
	ID macro	**21.58 / 9.48**	**20.25 / 9.59**	**23.99 / 9.86**
	MMLU-Pro	68.57 / 46.43	65.71 / 46.43	68.57 / 47.86
	Safety	75.50 / 72.28	75.40 / 71.98	75.00 / 72.03
	TruthfulQA-MC	72.51 / 63.89	72.08 / 63.34	71.05 / 61.66
	GPQA	50.45 / 26.73	50.89 / 26.51	50.45 / 26.84
	OOD macro	**66.76 / 52.33**	**66.02 / 52.06**	**66.27 / 52.10**
	Overall macro	**44.17 / 30.91**	**43.13 / 30.83**	**45.13 / 30.98**

Seed ranges clarify the robustness of the comparison. Every ExPPO ID coverage run (24.18–26.60) exceeds every baseline run; the best baseline run is REPO-R at 23.99. OOD coverage ranges overlap for ExPPO (64.92–68.05), REPO-R (66.02–66.76), and GRPO-UR (56.64–67.21). F-GRPO spans 13.76–60.01, with its lowest run accounting for much of its variation. GRPO-UR's higher mean OOD response accuracy comes with a wider range (48.85–55.91) than ExPPO (51.46–53.60) or REPO-R (52.06–52.33).

J.3 Independent-Run Stability on Phi-3.5-mini

Table 13: Phi-3.5-mini performance across the original and two additional independently trained runs at the last checkpoint. Domain: In-Domain (ID) or Out-of-Domain (OOD); upper panel: any-success coverage; lower panel: mean verified-response share. Scores are in percentage points at $k = 16$ for AIME24/25, $k = 8$ for Minerva, and $k = 4$ otherwise. Avg.: unweighted four-task mean within each run, followed by the across-run mean $\pm$ sample standard deviation. Olympiad/TruthfulQA abbreviate OlympiadBench/TruthfulQA-MC. Bold: stronger mean per column.

Coverage

Domain	Method	AIME24	AIME25	Minerva	Olympiad	Avg.
ID	GRPO	7.78±5.09	2.22±1.92	33.09±13.78	16.15±7.40	14.81±6.89
	ExPPO	**14.44±5.09**	**8.89±3.85**	**44.85±2.24**	**27.80±1.11**	**24.00±0.62**
Domain	Method	GPQA	MMLU-Pro	Safety	TruthfulQA	Avg.
OOD	GRPO	41.00±6.19	58.57±6.23	70.53±4.37	64.57±2.05	58.67±3.07
	ExPPO	**49.78±3.72**	**71.90±5.02**	**75.63±0.67**	**71.93±3.98**	**67.31±3.25**

Response accuracy

Domain	Method	AIME24	AIME25	Minerva	Olympiad	Avg.
ID	GRPO	1.60±1.53	0.21±0.21	18.03±6.49	8.04±3.06	6.97±2.13
	ExPPO	**2.71±0.55**	**0.76±0.12**	**24.32±2.06**	**15.42±0.64**	**10.80±0.82**
Domain	Method	GPQA	MMLU-Pro	Safety	TruthfulQA	Avg.
OOD	GRPO	**26.94±1.18**	**49.40±2.58**	67.16±5.56	62.17±2.08	51.42±0.73
	ExPPO	25.39±1.12	47.62±1.35	**72.12±0.28**	**63.46±2.07**	**52.15±0.98**

ExPPO is stable across mathematical adaptation and OOD generalization. ExPPO raises mean coverage on all eight tasks and both domain averages. Its ID coverage varies little across training runs, while GRPO is more sensitive to the training trajectory. Response accuracy follows the same macro pattern: ExPPO gains on the mathematical suite and the OOD average, with complementary task-level strengths on the OOD suite.

J.4 Code-Reasoning Adaptation and OOD Generalization

We evaluate Phi-3.5-mini on a CodeReasoner-derived Python execution-reasoning set using the released data[2] (Tang et al., 2025). Table 5 reports the code and OOD-summary results; Table 14 gives the OOD task breakdown for the same evaluations.

[2] CodeReasoner-Dataset on Hugging Face

Table 14: OOD generalization after code adaptation, disaggregating Table 5. Upper panel: any-success prompt coverage at $k = 4$; lower panel: mean verified-response share. Scores are percentages. Initialization: fixed base model; trained entries: mean ± sample standard deviation across three independent runs at the last checkpoint. Bold: best trained mean.

