RAW EVALUATOR DISAGREEMENT CAN MISRANK TRAJECTORIES: INTERACTION-TARGETED AGENT RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Raw evaluator disagreement can be an unstable policy-learning target in multi-evaluator agent RL. It mixes evaluator-wide strictness with trajectory-specific evaluator dependence, so additive evaluator shifts can change trajectory-penalty rankings even when the interaction profiles are unchanged. We trace this instability to a row-specific cross term and show that, in a fully crossed block with unclipped scores, unrestricted additive shifts can reverse the raw ordering of any two distinct interaction profiles while leaving their interaction-energy ordering unchanged. Interaction-Suppressing Advantage (ISA) addresses this mismatch, relative to an interaction-targeted objective, by double centering the score block before assigning trajectory-level disagreement penalties. With a fixed common baseline and no clipping, its fixed-block advantage is invariant to evaluator-wide additive offsets. In 40 frozen-policy pilot blocks, raw and interaction-energy rankings disagree on 41.98% of trajectory pairs. Matched Raw-Energy/DC-Energy training under the same adaptive energy construction tests whether this target change affects learned policies: double centering lowers rollout discrepancy by 0.0241 (95% CI $[-0.0272, -0.0210]$), and a subsequent official-success evaluation of the frozen checkpoints yields a $+3.64$ percentage-point gain. In a separate 40-task Qwen3-8B study with ten seeds per arm, ISA reduces primary held-out global discrepancy by 27.5% relative to mean-minus-SD while improving held-out reward; with the interaction target fixed, a tuned comparison supports the adaptive penalty rule with a $+4.36$ percentage-point success gain. These results support a conditional design principle: when trajectory-specific evaluator dependence is the intended regularization target and a reliable fully crossed panel is available, remove shared evaluator shifts before forming trajectory-level disagreement penalties.

1 INTRODUCTION

A disagreement penalty is not only a diagnostic: it determines which trajectories a policy learns to avoid. In agent reinforcement learning, evaluators score tool-use and software-task trajectories (Yao et al., 2023; Qin et al., 2023; Jimenez et al., 2023; Luo et al., 2025) and can disagree on the same behavior (Mehta et al., 2026). When the intended regularization target is trajectory-specific evaluator dependence, raw within-trajectory dispersion can penalize the wrong quantity because it also contains evaluator-wide differences in strictness.

This mismatch changes rankings, not merely scales. Raw row energy decomposes into shared evaluator energy, trajectory-specific interaction energy, and a row-specific cross term. The shared term cancels in pairwise comparisons, but the cross term does not. Evaluator-wide additive shifts can therefore reverse raw penalty orderings while leaving the interaction profiles unchanged, and positive coefficient tuning cannot repair an ordering it preserves. We call this a disagreement-target mismatch relative to the stated interaction objective; total dispersion remains a valid alternative when shared evaluator differences are themselves part of the intended objective.

ISA changes the policy-learning target rather than the evaluators. Within each fully crossed scoring block, classical double centering removes row and evaluator main effects. The policy update retains centered mean reward and penalizes the remaining interaction energy, using the companion trajectories in the block to separate shared strictness from trajectory-specific departures (Figure 1).

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.

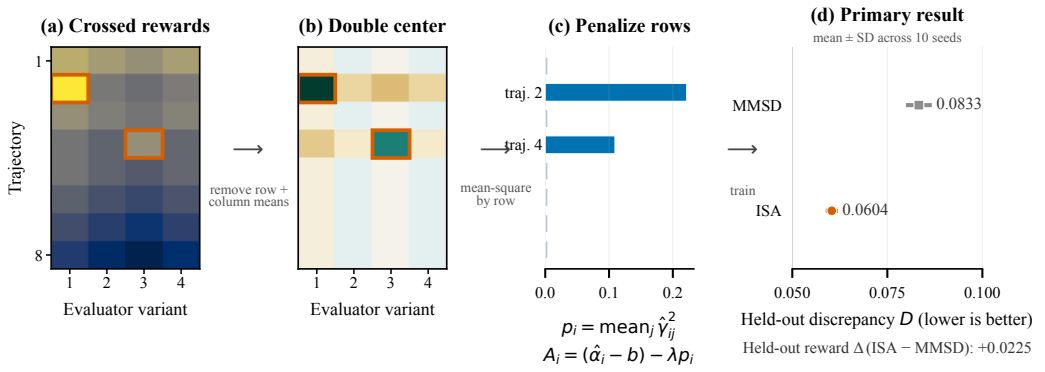


Figure 1: Full crossing makes shared evaluator shifts removable before forming a trajectory-level penalty. The crossed reward grid (a) enables this separation; double centering removes row and evaluator main effects, leaving an interaction residual (b). Outlined cells mark two schematic local interactions, whose rows receive larger interaction-energy penalties (blue bars in c). In the primary `T_MAIN`/P cohort, ISA has the lowest held-out-variant discrepancy among the primary arms and improves over mean-minus-SD while increasing held-out reward (d); error bars show sample SD over ten independently trained seeds. The residual and its row energy quantify trajectory-specific evaluator dependence.

ISA neither retrains evaluators nor estimates which evaluator is more reliable. Under evaluator-wide additive offsets, no clipping, and a fixed common baseline, its fixed-block advantage is invariant.

Evidence spans mechanism and learning: 40 frozen-policy blocks show 41.98% ranking disagreement; matched Raw-Energy/DC-Energy training lowers rollout discrepancy by 0.0241 (95% CI $[-0.0272, -0.0210]$), and frozen-checkpoint evaluation yields a $+3.64$ percentage-point success gain. Separate studies isolate the adaptive rule and assess complete-method transfer.

Contributions. (1) We identify a disagreement-target mismatch in agent RL: raw within-trajectory dispersion contains a row-specific cross term through which evaluator-wide additive shifts can change trajectory-penalty rankings. For fully crossed blocks with unclipped scores, we prove that unrestricted shifts can reverse the raw ordering of any two distinct interaction profiles while leaving their interaction energies unchanged. (2) We introduce *interaction-targeted policy learning* through ISA, which removes shared evaluator shifts by double centering and penalizes trajectory-specific interaction energy. Its fixed-block advantage is invariant to evaluator-wide additive offsets under the stated baseline and no-clipping conditions. (3) We test whether this target intervention changes learned policies through matched Raw-Energy/DC-Energy training and a subsequent frozen-checkpoint official-success evaluation. Separate experiments assess the adaptive penalty rule with the target fixed, complete-method transfer, and estimator scope. The results support the intervention in the tested regime but do not determine the optimal target estimator across block sizes, calibration quality, or drift.

2 Related Work

Uncertainty penalties in RLHF. Reward overoptimization separates measured reward from intended success (Amodei et al., 2016; Pan et al., 2022; Gao et al., 2023). UWO penalizes ensemble variance (Coste et al., 2024), UP-RLHF subtracts centered ensemble uncertainty (Zhai et al., 2024), and Bayesian/KL-DRO formulations connect mean, worst-case, and uncertainty-weighted objectives (Hahami et al., 2026). Ensembles can still share errors under policy-induced shift (Eisenstein et al., 2024). These methods motivate conservative learning from evaluator panels, but total within-row uncertainty does not distinguish evaluator-wide shifts from trajectory-specific interaction. ISA makes this distinction before constructing the policy-update penalty.

Calibration and interaction recovery. CHARM uses external Elo anchors to correct policy-source bias (Zhu et al., 2025); CBC applies two-way double centering to recover language–judge

interactions without labels (Mahmood et al., 2026); PEBS shrinks noisy per-rater calibration parameters toward a population estimate (Raj, 2026). Double centering is classical and is not our contribution. Our contribution is to identify its interaction residual as a conditional policy-learning target, derive the raw-ranking instability that motivates this choice, and test the target intervention through matched policy training. Current-block, historical, externally anchored, and shrinkage variants are compared separately as estimator choices.

