SHARING ENVIRONMENT RANDOMNESS: WHEN REWARD VARIANCE FALLS BUT GRADIENT RISK RISES

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-based reinforcement learning is widely used to train language model agents by comparing rewards across sampled trajectories. Sharing environmental randomness can make these comparisons less noisy, but it also reduces independent environment coverage under a fixed rollout budget. We construct counterexamples with identical reward distributions but opposite sharing effects on gradient risk, showing that reward statistics alone cannot determine when sharing is beneficial. In this paper, we introduce **Environment-Sharing Risk Analysis (ESRA)**, a framework for evaluating environment sharing through its effect on gradient estimation risk. We derive an exact finite-group risk criterion for raw leave-one-out gradients at a frozen policy, quantifying the trade-off between within-environment fluctuations and differences in mean gradients across environments. We test the criterion with independent gradient audits and controlled sharing interventions. Using Qwen3.5-35B-A3B-Base, we evaluate ESRA on HotpotQA and ALFWorld, covering multi-hop question answering and interactive decision-making. At matched total training cost, environment sharing alone improves native GRPO by 7.10 percentage points on average. Holding the raw-LOO estimation and calibration rules fixed, shared collection improves the two-task mean by 7.00 percentage points over independent collection.

1 INTRODUCTION

Reinforcement learning trains language model agents to retrieve information, use tools, and interact with external environments. GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024) compare returns for the same prompt; GiGPO (Feng et al., 2025) also aligns the task and initial environment state. Even under common initial conditions, external inputs encountered during interaction can obscure action quality and increase policy-gradient variance (Mao et al., 2019).

Here, *sharing environment randomness* means drawing a designated environmental input C once per trajectory group and reusing it across its members. Each member uses independent policy randomness and independently samples the remaining environmental inputs U_i conditional on C. This preserves each trajectory's marginal law while changing dependence within the group. In our HotpotQA protocol, calls with the same query and occurrence reuse a sampled tail-document subset, while document ordering remains independently sampled (Appendix C.1). Appendix A gives concrete favorable and adverse evidence draws (Table 7).

This design builds on common random inputs in policy evaluation (Ng & Jordan, 2000) and input-dependent baseline learning (Mao et al., 2019). Under a fixed rollout budget, however, each shared group covers fewer independent environmental realizations. If mean policy gradients vary across environments, this loss of coverage can outweigh the benefit of cleaner reward comparisons. When does sharing environmental randomness improve gradient estimation, rather than merely reduce reward variance?

We show that reward statistics alone cannot resolve this trade-off. We construct two environments with identical reward distributions in which sharing produces the same reduction in group-centered reward variance, yet has opposite effects on the mean-squared error of the raw leave-one-out (raw-

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.

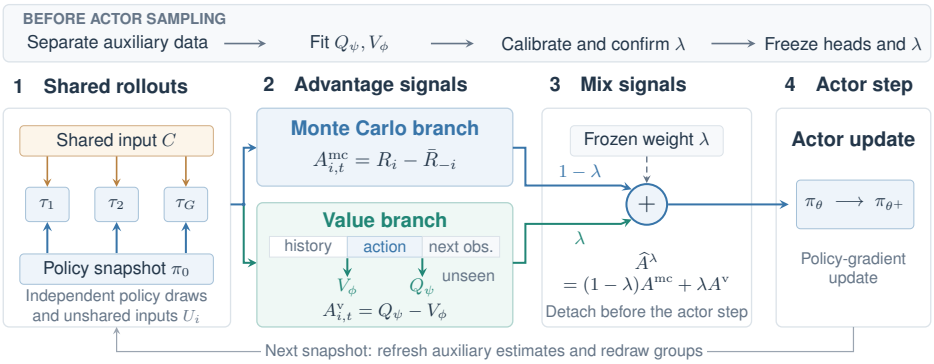


Figure 1: **Shared sampling with an optional calibrated raw-LOO branch.** Shared inputs C and independent policy draws generate each rollout group. Sharing-only uses the MC branch directly. In the calibrated extension, separate auxiliary data fit the value heads and calibrate and confirm λ before fresh actor sampling. MC and value advantages are mixed with frozen λ, detached, and used for one policy step. Value readouts precede the next observation; fitted heads may remain biased. Appendix E.1 gives the split-data procedure.

LOO) policy-gradient estimator. The difference arises from how rewards couple to policy-gradient directions, which reward distributions alone do not capture.

To characterize this missing information, we introduce **Environment-Sharing Risk Analysis (ESRA)**. For raw-LOO estimators at a frozen policy, we derive an exact finite-group risk decomposition separating variation in mean gradients across environments, within-environment fluctuations, and finite-group corrections. Under the sampling assumptions of our analysis, this yields a necessary and sufficient condition for sharing to reduce risk. The risk terms can be estimated from diagnostic trajectories and checked against independent gradient batches. We connect this analysis to training through controlled sharing comparisons under raw LOO and native GRPO. An optional calibrated value mixture extends the estimator comparisons within each sampling design.

Our contributions are threefold:

1. **Reward-only non-identifiability.** We establish that identical reward distributions can conceal opposite effects of sharing on raw-LOO gradient risk.
2. **An exact sharing criterion.** We derive a finite-group risk boundary quantifying the trade-off between within-environment fluctuations and variation in mean gradients across environments.
3. **Controlled training comparisons.** We isolate sharing effects under native GRPO and matched raw-LOO configurations at equal total training cost.

We evaluate Qwen3.5-35B-A3B-Base on HotpotQA and ALFWorld, covering multi-hop question answering and interactive decision-making. At matched total training cost, sharing alone improves native GRPO by **7.10 percentage points** on average across the two tasks. With the same raw-LOO estimation and calibration rules, shared collection improves the two-task average by **7.00 percentage points** over independent collection.

2 Related Work

Shared randomness in simulation and policy learning. Common random numbers couple stochastic evaluations through shared inputs; their variance benefits depend on the induced covariance (Glasserman & Yao, 1992). PEGASUS reuses scenarios across policy evaluations (Ng & Jordan, 2000). Mao et al. (2019) repeat an exogenous input sequence across rollouts to learn input-dependent baselines. These approaches motivate environmental coupling. ESRA examines its effect on group-relative gradient MSE at fixed rollout counts, where the loss of independent environment coverage also matters.

Group construction in language model and agent reinforcement learning. GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024) compare returns within prompt groups. GiGPO (Feng et al., 2025) starts from identical tasks and initial states and regroups repeated anchor states for credit assignment. RAGEN (Wang et al., 2025) studies multi-turn rollout design; its implementation assigns a common environment configuration and seed within each group (RAGEN Contributors, 2026). MetaFlow (Luo et al., 2026) evaluates candidate workflows on the same sampled problem instances before computing group-relative advantages. These designs share prompts, initial conditions, or evaluation instances. ESRA couples designated environmental inputs during interaction, holding individual trajectory marginals and estimator configurations fixed.

Gradient risk and rollout allocation. Zhou et al. (2026) characterize group-relative policy gradients using U-statistics, including their MSE and group-size effects. VIP (Nguyen et al., 2026) uses predicted variance to allocate a rollout budget across prompts. ESRA conditions the U-statistic representation on shared environmental inputs. Its exact finite-group boundary separates variation in mean gradients across environments from within-environment fluctuations and finite-group corrections. At a frozen policy, it determines when sharing lowers raw-LOO risk at matched rollout counts and group sizes (Theorem 2).

3 Environment Sharing: Reward Variance and Gradient Risk

3.1 Setting

A task specification $x \sim \mathcal{D}$ is fixed before reset; any initial observation o_0 is generated by the environment tape. A policy emits $a_t \sim \pi_\theta(\cdot \mid h_{t-1})$ and receives o_t, with $h_t = (x, o_0, a_1, o_1, \ldots, a_t, o_t)$. We study bounded-horizon, undiscounted terminal utility $R = r(\tau, z) \in [0, 1]$ with no explicit θ dependence; evaluation is a final observation. A random input tape $z \sim \mu_x$, independent of the policy's random numbers and of θ, determines responses conditional on actions:

$$o_t = \mathcal{E}(h_{t-1}, a_t, z), \qquad J(\theta) = \mathbb{E}_{x,z,\tau\sim\pi_\theta}[R]. \tag{1}$$

Sharing z does not force different actions to receive identical observations. All comparisons below fix a reference policy π_0, decoding, task x, and harness. Environment drift violates this fixed-distribution assumption.

3.2 Which reward variation can sharing remove?

Let C be the inputs shared within a group and U those sampled separately. We decompose reward variation in that order: across C, across U given C, then across policy draws. Write $z = (C, U)$ and $m(C, U) = \mathbb{E}_{\pi_0}[R \mid x, C, U]$. The inputs C, U may be dependent; all terms below fix x, π_0 and the harness. Iterated total variance gives

$$\begin{aligned} S := \mathrm{Var}(R \mid x) &= \sigma_C^2 + \sigma_U^2 + \sigma_\pi^2, \\ \sigma_C^2 &= \mathrm{Var}_C\big(\mathbb{E}_{U|C} m(C, U)\big), \quad \sigma_U^2 = \mathbb{E}_C \mathrm{Var}_{U|C} m(C, U), \\ \sigma_\pi^2 &= \mathbb{E}_{C,U} \mathrm{Var}_{\pi_0}(R \mid x, C, U). \end{aligned} \tag{2}$$

For $S > 0$, the shared-environment fraction is $\eta_C = \sigma_C^2/S$ and the total environment fraction is $\eta = (\sigma_C^2 + \sigma_U^2)/S$. The unshared fraction is their difference $\eta - \eta_C$. All use the same denominator and are undefined at $S = 0$. Sharing is harness-dependent. These fractions describe the ordered conditional variance decomposition.

3.3 Finite-group comparison of sampling designs

Assumption 1 (Matched marginals and conditional independence)**.** *Let $G \geq 2$. Independent collection draws (C_i, U_i) i.i.d.; partial blocking draws one C and independent $U_i \sim \mu_x(\cdot \mid C)$; full blocking shares (C, U). Policy random numbers are independent across members, so returns are conditionally i.i.d. given the shared variables. All designs preserve each rollout's marginal distribution: automatic for the constructions above, and an implementation requirement for a real harness; sampler checks are specified in Appendix C.1.*

Theorem 1 (Finite-group residual variance)**.** *Under Assumption 1, let $D_i = R_i - \bar{R}$. Group centering removes the shared conditional reward mean. For independent, partially blocked (pb) and fully blocked (fb) collection,*

$$\begin{aligned} \mathrm{Var}_{\mathrm{ind}}(D_i) &= (1 - 1/G)S, \\ \mathrm{Var}_{\mathrm{pb}}(D_i) &= (1 - 1/G)(\sigma_U^2 + \sigma_\pi^2), \\ \mathrm{Var}_{\mathrm{fb}}(D_i) &= (1 - 1/G)\sigma_\pi^2. \end{aligned} \tag{3}$$

For nonzero denominators, comparing the same G gives

$$\frac{\mathrm{Var}_{\mathrm{ind}}(D_i)}{\mathrm{Var}_{\mathrm{pb}}(D_i)} = \frac{1}{1 - \eta_C}, \qquad \frac{\mathrm{Var}_{\mathrm{ind}}(D_i)}{\mathrm{Var}_{\mathrm{fb}}(D_i)} = \frac{1}{1 - \eta}. \tag{4}$$

Proof sketch. Center conditional on the shared inputs, then average the remaining variance (Appendix B.1).

The leave-one-out (LOO) baseline uses the other members: $R_i - \bar{R}_{-i} = D_i/(1 - 1/G)$, with independent variance $S/(1 - 1/G)$. Only the matched-design ratios are G-invariant. These are reward identities; normalized, clipped or selected groups require separate analysis (Appendix B.4).