Method	GPQA	MMLU-Pro	Safety	TruthfulQA-MC
Coverage				
Initialization	38.17	58.57	74.50	66.37
GRPO	32.96±0.46	47.62±1.65	70.10±0.30	58.28±0.83
GRPO-H	37.57±0.90	59.52±3.60	72.37±0.40	72.56±0.97
ExPPO	**44.35±1.65**	**65.24±4.12**	**76.37±0.25**	**77.00±0.69**
Response accuracy				
Initialization	17.52	33.93	66.75	49.52
GRPO	14.55±0.23	30.60±3.38	61.88±0.65	38.65±0.93
GRPO-H	16.39±0.67	35.71±1.07	60.65±0.90	55.67±0.31
ExPPO	**21.73±0.90**	**44.64±4.02**	**72.51±0.14**	**65.22±0.77**

ExPPO improves code adaptation and OOD generalization. ExPPO finishes highest on both code metrics and every OOD column. Its higher OOD response accuracy shows that the generalization gain includes a larger share of correct responses across samples.

J.5 Per-Benchmark Checkpoint Trajectories

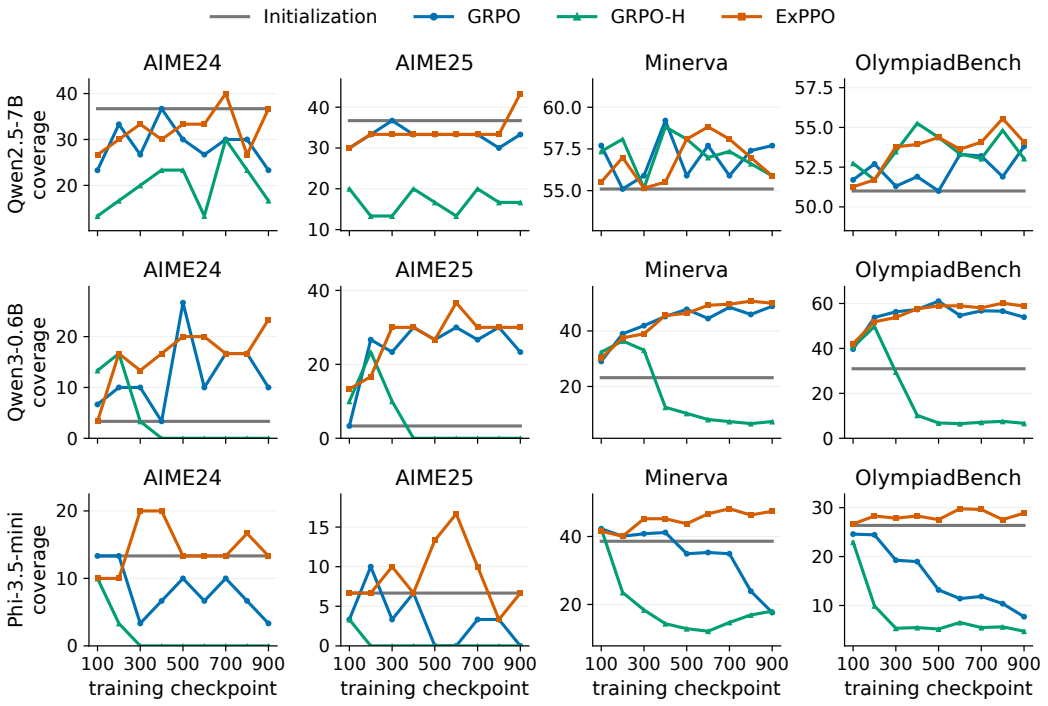

Figure 5: Per-task In-Domain (ID) coverage through the last checkpoint; disaggregation of Fig. 1. Rows: Qwen2.5-7B, Qwen3-0.6B, Phi-3.5-mini; columns: AIME24, AIME25, Minerva, OlympiadBench. Horizontal axis: checkpoint; vertical axis: coverage in percentage points at $k = 16, 16, 8, 4$, respectively. Gray lines: Initialization; colored curves: GRPO, GRPO-H, ExPPO.