Reward decomposition and invariance. Reward debiasing and decomposition address length effects, RLHF, and tool use (Zheng et al., 2025; Mao et al., 2025; Liu, 2026). ISA instead decomposes a fully crossed trajectory-by-evaluator panel. Its additive-shift invariance concerns the fixed-block advantage map, unlike representation-based invariance objectives (Ganin et al., 2016; Arjovsky et al., 2019). MMSD penalizes total row dispersion; matched Raw/DC training changes the residual target under a common adaptive energy construction; tuned DC-SD then compares penalty rules with that target fixed. Turn-level credit assignment is orthogonal to ISA's evaluator-axis decomposition; Appendix O gives broader context.

3 Interaction-Targeted Policy Learning

ISA replaces the raw row-centered residual with the interaction residual before forming the penalty, leaving the centered mean-reward term unchanged; the intervention corrects the target rather than changing the underlying evaluators.

3.1 Separating shared shifts from trajectory dependence

For a fixed task or prompt, let R_{ijk} be the score of trajectory $i \in [K]$ under evaluator $j \in [c]$ in replication $k \in [r]$. Every trajectory is scored by the same panel, with $r \geq 1$ scores per cell and cell mean $\bar{R}_{ij} = r^{-1} \sum_k R_{ijk}$. Dots denote arithmetic averages: $\bar{R}_{i\cdot}$, $\bar{R}_{\cdot j}$, and $\bar{R}_{\cdot\cdot}$ are the row, column, and grand means. The two-way model is

$$R_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{1}$$

with conditionally mean-zero noise and sum-to-zero constraints on α, β, and every row and column of γ. The observed row effect and double-centered residual are

$$\hat{\alpha}_i = \bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot}, \qquad \hat{\gamma}_{ij} = \bar{R}_{ij} - \bar{R}_{i\cdot} - \bar{R}_{\cdot j} + \bar{R}_{\cdot\cdot}. \tag{2}$$

The estimated column effect is $\hat{\beta}_j = \bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot}$. Using the other trajectories to estimate this shared effect distinguishes an evaluator that is generally strict from a trajectory-specific departure.

Why a shared shift can change row rankings. Define interaction energy $p_i = c^{-1} \sum_j \hat{\gamma}_{ij}^2$, main-effect energy $B = \|\hat{\beta}\|_2^2 / c$, and $p_i^{\mathrm{raw}} = c^{-1} \sum_j (\bar{R}_{ij} - \bar{R}_{i\cdot})^2$. Then

$$p_i^{\mathrm{raw}} = B + p_i + \frac{2}{c} \langle \hat{\beta}, \hat{\gamma}_i \rangle. \tag{3}$$

The shared term B cancels in pairwise penalty differences, but the row-specific cross term need not. Its average over trajectories is zero, giving $\overline{p^{\mathrm{raw}}} = B + \bar{p}$. Thus block-level orthogonality does not imply row-level orthogonality. Sample SD is a monotone transform of raw energy and preserves its penalty ordering. Appendix G proves the decomposition and defines the corresponding diagnostics.

3.2 Interaction-Suppressing Advantage

Raw row-centered residuals mix evaluator main effects with trajectory-specific interactions. ISA removes the evaluator main effect through double centering before forming the trajectory-level penalty. It retains the centered row effect and subtracts a scaled interaction-energy penalty:

$$A_i = (\hat{\alpha}_i - b) - \lambda p_i, \qquad p_i = \frac{1}{c} \sum_{j=1}^{c} \hat{\gamma}_{ij}^2, \qquad \lambda = \frac{\kappa}{\sqrt{q_\theta(P_+)}}, \tag{4}$$

where b is a common group baseline, $P_+ = \{p_i : p_i > 0\}$, and q_θ is the empirical quantile over positive penalties in the current block. We use $\kappa = 0.15$ and $\theta = 0.5$, and set $\lambda = 0$ when P_+

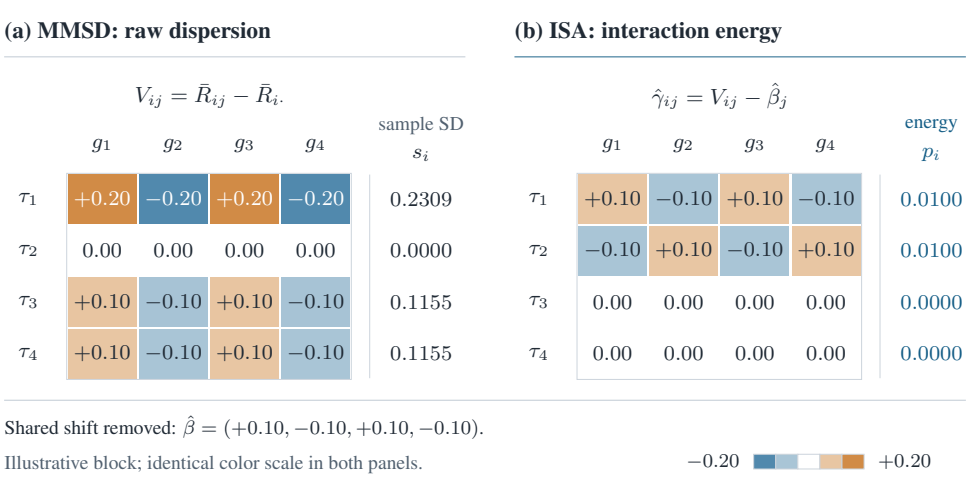


Figure 2: Double centering separates interaction from shared evaluator shifts. In this deterministic illustrative block ($K = c = 4$), the shared shift $\hat{\beta}_j = \bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot}$ reinforces τ_1's interaction and cancels τ_2's, so raw SDs differ although $p_1 = p_2 = 0.01$. Rows τ_3, τ_4 have nonzero raw SD but zero interaction. Cells share a signed color scale. The outputs are sample SD (denominator $c - 1$, reward units) and unweighted interaction energy (squared-reward units); their magnitudes are not directly comparable. The example illustrates residualization, not empirical performance or gaming labels.

is empty. The weight is recomputed for each block. The baseline cancels in pairwise advantage differences but is required for absolute advantages.

Algorithm 1 Interaction-Suppressing Advantage for one crossed block

Require: Trajectories $\tau_{1:K}$, evaluator variants $g_{1:c}$, replication count $r \geq 1$
Require: Common baseline b; $\kappa = 0.15$, $\theta = 0.5$
1: Score every trajectory under every variant r times and compute $\bar{R}_{ij}$
2: Compute $\hat{\alpha}_i$ and $\hat{\gamma}_{ij}$ using Eq. (2)
3: Set $p_i \leftarrow c^{-1} \sum_j \hat{\gamma}_{ij}^2$ and $P_+ \leftarrow \{p_i : p_i > 0\}$
4: Set $\lambda \leftarrow 0$ if $P_+ = \varnothing$; otherwise $\lambda \leftarrow \kappa / \sqrt{q_\theta(P_+)}$
5: Return $A_i \leftarrow (\hat{\alpha}_i - b) - \lambda p_i$ for every trajectory

Policy update. The deployed rule uses $b = 0$ and treats computed advantages as fixed coefficients for policy gradients. There is no subsequent advantage centering, SD normalization, or clipping. Each fresh group yields one on-policy update with a token-averaged loss and KL weight 0.01 relative to the initial policy, without PPO ratio clipping. Global gradient clipping at norm 1.0 is separate from these advantage operations; Raw/DC update-level diagnostics appear in Appendix I.9. Appendix F.2 specifies the complete loss and accumulation weights; Appendix B gives the local surrogate.