3.4 When sharing helps rewards but hurts gradients

A gradient weights reward by the trajectory's score: how policy parameters change its log probability. Thus reward fluctuations alone do not describe the resulting update directions. Fix x, θ_0 and let $s = \sum_t \nabla_\theta \log \pi_\theta(a_t \mid h_{t-1})|_{\theta_0}$ be the complete score, without length normalization. Assume differentiable, parameter-independent support, exchange of derivative and expectation, and $\mathbb{E}\|s\|^2 < \infty$; then $\mathbb{E}[s \mid C] = 0$.

Put $m_C = \mathbb{E}[R \mid C]$, $g_C = \mathbb{E}[sR \mid C]$, and $g_x = \mathbb{E}g_C$. Three moments describe gradient risk: β_C measures changes in the mean gradient across environments, v_C measures within-environment fluctuations of centred score-weighted returns, and u_C supplies the finite-group correction. Their exact forms are

$$\begin{aligned} &\beta_C = \mathbb{E}\|g_C - g_x\|^2, \quad v_C = \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(s(R - m_C) \mid C), \\ &u_C = \mathbb{E}\left[\mathbb{E}(\|s\|^2 \mid C)\mathrm{Var}(R \mid C) + \|g_C\|^2\right]. \end{aligned} \tag{5}$$

These expectations average over C.

Theorem 2 (Raw-LOO sampling risk)**.** *Under Assumption 1 and the score conditions above, collect P independent groups sharing C, with conditionally i.i.d. $(s_i, R_i) \mid C$, and average $s_i(R_i - \bar{R}_{-i})$ over $N = PG$ trajectories. Full sharing replaces C by (C, U). The mean is g_x and the MSE is*

$$\mathcal{R}_C(G, N) = \frac{G\beta_C + v_C + u_C/(G - 1)}{N}. \tag{6}$$

Let v_0, u_0 denote the moments for independent collection, with a constant shared variable. At the same N, G, sharing strictly lowers risk if and only if

$$G\beta_C + v_C - v_0 + \frac{u_C - u_0}{G - 1} < 0. \tag{7}$$

Sharing helps only when the reduction in the within-environment terms outweighs the variation in mean gradients. The proof uses the established U-statistic decomposition (Zhou et al., 2026); Appendix B.2 gives details and the random-prompt extension.

Proposition 1 (Reward statistics do not identify the benefit of sharing)**.** *Let C, A be independent uniform signs at $\theta_0 = 0$, with $\pi_\theta(A) \propto e^{\theta A}$ and score $s = A$. The rewards $R_+ = \frac{1}{2} + 0.1C + 0.2A$ and $R_\times = \frac{1}{2} + 0.1C + 0.2AC$ have the same (C, R) law and $(\sigma_C^2, \sigma_U^2, \sigma_\pi^2) = (0.01, 0, 0.04)$. Sharing C reduces their group-centred reward variance by 20%. Yet at $G = 2$ it changes gradient MSE from $0.10/N$ to $0.08/N$ for R_+, and to $0.16/N$ for $R_\times$.*

The reward labels can switch between actions without changing the reward distribution, yet their association with the policy score changes. Thus even a perfectly measured η_C cannot replace equation 7. Appendix B.3 also gives a nonzero-gradient harmful case and a continuous transition at fixed reward components.

4 ESRA: FROM THE RISK BOUNDARY TO PRACTICAL COMPARISONS

ESRA compares independent and shared collection while holding the gradient estimator fixed. Sharing-only GRPO+B retains the native advantages and update rule. Optional value estimation and calibration extend the comparisons within each sampling design. In the tables, +ESRA denotes this full shared configuration (Figure 1); raw-LOO and native-objective implementations are labelled separately. The sharing design is declared before evaluation.

4.1 SHARED INPUTS AND MARGINAL PRESERVATION

For a fixed task x, split environmental inputs into a shared component C and a trajectory-specific component U. Partial sharing draws

$$C \sim \mu_x^C, \qquad C_1 = \cdots = C_G = C, \qquad U_i \overset{\text{iid}}{\sim} \mu_x(\cdot \mid C). \tag{8}$$

Each member has its own reset environment state and independent policy random numbers; responses still depend on its own actions. Resample C across groups and updates. Independent collection instead draws a separate C_i for each member. With the same conditional response law and input distributions, either design gives

$$p_{\text{sh}}(\tau_i \mid x) = \int p_{\pi_0}(\tau_i \mid x, c, u)\, \mu_x(du \mid c)\, \mu_x^C(dc) = p_{\text{ind}}(\tau_i \mid x). \tag{9}$$

Full sharing also couples U, preserving marginals. Appendix C.1 defines C as retrieval-content selection (HotpotQA) or object-ID maps (ALFWorld), with independent list ordering under partial sharing, and specifies reset/replay checks. Inputs ineffective after fixing x yield zero population risk difference.

4.2 SHARING-ONLY AND OPTIONAL VALUE ESTIMATION

Raw LOO uses sampled returns without a critic. The optional value branch predicts the action's contribution before the next observation, using independently fitted heads frozen during actor construction. Its advantage signals are

$$A_{i,t}^{\text{mc}} = R_i - \bar{R}_{-i}, \qquad A_{i,t}^{\text{v}} = Q_\psi(h_{i,t-1}, a_{i,t}) - V_\phi(h_{i,t-1}). \tag{10}$$

The Q readout follows the action and the V readout precedes it, using h_0 for the first action. Appendix E.1 specifies the fitting and observation boundaries. Let $s_{i,t} = \nabla_\theta \log \pi_\theta(a_{i,t} \mid h_{i,t-1})|_{\theta_0}$; averaging $\sum_t s_{i,t} A_{i,t}^{\text{mc}}$ and $\sum_t s_{i,t} A_{i,t}^{\text{v}}$ over the batch gives Y and Z. Under score regularity, $\mathbb{E}Y = g = \nabla J(\theta_0)$ and $\mathbb{E}Z = g + d$, where d is the score-weighted Q approximation bias (Appendix E.1).

Optional mixtures. The extension uses

$$\hat{A}_{i,t}^{\lambda} = (1-\lambda) A_{i,t}^{\text{mc}} + \lambda A_{i,t}^{\text{v}}. \tag{11}$$

Setting $\lambda = 0$ gives raw MC under the declared sampling design; $\lambda = 1$ uses values only, and intermediate mixtures retain bias λd. Fixed mixtures select one weight on train-validation. Calibrated mixtures fit a weight per snapshot using independent fitting and confirmation data, followed by fresh actor data; failed calibration uses $\lambda = 0$. Appendix E.1 gives the risk calculation and fitting procedure. All auxiliary work is charged to the comparison budget; cost measurement is specified in Appendix E.3.

4.3 CONNECT THE RAW REFERENCE TO PRACTICAL OBJECTIVES

For a concrete reference within this design space, use

$$\hat{L}_\lambda(\theta) = \frac{1}{PG} \sum_{x \in \mathcal{B}} \sum_{i,t} \rho_{i,t}(\theta)\, \text{sg}\left[\hat{A}_{i,t}^{\lambda}\right], \tag{12}$$

where $P = |\mathcal{B}|$ and $\rho_{i,t} = \pi_\theta(a_{i,t} \mid h)/\pi_0(a_{i,t} \mid h)$. At θ_0, $\nabla \hat{L}_\lambda = (1-\lambda)Y + \lambda Z$; $\lambda = 0$ recovers raw LOO under the current design. This equation specifies the gradient used in the reference audit.

Connection and perturbation. At θ_0, standard PPO clipping (Schulman et al., 2017) is locally inactive for a finite batch and positive clipping radius (Appendix B.4); normalization and length weights can still change the gradient. The following bound makes this change explicit at the same frozen policy and reward-gradient target.

Proposition 2 (Stability of the sharing-risk sign)**.** *For $d \in \{C, 0\}$, let Y_d be the raw-LOO batch gradient of Theorem 2, $\widetilde{Y}_d = Y_d + E_d$ the implemented gradient on the same data, and $\epsilon_d^2 = \mathbb{E}\|E_d\|^2 < \infty$. Define $\widetilde{\mathcal{R}}_d = \mathbb{E}\|\widetilde{Y}_d - g\|^2$ and $\Delta = \mathcal{R}_C - \mathcal{R}_0$. Then*

$$\left|\widetilde{\mathcal{R}}_C - \widetilde{\mathcal{R}}_0 - \Delta\right| \le \Gamma := \sum_{d\in\{C,0\}} \left(2\sqrt{\mathcal{R}_d}\,\epsilon_d + \epsilon_d^2\right). \tag{13}$$

Thus $|\Delta| > \Gamma$ preserves the sign of the sharing effect.

Appendix B.4 proves the bound and decomposes standardization into covariance and bias corrections to the exact raw risk. This is a sufficient population condition; its empirical coverage is specified for a separate audit, and it does not assert later-update equivalence. Native adapters retain their base objectives (Table 10); Appendix E.1 specifies the implementation. All diagnostics and failed calibration attempts count toward total cost under Appendix C.

5 Experiments

We test the risk criterion, isolate sharing effects, and follow the same diagnostic conditions into GRPO continuation. Full-training comparisons then measure sharing gains and optional estimator extensions. Cost measurement is specified in Appendix E.3.

5.1 Tasks, sharing protocols, and budgets

The core model is Qwen3.5-35B-A3B-Base (Qwen Team, 2026a) (35B total/3B active), trained on a text-only 50:50 mixture of HotpotQA (Yang et al., 2018) and ALFWorld (Shridhar et al., 2021), with five training seeds. HotpotQA uses a fixed local Wikipedia index for tool-based retrieval; ALFWorld uses text interaction with held-out games. Task conditions are fixed within each sharing contrast; policy sampling remains independent. Environment settings and gradient objectives are identified separately: GRPO retains its original advantage normalization and clipping, while raw LOO uses the theoretical estimator. Core HotpotQA and ALFWorld runs use REF-H and REF-A, respectively:

Protocol / task	Shared C	Independent U (partial)
REF-H / HotpotQA	Tail-passage selection	Returned-passage order
REF-A / ALFWorld	Object-instance ID map	Unordered-list order

Main shared contrasts use partial sharing; full sharing is confined to labelled mechanism conditions. Appendix C.1 specifies sampling and reset/replay; Appendix C.2 maps results to protocols and sharing modes.

Scores are single-attempt success/accuracy. HotpotQA dev and ALFWorld unseen are the core tests; WebShop (Yao et al., 2022), WebArena (Zhou et al., 2024), τ-bench (Yao et al., 2024) and BFCL v3 multi-turn (Patil et al., 2025) evaluate transfer of the same checkpoints. MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023) and AIME'25 are reasoning controls. Actor-token and total-cost comparisons use separate caps, charging diagnostics, critic fitting, calibration and resets to total training cost. Evaluation cost is logged separately. Splits, the configuration matrix and paired inference are in Appendix C; mechanism protocols are in Appendix D.

5.2 Predicting sharing risk on held-out agent batches

Appendix B.3 proves reward-only non-identifiability. A condition fixes the task law, training seed, checkpoint, group size and sharing implementation. Independent diagnostic data predict the raw-gradient risk difference before audit outcomes are seen:

$$\widehat{\Delta}_C = \frac{G\widehat{\beta}_C + \widehat{v}_C - \widehat{v}_0 + (\widehat{u}_C - \widehat{u}_0)/(G-1)}{N}. \tag{14}$$

Coefficients are averaged over the same prompt law; its heterogeneity term cancels between designs only at matched N, G. A negative interval predicts benefit, a positive interval predicts harm, and an interval crossing zero is unresolved. Fresh complete raw-gradient batches audit the sign using the reference-free variance identity in Appendix D.1.

