POLYMORPHIC REINFORCEMENT LEARNING

Przemek Chojecki
ulam.ai

Abstract. Polymorphic reinforcement learning trains across executable families with changing environments, rules, and interfaces, rather than only changing the wording of a fixed task. We distinguish decision-preserving transformations from mechanism changes that require identifying the active rules and conditioning the policy on them. Transport and information bounds delimit valid reuse, while exact KL-regularized objectives quantify the cost of imposing a shared policy. For independent episodes from different variants, we derive the exact finite-group reward gradient and prove that count-dependent normalization preserves mean-success alignment for every shared-parameter policy family if and only if its nontrivial normalizers are equal. Exact stochastic calculations show that more mixed-reward groups and lower gradient variance can coexist with reduced success; existing implementations and published LLM comparisons establish the empirical scope of the analysis.

1. Introduction

The central problem in polymorphic reinforcement learning (PRL) is *learning to act when environments or their operative rules change*. In a family $M_{b,v}$, a shared base specification b supports variants v with different transition rules, observations, rewards, permissions, or action interfaces. Changing a rule can change the optimal action; invariance to that change can therefore be wrong. Representation changes are a controlled special case, useful for separating recognition of an unchanged problem from adaptation to a changed one. A collection of unrelated benchmark tasks is not, by itself, such a family.

For language-model post-training, two questions must be separated: what environmental structure is reusable, and what objective the parameter update actually improves. Contextual and meta-RL formalize adaptation across mechanisms; abstractions and symmetries formalize decision-preserving changes [2, 4, 7, 13–16]. Neither environmental equivalence nor additional rule variation guarantees a beneficial shared-parameter update. Our main results identify the exact heterogeneous-group potential (Theorem 4.3), characterize universal mean-success alignment (Theorem 4.4), and separate mixed rewards, variance, and improvement at equal rollout budgets (Proposition 4.5). These results apply to independent interactive episodes with fixed binary verifiers, not only to rephrased questions.

TA-GRPO and PrAg-PO are related prompt-augmentation methods, whereas EnvHarness transforms interactive environments [42, 58, 59]. The homogeneous finite-group identity is due to Bay and Yearick; its implicit-objective interpretation and other normalization analyses are established [55, 60–62, 64]. Stratified GRPO groups trajectories by their realized structure, unlike our fixed pre-sampling quotas [66]. Gibbs optimization, logarithmic pooling, and the transport and statistical bounds below are supporting established tools, not claimed as new general-purpose results. All new computations use finite policies; published LLM scores are reanalyzed rather than independently reproduced training runs.

2. Changing environments and rules

Let $M = (S, A, O, \mu, K)$ be a finite partially observable Markov decision process (POMDP), with initial law $\mu(s, o)$ and joint next-state, observation, and reward kernel $K(s', o', r \mid s, a)$. Rewards lie in a finite subset of $[0, 1]$. A policy acts from $h_t = (o_0, a_0, r_0, \ldots, a_{t-1}, r_{t-1}, o_t)$; private training rewards are omitted from its input. Invalid actions have explicit outcomes, and termination uses an absorbing state. Write $J_M^U(\pi) = \mathbb{E}_M^\pi U(\tau_H)$ for a specified utility $U \in [0, 1]$

Date: September 7, 2026.

on H transitions, and $J_M(\pi) = \mathbb{E}^\pi_M \sum_{t\geq 0} \gamma^t r_t$ for discounted return, $0 < \gamma < 1$. Finite alphabets keep the proofs elementary; continuous implementations require measurable counterparts.

Definition 2.1 (Operational polymorphic family)**.** A polymorphic family $\{M_{b,v}\}$ consists of executable decision processes generated from shared base specifications b and variant descriptors v, with documented operations that change decision rules, task structure, observation or action interfaces, or representations. Each operation specifies what it retains and what it changes; the relationship is implemented by a generator, interpreter, transformation layer, or alignment mechanism.

A learning problem additionally fixes a distribution $\mathcal{D}$ over (b, v), a policy class, and an interaction budget, with objective $\sup_\pi \mathbb{E}_{\mathcal{D}} J^{U_{b,v}}_{M_{b,v}}(\pi)$. A hidden v is accessible only through the prescribed observations. A multi-episode adaptation sequence, including retained memory, is one meta-episode. Independent resets of these sequences, rather than successive dependent trials, are the sampling units in our group analysis. Deployment under $\mathcal{D}_{\text{test}} \neq \mathcal{D}$ is a separate objective.

Mechanism variation changes action consequences, observations, rewards, or permissions; *task-structural variation* composes goals and dependencies; *interface variation* changes their low-level execution. Representation variation changes encodings without necessarily changing the task. A random seed or a list of unrelated tasks does not document any of these relationships. Conversely, rules may change between episodes or during interaction: including the current rule configuration in state yields a stationary kernel even when the agent changes operative rules. The distinction is relative to the declared factorization, since a fixed episode descriptor can also be included in state as (s, v).

Rule-changing variants generally require a context-conditioned model, not one invariant controller. For a finite known family, the posterior $b_t(v, s) = \mathbb{P}(v, s_t = s \mid h_t)$ is sufficient; for observed rewards its update is $b_{t+1}(v, s') \propto \sum_s b_t(v, s) K_v(s', o_{t+1}, r_t \mid s, a_t)$. Private rewards are marginalized out. This representation retains uncertainty about the mechanism but presupposes its prior and kernels. If v is unobservable, an omniscient solution is not an executable policy.

The cost of ignoring changed rewards is already exact in one decision. For common actions and weights $w_v > 0$, $\sum_v w_v = 1$, forcing one distribution incurs loss $\Delta = \sum_v w_v \max_a r_v(a) - \max_a \sum_v w_v r_v(a) \geq 0$. Linearity shows that equality holds precisely when the variants share a maximizing action. Opposite binary rewards with equal weights give conditioned value 1 but tied value $1/2$. With weights $1 - \lambda, \lambda$, uniform reference and KL coefficient β, the tied first-action probability is $\sigma((1 - 2\lambda)/\beta)$, where $\sigma(z) = (1 + e^{-z})^{-1}$. This is a counterexample to incorrect invariance, not an additional environment proposal.

Unknown rules also impose an information limit. Write $\mathrm{TV}(p, q) = \frac{1}{2} \sum_y |p(y) - q(y)|$. Let $v \in \{0, 1\}$ have equal prior probabilities. A diagnostic policy η collects an observable transcript with law P^η_v, including its actions and observed rewards, before selecting a label; $\mathcal{E}_B$ permits at most B interactions.

Proposition 2.2 (Information-limited adaptation)**.** *The smallest terminal identification error is*

$$\inf_{\eta \in \mathcal{E}_B, \widehat{v}} \mathbb{P}(\widehat{v} \neq v) = \frac{1 - \sup_{\eta \in \mathcal{E}_B} \mathrm{TV}(P^\eta_0, P^\eta_1)}{2}. \tag{2.1}$$

Applying any variant-independent randomized transformation to a transcript cannot increase the total variation of its two laws. In particular, observationally indistinguishable variants remain indistinguishable after any amount of such augmentation.

Proof. Fix η and abbreviate $p_v(h) = P^\eta_v(h)$. Choosing the more likely label gives error $\frac{1}{2} \sum_h \min\{p_0(h), p_1(h)\} = \frac{1}{2}(1 - \mathrm{TV}(p_0, p_1))$. Randomization cannot improve this pointwise optimum; optimizing η proves (2.1). For a variant-independent Markov kernel $Q(z \mid h)$, set $d(h) = p_0(h) - p_1(h)$. The triangle inequality and the unit row sums of Q give $\sum_z |\sum_h d(h) Q(z \mid h)| \leq \sum_h |d(h)|$, proving contraction of total variation. □

Identification is required only when the unresolved mechanisms demand different decisions. Rephrasing existing evidence cannot distinguish statistically identical mechanisms, although

it can alter a bounded model's computation. Small *known* rule changes permit approximate transfer: Theorem A.4 gives utility error at most $\epsilon_0 + H\epsilon$ for initial-law error ϵ_0 and uniform joint-kernel error ϵ in total variation. Large or unobservable changes have no such guarantee. Exact representation and interface relations are treated separately in Appendix A.

Table 1 lists implemented relationships, not proposed extensions. Its first eight entries change rules, task structure, or interaction contracts; the final two supply aligned-interface and representation controls. The classification uses Definition 2.1, not a claim that the original authors used the term PRL or proved exact equivalence.

Table 1: Existing polymorphic families: environment and rule changes, followed by controlled changes of realization.