3.3 Predictions under evaluator-wide shifts

Proposition 1 (Raw-ranking instability under evaluator shifts)**.** *Fix a fully crossed cell-mean matrix and two trajectories i, k with $\hat{\gamma}_i \neq \hat{\gamma}_k$. If evaluator-wide additive shifts are unrestricted and scores are not clipped, there exist shifts under which $p_i^{\text{raw}} > p_k^{\text{raw}}$ and shifts under which $p_i^{\text{raw}} < p_k^{\text{raw}}$. In contrast, p_i, p_k, and their ordering are unchanged.*

Proof. Let $v = \hat{\gamma}_i - \hat{\gamma}_k$, whose entries sum to zero. After a centered evaluator shift d, Eq. (3) gives $p_i^{\text{raw}} - p_k^{\text{raw}} = p_i - p_k + 2\langle \hat{\beta} + d, v \rangle / c$. Choosing $d = tv$ makes this difference affine in t with nonzero slope $2\|v\|_2^2/c$, so either sign is attainable. Double centering removes d, leaving both interaction energies unchanged. □

Proposition 2 (Additive evaluator-shift invariance)**.** *Fix a fully crossed cell-mean matrix $\bar{R}$ and the common baseline b. For any $\delta \in \mathbb{R}^c$, let $\bar{R}'_{ij} = \bar{R}_{ij} + \delta_j$. If no score clipping or other nonlinear score transformation is applied, Eq. (4) gives $A_i(\bar{R}', b) = A_i(\bar{R}, b)$ for every trajectory, including blocks with zero interaction.*

Proof. Each row mean and the grand mean increase by $\bar{\delta}$, leaving $\hat{\alpha}_i$ unchanged. Column j's mean increases by δ_j, so double centering cancels the added offsets and $\hat{\gamma}'_{ij} = \hat{\gamma}_{ij}$. Thus p_i, P_+, its quantile, and λ are unchanged, including the $P_+ = \varnothing$ branch. Holding b fixed completes the result. □

This property concerns the fixed-block score-to-advantage map, not arbitrary evaluator changes or invariance of an entire training process. Raw dispersion generally changes under the same column offsets through the cross term in Eq. (3). Under positive proportional reward rescaling, the penalty subtraction in ISA scales linearly with reward when b scales accordingly; Appendix B covers the units and degenerate blocks.

Bounded score-feasible shifts. Proposition 1 is an unrestricted existence result. For observed scores in $[0, 1]$, a no-clipping audit instead restricts each additive shift to $d_j \in [L_j, U_j]$, where $L_j = -\min_i \bar{R}_{ij}$ and $U_j = 1 - \max_i \bar{R}_{ij}$. For a pair (i, k), let $v = \hat{\gamma}_i - \hat{\gamma}_k$ and $\Delta_0 = p_i^{\mathrm{raw}} - p_k^{\mathrm{raw}}$. The shifted raw difference is

$$\Delta(d) = \Delta_0 + \frac{2}{c}\langle d, v\rangle. \tag{5}$$

Its extrema over $[L, U]$ are obtained coordinatewise, so both raw orderings are feasible exactly when the minimum is negative and the maximum is positive. Section 4.2 reports a frozen-block audit of this condition and the smallest ℓ_∞ radius that reaches a tie; when both signs are feasible, strict reversal occurs arbitrarily beyond that boundary. Raw/DC disagreement on the unshifted matrix is a separate diagnostic and is not a shift-induced flip rate.

The fixed-block prediction does not imply a policy improvement by itself. The learning hypothesis is that removing this nuisance contribution improves training when trajectory-specific evaluator dependence is the intended regularization target; Section 4.3 tests it by changing the residual input under a matched adaptive energy rule. Replication noise and companion composition remain estimation-error sources, quantified in Appendix B.2 and Section 4.7 respectively.

4 Experiments

Experiments test ranking disagreement, the matched residual-target intervention, the adaptive rule at a fixed target, and transfer and estimator scope. Cohorts, endpoints, selection roles, and official-success evaluations remain separate.

4.1 Experimental setup

Tasks and training. The primary Qwen3-8B study uses `T_MAIN`: 40 tasks with $15/10/15$ from tau-bench, SWE-bench-lite, and Toolathlon. Six arms use ten seeds, 240 updates, $K = 14$, and eight evaluation rollouts per task and seed; five crossed arms use $c = 4$, while GRPO uses A1 only. A disjoint `T_ABLATION` has the same quotas. P and Q are the respective native held-out aggregation rules; the sensitivity analysis evaluates both task sets under Q. A1–A4 train, A5–A6 are held out, and B/C are independently implemented external measurements. Official labels are reserved for reference-target calibration or confirmation evaluation and never train current-block ISA. Appendices F, L, and L.3 give full roles and task relationships.

Baselines. DR averages the four evaluators; DR+ weights them inversely to past marginal reward variance; Pess-min uses the minimum score. The primary comparator is mean-minus-SD (MMSD), $\bar{R}_{i\cdot} - 0.50\,\mathrm{sd}_j(\bar{R}_{ij})$, which penalizes raw dispersion at the same grading budget. ISA uses Eq. (4) with $\kappa = 0.15$ and $\theta = 0.5$. GRPO uses $c = 1$ and is a lower-grading-budget reference, not a matched-budget comparator. The MMSD coefficient sweep tests sensitivity to the fixed 0.50 choice.

Endpoints and inference. Evaluation scores each generated rollout with the training and held-out panels; endpoints use protocol-specific aggregated rewards rather than the training interaction penalty. Let v denote the held-out aggregation rule. Within a seed, $d_{tm}^{(v)}$ is the training-panel reward minus the held-out reward aggregated by rule v, with task mean $\bar{d}_t^{(v)}$ and overall mean $\bar{d}^{(v)}$. For T tasks and M rollouts per task, define

$$D^{\mathrm{global},(v)} = |\bar{d}^{(v)}|, \qquad D^{\mathrm{task},(v)} = \frac{1}{T}\sum_t |\bar{d}_t^{(v)}|, \qquad D^{\mathrm{rollout},(v)} = \frac{1}{TM}\sum_{t,m} |d_{tm}^{(v)}|.$$

We suppress v within a fixed-protocol table: P is native to the primary study and Q to the ablation and unified-Q sensitivity analysis. Appendix I.1.1 defines protocol-specific panel aggregation.

The primary cohort uses D^{global}; extensions and the residual-target ablation also use finer-grained endpoints with less cancellation. Method effects are compared within cohort and protocol. Unless stated otherwise, tables report mean$\pm$sample SD and ISA-minus-comparator contrasts use 95% paired-t intervals over training seeds, not individual rollouts. The between-cohort policy comparison instead bootstraps the independent seed groups. The primary MMSD comparison and -0.01 held-out-reward safeguard were specified before outcome inspection; DR, DR+, and Pess-min form a Bonferroni-corrected family of three secondary contrasts.

4.2 Do raw and interaction penalties rank trajectories differently?

Fixed-block diagnostic. The cross term in Eq. (3) can change penalty ordering even when the underlying interactions remain fixed. In the 40 pilot blocks, using all 15 trajectories per task including the planted controls, the ratio of main-effect energy to mean interaction energy, $B/\bar{p}$, has median 7.54 and interquartile range $[6.75, 7.90]$. Raw and double-centered energies reverse $1,763$ of $4,200$ pairwise orderings (41.98%; no ties). Figure 3 shows the task-level distributions. The high ratio establishes a substantial contribution to average raw energy, and the reversals establish an effect on which trajectories receive larger penalties; neither statistic identifies gaming or estimates a trained-policy benefit. Appendix G gives the full decomposition, outlier analysis, and the weak descriptive association with downstream performance.

Bounded-shift audit. Forty $K = 14, c = 4$ frozen blocks yield $3{,}640$ unordered trajectory pairs. All pairs admit both raw orderings somewhere in their score-feasible additive-shift boxes. At the smaller budget $\|d\|_\infty \le 0.05$, 54.81% admit a reversal (task-bootstrap 95% CI $[51.26, 58.60]\%$); the median boundary radius is 0.04524. The shift is optimized separately for each pair: these rates do not imply a single common shift reverses that fraction simultaneously. On the unshifted matrices, raw/DC rankings disagree for $1{,}480/3{,}640$ pairs (40.66%). This audit is distinct from the separate $K = 15$ pilot with planted controls (41.98% disagreement). Appendix H gives the full budget curve and analysis definitions.