Table 1 compares the full criterion, reward-only predictions and the $-\beta$ ablation. The full criterion agrees with audited signs in 97/120 primary conditions, versus 64/120 for reward-only. Complete sign matrices retain mistaken and unresolved predictions (Appendix D.1). Table 3 follows these same diagnostic strata through reward variance, raw-gradient risk and GRPO continuation.

Figure 2 in the appendix gives the fixed-N group-size sweep, including a HotpotQA sign crossing and an ALFWorld no-crossing case.

Table 1: Held-out raw-gradient risk prediction. Primary task settings; 60 conditions per task from five training seeds. Coverage, agreement and missed harm are percentages; gap MAE uses 10^{-3} units. Brackets give 95% intervals.

Task	Predictor	Coverage	Agreement	Missed harm	Gap MAE
HotpotQA	Full criterion	86.7 [76, 94]	81.7 [70, 90]	6.2 [0, 24]	0.38 [0.27, 0.51]
	Reward-only	85.0 [74, 93]	55.0 [42, 67]	75.0 [49, 92]	1.62 [1.21, 2.10]
	Criterion $-\beta$	90.0 [79, 97]	70.0 [57, 81]	56.2 [30, 80]	1.08 [0.78, 1.43]
ALFWorld	Full criterion	85.0 [73, 93]	80.0 [67, 89]	4.8 [0, 20]	0.42 [0.30, 0.57]
	Reward-only	85.0 [73, 93]	51.7 [39, 64]	76.2 [54, 91]	1.85 [1.36, 2.42]
	Criterion $-\beta$	85.0 [73, 93]	63.3 [50, 75]	52.4 [31, 73]	1.29 [0.92, 1.72]

5.3 Isolating sharing from value estimation

Cross independent/shared collection with raw MC, a validation-fixed mixture, and learned-only estimation. Table 2 separates two outcomes. Panel A fixes the policy, value heads, mixture weight, prompt law and N, G, and audits gradient MSE using independent references. Panel B starts from matched policy and optimizer states, then refits heads as training evolves under a common total-cost cap. Incoming-checkpoint cost is identical and disclosed separately.

For the fixed-head, per-trajectory learned branch, the population shared-minus-independent risk is nonnegative under matched marginals and conditional independence. The positive gap in Table 2A satisfies this check (Appendix D.2). A degenerate shared input provides a zero-effect control with identical sampling laws. Panel B separately measures evolving policies; Table 3 follows the diagnostic strata into GRPO continuation.

Table 2: Sharing effects. A: frozen-policy risk in 10^{-3} units at common N, G. B: task differences after matched-total-cost continuation, in percentage points. $I = \Delta_{\text{mixture}} - \Delta_{\text{MC}}$.

A. Frozen policy and heads; same sampling budget			
Estimator/contrast	Independent risk	Shared risk	Δ risk [95%]
Raw MC	10.0	7.5	-2.5 [$-3.2, -2.0$]
Fixed mixture	8.2	5.9	-2.3 [$-3.1, -1.7$]
Learned-only	6.5	6.8	$+0.3$ [0.14, 0.48]
Interaction I	n/a	n/a	0.2 [-0.4, 0.8]
Degenerate shared input (MC)	10.0	10.01	$+0.01$ [-0.12, 0.14]
B. Evolving policies; matched total continuation cost			
Estimator/contrast	Δ HotpotQA	Δ ALFWorld	Δ aggregate [95%]
Raw MC	3.0 [2.2, 3.8]	5.0 [3.8, 6.2]	4.0 [3.1, 4.9]
Fixed mixture	3.0 [2.3, 3.7]	4.0 [3.0, 5.0]	3.5 [2.7, 4.3]
Learned-only	1.4 [0.8, 2.0]	1.8 [1.0, 2.6]	1.6 [1.0, 2.2]
Interaction I	0.0 [-0.7, 0.7]	-1.0 [-2.1, 0.1]	-0.5 [-1.2, 0.2]

5.4 From risk predictions to GRPO learning

Table 3 joins the diagnostic and continuation comparisons by the same task/seed/checkpoint/group-size/sharing condition. Diagnostics fix the strata; fresh batches audit raw-gradient risk, and separate rollouts train GRPO and GRPO+B from cloned policy and optimizer states. GRPO advantages and

training settings are common; only environment sharing changes. Both arms pay the diagnostic and training costs under one continuation cap (Appendix D.3).

The harmful strata show lower reward variance, higher audited gradient risk and lower continuation scores on both tasks. Beneficial strata improve GRPO scores, while unresolved strata have intervals crossing zero. These are stratum-level learning outcomes; the raw-risk criterion itself applies at the frozen policy. On the same frozen batches, sign agreement is 97/120 for raw LOO, 95/120 after normalization and 92/120 for GRPO (Table 12).

Table 3: From predicted risk to GRPO continuation under the same sharing conditions. Strata are fixed by diagnostic intervals. All changes are shared minus independent; variance and risk use 10^{-3} units, and score is GRPO+B minus GRPO in percentage points at matched total continuation cost. Brackets give 95% intervals; n counts conditions.

Task	Stratum	n	$\Delta\mathrm{Var}(D)$	Predicted Δ raw risk	Audited Δ raw risk	GRPO Δ score
HotpotQA	Benefit	37	-12 $[-15, -9]$	-4.15 $[-4.75, -3.55]$	-4.00 $[-4.70, -3.30]$	$+5.0$ $[3.8, 6.2]$
	Harm	15	-14 $[-18, -10]$	1.90 $[1.35, 2.45]$	2.10 $[1.45, 2.75]$	-2.0 $[-3.1, -0.9]$
	Unresolved	8	-7 $[-11, -3]$	0.08 $[-0.42, 0.58]$	-0.05 $[-0.61, 0.51]$	$+0.2$ $[-0.8, 1.2]$
ALFWorld	Benefit	31	-18 $[-23, -13]$	-7.80 $[-8.65, -6.95]$	-8.00 $[-9.00, -7.00]$	$+6.4$ $[4.9, 7.9]$
	Harm	20	-20 $[-26, -14]$	3.05 $[2.32, 3.78]$	3.20 $[2.36, 4.04]$	-2.8 $[-4.1, -1.5]$
	Unresolved	9	-9 $[-14, -4]$	-0.06 $[-0.64, 0.52]$	0.10 $[-0.57, 0.77]$	-0.1 $[-1.3, 1.1]$

Dense/MoE perturbation audit. The audit pairs the core MoE with dense Qwen3.5-27B, 9B-Base and 4B models on the same REF-H/REF-A grid. It tests within-family sensitivity; 27B and 4B use post-trained checkpoints, so cross-model differences also reflect training stage. Table 4 summarizes agreement and margin coverage; Appendix D.4 specifies the audit.

Table 4: Dense/MoE audit. Agreement counts use all 120 conditions across both tasks. Margin coverage (%) is the fraction meeting the estimated native-gradient condition $|\Delta| > \Gamma$.

Qwen3.5 model	Raw LOO	Norm.-only	Native GRPO	Margin coverage
35B-A3B-Base (MoE)	97/120	95/120	92/120	80.0
27B (dense, post-trained)	102/120	100/120	98/120	75.0
9B-Base (dense)	92/120	91/120	88/120	75.0
4B (dense, post-trained)	72/120	81/120	83/120	70.0

5.5 Sharing gains and estimator extensions

At matched total training cost, GRPO+B isolates sharing with native advantages and improves the two-task mean over GRPO by +7.10 points (Appendix D.3). Holding the raw-LOO estimation and calibration rules fixed, shared collection improves the mean by +7.00 points (Table 13). The optional GRPO+ESRA extension adds value estimation and calibration, with a full-minus-B gain of +5.50 points. Table 5 places these GRPO contrasts before wider adapter and transfer comparisons. Holm correction (Holm, 1979) covers the five secondary adapter contrasts; transfer intervals are descriptive. MATH-500 uses the predeclared -0.5-point non-inferiority margin (Appendix C); AIME'25 is a separate reasoning control.

Table 5: Success/accuracy (%) at matched total training cost. HPQA: HotpotQA; ALFW: ALFWorld; WShop: WebShop; WArena: WebArena; BFCL-MT: BFCL v3 multi-turn; τ: τ-bench retail/airline macro mean. B adds sharing with the original GRPO advantages; +ESRA also uses values and calibration. Five-seed means $\pm$ std; stability: Appendix E.3.

	Training-domain tests		Held-out transfer environments				Reasoning controls	
Estimator	HPQA	ALFW	WShop	WArena	τ	BFCL-MT	MATH-500	AIME'25
PPO reference	$52.1_{\pm 2.9}$	$66.9_{\pm 3.1}$	$48.2_{\pm 1.4}$	$19.4_{\pm 1.5}$	$27.3_{\pm 1.3}$	$38.6_{\pm 1.2}$	$46.6_{\pm 1.6}$	$11.4_{\pm 1.9}$
GRPO	$54.0_{\pm 2.0}$	$68.0_{\pm 3.1}$	$45.0_{\pm 2.2}$	$20.0_{\pm 2.0}$	$30.0_{\pm 2.0}$	$38.0_{\pm 1.8}$	$62.0_{\pm 0.6}$	$15.0_{\pm 1.0}$
+B	$61.8_{\pm 2.8}$	$74.4_{\pm 1.9}$	$48.4_{\pm 1.8}$	$23.2_{\pm 1.7}$	$33.4_{\pm 1.6}$	$40.8_{\pm 1.4}$	$62.1_{\pm 0.55}$	$15.1_{\pm 0.95}$
+ESRA	$65.4_{\pm 1.7}$	$81.8_{\pm 1.5}$	$51.0_{\pm 1.5}$	$26.0_{\pm 1.4}$	$36.0_{\pm 1.4}$	$43.0_{\pm 1.2}$	$62.2_{\pm 0.5}$	$15.2_{\pm 0.9}$
RLOO	$55.4_{\pm 1.7}$	$68.7_{\pm 2.9}$	$46.3_{\pm 2.0}$	$21.4_{\pm 1.9}$	$31.3_{\pm 1.7}$	$39.4_{\pm 1.6}$	$62.3_{\pm 0.6}$	$15.3_{\pm 1.0}$
+ESRA	$60.1_{\pm 1.2}$	$76.0_{\pm 1.7}$	$52.4_{\pm 1.5}$	$27.4_{\pm 1.4}$	$37.4_{\pm 1.2}$	$44.4_{\pm 1.3}$	$62.9_{\pm 0.5}$	$15.6_{\pm 0.9}$
GSPO	$56.2_{\pm 1.9}$	$70.4_{\pm 2.5}$	$47.1_{\pm 1.7}$	$21.9_{\pm 2.1}$	$32.6_{\pm 1.9}$	$40.1_{\pm 1.5}$	$62.5_{\pm 0.55}$	$15.4_{\pm 0.95}$
+ESRA	$60.9_{\pm 1.1}$	$76.7_{\pm 1.5}$	$53.0_{\pm 1.4}$	$28.1_{\pm 1.2}$	$38.1_{\pm 1.3}$	$45.1_{\pm 1.1}$	$63.0_{\pm 0.5}$	$15.8_{\pm 0.85}$
DAPO	$57.1_{\pm 1.5}$	$70.6_{\pm 2.7}$	$48.2_{\pm 1.6}$	$23.1_{\pm 1.6}$	$33.1_{\pm 1.5}$	$41.3_{\pm 1.7}$	$62.6_{\pm 0.6}$	$15.5_{\pm 1.0}$
+ESRA	$61.4_{\pm 1.2}$	$76.8_{\pm 1.6}$	$53.9_{\pm 1.2}$	$28.9_{\pm 1.1}$	$38.9_{\pm 1.4}$	$45.9_{\pm 1.2}$	$63.1_{\pm 0.55}$	$15.9_{\pm 0.9}$
GiGPO	$57.8_{\pm 1.4}$	$73.4_{\pm 2.1}$	$48.6_{\pm 1.9}$	$23.6_{\pm 1.5}$	$34.4_{\pm 1.6}$	$41.8_{\pm 1.4}$	$62.7_{\pm 0.6}$	$15.6_{\pm 1.0}$
+ESRA	$61.6_{\pm 1.0}$	$78.4_{\pm 1.3}$	$54.2_{\pm 1.3}$	$29.2_{\pm 1.3}$	$39.6_{\pm 1.1}$	$46.4_{\pm 1.0}$	$63.2_{\pm 0.5}$	$16.1_{\pm 0.9}$
turn-PPO	$59.2_{\pm 1.3}$	$74.1_{\pm 2.2}$	$50.4_{\pm 1.5}$	$25.3_{\pm 1.7}$	$35.7_{\pm 1.4}$	$43.2_{\pm 1.6}$	$62.8_{\pm 0.5}$	$15.8_{\pm 0.9}$
+ESRA	$61.8_{\pm 0.9}$	$78.5_{\pm 1.4}$	$55.0_{\pm 1.1}$	$30.0_{\pm 1.0}$	$40.7_{\pm 1.2}$	$47.7_{\pm 0.9}$	$63.4_{\pm 0.5}$	$16.2_{\pm 0.85}$