System	Implemented change and evidence boundary
Alchemy [26, 27]	Causal structures and perceptual mappings are resampled between episodes; trials within an episode permit inference of the current chemistry. A meta-RL environment, not an invariance benchmark.
XLand [28, 29]	Compositional worlds, game objectives, and interaction rules; subsequent work studies within-task adaptation. Reported agents are non-language RL agents; AdA is not another environment.
XLand-MiniGrid [30, 31]	Executable rules and goals change object interactions and dependencies, rather than only layouts. Shared primitives do not imply a common optimal policy.
Kinetix [32]	Assemblies of objects, joints, motors, and thrusters change the controllable dynamics within one physics engine. The simulator's equations remain shared.
Baba Is AI [33, 34]	Movable word tiles encode rules the agent can alter during play. Published LLM evaluation diagnoses rule-manipulation failures, not RL-training gains.
TextWorld [35, 36]	A logic engine composes worlds and quests; separate grammar control regenerates their text. Grammar changes require referent preservation for exact equivalence.
NASimEmu [38, 39]	Generated network configurations change reachability and action consequences; a shared simulator/emulator interface supports empirical transfer, not certified equality of dynamics.
EnvHarness [42]	Contract wrappers change action/observation interfaces and Chain operations compose tasks while retaining base evaluators. The wrapped family qualifies, not each unmodified benchmark.
ALFWorld [37]	Explicitly aligned symbolic and embodied realizations of corresponding goals. Perception, navigation, and duration prevent automatic exact equivalence.
Multilingual Reasoning Gym [40, 41]	Parallel procedural problems use translated templates and adapted verifiers. A one-response representation control, not an interactive rule-changing world.

Rule generation, adaptation, and curriculum selection are distinct. RL^2, MAML, PEARL, and VariBAD learn adaptation procedures [13–16]; POET, PAIRED, prioritized level replay, and ACCEL generate or select challenges [17–20]. PolySkill is agent-side skill abstraction, not an environment [21]. Reasoning Gym supplies generator–verifier pairs, but generation alone does not establish rule variation [43]. Multilingual tasks and prompt augmentation remain controls for representational effects, not evidence of adaptation to changed transition rules.

3. Post-training objectives

Parameter post-training is distinct from prompting or fixed-weight adaptation [44, 46]. Sampling variants, tying their decision policies, pooling reward groups, and reusing trajectories are different interventions. We analyze KL-regularized reward maximization and specified score-function updates [45, 52–54].

First consider a one-response task. For input $x = x_{b,v}$, let $\mathcal{Y}_x$ be the finite support of complete responses under fixed masking, sampling, and stopping conventions. An autoregressive policy has $\pi_\theta(y \mid x) = \prod_t \pi_\theta(y_t \mid x, y_{<t})$. Let $q_x(y) > 0$ be a fixed reference, $R_x(y) \in [0, 1]$ a

parameter-independent reward, $\beta > 0$, and ρ a fixed input law. Define

$$F(\pi) = \mathbb{E}_{x\sim\rho}\big[\mathbb{E}_{\pi_x} R_x - \beta\,\mathrm{KL}(\pi_x\|q_x)\big], \qquad \mathrm{KL}(p\|q) = \sum_y p(y)\log\frac{p(y)}{q(y)}. \tag{3.1}$$

Logarithms are natural. Neural parameterization restricts this population objective; clipping, token averaging, and finite-group normalization need not produce its gradient. Zero reference mass excludes a response from every finite-KL policy. The following semantic reduction applies to *decision-preserving* variants, not arbitrary changed rules.

Theorem 3.1 (Semantic reduction and optimization gap)**.** *Let $s_x : \mathcal{Y}_x \to \mathcal{A}_b$ be a known surjection and suppose $R_x(y) = r_b(s_x(y))$. Write $p_x(a) = \sum_{s_x(y)=a} \pi_x(y)$ and $Q_x(a) = \sum_{s_x(y)=a} q_x(y)$. Then*

$$\mathrm{KL}(\pi_x\|q_x) = \mathrm{KL}(p_x\|Q_x) + \sum_a p_x(a)\,\mathrm{KL}(\pi_x(\cdot\mid a)\|q_x(\cdot\mid a)). \tag{3.2}$$

The unique pointwise optimizer has $p^*_x(a) = Q_x(a)e^{r_b(a)/\beta}/Z_x$, where $Z_x = \sum_a Q_x(a)e^{r_b(a)/\beta}$, and preserves $q_x(\cdot\mid a)$ within each semantic class. Its value is $F^* = \beta\mathbb{E}_x \log Z_x$, and every policy satisfies

$$F^* - F(\pi) = \beta\mathbb{E}_x\,\mathrm{KL}(\pi_x\|\pi^*_x). \tag{3.3}$$

If this gap is at most δ, then $\mathbb{E}_x\mathbb{E}_{\pi_x} R_x \geq \mathbb{E}_x\mathbb{E}_{\pi^*_x} R_x - \sqrt{\delta/(2\beta)}$.

Proof. Factor each response probability into its semantic marginal and its conditional probability within the corresponding class; splitting the logarithm gives (3.2). The conditional divergence is minimized at the reference conditional. Substituting $\log\pi^*_x = \log q_x + R_x/\beta - \log Z_x$ gives (3.3); nonnegativity and the equality case of KL prove pointwise optimality and uniqueness. For completeness, $\mathrm{TV}(p,q) \leq \sqrt{\mathrm{KL}(p\|q)/2}$ follows by coarsening to $A = \{y : p(y) \geq q(y)\}$ using the log-sum inequality. The resulting Bernoulli KL, as a function of its first parameter, has second derivative $1/[u(1-u)] \geq 4$, value zero and derivative zero at the second parameter, and hence is at least $2(p(A) - q(A))^2$. The log-sum inequality itself follows from Jensen's inequality for $-\log$. Bounded rewards differ in expectation by at most TV; averaging and applying Jensen again proves the final bound. □

The optimizer is the standard Gibbs distribution [52]. For binary success and reference probability $q \in (0,1)$, $p^* = \sigma(\mathrm{logit}\, q + 1/\beta)$, where $\mathrm{logit}\, q = \log[q/(1-q)]$; equally rewarded strategies retain their relative reference probabilities. For a neural class Π_Θ, define approximation error $a = F^* - \sup_{\pi\in\Pi_\Theta} F(\pi)$. An α-optimal member has gap at most $a + \alpha$, so reference success improves whenever $\mathbb{E}_x[p^*_x - Q_x(\text{success})] > \sqrt{(a+\alpha)/(2\beta)}$. The model's size does not establish either error bound.

Fix one base problem, positive variant weights w_v summing to one, common semantic rewards, and aligned action alphabets. Reference marginals Q_v can still differ across variants. Using reference conditionals within semantic classes leaves an objective over marginals. Requiring $p_v = p$ is an extra restriction, not a consequence of sampling variants.

Theorem 3.2 (The exact cost of semantic tying)**.** *Put $G(a) = \exp(\sum_v w_v \log Q_v(a))$ and $Z_v = \sum_a Q_v(a)e^{r(a)/\beta}$. Independently optimized variant policies attain $F^*_{\mathrm{sep}} = \beta\sum_v w_v \log Z_v$. A shared semantic policy instead attains*

$$p^*_{\mathrm{tie}}(a) = \frac{G(a)e^{r(a)/\beta}}{\sum_c G(c)e^{r(c)/\beta}}, \qquad F^*_{\mathrm{tie}} = \beta\log\sum_a G(a)e^{r(a)/\beta} \leq F^*_{\mathrm{sep}}. \tag{3.4}$$

If all $Q_v = Q$, tying has zero optimal-value cost. Moreover, for any policies p_v, their mixture $\bar{p} = \sum_v w_v p_v$, used in every variant, preserves average reward, weakly improves the regularized average objective, and raises minimum-variant reward to the original average reward. The objective-improvement conclusion requires matching references; the reward conclusions require a common reward.

Proof. For a shared p, the objective is $\sum_a p(a)r(a) - \beta \sum_a p(a) \log[p(a)/G(a)]$. Normalizing its exponential gives (3.4). The inequality also follows directly because the tied class is a subset of the separate class. With a common reference, the separate Gibbs optima coincide. For arbitrary policies, reward is linear in p and KL is convex in its first argument, so $\mathrm{KL}(\bar{p}\|Q) \le \sum_v w_v \mathrm{KL}(p_v\|Q)$. Every variant now has reward $\sum_v w_v \langle p_v, r\rangle$, which is at least the previous minimum. □

The geometric aggregation is logarithmic opinion pooling [63]; its exact objective penalty is as follows.

Corollary 3.3 (Reference disagreement and the sign of sharing)**.** *Under Theorem 3.2, let p_v^* denote the separately optimal semantic policies. Then*

$$F_{\mathrm{sep}}^* - F_{\mathrm{tie}}^* = -\beta \log \sum_a \prod_v p_v^*(a)^{w_v}. \tag{3.5}$$

The loss is zero exactly when all p_v^* coincide, equivalently all Q_v coincide. If $|\log Q_v(a) - \log Q(a)| \le \varepsilon$ for a common Q, the loss is at most $2\beta\varepsilon$. For two equally weighted binary variants, put $z_i = \operatorname{logit} Q_i(1) + 1/\beta$, $m = (z_1 + z_2)/2$, $d = (z_1 - z_2)/2$, $t = \tanh(m/2)$ and $u = \tanh(d/2)$. The tied success probability minus the average separate success probability is exactly

$$\sigma(m) - \tfrac{1}{2}[\sigma(m+d) + \sigma(m-d)] = \frac{tu^2(1-t^2)}{2(1-t^2u^2)}. \tag{3.6}$$

For unequal references its sign is therefore the sign of m.