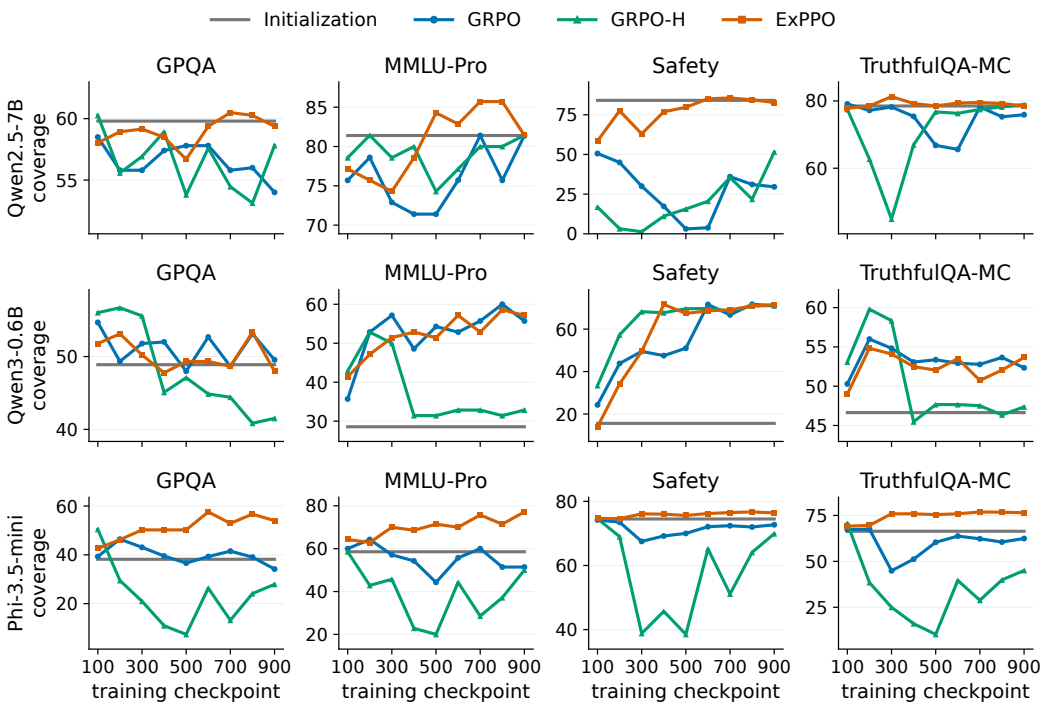

Figure 6: Per-task Out-of-Domain (OOD) coverage through the last checkpoint; disaggregation of Fig. 1. Rows: Qwen2.5-7B, Qwen3-0.6B, Phi-3.5-mini; columns: GPQA, MMLU-Pro, Safety, TruthfulQA-MC. Horizontal axis: checkpoint; vertical axis: coverage in percentage points at $k = 4$. Gray lines: Initialization; colored curves: GRPO, GRPO-H, ExPPO.

ExPPO is consistent while GRPO-H collapses on smaller backbones. Figures 5 and 6 recover the macro pattern while exposing task-specific variation. ExPPO finishes strongest on most panels, whereas GRPO-H loses mathematical coverage across Qwen3-0.6B and Phi-3.5-mini and becomes unstable on several OOD tasks.

J.6 DAPO-Based ExPPO

ExPPO extends to the DAPO recipe. We apply ExPPO to DAPO while retaining asymmetric clipping, dynamic sampling, token-level loss aggregation, and overlong-response shaping (Yu et al., 2025). Figure 7 shows DAPO's OOD macro declining through training, whereas ExPPO rises and finishes above its initialization.

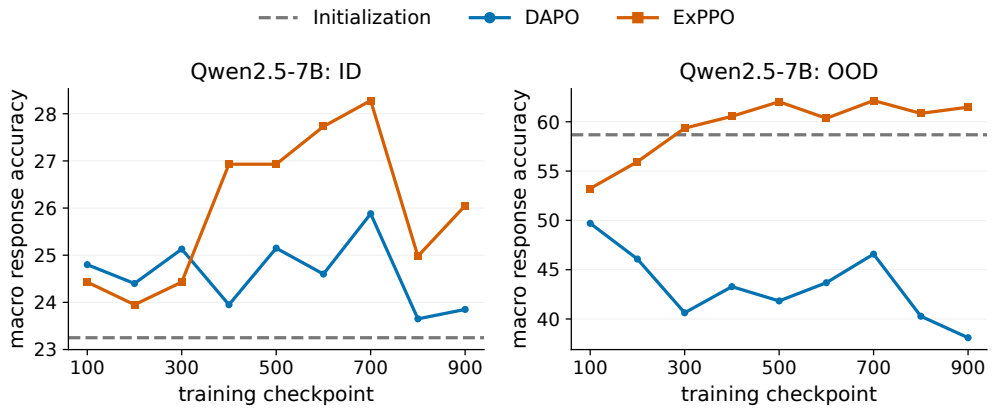

Figure 7: Qwen2.5-7B macro response-accuracy trajectories under DAPO. Left: unweighted mean over AIME24, AIME25, Minerva, OlympiadBench (In-Domain; ID); right: unweighted mean over GPQA, MMLU-Pro, Safety, TruthfulQA-MC (Out-of-Domain; OOD). Horizontal axis: checkpoint; vertical axis: response accuracy in percentage points. Gray dashed lines: Initialization; colored curves: DAPO, ExPPO.