Table 6: Paired gains and controls. HotpotQA/ALFWorld mean differences in percentage points, with paired 95% intervals. Top: +ESRA minus base at matched budgets; Holm p covers the five secondary total-cost contrasts (GRPO is primary). Bottom: matched total cost; B adds sharing only and raw LOO uses full calibration. MATH-500 uses a one-sided 95% lower bound (LCB), with a -0.5-point margin.

Budget / statistic	GRPO	RLOO	GSPO	DAPO	GiGPO	turn-PPO
Total cost	12.60 [11.0, 14.2]	6.0 [4.8, 7.2]	5.5 [4.3, 6.7]	5.25 [4.1, 6.4]	4.4 [3.3, 5.5]	3.5 [2.4, 4.6]
Actor tokens	14.00 [12.3, 15.7]	7.5 [6.2, 8.8]	6.9 [5.7, 8.1]	6.6 [5.4, 7.8]	5.6 [4.4, 6.8]	4.7 [3.5, 5.9]
Holm p (total cost)	n/a	0.002	0.004	0.006	0.010	0.020

Effect	Estimator	Contrast	Δ (pp)	95% interval
Sharing	GRPO	+B $-$ base	**+7.10**	[5.8,8.4]
	Raw LOO	Shared $-$ independent	**+7.00**	[5.7,8.3]
Value + calibration	GRPO	+ESRA $-$ +B	**+5.50**	[4.2,6.8]
MATH-500 control	GRPO	+ESRA $-$ base	+0.2	LCB -0.15; pass

Training stability. At matched total cost, GRPO+ESRA reduces seed std by 15% on HotpotQA and 50% on ALFWorld (Appendix E.3).

6 Conclusion

Environment sharing trades lower within-group reward noise for reduced independent environment coverage. Our matched-reward counterexample and exact raw-LOO criterion show how between-environment gradient heterogeneity and within-design score–reward moments determine this trade-off. At matched total cost, reported sharing gains on the HotpotQA/ALFWorld mean are 7.10 percentage points for native GRPO and 7.00 points when raw-LOO estimation and calibration rules are held fixed. Independent gradient audits and matched training comparisons connect the risk criterion to both beneficial and harmful sharing conditions. Optional value-calibration results extend this analysis with explicit auxiliary costs.

AI USE STATEMENT

No generative AI tools were used in conducting this research or in writing, editing, or preparing this manuscript. The authors take full responsibility for the entire content of the paper.

ETHICS STATEMENT

This study uses publicly available datasets and does not involve experiments with human participants. No empirical data were fabricated or falsified.

REPRODUCIBILITY STATEMENT

The core algorithm implementation and evaluation scripts are submitted with the paper in `ESRA_supplementary`. The core modules are located in `ESRA_supplementary/esra/`, including `sharing.py` for shared-input sampling, `risk.py` for the raw-LOO risk criterion, and `trainer.py` for the algorithmic update procedure. Configuration defaults are defined by `ESRAConfig` in `esra/trainer.py`; evaluation-specific parameters, random seeds, and repetition counts are specified in the scripts under `scripts/`. The package `README.md` documents the module structure, evaluation coverage, and execution commands. Assumptions and proofs are given in Appendix B, experimental settings and sharing protocols in Appendix C, independent risk prediction and audit procedures in Appendix D, and optional value estimation, calibration, and cost accounting in Appendix E.

REFERENCES

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. *arXiv preprint arXiv:2402.14740*, 2024.

ALFWorld Contributors. ALFWorld: Object-instance name randomization implementation. `https://github.com/alfworld/alfworld/blob/master/alfworld/agents/utils/misc.py`, 2026. Demangler source snapshot accessed September 23, 2026.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL `https://www.jstor.org/stable/4615733`.

Junbo Li, Peng Zhou, Rui Meng, Meet P. Vadera, Lihong Li, and Yang Li. Turn-PPO: Turn-level advantage estimation with PPO for improved multi-turn RL in agentic LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.328. URL `https://aclanthology.org/2026.findings-eacl.328/`.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Gan Luo, Zihan Qin, Bin Dong, and Wotao Yin. From search to synthesis: Training LLMs as zero-shot workflow generators. *arXiv preprint arXiv:2606.30704*, 2026.

Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. In *International Conference on Learning Representations (ICLR)*, 2019.

Andrew Y. Ng and Michael I. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000.

Hieu Trung Nguyen, Bao Nguyen, Wenao Ma, Yuzhi Zhao, Ruifeng She, and Viet Anh Nguyen. Adaptive rollout allocation for online reinforcement learning with verifiable rewards. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.01601v3.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Qwen Team. Qwen3.5-35B-A3B-Base model card. https://huggingface.co/Qwen/Qwen3.5-35B-A3B-Base, 2026a. Official pre-trained Base model card.

Qwen Team. Qwen3.5-9B-Base model card. https://huggingface.co/Qwen/Qwen3.5-9B-Base, 2026b.

Qwen Team. Qwen3.5-27B model card. https://huggingface.co/Qwen/Qwen3.5-27B, 2026c. Official post-trained dense model card.

Qwen Team. Qwen3.5-4B model card. https://huggingface.co/Qwen/Qwen3.5-4B, 2026d. Official post-trained dense model card.

RAGEN Contributors. RAGEN: Environment group configuration. https://github.com/mll-lab-nu/RAGEN/blob/main/config/base.yaml, 2026. Official implementation, accessed September 25, 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

John Schulman, Philipp Moritz, Sergey Levine, Michael I Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

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.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025. URL https://arxiv.org/abs/2504.20073v2.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022. URL https://arxiv.org/abs/2207.01206.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

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.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

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. *arXiv preprint arXiv:2603.01162*, 2026.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

A A HOTPOTQA CASE: FAVORABLE AND ADVERSE EVIDENCE DRAWS

We examine item `5ae1b2b9554299422ee99684` from the original HotpotQA (Yang et al., 2018) development set, using the official explorer sample.[1] The case compares two retrieval outcomes for the same question and how the returned evidence affects subsequent rollout decisions.

Question and evidence chain. *Who did President Franklin Roosevelt appoint that was responsible to transmit votes of the Electoral College to Congress?* The dataset answer is **Robert Digges Wimberly Connor**. The *National Archives and Records Administration* passage identifies the institution that transmits the votes (sentence 2); the *Robert Digges Wimberly Connor* passage identifies the first Archivist and Roosevelt's appointment (sentences 0 and 2; zero-based indices). Together they connect the responsibility to the appointed person.

Fixed query and candidate list. Use the question verbatim as the query at its first occurrence. For this case study, build a separate BM25 index (Robertson & Zaragoza, 2009) from the 975 title-deduplicated passages in the official 100-item sample: title plus all sentences, lowercase alphanumeric tokens, $k_1 = 1.5$, $b = 0.75$, and title-order tie breaking. Neither answers nor support labels enter ranking. Retain 20 candidates; always return ranks 1–5 and draw two distinct passages uniformly without replacement from ranks 6–20 (15 tail candidates), as in equation 24. The fixed top five are *United States presidential election*, *National Archives and Records Administration*, *United States presidential election, 1792*, *Roosevelt Franklin*, and *Harry S. Truman*. The critical appointment passage is at rank 9. The two draws below change one tail passage; rank 10 (*Emergency Banking Act*) is common to both.

Table 7: Two REF-H content draws for the same real question and fixed candidate list. Ranks refer to the case-study index. Both returns contain the same top five plus two tail passages.

Tail draw	Evidence exposed	Possible continuation
Favorable: $\{9, 10\}$	Connor's passage adds the appointment relation to the institution's responsibility already visible at rank 2.	Combining the two passages supports Connor as the answer without another retrieval call.
Adverse: $\{8, 10\}$	*James Farley* replaces Connor. Farley served as Postmaster General under Roosevelt; his passage does not establish the required Archives appointment.	The agent needs a follow-up query about Roosevelt's Archivist. Substituting Farley solely on the Roosevelt association yields a wrong answer.

Connection to sharing. At a matching query key, sharing C holds evidence availability fixed across the group, making reward comparisons less confounded by which trajectory obtained the appointment passage. Passage order and policy draws remain independent. The same coupling also makes the group share an omission: under the adverse draw, every trajectory must recover the missing relation through later actions. With one critical passage among 15 tail candidates, its single-draw inclusion probability is $2/15$ under either design. For eight trajectories issuing this same query, the probability that at least one sees it at this occurrence is $1 - (13/15)^8$ under independent draws, versus $2/15$ under sharing. These are sampling probabilities, not task-success rates. The case exposes both the comparability benefit and the coverage cost that motivate the gradient-risk criterion; it does not determine their net effect on training.

The companion `case_studies/hotpot_ref_h/` directory preserves the source snapshot, query, full ranked list, selected passages, a deterministic reproduction script and its verification log.

[1] Source: HotpotQA official development-sample JSON (official repository copy), accessed September 24, 2026 and rechecked September 26, 2026; dataset licensed under CC BY-SA 4.0. The companion directory includes the snapshot's SHA-256 and verification log.

B PROOFS, ASSUMPTIONS, AND COUNTEREXAMPLES

B.1 FINITE-GROUP RESIDUAL VARIANCE

Fix x, π_0 and the harness. Let W denote the shared variable: a constant for independent collection, C for partial blocking, and (C, U) for full blocking. Conditional on W, the returns are i.i.d. with mean m_W and variance v_W. Put $D_i = R_i - \bar{R}$. Then

$$\begin{aligned} \mathbb{E}[D_i \mid W] &= 0, \\ \mathrm{Var}(D_i \mid W) &= \left[(1-1/G)^2 + (G-1)/G^2\right] v_W = (1-1/G)v_W. \end{aligned} \tag{15}$$

Averaging over W gives $(1-1/G)\mathbb{E}v_W$; since $\mathbb{E}[D_i \mid W] = 0$, this is also the unconditional variance. Iterated total variance gives $\mathbb{E}v_W = S$, $\sigma_U^2 + \sigma_\pi^2$, or σ_π^2 respectively, proving equation 3. In particular, $\mathrm{Var}(D_i \mid C, U) = (1-1/G)\mathrm{Var}(R \mid C, U)$, not $(1-1/G)\sigma_\pi^2$ before averaging. Division proves equation 4 whenever its denominator is positive.