Proof. Substitute $p_v^*(a) = Q_v(a)e^{r(a)/\beta}/Z_v$ into (3.5). Weighted arithmetic–geometric mean gives a sum at most one, with equality precisely when the positive distributions coincide pointwise. The common reward and positive references make $Q_v \mapsto p_v^*$ invertible. A uniform ε perturbation of log probabilities changes each relevant log partition by at most ε, proving the stated $2\beta\varepsilon$ bound. Finally use $\sigma(z) = (1 + \tanh(z/2))/2$ and the addition identities for tanh; combining the two fractions gives (3.6). □

The mixture guarantee concerns one aligned family and assumes an executable shared policy. A fixed neural parameterization need not realize it. If implementation error is at most ε in TV in each variant, minimum reward is at least the former average minus ε. With identical references, variation cannot raise the unrestricted population optimum; any benefit must arise from estimation, optimization, restricted model capacity, or a different evaluation distribution.

For genuinely changing environments, optimize a causal policy conditioned on observed history. Exponentially tilting entire interactive trajectories would incorrectly allow the agent to change nature's transition law. With expected stage reward $\bar{r}(h,a) = \mathbb{E}[r_t \mid h, a]$, finite-horizon causal KL gives

$$V_t(h) = \beta \log \sum_a q(a \mid h) \exp\left(\frac{\bar{r}(h,a) + \mathbb{E}[V_{t+1}(h') \mid h, a]}{\beta}\right). \tag{3.7}$$

The boundary condition is conditional expected terminal utility; stage rewards are zero for terminal-only tasks. One-step Gibbs optimization and backward induction prove the recursion. The conditional expectation uses the active mechanism or its history-conditioned posterior. Transport requires matching environmental laws and reference action probabilities; different-duration interfaces additionally require Proposition A.3. Thus the common-reward sharing result is a special case, not a justification for ignoring rule changes.

4. Finite-group updates across variants

For a fixed environment variant, use the causal objective $F(\theta) = \mathbb{E}_{P_\theta} R - \beta \mathrm{KL}(P_\theta\|P_q)$. The trajectory density ratio between the two policies cancels the environment factors, giving $\mathrm{KL}(P_\theta\|P_q) = \mathbb{E}_{P_\theta} \sum_t \log[\pi_\theta(y_t \mid h_t)/q(y_t \mid h_t)]$, summed only over agent-generated tokens. Let this log ratio be $L_\theta(\tau)$ and the policy score be $S_\theta(\tau) = \sum_t \nabla_\theta \log \pi_\theta(y_t \mid h_t)$. On a fixed finite support with positive differentiable policy probabilities,

$$\nabla F(\theta) = \mathbb{E}[(R - \beta L_\theta - b(x))S_\theta], \qquad \mathbb{E}[S_\theta \mid x] = 0. \tag{4.1}$$

The baseline is held fixed in differentiation. Differentiating L_θ contributes $-\beta \mathbb{E} S_\theta = 0$; local KL surrogates need not recover this full gradient [56]. Rule-changing actions are allowed: the rule configuration is state, while the environment kernel itself is independent of the model parameters.

Proposition 4.1 (A test for improvement or negative transfer)**.** *Let J be a target score with L-Lipschitz gradient on all update segments. At fixed θ, take $\theta^+ = \theta + \eta\widehat{g}$, with $g = \mathbb{E}\widehat{g}$, covariance Σ, and $g_T = \nabla J(\theta)$. Then*

$$\left|\mathbb{E}[J(\theta^+) - J(\theta)] - \eta g_T^\top g\right| \le \frac{L\eta^2}{2}(\|g\|^2 + \operatorname{tr}\Sigma). \tag{4.2}$$

A positive lower endpoint certifies improvement; a negative upper endpoint certifies deterioration. For a fixed linear preconditioner P, replace g by Pg and Σ by $P\Sigma P^\top$.

Proof. The Lipschitz-gradient assumption bounds the absolute Taylor remainder by $L\|\eta\widehat{g}\|^2/2$. Take expectations and use $\mathbb{E}\|\widehat{g}\|^2 = \|g\|^2 + \operatorname{tr}\Sigma$. Apply the same argument to $P\widehat{g}$. □

For a target different from the training objective, $g_T^\top g$ can be negative even under equivalent rules. Adaptive preconditioning, momentum, and clipping must enter the actual update law, not be treated as a fixed matrix.

For group-relative policy optimization (GRPO), take $k \ge 2$ independent episodes of one variant, binary terminal rewards R_i, and policy scores S_i. A one-response task is a special case. Write $p = \mathbb{E}R_i \in (0,1)$ and $\bar{R} = k^{-1}\sum_i R_i$. The reward update is $\widehat{g}_k = k^{-1}\sum_i (R_i - \bar{R})S_i/\sqrt{\bar{R}(1-\bar{R})}$, defined to be zero when all rewards coincide. This uses population standard deviation without an epsilon, fixed advantages during differentiation, and sequence-score sums; KL, clipping, and length weighting are excluded. The following homogeneous identity is established [55, 56, 60, 61].

Proposition 4.2 (Finite-group GRPO under binary rewards)**.** *For $M \sim \mathrm{Binomial}(k,p)$,*

$$\mathbb{E}\widehat{g}_k = c_k(p)\nabla p, \qquad c_k(p) = \frac{\mathbb{E}\sqrt{M(k-M)}}{kp(1-p)}, \qquad \mathbb{P}(0 < M < k) = 1 - p^k - (1-p)^k. \tag{4.3}$$

Without standard-deviation normalization, group-mean centering gives $(1-1/k)\nabla p$; a leave-one-out baseline restores ∇p. At $k = 2, 3$, $c_2(p) = 1$ and $c_3(p) = \sqrt{2}$; as $k \to \infty$, $c_k(p) \to 1/\sqrt{p(1-p)}$. A group made exclusively of reward-preserving copies of one sampled trajectory has zero reward-gradient contribution, irrespective of its number of surface forms.

Proof. The score identities give $\mathbb{E}[S \mid R = 1] = \nabla p/p$ and $\mathbb{E}[S \mid R = 0] = -\nabla p/(1-p)$. Conditional on $M = m \in \{1, \dots, k-1\}$, the advantages are $\sqrt{(k-m)/m}$ for successes and $-\sqrt{m/(k-m)}$ for failures. Their expected score sum divided by k is $\sqrt{m(k-m)}\nabla p/[kp(1-p)]$, proving (4.3). Independence makes $\mathbb{E}[\bar{R}S_i] = \nabla p/k$, proving the centering result and its leave-one-out correction. The $k = 2, 3$ identities follow by substituting the possible nontrivial counts. Bounded convergence applied to $\sqrt{(M/k)(1-M/k)}$ proves the limit. Identical rewards give zero centered advantages, proving the last assertion. □

Across variants, the mean update is $\mathbb{E}_v[c_k(p_v)\nabla p_v]$, not generally $\nabla \mathbb{E}_v p_v$. Its large-group limit optimizes $\mathbb{E}_v[2\arcsin\sqrt{p_v}]$ [56, 61]. For a shared normalization group, however, episodes from different rules or interfaces need not have identical reward probabilities. The next theorem permits these differences, using each episode's own fixed verifier and the score of its actual generating policy.

Theorem 4.3 (Heterogeneous groups and their exact potential)**.** *Fix k environment variants, allowing repetitions, and independently sample trajectories $\tau_i \sim P_{\theta,i}$. Each variant has parameter-independent dynamics, a fixed binary verifier $R_i(\tau_i)$, success probability $p_i \in (0,1)$, and score $S_i = \nabla_\theta \log P_{\theta,i}(\tau_i)$, equal to the sum of its agent-action log-probability gradients. Assume differentiable positive probabilities on fixed finite trajectory supports. Set $M = \sum_i R_i$. Fix positive constants $d_1, \dots, d_{k-1}$ and define $A_i = (R_i - M/k)/d_M$ when $0 < M < k$ and $A_i = 0$*

otherwise; set $\widehat{g} = k^{-1}\sum_i A_i S_i$*, with advantages held fixed in differentiation. For* $0 \le m \le k-1$*, define*

$$a_m = \frac{k-1-m}{kd_{m+1}} + \frac{m}{kd_m}, \qquad h_i(p_{-i}) = \frac{1}{k}\mathbb{E}[a_{M_{-i}}], \qquad M_{-i} = \sum_{j\neq i} R_j, \tag{4.4}$$

where terms with zero numerator at an undefined endpoint are zero. Then

$$\mathbb{E}\widehat{g} = \sum_i h_i(p_{-i})\nabla p_i = \nabla_\theta \Phi(p_1,\dots,p_k), \qquad \Phi(p) = \frac{1}{k}\mathbb{E}[H(M)], \tag{4.5}$$

where $H(0) = 0$ *and* $H(m+1) - H(m) = a_m$*. The expectation defining* Φ *uses independent Bernoulli variables of parameters* p_i*; it is an explicitly computable multiaffine polynomial. In particular, population-standard-deviation normalization gives* $a_m = \sqrt{(k-1-m)/(m+1)} + \sqrt{m/(k-m)}$*.*

Proof. The score identities are $\mathbb{E}S_i = 0$ and $\mathbb{E}[R_i S_i] = \nabla p_i$. Conditional on the other rewards, $M_{-i} = m$, independence gives $\mathbb{E}[A_i S_i \mid R_{-i}] = (A_i(1) - A_i(0))\nabla p_i = a_m \nabla p_i$. Averaging proves the first equality in (4.5). Differentiating the finite Bernoulli product expectation with respect to p_i gives $\partial_i \Phi = k^{-1}\mathbb{E}[H(M_{-i}+1) - H(M_{-i})] = h_i$. The chain rule proves the second equality. Substitution of $d_j = \sqrt{j(k-j)}/k$ proves the final formula. □

Here M_{-i} is Poisson–binomial; its probabilities follow by convolving $(1-p_j) + p_j z$. Choosing d_j accommodates Bessel correction, a fixed epsilon, or no normalization. The result excludes dependent episodes, differentiated advantages, adaptive within-group selection, active clipping, length weighting, and scoring under a different variant. Equal rewards or rules do not imply these sampling conditions.