4.3 Does the target correction improve learned policies?

Matched residual-target training comparison. A separate five-arm, ten-seed study on `T_ABLATION` (the same $15/10/15$ quotas but no task-ID overlap with `T_MAIN`) compares raw row-centered and double-centered (DC) residuals under SD and energy penalties (Table 1). The main mechanism contrast is DC-Energy/ISA versus Raw-Energy under the same adaptive energy family, using D^{rollout} as its primary endpoint. The residual input changes while the functional form of the adaptive energy penalty is held fixed; its block-dependent weight is recomputed for each input. This tests the complete residual-target intervention under the deployed rule.

Relative to Raw-Energy, DC-Energy/ISA gives a D^{rollout} difference of -0.0241 (95% CI $[-0.0272, -0.0210]$) and raises held-out reward by 0.0197. The D^{global} contrast is inconclusive: -0.0031 (95% CI $[-0.0079, 0.0017]$). Lower task discrepancy in the energy comparison and lower rollout discrepancy for DC-SD versus Raw-SD support the residual-target intervention. These are within-ablation effects; the native primary/ablation discrepancy scales also differ in task composition and held-out aggregation rule (Appendix L).

Official-success follow-up. On the fixed 90-task `N090` task set, ten paired frozen Raw/DC checkpoints yield 57.40% versus 61.04% success, a $+3.639$ percentage-point difference (95% bootstrap

Table 1: Residual-target ablation on T_ABLATION under aggregation Q (ten seeds; mean±SD). The primary contrast is DC-Energy/ISA versus Raw-Energy on D^{rollout}; all endpoints use the same rollouts.

Arm	$D^{\mathrm{rollout}} \downarrow$	$D^{\mathrm{task}} \downarrow$	$D^{\mathrm{global}} \downarrow$	$R_{\mathrm{out}} \uparrow$
Raw-SD	0.0894 ± 0.0047	0.0890 ± 0.0047	0.0156 ± 0.0071	0.6494 ± 0.0011
DC-SD	0.0707 ± 0.0028	0.0697 ± 0.0031	0.0081 ± 0.0068	0.6696 ± 0.0012
Raw-Energy	0.0846 ± 0.0030	0.0842 ± 0.0030	0.0107 ± 0.0050	0.6602 ± 0.0014
DC-Energy / ISA	0.0605 ± 0.0030	0.0590 ± 0.0035	0.0076 ± 0.0044	0.6799 ± 0.0011
MMSD 0.50	0.0803 ± 0.0034	0.0794 ± 0.0035	0.0125 ± 0.0062	0.6592 ± 0.0011

CI $[+1.027, +6.292]$). This subsequent evaluation supports the target intervention but was not initially preregistered; evidence-separation and gradient diagnostics are in Appendices I.8 and I.9.

4.4 Does ISA's adaptive rule add value after target correction?

With double centering fixed, ISA uses adaptive squared interaction energy whereas DC-SD uses a fixed coefficient on its square root; this compares complete penalty rules, not isolated shape and normalization effects. Development selection from $\{0.075, 0.15, 0.30, 0.60\}$ chose 0.15 for both rules, followed by ten paired confirmation seeds on the remaining 30 T_MAIN tasks. Appendix J.1 gives the selection and evidence-separation protocol.

Table 2 reports the confirmation results. ISA lowers rollout-level discrepancy from 0.0760 to 0.0620, a relative reduction of 18.4%, while increasing held-out reward from 0.6650 to 0.6710. The paired rollout-discrepancy interval excludes zero, and the one-sided lower bound for the reward difference is $+0.0043$, above the -0.01 safeguard.

Does official task completion also improve? The $N0_{\mathrm{confirm}}$ evaluation on the shared N090 task set with the same ten paired seeds gives $4234/7200$ successes for ISA and $3920/7200$ for DC-SD, a $+4.36$ percentage-point difference. A 10,000-draw bootstrap resamples paired seeds and tasks within benchmark blocks, giving a 95% interval of $[+2.35, +6.36]$ percentage points. N0 directly applies the official binary rules on 30 tasks per benchmark, independently of auxiliary-score thresholds (Appendix J.2). Both arms already remove evaluator main effects, so this success gain belongs to the adaptive-rule comparison; it is not an isolated effect of residualization.

Table 2: Fixed-target adaptive-rule comparison (ten paired seeds). Coefficients are selected on ten development tasks before evaluation on 30 confirmation tasks; official binary success is measured on 90 tasks. Differences are ISA minus DC-SD.

Endpoint	DC-SD	ISA	Difference
$D^{\mathrm{rollout}} \downarrow$	0.0760 ± 0.0040	$\mathbf{0.0620 \pm 0.0035}$	-0.0140 $[-0.0172, -0.0108]$
$R_{\mathrm{out}} \uparrow$	0.6650 ± 0.0045	$\mathbf{0.6710 \pm 0.0040}$	$+0.0060$ $[+0.0039, +0.0081]$
Task success ↑	$3920/7200$ (54.44%)	$\mathbf{4234/7200}$ (58.81%)	$+4.36$ pp $[+2.35, +6.36]$

Appendices I.4.3, I.4.2, and I.4.4 report supporting evaluator-shift, rescaling, and penalty-assignment controls.

4.5 Do the complete-method gains transfer?

Same-family evaluator transfer. Under the primary cohort's native aggregation P, ISA reduces D^{global} by approximately 27.5% relative to MMSD, from 0.0833 to 0.0604 (Table 3). The paired difference is -0.02292 (95% CI $[-0.02579, -0.02005]$), with all ten seeds favoring ISA. All three Bonferroni-corrected secondary contrasts also favor ISA (Appendix I.1.1).

Held-out reward increases by 0.02247 (95% CI $[0.01935, 0.02559]$), satisfying the pre-specified -0.01 safeguard. Task- and rollout-level discrepancies also favor ISA, showing that the primary

Table 3: Same-family transfer on T_MAIN under native aggregation P (ten seeds; mean±SD). MMSD 0.50 is primary; all arms use $c = 4$ except GRPO.

Method	$D^{\text{global}} \downarrow$	Held-out reward ↑
GRPO ($c = 1$, reference)	0.1617 ± 0.0042	0.5724 ± 0.0057
DR	0.1329 ± 0.0052	0.6208 ± 0.0077
DR+	0.0923 ± 0.0033	0.6456 ± 0.0056
Pess-min	0.0841 ± 0.0035	0.6409 ± 0.0079
MMSD 0.50	0.0833 ± 0.0034	0.6517 ± 0.0059
ISA	$\mathbf{0.0604 \pm 0.0016}$	$\mathbf{0.6742 \pm 0.0042}$

improvement is not solely aggregate cancellation; full inference and per-seed results are in Appendices I.1.1, M, and I.1.2.

Coefficient sensitivity. In a separately trained cohort using aggregation P, ISA improves on MMSD across $\kappa_{\text{SD}} \in \{0, 0.25, 0.50, 0.75, 1.00\}$ under the frozen D^{task} endpoint. Post-hoc D^{global} recomputation leaves the 0.25 coefficient inconclusive. This is a robustness study, not the development-grid selection used for the DC-SD confirmation (Appendix I.3.1).

Unseen tasks. On 180 unseen tasks, with 60 each from tau-bench, SWE-bench, and Toolathlon and ten paired seeds, the ISA-minus-MMSD D^{task} difference is -0.01276 (95% CI $[-0.01556, -0.00980]$; 10,000-draw paired-seed and suite-stratified-task bootstrap). The held-out-reward safeguard passes; post-hoc D^{global} also favors ISA. These results extend coverage within the tested suites. Appendix I.3.1 reports the task-selection procedure, DR comparison, and post-hoc estimates.

Independent evaluator implementations. Policies trained with Family-A variants are evaluated on independently implemented families B and C. Relative to MMSD, ISA lowers D^{global} by 0.0230 and 0.0272, while increasing reward by 0.0206 and 0.0248; paired intervals exclude zero (Appendix I.3.2). The families share benchmark requirements, so implementation independence does not imply independent task semantics.