Table 15: Qwen2.5-7B DAPO response accuracy at the last checkpoint. Upper/lower panels: In-Domain (ID)/Out-of-Domain (OOD). Values: mean verified-response share in percentage points at $k = 16$ for AIME24/25, $k = 8$ for Minerva, and $k = 4$ otherwise. Avg.: unweighted four-task mean. Olympiad/TruthfulQA abbreviate OlympiadBench/TruthfulQA-MC. Initialization: fixed base model; bold: better trained method per column.

ID					
Method	AIME24	AIME25	Minerva	Olympiad	Avg.
Initialization	12.10	7.70	37.10	36.10	23.25
DAPO	10.42	7.92	**41.31**	35.78	23.85
ExPPO	**14.38**	**10.21**	40.85	**38.67**	**26.05**

OOD					
Method	GPQA	MMLU-Pro	Safety	TruthfulQA	Avg.
Initialization	32.80	62.50	73.00	66.40	58.68
DAPO	33.04	64.64	0.85	53.91	38.10
ExPPO	**36.10**	**71.79**	**72.03**	**66.01**	**61.48**

ExPPO restores OOD generalization under DAPO. In Table 15, ExPPO raises both split macros over DAPO, exceeds the initialization on both, and leads DAPO on seven of eight tasks. The largest recovery occurs on Safety, where DAPO loses most of its initial accuracy and ExPPO preserves it.

J.7 Controlled Multi-Answer Task

Exact verification makes correct-support breadth measurable. The controlled set is a deterministically generated synthetic evaluation panel with 64 fixed prompts, 4–46 exhaustively verified modes per prompt in the realized bank, and zero exact prompt overlap with DAPO-Math-17k. Each prompt specifies four positive denominations $d = (d_1, d_2, d_3, d_4)$ sampled from 2–13, a target T sampled from 24–70, and a maximum count of six. The model is instructed to return a JSON object of the form {`"counts":[c1,c2,c3,c4]`}. The unified final-answer parser selects the final-most literal JSON `"counts"` field in the response. A response is correct exactly when this field contains four integers satisfying

$$c_j \in \{0, 1, \ldots, 6\} \quad \text{for every } j, \qquad \sum_{j=1}^{4} c_j d_j = T. \tag{43}$$

Each distinct verified count vector is a semantic answer mode. For example, when $d = (4, 9, 10, 13)$ and $T = 41$, four valid modes are

$$(0, 2, 1, 1), \quad (1, 3, 1, 0), \quad (2, 0, 2, 1), \quad (3, 1, 2, 0). \tag{44}$$

This construction gives exact correctness and mode identity.

Every model–method pair receives the same prompt bank and $K = 128$ samples per prompt using temperature 1.0, top-$p = 1.0$, and top-$k = 0$. Each prompt contains the user task and JSON-output instruction and is rendered with the tokenizer's native template; Qwen3 thinking is disabled. Evaluation caps are 2,048 tokens for Qwen2.5-7B and Phi-3.5-mini and 8,192 tokens for Qwen3-0.6B. Qwen2.5-7B and Qwen3-0.6B are generated directly at these caps; the Phi-3.5-mini result is exact 2,048-token prefix re-verification of a common 16,384-token response bank.

For prompt q, let $n_{q,m}$ be the number of verified responses in correct mode m, $C_q = \sum_m n_{q,m}$, and $U_q = |\{m : n_{q,m} > 0\}|$. Fixed-budget observed correct-mode yield is $64^{-1} \sum_q U_q$. When $C_q > 0$, let $p_{q,m} = n_{q,m}/C_q$ and define

$$N_{\text{eff}}(q) = \exp\left(-\sum_m p_{q,m} \log p_{q,m} \right). \tag{45}$$

We set $N_{\text{eff}}(q) = 0$ for unsolved prompts, so the prompt mean is a fixed-budget effective correct-mode yield that combines discoverability and diversity. Table 16 supplements the main results with response counts and diversity-eligible prompt counts.

Table 16: Additional controlled-task diagnostics ($K = 128$; 64 prompts; last checkpoint). Correct: verified responses among 8,192 samples; $C \geq 2$: prompts with at least two correct responses, eligible for diversity evaluation. Qwen2.5-7B ExPPO uses the earlier evaluation underlying the GRPO/GRPO-H pairs in Table 2 and the GRPO-UR comparison in Table 18. Other rows match the main evaluation.