Theorem 4.4 (When normalization preserves mean-success alignment)**.** *Under Theorem 4.3, let* $J = k^{-1}\sum_i p_i$*. The first-order alignment inequality* $\nabla J^{\mathsf{T}} \mathbb{E}\widehat{g} \ge 0$ *holds for all variant/policy families allowed above and all interior probability vectors if and only if* $d_1 = \cdots = d_{k-1} = d$*. In that case* $\mathbb{E}\widehat{g} = (k-1)\nabla J/(kd)$*. For population standard deviation, this condition holds at* $k = 2, 3$*, with multipliers* $1, \sqrt{2}$*, respectively, and fails for every* $k \ge 4$*. With* $d = 1$*, replacing the group mean by the leave-one-out mean gives exactly* $\mathbb{E}\widehat{g} = \nabla J$ *even for heterogeneous variants.*

Proof. Distinct observed variant labels suffice for necessity. At a fixed interior p, arbitrary derivatives t_i are realized by the one-parameter policies $p_i(\theta) = \sigma(\operatorname{logit} p_i + \theta t_i/[p_i(1-p_i)])$ at $\theta = 0$. If h is not proportional to the all-ones vector $\mathbf{1}$, put $u = h - \bar{h}\mathbf{1}$ and take $t = \mathbf{1} - \lambda u$ with $\lambda > \sum_i h_i/\|u\|^2$. Then $\mathbf{1}^{\mathsf{T}} t = k > 0$ while $h^{\mathsf{T}} t < 0$. Universal alignment therefore requires all h_i to agree at every interior p. For distinct i, j, letting $N = \sum_{\ell \notin \{i,j\}} R_\ell$, independence gives $h_i - h_j = (p_j - p_i)\mathbb{E}[a_{N+1} - a_N]/k$. Hold $p_i \neq p_j$ and let the remaining probabilities approach zeros and ones with sum m. Continuity implies $a_{m+1} = a_m$ for every $0 \le m \le k-2$. Write their common value as C and $b_j = 1/(kd_j)$. The equations $(k-1-m)b_{m+1} + mb_m = C$ give $b_1 = C/(k-1)$ and, by induction, $b_j = C/(k-1)$ for every j. Hence the d_j are constant. Conversely constant d gives $a_m = (k-1)/(kd)$, proving the stated positive multiplier. Population standard deviations are constant over nontrivial counts precisely for $k = 2, 3$. For the last assertion, $R_i - \bar{R}_{-i} = k(R_i - \bar{R})/(k-1)$; alternatively independence makes every other response's reward a valid baseline for S_i. □

This characterizes universal *first-order* alignment, not convergence speed: variance and curvature still enter Proposition 4.1. Positive normalization results under restricted gradient interaction are compatible with it [62]. Removing normalization is an existing estimator choice, not a new algorithm.

Proposition 4.5 (Equal-rollout grouping comparisons)**.** *Fix* k *equally weighted variants of one base problem and spend* k *independent episode draws per group. Let a homogeneous group choose one variant uniformly and draw all episodes there; an iid-variant group choose each variant*

independently and uniformly; and a fixed-quota group draw once per variant. With $\bar{p} = k^{-1}\sum_i p_i$*, their probabilities of containing both reward values satisfy*

$$\underbrace{\frac{1}{k}\sum_i [1 - p_i^k - (1-p_i)^k]}_{u_{\mathrm{hom}}} \le \underbrace{1 - \bar{p}^k - (1-\bar{p})^k}_{u_{\mathrm{iid}}} \le \underbrace{1 - \prod_i p_i - \prod_i (1-p_i)}_{u_{\mathrm{quota}}}. \tag{4.6}$$

For population-standard-deviation normalization with the actual episode-policy scores, the iid-variant group's expected update is $c_k(\bar{p})\nabla\bar{p}$*. In contrast, the fixed-quota group need not be positively aligned with* $\nabla\bar{p}$ *when* $k \ge 4$*.*

Proof. Conditioning on the sampled variant gives u_{hom}; iid joint draws (V_i, τ_i) have iid Bernoulli rewards of mean $\bar{p}$; independence without identical probabilities gives u_{quota}. Convexity of z^k proves the first inequality and the arithmetic–geometric mean inequality proves the second. For iid joint draws the variant probability is fixed in differentiation, so $\nabla \log \mathbb{P}(V_i, \tau_i) = S_i$. Proposition 4.2 applies to this joint sample space. The negative assertion follows from Theorem 4.4 and is instantiated in Table 2. □

The iid result is conditional on one base; across bases, $c_k(\bar{p}_b)$ reweights their objectives. Mixed rewards also need not yield a nonzero gradient, since score-weighted advantages can cancel.

Original-prompt scoring is a different, response-level intervention. TA-GRPO pools rewards from rephrasings but scores responses using ratios conditioned on the original prompt [59]. Its sampling mismatch requires a separate bound.

Proposition 4.6 (Anchored scores under approximate sampling equivalence)**.** *Let* P_0 *be the current anchor-prompt response law,* $B_1, \dots, B_k$ *independent response laws on the same finite alphabet, with* $P_0(y) > 0$ *for every symbol, and* R *a common binary verifier. Write* $p_0 = \mathbb{E}_{P_0} R \in (0,1)$ *and* $T(y) = \nabla \log P_0(y)$*, and assume* $\sup_y \|T(y)\| \le S$*. Use population-standard-deviation advantages and anchored scores to form* $\widehat{g}_B = k^{-1}\sum_i A_i T(Y_i)$*, with* $Y_i \sim B_i$*. Put* $D = \min\{1, \sum_i \mathrm{TV}(B_i, P_0)\}$*. Then*

$$\|\mathbb{E}\widehat{g}_B - c_k(p_0)\nabla p_0\| \le 2SD. \tag{4.7}$$

Consequently $c_k(p_0)\|\nabla p_0\| > 2SD$ *suffices for strict positive alignment with anchor success. When the response alphabet has exactly two symbols, one successful and one unsuccessful, the stronger exact identity is* $\mathbb{E}\widehat{g}_B = \mathbb{E}[\sqrt{\bar{R}(1-\bar{R})}]\nabla \operatorname{logit} p_0$*, regardless of the success probabilities of the* B_i*.*

Proof. Consider the same group statistic as a function of k responses under $\prod_i B_i$ and under $P_0^{\otimes k}$. Its norm is at most S: in a mixed group, $k^{-1}\sum_i A_i^2 = 1$, so Cauchy–Schwarz gives $k^{-1}\sum_i |A_i| \le 1$. Product coupling bounds the total variation of the two group laws by D. Duality with unit vectors therefore bounds the difference of their expected statistics by $2SD$. Proposition 4.2 supplies the target expectation. Taking the inner product with ∇p_0 proves the alignment claim. For two response symbols, $T(Y_i) = (R_i - p_0)\nabla \operatorname{logit} p_0$; summing the centered products gives $\bar{R}(1-\bar{R})$, and division by the standard deviation proves the identity. □

The two-string identity does not cover an LLM's many distinct rewarded outputs. The TV distances and score bound are not measured in the cited runs, and the general bound can be vacuous. It concerns only the unclipped anchored reward term; semantic equivalence does not imply likelihood equivalence [65].

For an exact history/action bijection T_v, transporting a rollout of π yields a rollout of $T_v\pi$, not necessarily of the same model queried in variant v. With target support covered,

$$\mathbb{E}_{P_v^\pi} f = \mathbb{E}_{P_v^{T_v\pi}}[W_v f], \qquad W_v = \prod_t \frac{\pi(a_t \mid h_t)}{(T_v\pi)(a_t \mid h_t)}. \tag{4.8}$$

Sequential density ratios prove the identity by canceling matched environment factors. Nonlinear group statistics need the entire group's ratio. Unweighted reuse is generally biased, while correct

weights may have large second moments [49, 50]. Under *genuine rule changes*, environmental factors need not cancel, so this policy-only ratio does not justify reusing old-rule trajectories.

Token averaging also changes weights. With fixed variant lengths ℓ_v, averaging a sequence score over tokens divides its mean by ℓ_v; adding inverse-length sampling produces relative weights $1/\ell_v^2$. If length depends on the response, even a prompt-only baseline can fail because $\mathbb{E}[S/|y| \mid x]$ need not vanish [57]. These loss conventions differ from exact KL, which is invariant under a bijection of complete strings with transported distributions.