4.6 Should interaction use current-block or reference offsets?

The reference-target comparison evaluates current-block, historical, externally anchored, and shrinkage targets in a 20-paired-seed T_MAIN study with $K = 14$, $c = 4$, $r = 4$, 240 updates, and a shared energy-penalty rule. Against shrinkage, current ISA lowers rollout discrepancy by 0.00423 (95% CI $[-0.00488, -0.00358]$) and raises reward by 0.00270 (one-sided lower bound $+0.00211$); historical and external-anchor targets also have higher mean discrepancies. Appendices H and H.1 give grading, calibration, label-access, and coverage details.

4.7 Where does interaction targeting remain incomplete?

Companion-composition audit. Across 40 tasks, eight fixed anchors are combined with six resampled companions in 100 draws from 64 trajectories. Relative to the full pool, median Spearman agreement is 0.929 (95% CI $[0.893, 0.952]$), ICC is 0.923, sign consistency is 91.88%, top-two overlap is 80.84%, and median anchor CV is 0.220. Rankings are stable here, but penalties and advantages remain composition-dependent.

Does the dominant-companion mode grow during training? Across frozen checkpoints at updates $0/60/120/180/240$, the extreme exact-13 composition remains below 1% (largest bootstrap upper bound 0.723%) without detectable growth; ranking and sign-stability lower bounds remain high (Appendix J.3). We therefore retain high-prevalence amplification as a boundary rather than an observed training regime.

Evaluator-shared errors. Interaction is narrower than gaming. Variant-specific planted patterns have raw-reward win rate 0.0441 under ISA versus 0.3280 under DR, whereas evaluator-shared pat-

terns remain high (0.6782 versus 0.7121) and receive only $1.09\times$ the reference median penalty. A 300-trajectory human check gives penalty AUC 0.6517 for gaming labels (Appendices I.6.2 and I.6.1). Thus ISA is an evaluator-dependence regularizer, not a gaming detector.

5 Discussion

ISA changes which disagreement the policy learns to avoid: double centering removes shared evaluator strictness before assigning trajectory-level costs. Matched Raw/DC training and its frozen-checkpoint success follow-up support this intervention beyond a fixed-score diagnostic, but do not establish interaction as a universally optimal uncertainty target.

DC-SD isolates the penalty rule; transfer studies test the complete method; the reference-target comparison assesses alternative estimators; and the P/Q sensitivity analysis separates aggregation-rule, task-set, and policy-cohort differences descriptively. Reused tasks do not make these effects additive or independent replications. ISA requires a reliable crossed panel, and interaction may reflect legitimate specialization while missing evaluator-shared errors. P/Q aggregation sensitivity, longitudinal companion resampling, and bounded-floor retraining test these assumptions (Appendices L.2, J.4, and J.6). Composition, near-zero energies, evaluator dominance, nonlinear transformations, and drift remain open limitations.

6 Conclusion

Raw evaluator disagreement can change trajectory-penalty rankings through a cross term between evaluator-wide strictness and trajectory-specific interaction. ISA removes this term before the policy update and trains against interaction energy instead. Matched residual-target training and its frozen-checkpoint success follow-up support this target intervention in the tested regime; separate adaptive-rule and transfer studies support the complete method. The broader lesson is conditional but actionable: decide which disagreement should influence learning before choosing how strongly to penalize it.

Ethics and AI Use Statement

The study includes a blinded human-label comparison on 300 trajectories and separate continuous human reference scores used only to calibrate aggregation Q. The interaction penalty should be paired with task-success and safety evaluation. LLM tools assisted methodology and experiment design, implementation scaffolding, statistical checking, literature search, and manuscript drafting and editing. The authors remain responsible for the manuscript.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

Tim Brys, Ann Nowé, Daniel Kudenko, and Matthew Taylor. Combining multiple correlated reward and shaping signals by measuring confidence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 2014. doi: 10.1609/aaai.v28i1.8998.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011. doi: 10.1145/1970392.1970395.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. doi: 10.1137/090761793.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. arXiv preprint arXiv:1104.0354, 2011.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475, 2024.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, D. J. Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *Conference on Language Modeling*, 2024.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. arXiv preprint arXiv:2505.10978, 2025. NeurIPS 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. arXiv preprint arXiv:2210.10760, 2023.

Yiran Guo, Yang Xu, et al. Segment policy optimization: Effective segment-level credit assignment in reinforcement learning for large language models. arXiv preprint arXiv:2505.23564, 2025. NeurIPS 2025.

Ely Hahami, Yoel Zimmermann, Ray Zhou, and Jack Benarroch Jedlicki. A unifying lens on reward uncertainty in RLHF. arXiv preprint arXiv:2606.09073, 2026.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? arXiv preprint arXiv:2310.06770, 2023.

Sayash Kapoor, Benedikt Stroebl, et al. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. arXiv preprint arXiv:2510.11977, 2025.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

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. arXiv preprint arXiv:2512.17008, 2026. EACL 2026 Findings.

Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, Junteng Liu, Zhaochen Su, Yiyang Guo, Fan Zhou, Lueyang Zhang, Juan Michelini, Xingyao Wang, Xiang Yue, Shuyan Zhou, Graham Neubig, and Junxian He. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. arXiv preprint arXiv:2510.25726, 2025.

Sicong Liao, Zhi Chen, and Yaohua Tang. TCPO: Turn-level credit policy optimization. arXiv preprint arXiv:2608.01667, 2026.

Pengbo Liu. ToolRLA: Multiplicative reward decomposition for tool-integrated agents. arXiv preprint arXiv:2603.01620, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv preprint arXiv:2303.16634, 2023.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent Lightning: Train ANY AI agents with reinforcement learning. arXiv preprint arXiv:2508.03680, 2025.

Alhasan Mahmood, Samir Abdaljalil, and Hasan Kurban. Rank reversal in multilingual LLM judges: A label-free double-centering calibrator. arXiv preprint arXiv:2608.22432, 2026.

Liyuan Mao, Haoran Xu, Amy Zhang, Weinan Zhang, and Chenjia Bai. Information-theoretic reward decomposition for generalizable RLHF. arXiv preprint arXiv:2504.06020, 2025.

Sushant Mehta, Logan Ritchie, Liudas Panavas, and Edwin Chen. Cross-benchmark generalization in long-horizon agents. arXiv preprint arXiv:2608.00181, 2026. Accepted at COLM 2026 Workshop on Agent Behavior.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. arXiv preprint arXiv:2201.03544, 2022.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. arXiv preprint arXiv:2307.16789, 2023.

Arnav Raj. PEBS: Per-rater empirical-Bayes shrinkage for RLHF reward-model calibration. ICML Workshop on Pluralistic Alignment, 2026.

Bo Ren, Yirong Mao, Yi Yang, and Wenhui Que. IAPO: Influence-aware policy optimization for credit assignment in multi-turn service agents. arXiv preprint arXiv:2608.24588, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. arXiv preprint arXiv:1908.08342, 2019.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

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.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward LoRA ensembles. arXiv preprint arXiv:2401.00243, 2024.

Congmin Zheng, Jiachen Zhu, Jianghao Lin, Xinyi Dai, Weiwen Liu, Haoxuan Li, Yong Yu, Weinan Zhang, and Mengyue Yang. CoLD: Counterfactually-guided length debiasing for process reward models in mathematical reasoning. arXiv preprint arXiv:2507.15698, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. arXiv preprint arXiv:2306.05685, 2023.

Yuyao Zheng, Haipeng Sun, Junwei Bao, Lemao Liu, Hongfei Jiang, Yang Song, and Dejing Dou. PGPO: Potential-guided policy optimization for multi-turn agentic tasks. arXiv preprint arXiv:2609.02236, 2026.

Xiao Zhu, Chenmien Tan, Pinzhen Chen, Rico Sennrich, Huiming Wang, Yanlin Zhang, and Hanxu Hu. CHARM: Calibrating reward models with chatbot arena scores. arXiv preprint arXiv:2504.10045, 2025.