B.2 EXACT RAW-LOO GRADIENT RISK

Fix a prompt, policy, decoding rule and harness. Rewards are parameter-independent functions of the trajectory and tape. Scores include every stochastic policy action, including termination; there is no random trajectory-length normalization. Bounded horizon and the score regularity stated in §3.4 justify differentiation of the conditional trajectory law. Thus $\mathbb{E}[s \mid C] = 0$ and g_x is the conditional policy gradient. Full blocking replaces C by (C, U).

For one group put $X_i = R_i - m_C$. Its raw LOO gradient is the sample cross-covariance

$$Y_C = \frac{1}{G}\sum_i s_i(R_i - \bar{R}_{-i}) = \binom{G}{2}^{-1} \sum_{i<j} h_{ij}, \qquad h_{ij} = \tfrac{1}{2}(s_i - s_j)(X_i - X_j). \tag{16}$$

This representation uses the Hoeffding decomposition (Hoeffding, 1948); see Zhou et al. (2026) for related group-relative gradient analysis. Conditional on C, the kernel mean is g_C, its first projection is $h_1(s, X) = (sX - g_C)/2$, and its degenerate second projection is $h_2(1, 2) = -(s_1X_2 + s_2X_1)/2$. Let $v(C)$ and $u(C)$ denote the expressions inside the outer expectations defining v_C, u_C. Independence gives

$$\mathbb{E}[\|h_1\|^2 \mid C] = \frac{v(C)}{4}, \qquad \mathbb{E}[\|h_2\|^2 \mid C] = \frac{u(C)}{2}, \tag{17}$$

since $\mathbb{E}[\langle s_1, s_2\rangle X_1X_2 \mid C] = \|g_C\|^2$. Conditional projection orthogonality yields

$$\mathbb{E}[\|Y_C - g_C\|^2 \mid C] = \frac{v(C)}{G} + \frac{u(C)}{G(G-1)}. \tag{18}$$

Total variance adds β_C; averaging P independent groups divides the result by P and proves equation 6. For independent collection, take the shared variable to be constant and independently resample each member's environment from the original marginal. Its $\beta_0 = 0$, which proves equation 7 at common N, G.

Random tasks and estimator scope. When each group samples its own task x, risk relative to the global gradient additionally contains $\mathrm{tr}\,\mathrm{Cov}_x(g_x)/P$, and all fixed-prompt coefficients are averaged over x. This extra term cancels only for comparisons with the same prompt law and P. It must be retained when comparing budgets or group sizes. The estimator scope is specified in §4.3; unbounded stopping additionally requires score and gradient integrability.

B.3 MATCHED REWARD LAWS WITH OPPOSITE SHARING EFFECTS

For Proposition 1, consider the broader family

$$R = \tfrac{1}{2} + 0.1C + bA + eAC, \qquad b^2 + e^2 = 0.04, \quad b, e \geq 0, \tag{19}$$

with C a fair sign independent of $A \sim e^{\theta A}/(2\cosh\theta)$, evaluated at $\theta = 0$. Then $s = A$ and $R \in [0, 1]$, since $0.1 + b + e \leq 0.1 + \sqrt{0.08} < 0.5$. Every member satisfies $m_C = 0.5 + 0.1C$,

$S = 0.05$, and $(\sigma_C^2, \sigma_U^2, \sigma_\pi^2) = (0.01, 0, 0.04)$. Consequently, blocking reduces the variance of D_i by 20% for all $G \geq 2$.

The two endpoints $(b, e) = (0.2, 0)$ and $(0, 0.2)$ also have the same joint (C, R) distribution: given either C, the two reward values differ only by swapping the action labels. Their marginal rewards are $0.2, 0.4, 0.6, 0.8$, each with probability $1/4$. The intervening family shares the stated reward moments while its reward distributions vary. Direct evaluation gives

$$g_C = b + eC, \quad g_x = b, \quad \beta_C = e^2, \quad v_C = 0, \quad u_C = 0.08,$$
$$v_0 = 0.01 + e^2, \qquad u_0 = 0.09 - e^2. \tag{20}$$

At $G = 2$, therefore, $N\mathcal{R}_0 = 0.10$ and $N\mathcal{R}_C = 0.08 + 2e^2$. The sign changes at $e^2 = 0.01$ even though $\eta_C = 0.2$ throughout. At this boundary $g_x = \sqrt{0.03} > 0$; the harmful case $(b, e) = (0.1, \sqrt{0.03})$ also has nonzero gradient and $N\mathcal{R}_C = 0.14$. The endpoint risks $0.08/N$ and $0.16/N$ prove Proposition 1; the common centred reward variance is 0.025 before sharing and 0.020 after sharing at $G = 2$.

B.4 Perturbation from raw LOO to implemented gradients

Fix the policy, prompt law, trainable parameters, gradient metric and N, G in both designs. All risks below use the same raw reward-gradient target g. Theorem 2 provides $\mathcal{R}_d$; random-prompt risks include the common prompt-heterogeneity term from Appendix B.2. Gradients with normalization, length weights or filtering are treated as perturbations with finite second moments, without assuming they are unbiased.

Proof of Proposition 2. Writing $\widetilde{Y}_d - g = (Y_d - g) + E_d$ and expanding the square gives the exact identity

$$\widetilde{\mathcal{R}}_d - \mathcal{R}_d = 2\mathbb{E}\langle Y_d - g, E_d\rangle + \mathbb{E}\|E_d\|^2 =: T_d. \tag{21}$$

Cauchy–Schwarz bounds the cross term by $|\mathbb{E}\langle Y_d - g, E_d\rangle| \leq \sqrt{\mathcal{R}_d}\,\epsilon_d$. Applying the triangle inequality to $T_C - T_0$ proves equation 13. If $\Delta < -\Gamma$, then $\widetilde{\mathcal{R}}_C - \widetilde{\mathcal{R}}_0 < 0$; if $\Delta > \Gamma$, the difference is positive. Equality or $|\Delta| \leq \Gamma$ gives no sign conclusion. No independence between Y_d and E_d is used, so random-normalizer dependence is retained. Deterministic rescaling can be handled by replacing Y_d, g by aY_d, ag for one common, independently fixed $a > 0$; the experimental audit uses $a = 1$ to retain the existing target.

Exact group-standardization correction. Let $y_{d,j} = G^{-1}\sum_i s_{j,i}(R_{j,i} - \bar{R}_{j,-i})$ be the raw gradient of group j. For normalization alone, with complete scores and unchanged trajectory weights, define the implemented detached scalar $c_{d,j}$. For advantages $(R_{j,i} - \bar{R}_j)/(\hat{\sigma}_{d,j} + \varepsilon)$,

$$Y_d = \frac{1}{P}\sum_j y_{d,j}, \qquad c_{d,j} = \frac{1 - 1/G}{\hat{\sigma}_{d,j} + \varepsilon}, \qquad E_d = \frac{1}{P}\sum_j e_{d,j}, \quad e_{d,j} = (c_{d,j} - 1)y_{d,j}. \tag{22}$$

Other stabilizer conventions use their actual $c_{d,j}$ and recorded zero-variance rule. Assume the P groups are i.i.d. and $\mathbb{E}\|e_{d,1}\|^2 < \infty$. Let $m_d = \mathbb{E}e_{d,1}$, $q_d = \operatorname{tr}\operatorname{Cov}(e_{d,1})$ and $k_d = \operatorname{tr}\operatorname{Cov}(y_{d,1}, e_{d,1})$. Since $\mathbb{E}y_{d,1} = g$,

$$\epsilon_d^2 = \frac{q_d}{P} + \|m_d\|^2, \qquad T_d = \frac{2k_d + q_d}{P} + \|m_d\|^2. \tag{23}$$

To verify this, independent groups give $\operatorname{tr}\operatorname{Cov}(Y_d, E_d) = k_d/P$ and $\mathbb{E}\|E_d\|^2 = q_d/P + \|m_d\|^2$; substitute in equation 21. Thus, for fixed prompts, the normalized risk is the exact raw expression $[G\beta_d + v_d + u_d/(G-1)]/N$ plus T_d. The detached normalizer contributes both gradient covariance and squared bias. Its mean is $m_d = \mathbb{E}[(c_{d,1} - \mathbb{E}c_{d,1})(y_{d,1} - g)] + (\mathbb{E}c_{d,1} - 1)g$. A conservative bound follows from Jensen: $\epsilon_d^2 \leq \mathbb{E}[(c_{d,1} - 1)^2\|y_{d,1}\|^2]$. For native length weights or other transformations, use their actual full-batch residual E_d; the general proposition still applies, while this group-scalar identity need not.

Local clipping equivalence at the reference policy. Fix a finite batch, old policy and advantages; assume differentiable probabilities on positive, parameter-independent support. At θ_0, every likelihood ratio equals one and lies strictly inside any positive-width PPO clipping interval. Continuity and batch finiteness give a common neighbourhood in which clipped and unclipped branches coincide, regardless of advantage sign. Their policy-surrogate gradients therefore agree. With the weights in equation 12, $\lambda = 0$ and complete scores, this gives Y for the policy surrogate; auxiliary losses are handled separately. Later-epoch clipping is audited in Appendix D.3.

C CORE EXPERIMENTAL SETTINGS AND SHARING PROTOCOLS

Model and splits. The core model is `Qwen/Qwen3.5-35B-A3B-Base` (Qwen Team, 2026a) (35B total/3B active language-model parameters, pre-trained, Apache-2.0). All arms use the same initialization, tokenizer and text serialization; visual parameters are frozen. Training samples a 50:50 mixture of HotpotQA train and ALFWorld train games with replacement. Within each family, reserve disjoint 10% subsets for model selection and calibration evaluation, split by source article/game identity. HotpotQA dev and ALFWorld unseen games are held out. Seven estimator bases and six adapters use five seeds each; sharing-only and raw-LOO ablations are additional runs. The benchmark sweep reuses each run's selected checkpoint.

Evaluation. HotpotQA uses normalized final-answer exact match and ALFWorld binary goal completion. Transfer evaluations use the same checkpoints without transfer-test tuning: WebShop's 500 test instructions, full catalog and text interface (success $r = 1$); WebArena's 812 tasks, reset local website snapshots and textual accessibility trees; original τ-bench v1's 115 retail/50 airline tasks, single-attempt state/output success, 30-action limit and agent/user temperatures 0/1; and BFCL v3 multi-turn's four 200-conversation categories, official state/response checks and 20-step guard. τ domains and BFCL categories are macro-averaged within each seed. Official task evaluators, parser/judge settings and action limits are fixed across methods. MATH-500 and AIME'25 use fixed answer checkers.[2] Evaluation costs, including user/judge calls, are separate from training costs.

Budgets and selection. Select among eight configurations per base using equal train-validation tuning compute. The actor-token arm caps actor policy tokens at 10^8. The separate total-cost arm uses a cap locked by a GRPO pilot reaching that token budget. Count actor updates, discarded samples, critic fitting, calibration/reference batches, environment/reset work and failed gates. Measure resource use and convert separately billed resources to GPU-h equivalents using the accounting rules in Appendix E.3. Use the latest completed snapshot within the cap and record unused budget or overshoot. No test score enters selection. Base-specific objective settings are retained (Table 10). The actor uses AdamW (Loshchilov & Hutter, 2019).