5. Computations and published evidence

A finite-policy witness isolates grouping at equal rollout counts. Let $p_1(\theta) = p_2(\theta) = p_3(\theta) = \sigma(\log(1/9) + \theta)$ and $p_4(\theta) = \sigma(-53\theta/50)$, with common binary success reward. At $\theta = 0$, $p = (0.1, 0.1, 0.1, 0.5)$, mean success is $J = 0.2$, and $J' = 1/800$. The single-token two-symbol policies have scores $S_i = a_i(R_i - p_i)$, $a = (1, 1, 1, -53/50)$. Each design uses four independent completions, population-standard-deviation normalization, and update $\theta^+ = 10^{-4}\widehat{g}$, without KL, clipping, or length weights. Since even identical rules admit failure, rule variation alone cannot guarantee ascent.

Table 2. Exact stochastic updates with four responses. The last column is $10^{10}\mathbb{E}[J(\theta^+) - J(0)]$, not percentage points.

Group design	u	$\mathbb{E}\widehat{g}$	$\mathbb{E}\widehat{g}^2$	$10^{10}\mathbb{E}\Delta J$
Homogeneous	0.4766	+0.007032329	0.103293750	+9.069304
Iid variants	0.5888	+0.002072243	0.049379766	+2.723630
Fixed quota	0.6350	−0.001228680	0.031611469	−1.450498

Complete enumeration at 80-decimal precision evaluates J *after every realized update*, not at the mean update. The global curvature bound $L = (3 + (53/50)^2)/(24\sqrt{3})$ gives certified intervals for $10^{10}\mathbb{E}\Delta J$ of $[8.27808, 9.30275]$, $[2.34538, 2.83523]$, and $[-1.69264, -1.37905]$. Exact rational arithmetic with a squared rational enclosure of $\sqrt{3}$ certifies their signs. Subtracting squared means gives the same strict variance ordering. The small step certifies signs; its effect size is not an estimate of practical LLM degradation.

The reproducibility code checks 560 heterogeneous-group configurations, sixty three-action score configurations, 1,000 pooling cases, and 108 negative-alignment witnesses for $k = 4, \ldots, 12$. Maximum residuals are 8.33×10^{-17} for the heterogeneous gradient, 5.56×10^{-16} for its potential, and 8.4×10^{-16} for pooling. Appendix B gives additional computed gains and losses. Numerical checks test implementation; proofs establish the general claims.

EnvHarness reports GRPO training of Qwen3-8B-base with transformed interactive environments. Table 3 recomputes its four nonredundant differences. The comparison fixes environment seed zero and reports no independent-training-run dispersion, so the OOD decrease has no inferable significance level [42].

Table 3. Published EnvHarness RL comparison. Success differences are percentage points; the score difference is in score units.

Metric	Original	Transformed	Difference
ALFWorld in-distribution success	81.4	87.9	+6.5
ALFWorld out-of-distribution success	89.6	88.8	−0.8
WebShop score	75.6	79.2	+3.6
WebShop success	66.0	67.4	+1.4

PrAg-PO changes prompts and format rewards, not interactive transition rules. Table 4 recomputes macro-selected checkpoints against DAPO; macro means weight benchmarks equally

and micro means weight questions equally. Reported deviations concern five inference rounds, not training seeds. Format rewards and checkpoint selection confound a pure augmentation effect. Despite macro gains, AMC decreases by 1.45 points for Qwen2.5-Math and Olympiad by 1.18 for Qwen3 [58].

TABLE 4. Published PrAg-PO comparisons against DAPO. Entries are original → augmented mean accuracies; changes are percentage points.

Model	Step	Macro means	Change	Micro means	Change
Qwen2.5-Math-1.5B	2720	44.12 → 45.18	+1.06	49.62 → 50.92	+1.30
DeepSeek-R1-Distill-Qwen-1.5B	1100	54.41 → 58.23	+3.82	57.56 → 60.94	+3.38
Qwen3-1.7B	1520	53.88 → 56.18	+2.30	60.73 → 60.69	−0.04

TA-GRPO v5 includes matched 32-response controls. Table 5 includes all twelve Figure 6 comparisons: macro improvements are 2.8383 and 3.1750 points. These are pass@32, not pass@1, and matched responses do not establish equal FLOPs or transformation cost. No independent-training-run intervals are supplied. Its Appendix C changes both normalization and anchoring, so it does not isolate either ingredient [59].

TABLE 5. Matched-response TA-GRPO comparisons (32 responses per group). DS: GRPO with dynamic sampling. Values are pass@32 percentages; differences are percentage points.

	Qwen3-1.7B			Qwen3-4B		
Benchmark	DS	TA	Δ	DS	TA	Δ
AMC	79.42	83.74	+4.32	84.97	88.08	+3.11
OlympiadBench	68.51	70.45	+1.94	72.88	74.53	+1.65
AIME24	43.88	47.42	+3.54	50.51	57.37	+6.86
AIME25	35.91	37.14	+1.23	43.82	45.92	+2.10
Minerva	52.37	53.07	+0.70	59.54	60.81	+1.27
GPQA-Diamond	71.22	76.52	+5.30	80.79	84.85	+4.06

EnvHarness supplies evidence about changed interaction contracts; the other comparisons concern response-level augmentation. None measures the general benefit of learning changing transition rules, estimates our sufficient conditions, or isolates all causes of its reported gains. The reanalysis supports neither a pooled effect size nor a significance claim.

6. Verification, transfer, and retention

Rule changes must update the success criterion and relevant constraints. For true utility U, verifier $\widehat{U} \in [0,1]$, and $\sup_{\pi\in\Pi} |\mathbb{E}_\pi U - \mathbb{E}_\pi \widehat{U}| \le \xi$, an α-optimal verifier policy has true regret at most $2\xi + \alpha$, by inserting the two verifier expectations. Uniform accuracy over optimized policies matters; ordinary held-out accuracy is insufficient [48]. A protected endpoint checker does not certify intermediate permissions or delayed outcomes. Reward shaping and runtime shielding have separate assumptions [23, 47].

If a binary verifier independently flips labels at rate e, its true-success margin is $1-2e$, giving optimized success $\sigma(\operatorname{logit} q + (1-2e)/\beta)$. For $q = 0.2$, $\beta = 0.2$, and $e = 0, 0.4, 0.5, 0.6$, this equals $0.973756, 0.404610, 0.2, 0.084224$. In the last case verifier score rises from 0.56 to 0.583155 while true success falls. False-positive and false-negative rates a, b replace the margin by $1-a-b$. These fixed-rate calculations do not model adversarially exploited errors.

Let ρ be the training-input law and ν a retained-task law; P_π^x, P_q^x are the two trajectory laws in the same environment, with actual stopping and interaction conventions.

Proposition 6.1 (A coverage-dependent retention bound)**.** *Suppose $\nu(x) \le C\rho(x)$ for all x and $\mathbb{E}_\rho \mathrm{KL}(P_\pi^x \| P_q^x) \le \varepsilon$. For any retained utility $U_x \in [0,1]$,*

$$\left|\mathbb{E}_{x\sim\nu}[\mathbb{E}_{P_\pi^x} U_x - \mathbb{E}_{P_q^x} U_x]\right| \le \min\{1, \sqrt{C\varepsilon/2}\}. \tag{6.1}$$

Without coverage, zero training KL can coexist with unit degradation on an omitted prompt.

Proof. Bound each expectation difference by trajectory TV, use the TV–KL inequality proved above, apply Jensen, and then use $\mathbb{E}_\nu \mathrm{KL} \le C\mathbb{E}_\rho \mathrm{KL}$. For the final assertion, let ρ assign zero probability to a prompt on which the reference always succeeds and the new policy always fails; let the two policies agree on the training support. Concentrating ν on the omitted prompt gives the claim. No finite KL on that omitted prompt is assumed. □

Even $C = 1$, $\varepsilon = 0.02$ gives a bound of 0.1, not unchanged behavior. Coverage of new variants does not protect unrelated skills or ensure zero safety violations.

Cybersecurity and business logs also lack unobserved interventions. If the logged action always returns 1/2, but an unseen alternative returns either 1 or 0 in two indistinguishable models, choosing it with probability p gives regrets $(1-p)/2$ and $p/2$. Worst-model regret is at least 1/4, regardless of relabeled sample count. NASimEmu's simulation/emulation relation does not identify arbitrary production dynamics; changed permissions, deadlines, and opponent responses are mechanisms, not nuisance encodings [38, 49, 50]. Game-rule changes have the same identification issue when observations cannot resolve their decision consequences (Proposition 2.2).

Evaluation must separate base specifications and their complete transformation descendants across partitions. Uncertainty must distinguish inference randomness, independent worlds, and independent training runs [51]. Comparisons must identify changes in models, optimizers, verifiers, memory, rollout count, and compute. The group theorem covers independent finite episodes with fixed binary rewards and stated score conventions, not adaptive within-group sampling, nonbinary normalization, or multiple clipped epochs.

For an LLM facing changing environments, the required response is conditional adaptation, not indiscriminate invariance. Exact relations justify specified reuse; unresolved rule changes limit identification; forced sharing and group normalization can lower the intended utility. The proved improvement conditions and counterexamples distinguish these mechanisms without treating task diversity, mixed rewards, or lower variance as substitutes for higher success.

Appendix A. Transport and statistical limits

MDP equivalence, homomorphisms, and bisimulation quantify decision-preserving structure [1–3]. Symmetry-based architectures and augmentation exploit known transformations; options handle temporally extended interfaces [4–6]. Contextual MDPs, domain randomization, and RL-generalization studies instead permit changed mechanisms [7–9]. Goal-conditioned values, successor features, and hindsight relabeling exploit particular shared dynamics, not equality of optimal policies across rewards [10–12]. The results here are consequences of these established tools, coupling, and statistical inequalities.