Model	Method	Correct / 8,192	$C \geq 2$ / 64
Qwen2.5-7B	GRPO	528	51
	GRPO-H	793	55
	ExPPO	2,608	64
Qwen3-0.6B	GRPO	8,063	64
	GRPO-H	2	1
	ExPPO	8,007	64
Phi-3.5-mini	GRPO	74	13
	GRPO-H	41	7
	ExPPO	185	24

To isolate the distribution among correct responses, for prompts with $C_q \geq 2$ we compute correct-pair diversity

$$D_q = 1 - \frac{\sum_m n_{q,m}(n_{q,m} - 1)}{C_q(C_q - 1)}. \tag{46}$$

This is the probability that two verified-correct responses sampled without replacement from the observed bank have different semantic modes. For each ExPPO–baseline comparison, we average D_q equally over the common prompts where both methods have at least two correct responses. The intervals use 10,000 paired prompt-bootstrap resamples. The main table reports the Qwen2.5-7B GRPO, GRPO-H, F-GRPO, and REPO-R pairs and the Qwen3-0.6B GRPO pair; Table 17 gives the remaining core comparisons, and Table 18 adds GRPO-UR.

Table 17: Additional core-baseline correct-pair comparisons from the evaluations in Table 16; other pairs appear in Table 2. Shared: prompts with at least two correct responses per method. D: mean distinct-mode probability of correct pairs on that shared set; ΔD: ExPPO minus baseline; CI: paired prompt-bootstrap 95% interval. D uses a 0–1 scale, with differences and intervals in the same units, versus the main table's 0–100 scale. The one-prompt Qwen3-0.6B comparison is reported as a point estimate.

Model	Baseline	Shared / 64	Baseline D	ExPPO D	ΔD	CI
Qwen3-0.6B	GRPO-H	1	0.0000	0.7531	$+0.7531$	–
Phi-3.5-mini	GRPO	6	0.0370	0.1620	$+0.1249$	$[-0.0741, +0.3768]$
	GRPO-H	4	0.5701	0.1310	-0.4391	$[-0.8125, -0.1131]$

Table 18: Qwen2.5-7B controlled-task comparison with GRPO-UR (Rewarding the Unlikely) at the last checkpoint ($K = 128$; 64 prompts). A: Accuracy: verified-response share (0–100); Solved: prompts with a success. Observed/effective yields: prompt means of distinct correct modes/exponentiated mode entropy, with zero for unsolved prompts. Correct/prompt: total verified successes divided by 64, including unsolved prompts. Bold: higher mean. B: Shared: prompts with at least two successes per method; D: mean distinct-mode probability of correct pairs on that set; ΔD: ExPPO minus baseline; in probability units. ExPPO uses the earlier evaluation; the updated F-GRPO/REPO-R comparisons appear in Table 2.

A. Fixed-budget correctness and correct-mode yield

Method	Accuracy	Solved / 64	Observed yield	Effective yield	Correct / prompt
GRPO-UR	17.77	64	3.0000	2.3021	22.75
ExPPO	**31.84**	64	**3.9688**	**2.7277**	**40.75**

B. Conditional correct-pair diversity

Baseline	Shared / 64	Baseline D	ExPPO D	ΔD
GRPO-UR	63	0.4545	0.5113	$+0.0569$

Higher mode yield also holds against GRPO-UR. ExPPO raises accuracy and both correct-mode yields over GRPO-UR.

J.8 Common-Checkpoint Continuation and Normalized Credit Allocation

Continuation from a shared trained policy. We evaluate one-update continuations of Qwen2.5-7B on the controlled coin-vector task. GRPO, ExPPO, and GRPO-H start from the identical ExPPO step-900 checkpoint. The study uses 32 update prompts and 32 held-out prompts. Held-out evaluation uses $K = 128$ responses per prompt and a direct 2,048-token generation cap. Every evaluated policy reaches at least eight correct responses on each of the 32 held-out prompts. Table 19 reports held-out accuracy and conditional correct-pair diversity D, the probability that two correct responses represent different solutions. Paired 95% confidence intervals use 10,000 prompt-bootstrap resamples and are unadjusted for multiple comparisons.

Table 19: Qwen2.5-7B common-checkpoint continuation on 32 held-out prompts. All three update arms start from the same ExPPO-900 policy. Panel A reports accuracy in percent and diversity as a probability. Panel B reports ExPPO-minus-baseline differences in percentage points for both metrics.