Statistical units. Average attempts within each task, then tasks within each run; report means and standard deviations across five training seeds. Four evaluation attempts per task (64 for AIME'25) estimate single-attempt success. Hierarchical paired bootstrap uses 10,000 replicates, resampling the five seed pairs and task IDs while preserving nested environment/policy draws and domain/category structure. A seeds-only bootstrap conditions on the fixed test set. The primary contrast is GRPO+ESRA minus GRPO on the equal-weight HotpotQA/ALFWorld mean. Holm correction covers the five secondary total-cost adapter contrasts; other paired intervals are descriptive. The MATH-500 non-inferiority margin is -0.5 percentage points, tested by the one-sided 95% lower bound.

C.1 SHARED INPUTS, SAMPLING AND RESET

Controlled sharing protocols. HotpotQA and ALFWorld use REF-H and REF-A, respectively, following common-random-input policy comparisons (Ng & Jordan, 2000). Independent and shared

[2]Public AIME'25 problem and solution references: 2025 AIME I and 2025 AIME II, hosted by LIVE by Po-Shen Loh with MAA permission. These identify contest content; the archived run records do not identify the evaluated I/II selection or data-file revision.

Table 8: Core training and audit settings. Base-specific objective settings are retained.

Setting	Default	Comparison settings
Training seeds	$0, 1, 2, 3, 4$	paired across core arms
Actor prompts / group size	$P = 8, G = 8$	$G \in \{4, 8, 16\}$ in the risk grid
Optimizer / actor learning rate	AdamW / 10^{-6}	$\{0.5, 1, 2\} \times 10^{-6}$
Action tokens / full context	8K / 32K	same limits for each A/B pair
Training turn limit / discount	16 / 1	evaluation limits are benchmark-specific
Actor-token cap, rollout arm	10^8 per run	actor policy tokens only
Total-cost cap, total arm	pilot-locked	GPU-h equivalents; all stages charged
Value heads / head learning rate	two-layer MLP / 10^{-4}	frozen policy trunk during head fit
Pilot / confirmation repeats	8 / 16	pilot $\{4, 8, 16\}$
Reference MC batch size	$2PG$	twice candidate batch
Root anchor weight	1	$\{0, 0.1, 1\}$
Calibration refresh	every policy snapshot	calibrated extension only
Evaluation attempts per task	4; AIME 64	non-AIME $\{4, 8, 16\}$ precision study
Bootstrap replicates	10,000	hierarchical paired resampling

arms have the same environment marginals. Main contrasts use partial sharing; labelled mechanism conditions also test full sharing (Appendix C.2).

REF-H: retrieval-content selection. Fix the question, corpus/index revision, deterministic retriever, query normalization, passage truncation, tool parser and answer checker in x. Resolve tied retrieval scores by immutable passage ID and retain at most 20 unique candidates. For a normalized query q, let L_q be this ordered list, A_q its first $\min(5, |L_q|)$ passages, and $T_q = L_q \setminus A_q$ its tail. At occurrence k of q within a trajectory, draw

$$B_{q,k} \sim \mathrm{Unif}\{B \subseteq T_q : |B| = r_q\}, \qquad r_q = \min(2, |T_q|), \qquad S_{q,k} = A_q \cup B_{q,k}. \tag{24}$$

The shared tape C supplies these subset draws, independently across distinct (q, k) keys. Conditional on the selected set, U supplies an independent uniform permutation of $S_{q,k}$. An empty candidate list returns an empty result; when the tail has at most two passages it is retained in full. Answer labels and support annotations do not enter selection. All arms use the same passage text, truncation, return-count rule and tool-call limits.

Replay keys contain protocol version, task ID, normalized query, candidate-list digest and the trajectory-local occurrence k, starting at one. Query normalization matches the fixed retriever. Under partial sharing, matching keys reuse content selection with independent ordering; distinct queries retain their own candidates. Each trajectory maintains its own counters.

REF-A: object-instance naming. Fix the game file, physical initial state, physical goal, transition engine, action interface and success checker in x. For every object or receptacle class t with m_t instances, list the physical IDs in canonical order and draw independently

$$\sigma_t \sim \mathrm{Unif}(\mathcal{S}_{m_t}), \qquad C = (\sigma_t)_t, \tag{25}$$

where $\mathcal{S}_m$ is the set of permutations of m indices. A class with one instance has the identity mapping. Apply each map consistently to the initial description, goals, observations and any displayed action candidates; invert it before executing an action. The map persists for the whole episode, including objects first observed later. Goal satisfaction, object locations and affordances are evaluated on physical state. This makes the intervention an invertible change of presentation, motivated by ALFWorld's optional object-name randomization (ALFWorld Contributors, 2026).

The tape U independently permutes each explicitly unordered object/receptacle listing. Temporal sequences, action arguments and other semantically ordered text retain their order. A lookup key contains the canonical physical state, canonical action (a distinguished reset action initially), listing-field ID and that event's trajectory-local occurrence count. Construct keys before alias rendering. Draw a uniform permutation of the physical-ID-sorted list, then render aliases. Empty or singleton lists are unchanged. The wrapper randomizes presentation while preserving the underlying physical state and transition rules.

Sampling, reset and replay contract. Tapes are sampled independently of policy parameters and policy RNGs before interaction. Lazy evaluation samples each unvisited key once with independent base draws and replays it thereafter. Uniform sampling can use partial/full Fisher–Yates shuffles and unbiased bounded-integer draws. Content, order and policy streams have separate namespaces; rewards, generated actions and other members' RNG states never seed them. Action-dependent keys select tape coordinates.

For update b, group g and member i, allocate tape identities as follows. Prefix every identity by the run/stage namespace and protocol version; policy streams always include (b, g, i).

Mode	Content/alias tape identity	Order tape identity
Independent	(b, g, i)	(b, g, i)
Partial sharing	(b, g)	(b, g, i)
Full sharing	(b, g)	(b, g)

Different identities and namespaces have independent streams, even for equal index tuples. Refresh tapes per group/update; diagnostic, audit, calibration and actor namespaces are disjoint. Task schedules and group-membership rules are matched across arms.

Every member receives its own environment instance, initial-state copy, conversation history, inventory, query/event counters and wrapper state. Only immutable inputs and keyed tape values are shared. Apply aliases before producing the reset observation. Replaying a fixed tape and action sequence must reproduce the same observations, transitions and reward regardless of other members' paths or worker scheduling. Invalid actions, empty retrievals and episode limits use a common deterministic rule and remain in the rollout record. An infrastructure failure is recorded as a failed attempt with its cost; retries reuse the same tape and policy-stream state under a fixed retry limit.

Connection to the risk assumptions. Write $\tau_i = F_x(C_i, U_i, \xi_i)$ for independent policy RNG ξ_i and common F_x across arms. Reusing tape identities preserves the subset/alias and conditional order laws, proving equation 9. Given C, the pairs (U_i, ξ_i) are i.i.d., establishing conditional independence; full sharing conditions on (C, U). The frozen policy, task and score regularity remain those of Theorem 2.

Implementation acceptance checks. Deterministic checks have zero mismatch tolerance: replay equality for a fixed tape/action script; unchanged traces when another member's actions or execution order change; identical physical reset states in REF-A; and exact alias/inverse round trips preserving physical transitions and reward. Stream isolation excludes a mutable global RNG. Disabling both wrappers and varying an unused shared seed must leave fixed-policy-stream traces unchanged.

The sampler-check specification uses a fixed suite of retrieval lists and game-state/listing fixtures, with 50,000 independent repetitions per fixture and mode. The targets are each tail passage's inclusion probability $r_q/|T_q|$, each physical instance's probability $1/m_t$ of receiving each alias, and each item-position probability $1/\ell$ in a list of length ℓ, conditional on its contents. The acceptance rule uses two-sided Clopper–Pearson intervals (Clopper & Pearson, 1934) with Bonferroni family coverage of 95% over the predeclared checks. The reference acceptance tolerance is 0.01 in absolute probability: each interval must lie within the target probability plus or minus that tolerance. Deterministic cases are checked exactly.

C.2 Result-to-protocol mapping

Table 9 links results to the REF-H/REF-A laws in Appendix C.1. Native GRPO uses its group-normalized objective within these harnesses. Every contrast matches task law, policy initialization and trajectory marginals.

Table 9: Core result-to-protocol assignments. H/A identify task-specific result families under REF-H/REF-A; I/P/F denote independent/partial/full sampling.

Result family	Results	Protocol modes and matched contrast
H-RISK / A-RISK	Tables 1, 3, 12, 11; Figure 2	P versus I and F versus I. Match task, seed, checkpoint, N, G and mode across prediction, audit and continuation; stage draws are disjoint.
H-FACT / A-FACT	Table 2	P versus I. Frozen-policy/head audit at matched N, G; warm-start continuation at matched total cost. The zero-effect control shares only an unused constant.
H-GRPO / A-GRPO	Tables 5, 6; Appendix E.3	Full training at matched total cost: GRPO uses I; GRPO+B and GRPO+ESRA use P with common group/reset rules. Only the latter adds values/calibration.
H-RAW / A-RAW	Table 13; raw-LOO contrast in Table 6	Full training at matched total cost: P for MC, learned-only and both mixtures. Independent calibration uses I with the same calibration rule.
H-TOKEN / A-TOKEN	Actor-token contrasts in Table 6	Separate full-training arm at 10^8 actor tokens; modes as above, with configuration-specific total costs.

C.3 Native-objective adapters

The bases are GRPO (Shao et al., 2024), RLOO (Ahmadian et al., 2024), GSPO (Zheng et al., 2025), DAPO (Yu et al., 2025), GiGPO (Feng et al., 2025) and turn-PPO (Li et al., 2026), with PPO (Schulman et al., 2017) as a reference. The PPO and turn-PPO branches use generalized advantage estimation (GAE) (Schulman et al., 2016). For a base $\mathcal{M}$, retain its advantage $A_{\mathcal{M}}$ and form $(1-\lambda)A_{\mathcal{M}} + \lambda\mathcal{P}_{\mathcal{M}}(A^{\mathrm{v}})$. The map $\mathcal{P}_{\mathcal{M}}$ broadcasts turn values to policy tokens for token losses, takes their policy-token-weighted mean for sequence losses, and retains turn resolution for turn losses. Apply the base's frozen positive reward scale and observation-token mask. Table 10 lists the retained objective components. With $\lambda = 0$ and sharing off, this recovers the base; sharing alone changes collection while preserving its objective. Calibration uses actual branch gradients at π_0 and the raw reference in Appendix E.1, including candidate bias. Each base retains its clipping, normalization and multi-epoch rules. Reward-shaping variants state their training utility; external success metrics remain fixed.

Table 10: Native-adapter definitions and retained objective components. Each base and its adapted counterpart use the same settings for these components.

Base	Base advantage / resolution	Retained settings
GRPO	group-standardized outcome / tokens	token ratio and clipping
RLOO	leave-one-out / sequence or tokens	reference implementation ratio
GSPO	group-standardized outcome / sequence	sequence ratio and clipping
DAPO	implemented outcome / tokens	dynamic sampling, clipping, penalties
GiGPO	episode + anchor signal / steps	anchor construction and weights
turn-PPO	implemented GAE / turns	turn ratio and GAE settings
PPO reference	learned-baseline GAE	independent comparator, no adapter
Reference ESRA	raw LOO + causal value / turns	equation 12, one update per snapshot

With sharing off and $\lambda = 0$, deterministic FP32 endpoint checks use the tolerance $10^{-6}\max(1, \|\mathrm{base}\|_\infty)$ for advantages and pre-optimizer gradients. Across the raw reference and six adapters, the reported maximum residuals are 3×10^{-7} and 6×10^{-7}, respectively.