Constraints belong to the trajectory semantics. A finite monitor $q_{t+1} = u(q_t, s_t, a_t, s_{t+1})$ with an absorbing violation set can be included in hidden state [22, 23]. The single-agent results hold with fixed opponent policies, not automatically for independently learning opponents. Invertible state/action/observation maps form a groupoid across spaces, whereas lossy observations, changed rules, and task compositions need not be invertible.

Theorem A.1 (Exact transport)**.** *Suppose M and M' have bijections $\phi : S \to S'$, $\psi : A \to A'$, and $\chi : O \to O'$, identical reward alphabets and time units, and*

$$\mu'(\phi(s), \chi(o)) = \mu(s,o), \qquad K'(\phi(x), \chi(y), r \mid \phi(s), \psi(a)) = K(x, y, r \mid s, a). \tag{A.1}$$

Let T apply these maps to histories and leave rewards unchanged. Then $\pi'(\psi(a) \mid Th) = \pi(a \mid h)$ transports the full trajectory law. Consequently discounted values agree, and finite-horizon

utilities and constraint probabilities agree whenever their evaluation functions are transported. The correspondence preserves optimality over the transported policy class.

Proof. The initial laws agree under the maps. If the finite trajectory laws agree up to a decision, the policy identity transports the next action law and (A.1) transports the conditional state, observation, and reward law. Induction proves agreement of all finite-dimensional distributions. Expectations of transported bounded finite-trajectory functions therefore agree. The discounted tail is at most $\gamma^H/(1-\gamma)$, so discounted values agree by passage to the limit. An eventual monitor violation is an increasing union of finite-horizon events, and hence its probability also agrees. The inverse bijections transport every competing policy back, proving the optimality claim. □

A state-dependent action decoder must be computable from observed history; evaluator knowledge does not give the agent that decoder. Exact maps also need not preserve bounded-context or wall-clock budgets. For a finite fully observed MDP with a finite symmetry group, an equivariant optimal stationary policy exists: Bellman uniqueness makes Q^* equivariant, and uniform choice among maximizing actions respects the bijection of maximizing sets. This does not prove learnability in a fixed transformer or preservation of sequential value under arbitrary policy averaging [4].

For a fully observed MDP, let $f : S \to Z$ and $g_s : A(s) \to B(f(s))$ be surjective maps onto a finite abstract MDP $\bar{M}$, with nonempty action sets. Suppose expected rewards and transition kernels obey

$$r(s,a) = \bar{r}(f(s), g_s(a)), \qquad \sum_{x:f(x)=z'} P(x \mid s,a) = \bar{P}(z' \mid f(s), g_s(a)). \tag{A.2}$$

Proposition A.2 (Common quotient control)**.** *Under* (A.2)*, $V^*_M(s) = V^*_{\bar{M}}(f(s))$. An abstract stationary policy lifts with equal value through any stationary state-dependent distribution over concrete preimages of its chosen abstract action. Thus every member of a family admitting these homomorphisms onto the same $\bar{M}$ admits the same abstract optimal controller, together with its own action realization.*

Proof. For $Lv = v \circ f$, substituting (A.2) into the Bellman optimality operators gives $T_M Lv = LT_{\bar{M}}v$: next states aggregate by f, and surjectivity of g_s makes maximization over concrete actions identical to maximization over abstract actions. Both operators are γ-contractions, so their unique fixed points satisfy the claimed identity. The same calculation for the policy-evaluation operators proves the lifting statement, because all preimages have the same expected reward and aggregate transition law. □

This expected-value statement need not preserve reward variance or authorization events. In a POMDP, a latent quotient is usable only through an accessible sufficient statistic. For history compression $z_t = e(h_t)$, the next-reward and next-compressed-state law must depend on history only through z_t for every action; similar embeddings do not establish that condition. Mechanism-changing variants generally require $\bar{M}_v$, not one common quotient.

For temporal realization, let a finite boundary state z be sufficient for each option's induced law. An option is a fixed primitive policy with a termination rule [6], of duration $\tau \geq 1$ finite almost surely. With finite nonempty boundary action sets, define

$$\bar{r}(z,b) = \mathbb{E}\left[\sum_{j=0}^{\tau-1} \gamma^j r_j \mid z, b\right], \qquad K_\gamma(z' \mid z,b) = \mathbb{E}[\gamma^\tau \mathbf{1}\{z_\tau = z'\} \mid z,b]. \tag{A.3}$$

Proposition A.3 (Interface realization)**.** *Two implementations with identical boundary action sets, $\bar{r}$, K_γ, and initial boundary distributions have identical discounted values for all stationary boundary policies and identical optimal boundary-policy values. This need not equal optimal primitive-policy value, preserve internal constraint events, or imply equivalence under a fixed physical deadline.*

Proof. The boundary Bellman operator is $Tv(z) = \max_b\{\bar{r}(z,b) + \sum_{z'} K_\gamma(z' \mid z,b)v(z')\}$. Its contraction modulus is at most γ, since each kernel row sums to $\mathbb{E}[\gamma^\tau \mid z,b] \le \gamma$. Equality of the evaluation and optimality operators therefore implies equality of their fixed points. Primitive policies can make additional internal decisions; internal events are not determined by boundary reward expectations; and a single discounted kernel does not determine the full duration distribution. These facts explain the qualifications. □

Deadlines require remaining time and duration laws; internal permissions require monitors. Additional information or interruptibility changes the policy class. Information that can be discarded for free cannot reduce attainable utility, by simulation of the poorer experiment [24]; it need not be free for a bounded-context model.

For approximate transfer, use known relabelings to put two POMDPs on common alphabets and time units. Unlike a pure symmetry, this permits genuinely changed joint transition, observation, and reward laws.

Theorem A.4 (Approximate experiment transport)**.** *Suppose* $\mathrm{TV}(\mu, \widetilde{\mu}) \le \epsilon_0$ *and* $\mathrm{TV}(K(\cdot \mid s,a), \widetilde{K}(\cdot \mid s,a)) \le \epsilon$ *for all* s,a*, where the latter bound concerns the joint next-state, observation, and reward law. For any common history-dependent policy and any common utility* $U(\tau_H) \in [0,1]$*,*

$$|J_M^U(\pi) - J_{\widetilde{M}}^U(\pi)| \le d_H, \qquad d_H = \min\{1, \epsilon_0 + H\epsilon\}. \tag{A.4}$$

If $\widetilde{\pi}$ *is* α*-optimal in* $\widetilde{M}$ *over a common policy class, its regret in* M *is at most* $\min\{1, 2d_H + \alpha\}$*. The same first bound applies to finite-horizon violation probabilities when the monitor and its interpretation are included in the aligned trajectory semantics.*

Proof. Maximally couple the initial state–observation pairs. Until the first disagreement, histories coincide, so couple actions identically and maximally couple the next joint outputs. The probability of any disagreement is at most $\epsilon_0 + H\epsilon$ by the conditional union bound. Utilities coincide off that event and differ by at most one on it, proving (A.4). For any competing policy insert its value in $\widetilde{M}$ and the two values of $\widetilde{\pi}$; the two model differences cost at most d_H each and approximate optimality costs at most α. Taking the supremum over competitors gives the regret bound. A violation indicator is another bounded trajectory utility. □

For a sum of H unit-bounded rewards the direct bound is Hd_H, not d_H. For aligned fully observed discounted MDPs with reward error ϵ_r and transition-TV error ϵ_p, Bellman subtraction instead gives $\|V_M^\pi - V_{\widetilde{M}}^\pi\|_\infty \le \epsilon_r/(1-\gamma) + \gamma\epsilon_p/(1-\gamma)^2$. This uses $|(p-q)v| \le \mathrm{TV}(p,q)/(1-\gamma)$ for $0 \le v \le (1-\gamma)^{-1}$. Latent-transition agreement alone is insufficient in a POMDP: changed observations can conceal a decision-relevant variable. Neither bound estimates its error parameters from logs.

For generalization, draw n iid base specifications X_i, each with its complete transformation and rollout bundle. Let a fixed bundle loss be $L_i(\pi) \in [0,1]$, for a finite policy class Π of size N fixed in advance. Define $R(\pi) = \mathbb{E}L_i(\pi)$ and $\widehat{R}_n(\pi) = n^{-1}\sum_i L_i(\pi)$.

Proposition A.5 (Generalization across independent specifications)**.** *For* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\sup_{\pi \in \Pi} |R(\pi) - \widehat{R}_n(\pi)| \le \sqrt{\frac{\log(2N/\delta)}{2n}}. \tag{A.5}$$

The empirical risk minimizer's excess population risk is at most twice this bound. Generating m *correlated variants per specification does not, by itself, justify replacing* n *by* nm*.*

Proof. Hoeffding's inequality [25] bounds the deviation for each policy by $2\exp(-2n\varepsilon^2)$. A union bound over N policies proves (A.5); adding and subtracting the empirical risks of empirical and population minimizers proves the excess-risk statement. If all variants in a bundle are identical, there are exactly n independent losses regardless of m, disproving an automatic nm replacement. □

Conditionally iid variants with loss mean $a_\pi(X)$ and variance $b_\pi(X)$ give bundle-mean variance $\mathrm{Var}(a_\pi(X)) + \mathbb{E}[b_\pi(X)]/m$. More variants reduce within-base noise, not between-base uncertainty. A valid invariant class may reduce complexity [5], but a rule change can introduce approximation bias $\inf_{\Pi_{\mathrm{inv}}} R - \inf_{\Pi} R > 0$. Unknown decoders count toward complexity. This finite-class bound does not cover unrestricted transformers, adaptively selected worlds, or unsupported off-policy evaluation. Independent evaluation after an adaptive curriculum remains possible.