A. Held-out policy outcomes

Policy	Accuracy (%)	Diversity D
Shared ExPPO-900 checkpoint	33.74	0.6404
+ GRPO update	35.64	0.6135
+ ExPPO update	36.52	0.6208
+ GRPO-H update	34.45	0.6001

B. Paired differences (percentage points)

Comparison	Accuracy	Diversity
ExPPO − GRPO	+0.88	+0.73
ExPPO − GRPO-H	+2.08	+2.07

All three updates increase observed held-out accuracy and reduce conditional diversity relative to the shared checkpoint. ExPPO has the highest post-update point estimates for both metrics and the smallest observed diversity decrease. The paired diversity intervals span zero. These outcomes describe local continuation behavior; Section 6.4 evaluates the final policies produced by the complete training schedules.

Credit allocation after normalization. We audit the final sequence advantages $\widetilde{A}_i$ from Eq. (9). For a response category $\mathcal{C}$, the recorded change is

$$\Delta_{\mathcal{C}} = \sum_{i \in \mathcal{C}} (\widetilde{A}_i - A_i).$$

The sum runs over the audited update responses and measures scalar credit before token aggregation. High- and low-surprisal categories use $u_i > 0$ and $u_i < 0$, respectively, with u_i defined by Eq. (6). Table 20 shows that all four aggregate directions persist after the shared normalization.

Table 20: Post-normalization coefficient audit on the Qwen2.5-7B update responses. Values are category sums of $\widetilde{A}_i - A_i$ in sequence-advantage units. Surprisal classes are relative to each prompt group. Allocation effects describe the aggregate scalar coefficients used by the policy loss.

Response	Surprisal class	$\Delta_{\mathcal{C}}$	Allocation effect
Correct	High ($u_i > 0$)	+6.245	Stronger reinforcement
Correct	Low ($u_i < 0$)	−3.113	Reduced reinforcement
Incorrect	High ($u_i > 0$)	+8.690	Softer penalty
Incorrect	Low ($u_i < 0$)	−5.557	Stronger penalty

The displayed category totals give signed changes of $+3.132$ for correct responses and $+3.133$ for incorrect responses. With verifier polarity preserved, the corresponding absolute-advantage-mass change over these four categories is approximately $3.132 - 3.133 = -0.001$, consistent with shared normalization at the reported precision. Their signed sum changes by $+6.265$. This audit characterizes the executed coefficient allocation;

J.9 Matched-Input OOD Representations

We compare Base with step-900 GRPO, GRPO-H, and ExPPO on GPQA (448 questions), MMLU-Pro (70), Safety (1,000), and TruthfulQA-MC (684). Each checkpoint receives identical questions and teacher-forced reference answers containing only the boxed correct option label. We average final-layer hidden states over answer tokens (layer 28 for Qwen; layer 32 for Phi). Within each backbone–benchmark panel, PCA is jointly fitted to the ℓ_2-normalized representations of the same 64 questions from all four checkpoints. Figures 8 to 10 show the first two components, with explained variance on the axes. Legend values report mean cosine distances to Base over the complete benchmark, computed before dimensionality reduction. These measure representation proximity, not accuracy, diversity, or policy KL; the projection can omit differences.

ExPPO has the smallest mean cosine distance among trained methods in 10 of 12 backbone–benchmark comparisons, excepting both Qwen Safety panels. This result is specific to the boxed-label probe: full-answer-text controls change the ExPPO–GRPO ordering for Qwen3. For Phi, Base is the released reference model; its identity with the original training initialization has not been established.