D Core mechanism and learning protocols

All experiments use the splits, paired seeds, evaluation IDs, hierarchical bootstrap and cost accounting in Appendix C, with the sampling/reset rules in Appendix C.1. Below we specify experiment-specific controls.

D.1 INDEPENDENT RISK PREDICTION AND AUDIT

Conditions and independent splits. Use the core model, five training seeds, early/late checkpoints fixed by training-cost fractions, $N = 64$ trajectories and $G \in \{4, 8, 16\}$. Each checkpoint fixes a uniform law over 16 diagnostic task IDs; batches independently sample prompts from that law. Partial/full versus independent sharing gives 60 conditions per task under REF-H/REF-A. Compare designs at the same N, G, policy and prompt law. Diagnostics and audits have disjoint environment and policy draws.

Prediction data. For each fixed prompt draw $K = 8$ independent shared-input values. At each value collect three conditionally independent cohorts a, b, c, each initially containing $n = 8$ complete trajectories. Use disjoint environment/policy randomness for the held-out audit. Denote cohort reward means by $\bar{R}_j^a, \bar{R}_j^b$, means of sR by $\bar{g}_j^a, \bar{g}_j^b$, the mean of $\|s\|^2$ in cohort a by q_j^a, and unbiased reward sample variance in cohort b by $S_{R,j}^{2,b}$. Independent-cohort products avoid squaring a noisy estimate of g_C:

$$\widehat{\beta}_C = \frac{1}{K}\sum_j \langle \bar{g}_j^a, \bar{g}_j^b \rangle - \frac{1}{K(K-1)}\sum_{j\neq k} \langle \bar{g}_j^a, \bar{g}_k^b \rangle, \tag{26}$$

$$\widehat{v}_C = \frac{1}{K}\sum_j \left\{ \frac{1}{n}\sum_{i\in c_j} \|s_i\|^2 (R_i - \bar{R}_j^a)(R_i - \bar{R}_j^b) - \langle \bar{g}_j^a, \bar{g}_j^b \rangle \right\}, \tag{27}$$

$$\widehat{u}_C = \frac{1}{K}\sum_j \left\{ q_j^a S_{R,j}^{2,b} + \langle \bar{g}_j^a, \bar{g}_j^b \rangle \right\}. \tag{28}$$

These estimators target the moments in equation 5 in expectation under the declared conditional sampling law. Do not clamp negative estimates before computing the risk contrast. For v_0, u_0, collect three independent cohorts from the original marginal environment law, resampling the environment for each trajectory, and use the same cohort formulas; $\beta_0 = 0$ by definition. Full sharing replaces C by (C, U). Average fixed-prompt estimates over the declared prompt law before applying equation 14. Finite cohorts affect precision; repeat the entire analysis at doubled K or n on a subset selected before examining outcomes.

Held-out audit. Collect independent pairs of complete raw-gradient batches $X_{d,k}, X'_{d,k}$ for each design d. Each batch has N trajectories and targets the same global gradient for the declared prompt law. Use

$$\widehat{\mathcal{R}}_d = \frac{1}{2L}\sum_{k=1}^{L} \|X_{d,k} - X'_{d,k}\|^2, \qquad \widehat{\Delta}_d^{\text{audit}} = \widehat{\mathcal{R}}_d - \widehat{\mathcal{R}}_0. \tag{29}$$

Initialize $L = 16$ independent pairs. A validation-only precision pilot fixes the final repeat count from $\{16, 32, 64\}$ before any held-out signs are examined. Predictions and audit outcomes are locked separately. Intervals strictly below/above zero classify benefit/harm at the 95% level; zero-crossing intervals are unresolved. Unbiased raw gradients use this reference-free audit; learned or normalized candidates use the independent two-reference audit in equation 30. Full-gradient dot products are accumulated in parameter blocks.

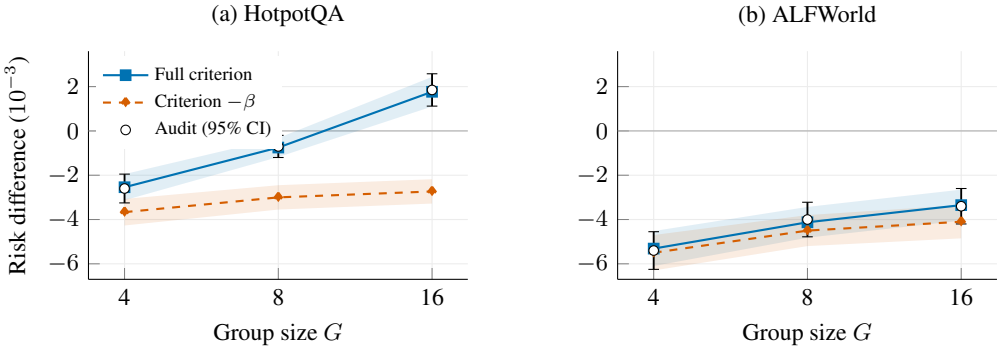
Absolute-risk audit for biased candidates. For an update gradient X_k, draw unbiased references H_k, H'_k independent of each other and of X_k at the same frozen policy and prompt law. Then

$$\widehat{\mathcal{R}}(X) = \frac{1}{K}\sum_{k=1}^{K} \langle X_k - H_k,\, X_k - H'_k \rangle \tag{30}$$

has expectation $\mathbb{E}\|X - g\|^2$, since the independent reference errors have zero cross expectation. Reuse the reference pair across candidates for paired differences, independently of weight fitting and confirmation. Report signed estimates with bootstrap intervals and charge both reference batches to diagnostics.

Comparators and metrics. The full criterion uses equation 14; the $-\beta$ control removes $G\widehat{\beta}_C/N$ from the same moment estimates and recomputes intervals within each bootstrap draw. The reward-only predictor uses reward summaries, η_C and G; features, capacity, threshold and magnitude calibration are fitted on separate validation conditions, holding out entire checkpoints. No audit sign enters fitting or condition selection. For each attempted condition, 95% intervals determine benefit, harm or unresolved. Coverage is the resolved-prediction fraction; agreement counts matching resolved signs over all attempts; missed harm counts predicted benefit among audit-confirmed harms. Gap MAE measures predicted versus audited risk differences. A zero denominator is reported as n/a. Table 11 retains all primary sign counts, including errors and unresolved cases. Replay and marginal-sampler checks follow Appendix C.1.

Group-size boundary. At fixed $N = 64$, reuse each seed's early policy and one validation-fixed partial-sharing law across $G \in \{4, 8, 16\}$. Reuse the diagnostic moment estimates, but match each audit's independent comparator in G and prompt law. Figure 2 combines five-seed means and intervals, including a sign crossing and a no-crossing case. Seed-averaged $(\beta, \Delta v, \Delta u)$ are $(0.018, -0.160, -0.224)$ for HotpotQA and $(0.003, -0.240, -0.336)$ for ALFWorld. Conditional reward-variance gaps $-0.032/-0.048$ become centered gaps after multiplying by $(G-1)/G$.



$\Delta\mathrm{Var}(D)$	$G = 4$	$G = 8$	$G = 16$
HotpotQA	$-0.024\,[-0.030, -0.018]$	$-0.028\,[-0.035, -0.021]$	$-0.030\,[-0.038, -0.022]$
ALFWorld	$-0.036\,[-0.045, -0.027]$	$-0.042\,[-0.052, -0.032]$	$-0.045\,[-0.056, -0.034]$

Figure 2: Risk boundary at fixed $N = 64$: shared minus independent raw-LOO risk, in 10^{-3} units. Policy and partial-sharing law are fixed within each seed; points average five seeds. Shaded prediction envelopes and audit bars give paired 95% intervals at tested group sizes; connecting lines guide the eye. Below: centered reward-variance gaps in squared reward units, with 95% intervals. Full prediction tracks the HotpotQA sign change despite falling reward variance; the $-\beta$ control misses it.

Complete prediction outcomes. Table 11 gives 97/120 confirmed full-criterion signs versus 64/120 for reward-only, using the agreement definition in Appendix D.1. The full criterion predicts benefit for two audited harms and harm for two audited benefits; these errors remain in the original strata. The equal-task continuation differences for benefit, harm and unresolved strata are $+5.70\,[4.6, 6.8]$, $-2.40\,[-3.3, -1.5]$ and $+0.05\,[-0.8, 0.9]$. The condition-count-weighted raw-risk differences within tasks are $-1.9483/-3.0517$ in 10^{-3} units, averaging to -2.50 in Table 12.

Table 11: Complete primary sign matrices (60 conditions per task). Each tuple lists audited benefit/harm/unresolved counts; columns fix the predicted sign.

Task	Predictor	Predicted benefit	Predicted harm	Predicted unresolved
HotpotQA	Full criterion	(35,1,1)	(1,14,0)	(2,1,5)
	Reward-only	(31,12,4)	(2,2,0)	(5,2,2)
	Criterion $-\beta$	(36,9,3)	(0,6,0)	(2,1,3)
ALFWorld	Full criterion	(30,1,0)	(1,18,1)	(2,2,5)
	Reward-only	(29,16,3)	(1,2,0)	(3,3,3)
	Criterion $-\beta$	(31,11,2)	(0,7,0)	(2,3,4)

D.2 Sharing and value-estimation attribution

The six cells in Table 2 cross independent/partial-sharing collection with raw MC, validation-fixed mixing and learned-only estimation. The frozen audit fixes the policy, N, G, prompt law, trainable parameters, independently fitted value heads and mixture weight across designs. Common independent reference pairs give paired risks including value bias via equation 30. The zero-effect control replaces the shared input by an unused constant while sampling effective environment and policy randomness independently; its population risk gap is zero, while the table reports the measured estimate and interval.

Fixed-head learned-only consistency check. At fixed x, π_0, Q_ψ, V_ϕ, define the per-trajectory vector $\ell_i = \sum_t s_{i,t}[Q_\psi(h_{i,t-1}, a_{i,t}) - V_\phi(h_{i,t-1})]$ and average it over $N = PG$ trajectories. Assume finite second moments, matched trajectory marginals and conditional independence within a shared-C group. Put $b_C^{\mathrm{v}} = \operatorname{tr} \operatorname{Cov}_C(\mathbb{E}[\ell_i \mid x, C])$. Then, relative to the same target gradient,

$$\mathcal{R}_C^{\mathrm{v}} - \mathcal{R}_0^{\mathrm{v}} = \frac{G-1}{N} b_C^{\mathrm{v}} \geq 0. \tag{31}$$

To see this, write $w = \mathbb{E}_C \operatorname{tr} \operatorname{Cov}(\ell_i \mid x, C)$. The shared average has variance $(w + G b_C^{\mathrm{v}})/N$, while the independent average has variance $(w + b_C^{\mathrm{v}})/N$. The common marginal mean gives identical squared bias, which cancels. For random prompts, average over the same prompt law and match the number of groups so that prompt heterogeneity also cancels. Full sharing replaces C by (C, U).

The learned-only row checks equation 31 using paired intervals under matched marginals, common frozen heads and independent references. Raw LOO contains a group-dependent return baseline and follows Theorem 2.

For continuation, clone policy and optimizer states and match initial heads, head architecture, fitting rules and tuning budgets. Subsequent head fits follow each evolving policy on independent data; MC spends saved auxiliary budget on actor updates within the common total cap. Report $\Delta_e = T_{S,e} - T_{I,e}$ and $I = \Delta_{\mathrm{mixture}} - \Delta_{\mathrm{MC}}$ with paired differences-of-differences intervals. Risk and success are averaged equally over tasks; their desirable signs are opposite.