Appendix B. Fixed-budget calculations

At fixed parameters, let Z be an unbiased gradient contribution, B a base specification, and each conditional draw independently sample a variant and rollout. Define $A = \operatorname{tr}\mathrm{Cov}(\mathbb{E}[Z \mid B])$ and $D = \mathbb{E}\operatorname{tr}\mathrm{Cov}(Z \mid B)$. Average m contributions for each of n independent bases.

Proposition B.1 (Variance and the replication budget)**.** *The mean-gradient estimator has trace covariance $A/n + D/(nm)$. For base cost c_0, per-draw cost $c_1 > 0$, and continuous budget $C = n(c_0 + c_1 m)$, this becomes*

$$V_C(m) = \frac{(A + D/m)(c_0 + c_1 m)}{C}, \qquad m_* = \max\left\{1, \sqrt{\frac{c_0 D}{c_1 A}}\right\} \quad (A, D > 0). \tag{B.1}$$

The optimizer is also clipped by the feasible budget and must be checked at integer allocations. If $c_0 = 0$ and $A > 0$, replication $m > 1$ strictly increases variance at fixed C. At a fixed base, however, equal-cost stratification over a finite uniform variant set removes the between-variant variance term relative to independently sampling variant labels.

Proof. Conditional independence leaves covariance $\mathrm{Cov}(\mathbb{E}[Z \mid B]) + \mathbb{E}\,\mathrm{Cov}(Z \mid B)/m$ for one bundle. Independent bundles divide this by n. Expanding the numerator in (B.1) leaves the m-dependent terms $Ac_1 m + Dc_0/m$; differentiating proves the formula. For a fixed base and K variants, let conditional gradient means and covariance traces be μ_v, d_v. With N independent mixture draws the trace is $(\overline{d} + \operatorname{tr}\mathrm{Cov}_v\,\mu_v)/N$. With exactly N/K independent draws per variant, N divisible by K, it is $\overline{d}/N$. The means are equal in both designs. □

The sampling law is fixed for each update. A parameter-dependent generator would add distribution-derivative terms. The stratification statement concerns an unbiased estimator with unchanged mean, unlike the nonlinear group normalization in Theorem 4.3.

The code records inputs and outputs in JSON/CSV (seed 20260907; Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0). On 120 randomized cases, numerical maximization and closed-form objectives differed by at most 1.61×10^{-15}. The semantic-KL residual was 4.44×10^{-16}; binary-group enumeration and a four-action score check agreed within 10^{-15}. A two-token autoregressive KL gradient agreed with central differences within 3.14×10^{-12}.

Table 6 evaluates tied and separate optima for two equally weighted variants with common binary reward and $\beta = 0.5$. Tying can raise or lower mean success while lowering the regularized optimum. For $(q_1, q_2) = (0.9, 0.1)$ it also reduces the first variant's success by 0.104388 despite the mean gain.

Table 6. Exact semantic-policy optima, expressed as probabilities. Objective loss is $F^*_{\mathrm{sep}} - F^*_{\mathrm{tie}}$.

q_1	q_2	p_1^*	p_2^*	p^*_{tie}	Change in mean	Objective loss
0.200	0.200	0.648786	0.648786	0.648786	0.000000	0.000000
0.900	0.100	0.985186	0.450853	0.880797	+0.162778	0.139421
0.100	0.001	0.450853	0.007342	0.072293	−0.156805	0.114170

With $q = 0.2$, $\beta = 0.5$, and objective gap $\delta \le 0.01$, Theorem 3.1 certifies success at least $0.648786 - 0.1 = 0.548786$. The gap must be established for the trained policy. Conversely, $q = 10^{-6}$ and $\beta = 0.2$ give optimum 0.000148391; 2,995,731 independent reference draws are required for a 95% chance of at least one success, since the miss probability is $(1 - q)^N$.

Shared-parameter interference is also exact. Take $p_1(\theta) = \sigma(\theta)$, $p_2(\theta) = \sigma(-2\theta)$, common binary reward, and uniform weighting. At $\theta = 0$, $g_T = 1/4$, $g = -1/8$, and the reference-KL gradient is zero. With $\eta = 0.2$ and $\beta = 0.1$, $\theta^+ = -0.025$: target success falls to 0.493750326 while the regularized mean objective rises to 0.503104335. Since $|\sigma''| \le 1/(6\sqrt{3})$, Proposition 4.1 certifies target change in $[-0.006280070, -0.006219930]$. This restricted-policy calculation is not an estimate of transformer effect size.

Table 7 evaluates eight independent responses, with the same entries for p and $1-p$. Reward-preserving copies of one trajectory instead give zero centered gradient.

TABLE 7. Exact on-policy, sequence-summed GRPO reward terms for $k = 8$. The last column uses a scalar logit parameter, so $\nabla p = p(1-p)$.

Success p	Informative-group probability	Multiplier $c_8(p)$	Expected logit gradient
0.01	0.077255	2.608862	0.025828
0.10	0.569533	2.327689	0.209492
0.50	0.992188	1.855656	0.463914

For Table 8, assume $A = 0.04$, $D = 0.21$, $c_1 = 1$, $C = 12{,}000$, and $n = \lfloor C/(c_0 + m) \rfloor$. Ten draws per base increase variance 2.44-fold when base generation is free, but reduce it 2.87-fold at base cost 20, whose continuous optimum is $\sqrt{105}$. A separate Monte Carlo check used $Z_{ij} = 0.2U_i + \sqrt{0.21}V_{ij}$ with independent symmetric signs. Across 18 allocations and 200,000 repetitions each, variances were within 0.626% of exact values; these moment inputs are not fitted LLM measurements.

TABLE 8. Fixed-budget gradient variance with assumed cost units. Replicas can help or hurt depending on amortizable base cost.

Base cost c_0	Draws per base m	Independent bases n	$10^4 \operatorname{tr}\operatorname{Cov}(\widehat{g})$
0	1	12000	0.208333
0	10	1200	0.508333
0	64	187	2.314505
20	1	571	4.378284
20	10	400	1.525000
20	64	142	3.047975

For transported binary behavior with success 0.9 and target success 0.1, importance weights are $1/9, 9$ and $\mathbb{E}W^2 = 73/9$. The normalized-weight effective-sample-size fraction converges to 0.123288; five independent factors give 2.84838×10^{-5}. This diagnoses weight degeneracy, not every statistic's variance. Separately, lengths $32, 128$ with uniform prompt sampling and per-response token averaging give relative weights $(0.8, 0.2)$; inverse-length sampling changes them to $(16/17, 1/17)$.

REFERENCES

[1] Robert Givan, Thomas Dean, and Matthew Greig, *Equivalence notions and model minimization in Markov decision processes*, Artificial Intelligence **147**(1–2), 163–223 (2003). doi:10.1016/S0004-3702(02)00376-4.

[2] Balaraman Ravindran, *An Algebraic Approach to Abstraction in Reinforcement Learning*, Ph.D. dissertation, University of Massachusetts Amherst (2004). Full text.

[3] Norm Ferns, Prakash Panangaden, and Doina Precup, *Metrics for Finite Markov Decision Processes*, UAI, 162–169 (2004). Archival preprint: arXiv:1207.4114 (2012).

[4] Elise van der Pol, Daniel E. Worrall, Herke van Hoof, Frans A. Oliehoek, and Max Welling, *MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning*, NeurIPS (2020). arXiv:2006.16908.