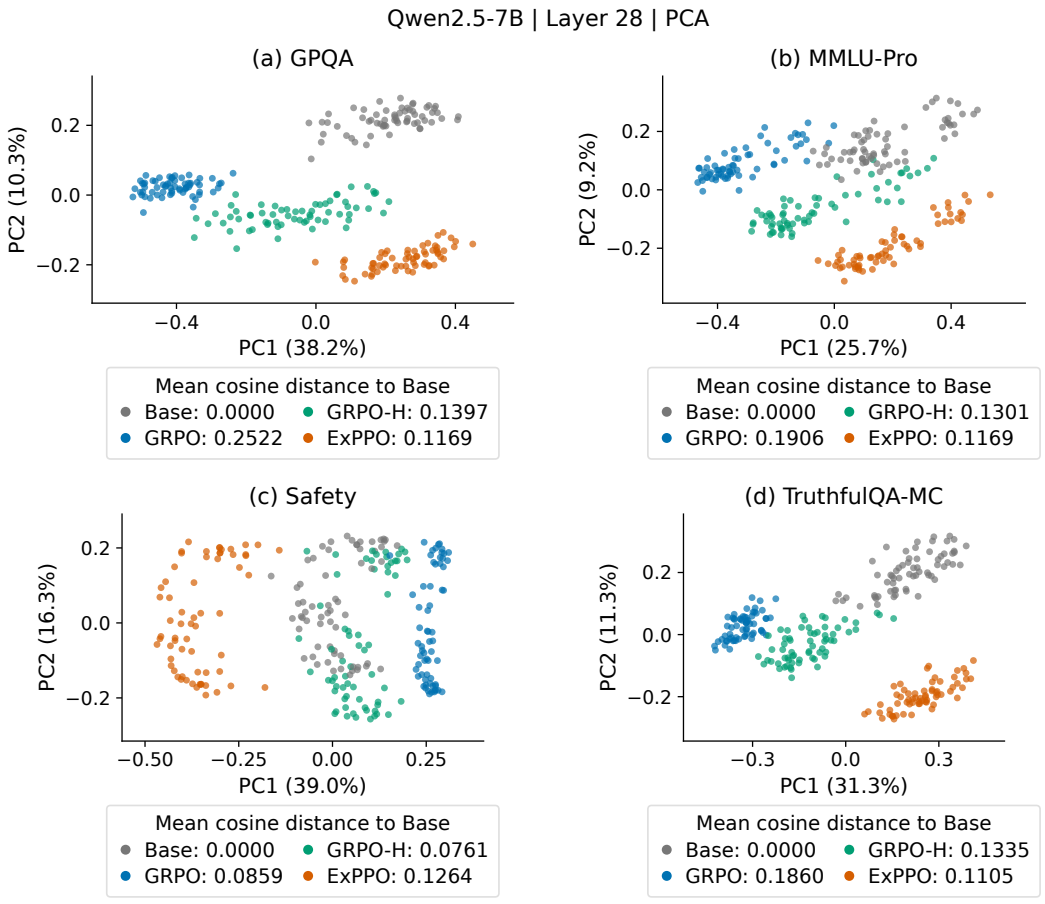

Figure 8: Qwen2.5-7B matched-input OOD representations. Each panel shows 64 shared questions for Base and three step-900 policies in a separately fitted PCA coordinate system. Axes report explained variance; legends report full-benchmark, original-space cosine distances to Base, not distances in the projection.