D.3 From raw risk to native GRPO learning

Prediction, audit and continuation match task law, seed, checkpoint, N, G and sharing mode, with disjoint draws at each stage. Diagnostic intervals lock strata before the joint reward-variance and raw-risk audit; Table 3 retains errors and unresolved conditions.

Frozen objectives. On the same audit trajectories compare complete-score raw LOO, normalization alone with otherwise unchanged weights, and the implemented native GRPO gradient. Trainable parameters and the raw reward-gradient target are common. Raw LOO uses equation 29; biased/normalized candidates use the two-reference audit in equation 30. At the reference policy, clipping is locally inactive; length weights, masks and group filters remain those of the native objective.

Continuation and full training. For each primary condition, clone the policy and complete optimizer state for GRPO and GRPO+B. Only environmental sharing changes: native advantages,

learning-rate schedule, clipping, masks, decoding and evaluation IDs are matched. A validation pilot locks a total cap covering diagnostics plus twenty baseline updates. Charge diagnostic allowances equally and include all reset and training costs; realized update counts may differ. Within-condition differences give task-specific stratum means and their equal-task mean under the shared inference protocol. Empty strata are n/a; prediction errors, risk increases and failures of lower risk to improve task performance remain in the analysis.

Table 12 also reports separate fixed-update probes: one, two or four within-batch PPO epochs, and twenty updates with fresh collections, evaluated on fresh paired trajectories. Matched-cost stratum results appear in Table 3. Full training separately compares GRPO, GRPO+B and GRPO+ESRA from the same initializer, task mixture and total-cost cap. Shared arms have identical group/reset rules; the full-minus-B contrast adds value estimation/calibration. Raw-LOO results are reported separately.

Table 12: Transfer of the diagnostic. Warm-start suite; frozen risk changes are in 10^{-3} units, reward changes in percentage points, and brackets are 95% intervals. Sign agreement counts all 120 primary conditions. Inapplicable quantities are n/a.

Objective/update	Δ risk	Agreement	Δ reward [95%]	KL (S/I)	Clip (S/I)
Raw LOO, frozen	-2.50 $[-3.10, -1.90]$	97/120	n/a	n/a	n/a
Normalization only	-1.90 $[-2.50, -1.30]$	95/120	n/a	n/a	n/a
Native GRPO, frozen	-1.60 $[-2.20, -1.00]$	92/120	n/a	n/a	n/a
Native GRPO, 1 epoch	n/a	n/a	$+0.9$ $[0.3, 1.5]$	.008/.009	.06/.07
Native GRPO, 2 epochs	n/a	n/a	$+1.5$ $[0.7, 2.3]$	.014/.016	.10/.12
Native GRPO, 4 epochs	n/a	n/a	$+1.9$ $[0.9, 2.9]$	.024/.028	.17/.20
Native GRPO, 20 updates	n/a	n/a	$+3.0$ $[1.8, 4.2]$	.038/.045	.21/.25

D.4 Dense/MoE audit and perturbation attribution

Models and frozen comparison grid. The core MoE is paired with dense Qwen3.5-9B-Base (Qwen Team, 2026b), Qwen3.5-27B (Qwen Team, 2026c) and Qwen3.5-4B (Qwen Team, 2026d) from the official `Qwen` repositories. 35B/9B use Base weights; 27B/4B use post-trained weights, jointly varying size, architecture and starting stage within Qwen3.5. All runs are text-only with frozen visual parameters and use the 120-condition grid in Appendix D.1. Within each contrast, freeze checkpoint, trainable parameters, tokenizer/template, decoding, task IDs and reward rule. Splits and protocols are common across models; gradient coordinates and references are model-specific. The native-gradient margin coverage in Table 4 uses 96, 90, 90 and 84 of 120 conditions for 35B, 27B, 9B and 4B.

Paired objectives and independent stages. Reuse the diagnostic/audit separation in Appendix D.1. On each fresh complete batch, compute raw LOO Y, normalization-only Z^{norm} with the complete scores/weights otherwise unchanged, and the actual frozen native gradient Z^{native}. Record reward-std convention, stabilizer, zero-variance handling, token masks, length weights and group filters. Set $E^{\mathrm{norm}} = Z^{\mathrm{norm}} - Y$ and $E^{\mathrm{native}} = Z^{\mathrm{native}} - Y$ with no per-arm rescaling. The two-reference audit uses the common raw reward target g within each model; the mean squared residual estimates ϵ_d^2 without an extra reference. For the signed correction, independent H gives $\mathbb{E}[2\langle Y - H, E\rangle + \|E\|^2] = T_d$. Estimate this correction on independent held-out repeats. Stream gradient dot products; all replays, backward passes and references count toward audit cost.

Attribution and reporting. Attribution uses 3×3 audited-sign transitions (benefit/harm/unresolved) and both directions of prediction-correctness changes; 97/120 to 92/120 is a net loss of five agreements. Summaries include validation-fixed raw-margin/reward-std strata, ϵ_d^2 and signed $T_C - T_0$. Margin coverage is the descriptive fraction satisfying the plug-in condition

in equation 13. Retain all predeclared conditions, including inconclusive margins, failures and empty/zero-variance groups, under the common paired inference protocol.

Clipping is locally inactive at the reference policy (Appendix B.4). The one/two/four-epoch probes audit later-update clipping with fixed batches and logged ratios (Appendix D.3). The dense/MoE grid measures frozen-gradient sensitivity within Qwen3.5; matched-cost training outcomes are reported for the core model.

E Optional value estimation and calibration

This section specifies the value-based variants and cost accounting used in the main comparisons.

E.1 Estimator and calibration procedure

Value bias and mixing. Under the score assumptions of Appendix B.2, raw LOO satisfies $\mathbb{E}Y = g$. For independently fitted, frozen heads, the action-independent baseline $V_\phi(h)$ integrates to zero, while the value gradient Z has bias

$$\mathbb{E}Z = g + d, \qquad d = \mathbb{E}\left[\sum_t \{Q_\psi(h_{t-1}, a_t) - Q^0(h_{t-1}, a_t)\} s_t\right]. \tag{32}$$

A fixed mixture has bias λd. For the same update-sized sampling unit, minimizing $\mathbb{E}\|(1-\lambda)Y + \lambda Z - g\|^2$ over $[0,1]$ gives

$$\lambda^\star = \Pi_{[0,1]} \frac{\mathbb{E}\langle Y - g, Y - Z\rangle}{\mathbb{E}\|Z - Y\|^2}. \tag{33}$$

Use zero if the denominator vanishes. An independent unbiased reference H preserves the risk ordering because

$$\mathbb{E}\|(1-\lambda)Y + \lambda Z - H\|^2 = \mathcal{R}(\lambda) + \mathbb{E}\|H - g\|^2. \tag{34}$$

Head fitting. Read $Q_\psi(h_{t-1}, a_t)$ at the end of a_t and $V_\phi(h_{t-1})$ at the end of o_{t-1} (or the prompt at $t = 1$), before the next observation. At each snapshot, freeze the trunk and fit both heads to independent on-policy terminal returns by squared error: five passes over 256 prompt groups, with a disjoint validation portion for stopping. The root loss anchors $V_\phi(x)$ to critic-training group means with the weight in Table 8. Freeze and detach the heads before pilot/actor sampling; exclude their parameters from the actor optimizer. Raw MC skips fitting. Fixed mixtures select one global $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ on train-validation; all value-based arms refit heads at every actor snapshot.

Weight fitting and confirmation. Each of eight pilot repeats draws a PG-trajectory candidate batch for Y, Z and an independent $2PG$-trajectory raw reference H under the same prompt law and reward target. Use common trainable parameters, dropout off and the unpreconditioned Euclidean metric, streaming dot products in parameter blocks. Fit

$$\hat\lambda = \Pi_{[0,1]} \frac{\sum_k \langle Y_k - H_k, Y_k - Z_k\rangle}{\sum_k \|Z_k - Y_k\|^2}. \tag{35}$$

Use zero when the denominator is below $10^{-8}\max\{\sum_k \|Y_k\|^2, 10^{-12}\}$. On 16 fresh repeats with the same batch sizes, compute $d_k = \|(1-\hat\lambda)Y_k + \hat\lambda Z_k - H_k\|^2 - \|Y_k - H_k\|^2$. Accept only if $\bar d + 1.96\,\mathrm{SE}(d) < 0$ (normal approximation); otherwise use zero. Freeze the accepted weight for one update on fresh actor data. Unaffordable calibration uses MC fallback. Each refresh samples $(8+16)(1+2) = 72$ actor-sized trajectory batches, in addition to head fitting; all spent work, including failed gates, enters the cost accounting below.

E.2 Matched full-training controls

Table 13A compares estimators under the same partial-sharing design at $G = 8$. The independent calibrated arm changes only sharing. All other settings follow Appendix C, including paired seeds, initialization, tuning and total-cost caps. Value-based arms share head architecture and fitting frequency; MC spends saved auxiliary budget on actor updates. The shared/independent calibrated

contrast supports the +7.00-point sharing gain in the main text. These full-training runs are distinct from the warm-start attribution suite in Appendix D.2.

Table 13: Full-training calibration and sharing controls. Scores and two-task means are percentages; Δ is full shared calibration minus the row, in percentage points; brackets give 95% intervals. A fixes one shared design (S); B gives the matched independent arm (I). Five paired seeds at matched total cost; warm-start continuation is a separate suite (Table 2B).

Configuration	Design	HotpotQA	ALFWorld	Mean	Δ [95%]
A. Same-sharing calibration comparison					
Raw LOO, full calibration	S	66.7	83.2	74.95	reference
Raw LOO MC	S	$60.7_{\pm 1.6}$	$73.9_{\pm 2.2}$	67.30	$+7.65$ [6.2, 9.1]
Learned-only	S	$59.6_{\pm 1.4}$	$76.8_{\pm 1.9}$	68.20	$+6.75$ [5.4, 8.1]
Validation-fixed mixture	S	59.4	76.4	67.90	+7.05 [5.9,8.2]
B. Matched independent calibration					
Raw LOO, full calibration	I	58.9	77.0	67.95	+7.00 [5.7,8.3]

E.3 Cost measurement and training stability

Resource accounting. For disjoint intervals j of t_j seconds with n_j allocated GPUs of type $h(j)$, report

$$H_{\mathrm{GPU}} = \frac{1}{3600} \sum_j n_j t_j, \qquad C_{\mathrm{eq}} = \frac{\sum_j p_{h(j)} n_j t_j / 3600 + K_{\mathrm{CPU/env}} + K_{\mathrm{tool}}}{p_{\mathrm{ref}}}. \tag{36}$$

GPU-hour prices p_h, reference price p_{ref} and separately billed charges K use common rates fixed before comparison. Retain raw GPU-hours and wall time. Charge tuning, rollouts, actor updates, head fitting, diagnostics/references, calibration, resets and failed attempts. Count allocated device time once, including reserved idle time; concurrent work on the same GPU pool shares a charge, while separate pools accrue their own. Held-out evaluation is separate, as is the common incoming-checkpoint cost for warm-start runs.

Use the token and total-cost stopping/selection rules in Appendix C. Per seed, retain hardware/configuration IDs, actual cost, actor tokens/updates, calibration counts and unused budget or overshoot; summarize paired budget discrepancies. Sampled batch counts do not by themselves determine GPU-hours.

Training stability. At matched total cost, the five-seed std values for GRPO/GRPO+ESRA are 2.00/1.70 on HotpotQA and 3.00/1.50 on ALFWorld (Table 5), reductions of 15% and 50%. The two-task mean has std 2.20/1.32 (40% reduction), computed after averaging tasks within each seed.