[5] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee, *A Group-Theoretic Framework for Data Augmentation*, Journal of Machine Learning Research **21**(245), 1–71 (2020). Publisher record.
[6] Richard S. Sutton, Doina Precup, and Satinder Singh, *Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning*, Artificial Intelligence 112, 181–211 (1999). Full text.
[7] Assaf Hallak, Dotan Di Castro, and Shie Mannor, *Contextual Markov Decision Processes*, preprint (2015). arXiv:1502.02259.
[8] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel, *A Survey of Zero-shot Generalisation in Deep Reinforcement Learning*, Journal of Artificial Intelligence Research 76 (2023). arXiv:2111.09794.
[9] Josh Tobin et al., *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World*, IROS (2017). arXiv:1703.06907.
[10] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver, *Universal Value Function Approximators*, ICML, PMLR 37, 1312–1320 (2015). Full text.
[11] André Barreto et al., *Successor Features for Transfer in Reinforcement Learning*, NeurIPS (2017). arXiv:1606.05312.
[12] Marcin Andrychowicz et al., *Hindsight Experience Replay*, NeurIPS (2017). arXiv:1707.01495.
[13] Yan Duan et al., *RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning*, preprint (2016). arXiv:1611.02779.
[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine, *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*, ICML, PMLR 70, 1126–1135 (2017). Full text.
[15] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen, *Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables*, ICML, PMLR 97, 5331–5340 (2019). Full text.
[16] Luisa Zintgraf et al., *VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning*, ICLR (2020). arXiv:1910.08348.
[17] Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley, *Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions*, preprint (2019). arXiv:1901.01753.
[18] Michael Dennis et al., *Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design*, NeurIPS (2020). arXiv:2012.02096.
[19] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel, *Prioritized Level Replay*, ICML, PMLR 139, 4940–4950 (2021). arXiv:2010.03934.
[20] Jack Parker-Holder et al., *Evolving Curricula with Regret-Based Environment Design*, preprint (2022). arXiv:2203.01302.
[21] Simon Yu, Gang Li, Weiyan Shi, and Peng Qi, *PolySkill: Learning Generalizable Skills Through Polymorphic Abstraction*, preprint (2025). arXiv:2510.15863.
[22] Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith, *Reward Machines: Exploiting Reward Function Structure in Reinforcement Learning*, Journal of Artificial Intelligence Research (2022). arXiv:2010.03950.
[23] Mohammed Alshiekh et al., *Safe Reinforcement Learning via Shielding*, AAAI, 2669–2678 (2018). arXiv:1708.08611.
[24] David Blackwell, *Equivalent Comparisons of Experiments*, Annals of Mathematical Statistics **24**(2), 265–272 (1953). doi:10.1214/aoms/1177729032.
[25] Wassily Hoeffding, *Probability Inequalities for Sums of Bounded Random Variables*, Journal of the American Statistical Association **58**(301), 13–30 (1963). doi:10.1080/01621459.1963.10500830.
[26] Jane X. Wang et al., *Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents*, preprint (2021). arXiv:2102.02926.
[27] Google DeepMind, *dm_alchemy*, official software and documentation. `https://github.com/google-deepmind/dm_alchemy` (accessed 7 September 2026).
[28] Open Ended Learning Team et al., *Open-Ended Learning Leads to Generally Capable Agents*, preprint (2021). arXiv:2107.12808.
[29] Adaptive Agent Team et al., *Human-Timescale Adaptation in an Open-Ended Task Space*, preprint (2023). arXiv:2301.07608.
[30] Alexander Nikulin et al., *XLand-MiniGrid: Scalable Meta-Reinforcement Learning Environments in JAX*, NeurIPS Datasets and Benchmarks (2024). arXiv:2312.12044.
[31] XLand-MiniGrid contributors, *XLand-MiniGrid: Rules and Goals*, official software and documentation. `https://github.com/dunnolab/xland-minigrid` (accessed 7 September 2026).

[32] Michael Matthews, Michael Beukman, Chris Lu, and Jakob Foerster, *Kinetix: Investigating the Training of General Agents through Open-Ended Physics-Based Control Tasks*, ICLR (2025). arXiv:2410.23208.

[33] Nathan Cloos, Meagan Jens, Michelangelo Naim, Yen-Ling Kuo, Ignacio Cases, Andrei Barbu, and Christopher J. Cueva, *Baba Is AI: Break the Rules to Beat the Benchmark*, ICML Workshop on LLMs and Cognition (2024). arXiv:2407.13729.

[34] Nathan Cloos et al., *Baba Is AI*, official software and documentation. `https://github.com/nacloos/baba-is-ai` (accessed 7 September 2026).

[35] Marc-Alexandre Côté et al., *TextWorld: A Learning Environment for Text-based Games*, preprint (2018). arXiv:1806.11532.

[36] TextWorld contributors, *Game and GameOptions*, official API documentation. `https://textworld.readthedocs.io/en/stable/textworld.generator.game.html` (accessed 7 September 2026).

[37] Mohit Shridhar et al., *ALFWorld: Aligning Text and Embodied Environments for Interactive Learning*, ICLR (2021). arXiv:2010.03768.

[38] Jaromír Janisch, Tomáš Pevný, and Viliam Lisý, *NASimEmu: Network Attack Simulator & Emulator for Training Agents Generalizing to Novel Scenarios*, preprint (2023). arXiv:2305.17246.

[39] NASimEmu contributors, *Scenario specification*, official software documentation. `https://github.com/jaromiru/NASimEmu/blob/master/docs/SCENARIOS.md` (accessed 7 September 2026).

[40] Konstantin Dobler, Simon Lehnerer, Federico Scozzafava, Jonathan Janke, and Mohamed Ali, *Multilingual Reasoning Gym: Multilingual Scaling of Procedural Reasoning Environments*, preprint (2026). arXiv:2603.10793v1.

[41] Apple, *Multilingual Reasoning Gym*, official software and documentation. `https://github.com/apple/ml-multilingual-reasoning-gym` (accessed 7 September 2026).

[42] Chengsong Huang et al., *EnvHarness: Awakening Static Worlds for Agent Learning*, preprint, 20 August 2026. arXiv:2608.19880v1.

[43] Zafir Stojanovski et al., *REASONING GYM: Reasoning Environments for Reinforcement Learning with Verifiable Rewards*, preprint (2025). arXiv:2505.24760.

[44] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao, *ReAct: Synergizing Reasoning and Acting in Language Models*, ICLR (2023). arXiv:2210.03629.

[45] DeepSeek-AI et al., *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, preprint (2025), revised 4 January 2026. arXiv:2501.12948v2.

[46] Michael Laskin et al., *In-context Reinforcement Learning with Algorithm Distillation*, ICLR (2023). arXiv:2210.14215.

[47] Andrew Y. Ng, Daishi Harada, and Stuart Russell, *Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping*, ICML, 278–287 (1999). Full text.

[48] Monte MacDiarmid et al., *Natural Emergent Misalignment from Reward Hacking in Production RL*, preprint (2025). arXiv:2511.18397.

[49] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu, *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems*, preprint (2020). arXiv:2005.01643.

[50] Nan Jiang and Lihong Li, *Doubly Robust Off-policy Value Evaluation for Reinforcement Learning*, ICML, PMLR 48, 652–661 (2016). Full text.

[51] Rishabh Agarwal et al., *Deep Reinforcement Learning at the Edge of the Statistical Precipice*, NeurIPS (2021). arXiv:2108.13264.

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn, *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*, NeurIPS (2023). arXiv:2305.18290v3.

[53] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker, *Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs*, ACL (2024). Full text.

[54] Zhihong Shao et al., *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*, preprint (2024). arXiv:2402.03300v3.

[55] Youssef Mroueh, *Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification*, preprint (2025), revised 20 October 2025. arXiv:2503.06639v4.

[56] Jiarui Yao, Ruida Wang, Hao Bai, and Tong Zhang, *Future-KL Regularized GRPO: Process-Level Credit Assignment from f-Divergence Regularization*, preprint (2026), revised 23 May 2026. arXiv:2601.10201v2.

[57] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin, *Understanding R1-Zero-Like Training: A Critical Perspective*, preprint (2025), revised 6 October 2025. arXiv:2503.20783v2.

[58] Wenquan Lu, Hai Huang, Enqi Liu, and Randall Balestriero, *PrAg-PO: Prompt Augmented Policy Optimization for Robust and Diverse Mathematical Reasoning*, preprint (2026), revised 8 May 2026. arXiv:2602.03190v3.

[59] Khiem Le, Phuc Nguyen, Youssef Mroueh, Chi-Heng Lin, Shangqian Gao, Ting Hua, and Nitesh V. Chawla, *Transformation-Augmented GRPO for Enhancing Exploration in Reasoning of Large Language Models*, preprint, revised 19 May 2026. arXiv:2601.22478v5.

[60] Yong Yi Bay and Kathleen A. Yearick, *GRPO, Dr. GRPO, and DAPO Are Three Operations on One Number: The Group-Standard-Deviation Identity*, preprint, 30 June 2026. arXiv:2607.00152v1.

[61] Christos Thrampoulidis, Sadegh Mahdavi, and Wenlong Deng, *Advantage Shaping as Surrogate Reward Maximization: Unifying Pass@K Policy Gradients*, Transactions on Machine Learning Research (2026). arXiv:2510.23049v3.

[62] Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang, *Why GRPO Needs Normalization: A Local-Curvature Perspective on Adaptive Gradients*, preprint (2026). arXiv:2601.23135v1.

[63] Eric Neyman and Tim Roughgarden, *No-Regret Learning with Unbounded Losses: The Case of Logarithmic Pooling*, preprint (2022), revised 10 October 2023. arXiv:2202.11219v2.

[64] Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, Shijin Gong, and Chengchun Shi, *Demystifying Group Relative Policy Optimization: Its Policy Gradient is a U-Statistic*, preprint, revised 22 March 2026. arXiv:2603.01162v3.

[65] Chaorui Yao, Yanxi Chen, Yuchang Sun, Yushuo Chen, Wenhao Zhang, Xuchen Pan, Yaliang Li, and Bolin Ding, *Group-Relative REINFORCE Is Secretly an Off-Policy Algorithm: Demystifying Some Myths About GRPO and Its Friends*, ICLR (2026). arXiv:2509.24203v2.

[66] Mingkang Zhu, Xi Chen, Bei Yu, Hengshuang Zhao, and Jiaya Jia, *Stratified GRPO: Handling Structural Heterogeneity in Reinforcement Learning of LLM Search Agents*, ICML, PMLR 306 (2026). Author's conference manuscript.