A Appendix Roadmap and Evidence Index

The appendix is organized around the claims a reviewer may wish to verify. Method definitions and estimation assumptions come first, followed by proofs, implementation details, ranking diagnostics, matched training evidence, transfer studies, and cohort separation. Table 4 gives a short verification map; Table 36 later records the full cohort and endpoint separation.

Fast verification path. The shortest path through the evidence is: the row-wise cross-term identity in Appendix G; bounded-shift and companion-composition diagnostics in Appendix H; the matched Raw-Energy/DC-Energy comparison in Appendix I.2; the official-success follow-up in Appendix I.8; and the fixed-target adaptive-rule comparison in Appendices J.1–J.2. Implementation details and statistical scope are documented in Appendices F and L.3.

Table 4: Reviewer-oriented map from claims to the evidence needed to assess them.

Question	Evidence to verify	Location
Why can raw disagreement misrank trajectories?	Row-wise cross-term decomposition, unrestricted-shift result, observed reversals, and bounded feasible shifts	Appendices G, H
What target does ISA optimize?	Dense interaction penalty, normalization, fixed-block invariance, and finite-replication scope	Appendices B, C, D
Does target correction change learned policies?	Matched Raw-Energy/DC-Energy training and subsequent official binary success	Appendices I.2, I.8
Does the adaptive rule add beyond double centering?	Grid-selected DC-SD/ISA comparison with the interaction target fixed	Appendices J.1, J.2
How broadly do the gains transfer?	Primary, unseen-task, evaluator-family, coefficient, and evaluator-count results	Appendices I.1.1, I.3.1, I.3.2, I.5
How dependent is the target on block composition and references?	Companion resampling and current/historical/external/shrinkage comparisons	Appendices H, H.1
What implementation and statistical details are reported?	Training loss, evaluator roles, per-seed values, task intersections, sample sizes, and reproducibility limits	Appendices F, M, L.3

B Detailed Method and Advantage Construction

ISA uses the complete crossed block to distinguish shared evaluator shifts from trajectory-specific departures. This section defines the residual and its normalization, shows how the resulting penalty changes a pairwise advantage, and characterizes through row averaging the quantities that scores alone leave unidentified. Sections I.4.2 and I.7.1 evaluate alternative normalizations.

B.1 Residual sensitivity and scope

Proposition 3 (Interaction energy bounds same-panel residual score changes)**.** *Fix an observed crossed reward block and trajectory i. Let $\hat{\gamma}_i = (\hat{\gamma}_{i1}, \ldots, \hat{\gamma}_{ic})$ and $p_i = c^{-1}\sum_j \hat{\gamma}_{ij}^2$, as in Eq. (4). For any two distributions w, w' over the same c evaluator variants,*

$$\left|\sum_{j=1}^{c}(w_j - w'_j)\hat{\gamma}_{ij}\right| \le \sqrt{c}\,\|w - w'\|_2\,\sqrt{p_i}. \tag{6}$$

Under a uniform draw from the panel, $\mathbb{E}_j|\hat{\gamma}_{ij}| \le \sqrt{p_i}$.

The proof is a direct application of Cauchy–Schwarz (Appendix D.4).

The bound controls the observed residual component on the same evaluator support. It does not bound the shared main-effect change $(w - w')^\top \hat{\beta}$ in the raw score or establish transfer to new evaluators. Reward increments shared across all evaluators contribute no interaction, even though other variation can give the same trajectory a nonzero penalty. Nor does the bound identify why the observed residual dependence occurs.

Under balanced full crossing and conditionally mean-zero scoring noise, the row and interaction estimators are unbiased for their panel-relative ANOVA quantities (Appendices C and D). The deployed energy is a plug-in quantity: stochastic grading noise contributes estimator variance and can affect penalty rankings. Interaction may reflect exploitation, legitimate specialization, or semantic differences; increments shared across evaluators enter the row effect. These limits motivate the separate transfer, task-success, and shared-pattern evaluations below.

B.2 Finite-replication error and target choice

Residualization removes an identified nuisance component but does not make the residual noiseless. Let $p_i^\star = c^{-1} \sum_j \gamma_{ij}^2$ denote latent interaction energy. Under the independent homoskedastic replication model used in Appendix C, Theorem 2 implies

$$\mathbb{E}[p_i] = p_i^\star + \nu, \qquad \nu = \frac{\sigma^2}{r} \left(1 - \frac{1}{K}\right) \left(1 - \frac{1}{c}\right). \tag{7}$$

Thus double centering addresses contamination by shared evaluator shifts, whereas replication controls one component of estimation error in the chosen target. With $r \geq 2$, pooled within-cell residual variance gives an implementable $\hat{\nu}$, and $[p_i - \hat{\nu}]_+$ is a noise-corrected diagnostic. Under heterogeneous cell noise, the correction may become row dependent and change rankings, although evaluator-only heteroskedasticity need not do so. The centering operators determine the required variance propagation.

Equation (7) is conditional on a fixed trajectory block and quantifies replication noise, not finite-K composition error. Because $\hat{\beta}_j$ is estimated from the sampled trajectories, replacing the other members of a rollout group can change a fixed trajectory's $\hat{\gamma}_i$, penalty rank, and advantage. Our composition audit fixes eight anchor trajectories per task and replaces six companions from a frozen 64-trajectory pool (Section 4.7). Its stability measures are task-aggregated diagnostics, not policy-level efficacy tests. Positive-part noise correction is likewise a diagnostic: truncation and the block-adaptive normalization do not preserve unbiasedness of the resulting advantage.

The same calibration pool supports a matched-budget comparison among current-block, historical, externally anchored, and shrinkage residual targets. The external target uses evaluator-specific affine maps fitted to independent official utility, avoiding the algebraic equivalence between a purely additive external offset and the historical correction. The shrinkage weight is selected only within the calibration pool by task-grouped cross-validation. The reported ISA runs use the uncorrected current-block plug-in target. Section 4.6 reports the reference-target comparison; it tests the selected estimators in one regime rather than determining the optimal estimator as block size or drift changes.

B.3 Reward block and notation

For a fixed task, the policy samples K trajectories, each evaluated by the same c variants. Let R_{ijk} denote the score of trajectory i under evaluator j in replication k. In the balanced design, $\bar{R}_{ij} = r^{-1} \sum_{k=1}^{r} R_{ijk}$ is the cell mean. Dots indicate arithmetic averaging over the corresponding indices: $\bar{R}_{i\cdot}$, $\bar{R}_{\cdot j}$, and $\bar{R}_{\cdot\cdot}$ denote row, column, and grand means.

$$R_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i \in [K],\ j \in [c],\ k \in [r]. \tag{8}$$

The ANOVA parameterization imposes $\sum_i \alpha_i = \sum_j \beta_j = 0$ and zero row and column sums for γ. These parameters describe the chosen evaluator panel; external task utility and gaming labels are separate quantities (Table 5).

Assumption 1 (Full-crossing ANOVA)**.** *(A1) Model (8) holds with conditionally mean-zero cell noise and finite variance; (A2) every (i, j) cell is observed with the same replication count r; (A3) harness assignment and replication noise are independent of trajectory identity conditional on the*

Table 5: Observed ANOVA quantities and external quantities that the reward grid alone does not identify.

Quantity	Symbol	Interpretation
Row	$i \in [K]$	Trajectory (policy rollout)
Column	$j \in [c]$	Harness variant (evaluator configuration)
Grand mean	μ	Average reward across trajectories and variants
ANOVA row effect	α_i	Centered row-mean estimand for the chosen harness set
Column effect	β_j	Centered harness main-effect estimand
ANOVA interaction	γ_{ij}	Double-centered trajectory$\times$harness residual
Cell noise	ε_{ijk}	Grading stochasticity (zero for deterministic graders)
External utility	$U(\tau_i)$	Transferable task utility from an independent anchor
Exploitation label	Z_i or Z_{ij}	External gaming label from a separately specified adjudication process

fixed crossed design; (A4) the sum-to-zero constraints define the ANOVA parameterization. These assumptions identify ANOVA parameters, not U or Z.