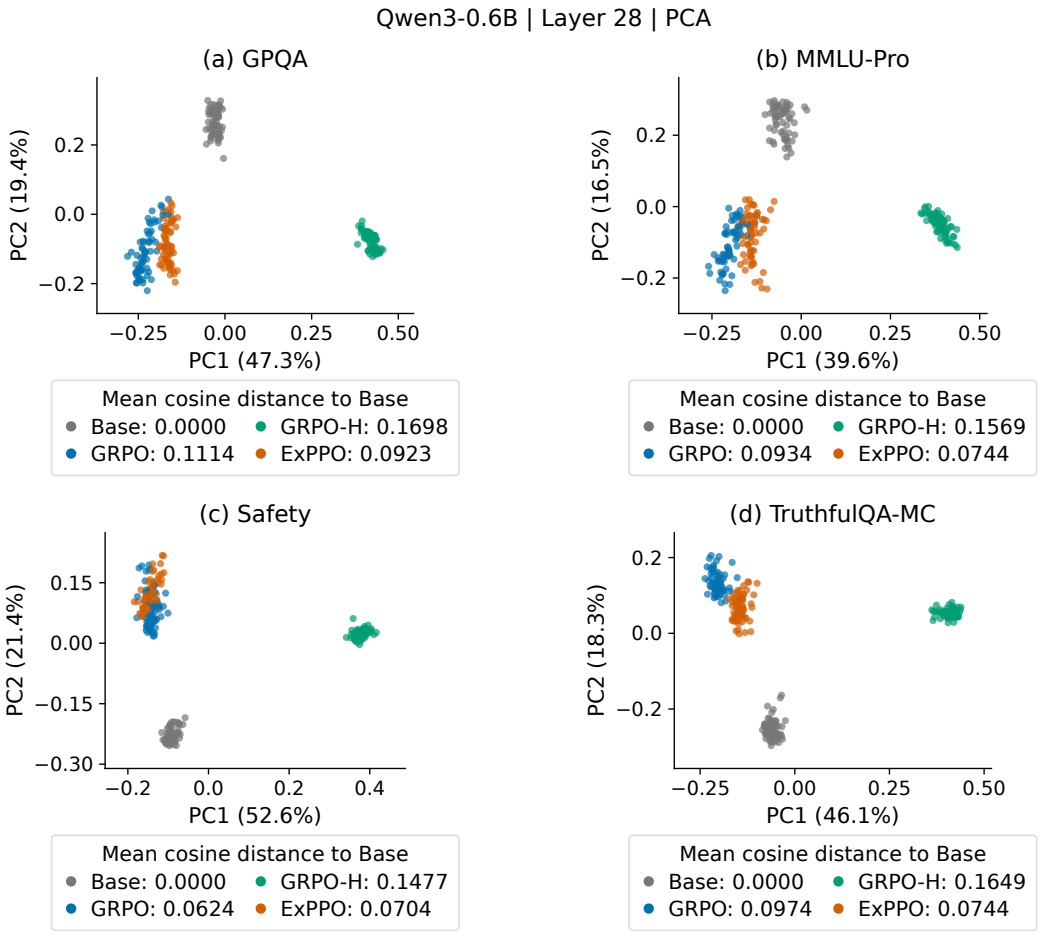

Figure 9: Qwen3-0.6B matched-input OOD representations, using the protocol in Appendix J.9. Each panel has its own PCA coordinates and 64 shared questions across four checkpoints. Legend values measure full-benchmark cosine distance to Base before projection; the relative ordering is sensitive to answer format.

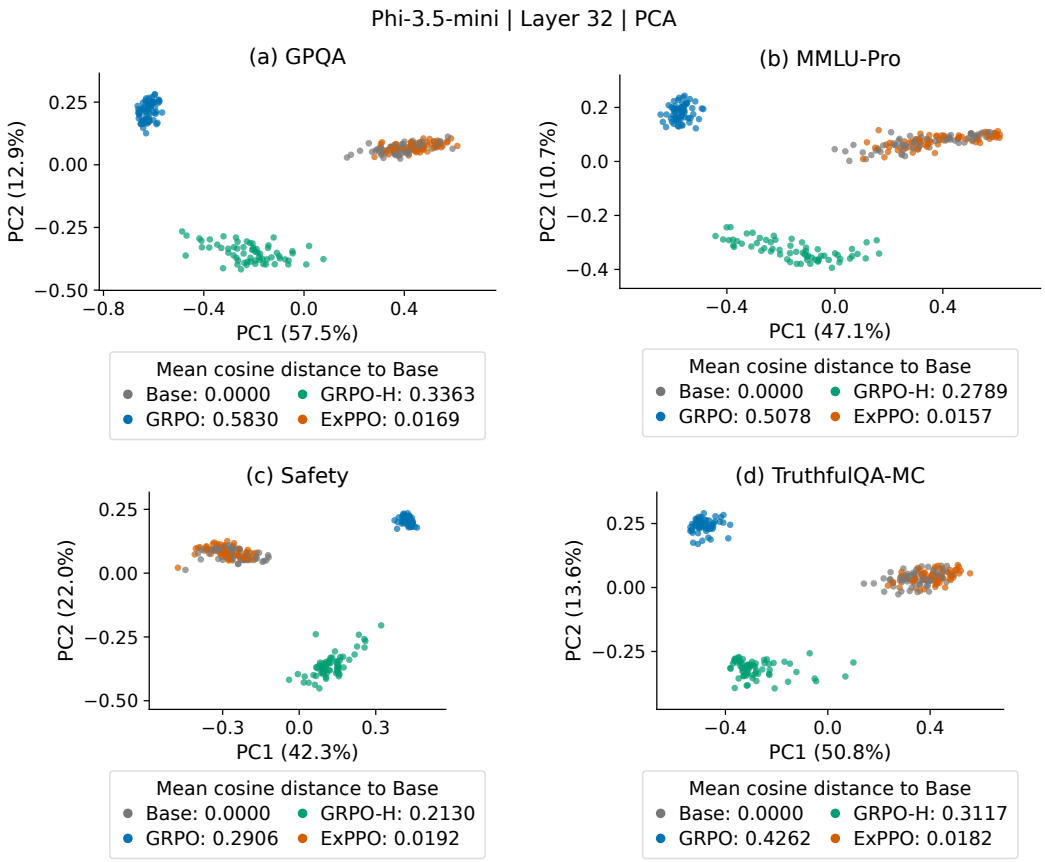

Figure 10: Phi-3.5-mini matched-input OOD representations, using the protocol in Appendix J.9. Each panel has its own PCA coordinates and 64 shared questions across four checkpoints. Legends report full-benchmark cosine distances to the released Base reference, whose identity with the training initialization is unverified. Proximity does not establish preserved accuracy or diversity.