B.4 Dense penalty and block-wise normalization

The row effect and interaction estimates are

$$\hat{\alpha}_i = \bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot}, \tag{9}$$

$$\hat{\gamma}_{ij} = \bar{R}_{ij} - \bar{R}_{i\cdot} - \bar{R}_{\cdot j} + \bar{R}_{\cdot\cdot}. \tag{10}$$

Double centering removes evaluator-wide shifts before measuring trajectory-specific residual energy. The dense penalty and final advantage are

$$p_i = \frac{1}{c}\sum_j \hat{\gamma}_{ij}^2, \qquad A_i = (\hat{\alpha}_i - b) - \lambda p_i, \tag{11}$$

where b is an additional common within-group baseline. Although it cancels in pairwise advantage differences, it affects the absolute advantages. The block-specific penalty weight is

$$\lambda = \frac{\kappa}{\sqrt{q_\theta(P_+)}}, \qquad P_+ = \{p_i : p_i > 0\}, \tag{12}$$

with $\kappa = 0.15$, $\theta = 0.5$, and $\lambda = 0$ for $P_+ = \varnothing$. The quantile is computed over positive penalties in the current block; λ is recomputed for each block.

Algorithm 2 Dense ISA advantage for one crossed block

Require: Complete cell-score matrix $\bar{R} \in \mathbb{R}^{K\times c}$; common baseline b; $\kappa = 0.15$, $\theta = 0.5$
Ensure: Advantages $A_1, \ldots, A_K$
1: Compute row means, column means, and the grand mean of $\bar{R}$
2: $\hat{\alpha}_i \leftarrow \bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot}$
3: $\hat{\gamma}_{ij} \leftarrow \bar{R}_{ij} - \bar{R}_{i\cdot} - \bar{R}_{\cdot j} + \bar{R}_{\cdot\cdot}$
4: $p_i \leftarrow c^{-1}\sum_j \hat{\gamma}_{ij}^2$; $P_+ \leftarrow \{p_i : p_i > 0\}$
5: **if** $P_+ = \varnothing$ **then**
6: $\quad\lambda \leftarrow 0$
7: **else**
8: $\quad\lambda \leftarrow \kappa/\sqrt{q_\theta(P_+)}$
9: **end if**
10: **return** $A_i = (\hat{\alpha}_i - b) - \lambda p_i$ for every i

Scale and degenerate blocks. The penalty p_i has squared-reward units and λ has inverse-reward units. Under positive proportional rescaling $\bar{R} \mapsto s\bar{R}$, $p_i \mapsto s^2 p_i$ and $\lambda \mapsto s^{-1}\lambda$, so the complete advantage scales by s when b also scales by s. This property applies to the fixed score-to-advantage

map; nonlinear score transformations and changes of evaluator fall outside its scope. If $K = 1$ or $c = 1$, double centering gives $\hat{\gamma} = 0$ and the output is $\hat{\alpha}_i - b$. With $K = 2$, the residual rows are opposites and their energies are equal, so the penalty leaves their advantage difference unchanged. These identities concern computed advantages and do not establish equivalence of complete training procedures.

Policy-gradient interpretation. For a sampled crossed block $\mathcal{B}$, we define the token-averaged local policy surrogate

$$\mathcal{J}_{\mathrm{ISA}}(\phi; \mathcal{B}) = \frac{1}{K} \sum_{i=1}^{K} \frac{\mathrm{sg}[A_i(\mathcal{B})]}{L_i} \sum_{t \in \mathcal{T}_i} \log \pi_\phi(a_{it} \mid h_{it}), \tag{13}$$

Here ϕ denotes policy parameters (θ remains the quantile level), $A_i(\mathcal{B})$ is defined in Eq. (11), $\mathcal{T}_i$ indexes policy-generated tokens, and $L_i = |\mathcal{T}_i|$. Stop-gradient sg fixes the computed advantages; sampled trajectories, histories, and lengths are also held fixed. This surrogate is the negative policy term of Eq. (20), separate from its 0.01-weighted categorical KL penalty to the frozen initial policy. Setting $b = 0$ removes only the additional baseline: $\hat{\alpha}_i$ still subtracts the sample group mean, which, like p_i and λ, depends on $\mathcal{B}$. This block-dependent local update rule does not establish an unbiased gradient of a fixed population interaction-risk objective. Each fresh group receives one on-policy update, without subsequent advantage centering, SD normalization, advantage clipping, or PPO ratio clipping. Appendices F.2 and F specify the complete loss, accumulation weights, and training configuration.

B.5 Worked example: penalty magnitude and pairwise ranking

Consider a synthetic $K = 3$, $c = 4$ block with externally specified utilities $U = (0.6, 0.2, 0.5)$. By construction, trajectory τ_3 receives an additional reward $e = 0.6$ from evaluator 2, and is labeled as exploiting that evaluator. The values are illustrative scores rather than benchmark results.

	g_1	g_2	g_3	g_4	row mean
τ_1	0.60	0.60	0.60	0.60	0.600
τ_2	0.20	0.20	0.20	0.20	0.200
τ_3	0.50	**1.10**	0.50	0.50	**0.650**

The external ordering is $\tau_1 \succ \tau_3 \succ \tau_2$, whereas row averaging places τ_3 above τ_1 by 0.05. Double centering gives $\hat{\alpha} \approx (0.1167, -0.2833, 0.1667)$ and $\hat{\gamma}_{3\cdot} = (-0.10, 0.30, -0.10, -0.10)$. The dense penalties are $p = (0.0075, 0.0075, 0.0300)$, so the deployed rule gives $\lambda = 0.15/\sqrt{0.0075} \approx 1.7321$. Consequently,

$$A_3 - A_1 = 0.05 - \lambda(0.0300 - 0.0075) \approx 0.01103.$$

The penalty narrows the gap but does not reverse this pair. For a fixed externally labeled pair (e, h) with $A_e^{\mathrm{DR}} > A_h^{\mathrm{DR}}$ and $p_e > p_h$, Lemma 1 gives the tie threshold

$$\lambda^\star_{e,h} = \frac{A_e^{\mathrm{DR}} - A_h^{\mathrm{DR}}}{p_e - p_h}. \tag{14}$$

In this example, the threshold is $0.05/0.0225 \approx 2.2222$, above the deployed weight. The external labels explain the example but are not required by the pairwise algebra. The threshold is therefore a diagnostic of a fixed pair, rather than a training-time weight selector or evidence that interaction energy identifies gaming.

B.6 Row averaging and identifiability

Double-centered interaction has zero row sums and therefore averages to zero across evaluators. A $1/c$ dilution applies to an isolated raw reward increment, not to the interaction itself. To make this distinction explicit, consider the raw-score model

$$R_{ij} = \mu + U(\tau_i) + \beta_j + E_{ij} + \varepsilon_{ij}, \tag{15}$$

where E_{ij} is a harness-specific increment, interpreted as exploitation only with separate labeling evidence. For this raw-score model, write $\bar{R}_{i\cdot}$ for the evaluator average. Under $\sum_j \beta_j = 0$ and conditionally mean-zero noise,

$$\mathbb{E}_\varepsilon[\bar{R}_{i\cdot} \mid i] = \mu + U(\tau_i) + \bar{E}_{i\cdot}, \qquad \bar{E}_{i\cdot} = \frac{1}{c}\sum_j E_{ij}. \tag{16}$$

A single-cell increment e contributes e/c to the expected row mean. However, replacing $U(\tau_i)$ by $U(\tau_i) + a_i$ and E_{ij} by $E_{ij} - a_i$ leaves every score unchanged. Consequently, the reward grid cannot separate external utility from row-average reward increments, including increments caused by evaluator-shared exploits. The dilution identity does not resolve this non-identifiability.

At the conditional-mean level, double centering yields $\Gamma = H_K E H_c$, where H_K and H_c are centering matrices. For a single-cell increment, the corresponding interaction cell is $e(1-1/K)(1-1/c)$, but the transformation does not recover a general raw E. As noted above, interaction need not indicate exploitation, and the observed plug-in penalty includes grading noise. Appendix C gives the estimation properties, and Appendix I.6.1 examines the distinction from gaming labels.

C Estimation Properties and Scope of ISA

Full crossing identifies the panel-relative row effects and interaction residuals under the assumptions below. We distinguish these estimation results from the behavior of the squared plug-in penalty and from its effect on a fixed pair of advantages. Partial-observation and sparse-recovery extensions require additional assumptions. Appendix D gives the proofs, and Appendix E provides numerical illustrations.

C.1 Estimation under full crossing

Both results use Assumption 1. The variance formulas and consistency statements additionally assume independent, homoskedastic replication noise with variance σ^2, with K and c fixed as $r \to \infty$.

Theorem 1 (ANOVA row-effect estimation)**.** *The estimator in Eq. (9) is unbiased for α_i at every finite r. Under the noise conditions above,*

$$\mathrm{Var}(\hat{\alpha}_i) = \frac{\sigma^2}{cr}\left(1 - \frac{1}{K}\right),$$

and $\hat{\alpha}_i$ is consistent as $r \to \infty$.

Theorem 2 (Interaction estimation)**.** *The estimator in Eq. (10) is unbiased for γ_{ij} at every finite r and recovers it exactly for deterministic graders with $r = 1$. Under the noise conditions above,*

$$\mathrm{Var}(\hat{\gamma}_{ij}) = \frac{\sigma^2}{r}\left(1 - \frac{1}{K}\right)\left(1 - \frac{1}{c}\right),$$

and $\hat{\gamma}_{ij}$ is consistent as $r \to \infty$. Replication with $r \geq 2$ permits estimation of within-cell noise variance.

These are panel-relative ANOVA quantities. Under Eq. (15), the row effect equals $U(\tau_i) + \bar{E}_{i\cdot}$ up to a common centering constant. Full crossing therefore identifies the row effect and interaction, while separating external utility from an evaluator-shared increment still requires an external anchor or an additional restriction.

C.2 Implications for the deployed penalty

The deployed penalty $p_i = c^{-1}\sum_j \hat{\gamma}_{ij}^2$ is the squared energy of an estimated interaction row. Even when $\hat{\gamma}_{ij}$ is unbiased, its squared value contains estimator variance in expectation. Thus p_i is not generally unbiased for latent interaction energy, and heterogeneous grading noise can alter relative penalties. ISA uses this observed plug-in quantity without noise correction.

Proposition 3 interprets the observed penalty through a deterministic bound on residual score changes under reweighting of the same panel. This guarantee concerns the observed interaction row, rather than recovery of latent interaction energy. It does not cover evaluator-wide score shifts, external utility, or failure modes outside the observed panel.

C.3 Pairwise effect of penalization

Lemma 1 (Pairwise ranking threshold)**.** *Fix two trajectories i and ℓ, a penalty vector p, and row-mean advantages with $A_i^{\mathrm{DR}} > A_\ell^{\mathrm{DR}}$. For $A_j = A_j^{\mathrm{DR}} - \lambda p_j$ and $\lambda \geq 0$, if $p_i > p_\ell$, then*

$$A_i \leq A_\ell \quad \Longleftrightarrow \quad \lambda \geq \frac{A_i^{\mathrm{DR}} - A_\ell^{\mathrm{DR}}}{p_i - p_\ell}.$$

If $p_i \leq p_\ell$, no nonnegative λ reverses the pair.

For an externally labeled exploiter–honest pair (e, h), the threshold is $\lambda^\star_{e,h}$ in Eq. (14). The lemma isolates the effect of a penalty contrast for fixed scores and a fixed pair. It neither establishes that gaming trajectories receive larger penalties nor ensures that the deployed normalization reaches the threshold. Equal penalties cannot separate the pair, including when an evaluator-shared exploit has the same interaction energy as an honest trajectory. The statement is specific to the row-mean-plus-penalty rule and does not rule out Pess-min, MMSD, or other aggregators.

C.4 Conditional extensions

Neither extension below is used in the reported training procedure. Their additional premises are collected in Assumption 2 and discussed in Appendix D.5.

Assumption 2 (Conditional structural reductions)**.** *(A5) For the partial-observation extension in Section C.4, the observation set is generated by a specified random sampling design with positive inclusion probabilities, its bipartite observation graph is connected and sufficiently covered, main effects and interaction are jointly estimated with the same weights, Γ is rank-ρ and μ_0-incoherent, and the cited completion theorem's sample-size and noise conditions hold. (A6) For the sparse-demixing extension in Section C.4, $R = L + E$ satisfies the cited robust-PCA rank–sparsity transversality condition, including support concentration and the stated tuning of the convex program. Neither (A5) nor (A6) is established by our experiments.*

Partial observation. Matrix-completion results require a specified sampling design, adequate observation-graph coverage, and a jointly identifiable treatment of main effects and low-rank interaction, together with incoherence, noise, and sample-size conditions (Candès & Recht, 2009; Keshavan et al., 2010). The present work does not instantiate such an estimator and error theorem. We therefore claim neither a partial-observation recovery rate nor a grading-cost reduction; the experiments use full crossing.

Sparse demixing. An optional decomposition $R = L + E$ can be related to existing rank–sparsity demixing results under their support, geometry, and tuning conditions (Chandrasekaran et al., 2011). The sufficient condition considered here, $\mu(L)\xi(E) < 1/6$, is unmet on all 40 reported tasks (median product 2.48). This leaves the empirical regime uncertified by that result; it does not establish that recovery is impossible. The deployed dense penalty makes no sparse-support claim, and the archived thresholded heuristic inherits no recovery guarantee from this extension.

D Proofs and Reductions

D.1 Theorem 1: ANOVA row-effect identification

Proof. Under full crossing,

$$\mathbb{E}[\bar{R}_{i\cdot\cdot}] = \mu + \alpha_i + \tfrac{1}{c}\sum_j \beta_j + \tfrac{1}{c}\sum_j \gamma_{ij} = \mu + \alpha_i.$$

The expected grand mean is μ, hence $\mathbb{E}[\hat{\alpha}_i] = \alpha_i$. Under independent homoskedastic cell noise, direct covariance calculation gives $\mathrm{Var}(\hat{\alpha}_i) = \sigma^2(cr)^{-1}(1 - 1/K)$, which is finite for finite r and tends to zero only as $r \to \infty$.

This identifies the constrained ANOVA parameter. Under the raw model, centering Eq. (16) across rows yields

$$\alpha_i = \{U(\tau_i) + \bar{E}_{i\cdot}\} - K^{-1} \sum_{i'} \{U(\tau_{i'}) + \bar{E}_{i'\cdot}\}.$$

For arbitrary constants a_i, the transformation $U(\tau_i) \mapsto U(\tau_i) + a_i$, $E_{ij} \mapsto E_{ij} - a_i$ preserves every R_{ij}. Therefore U is not identified without an anchor or restriction on E. □

D.2 Theorem 2: Interaction estimation

Proof. For deterministic rewards, substituting Eq. (8) into the two-way contrast cancels μ, α_i, and β_j, leaving $\hat{\gamma}_{ij} = \gamma_{ij}$. For stochastic rewards, linearity and mean-zero noise give unbiasedness at any finite r. Under independent homoskedastic noise, the contrast variance is

$$\mathrm{Var}(\hat{\gamma}_{ij}) = \frac{\sigma^2}{r}\left(1 - \frac{1}{K}\right)\left(1 - \frac{1}{c}\right).$$

When $r = 1$, the saturated layout has no within-cell residual degrees of freedom, so σ^2 is not estimable from that layout alone. When $r \geq 2$, replication supplies within-cell residual degrees of freedom and makes σ^2 estimable. The estimator itself is finite-r unbiased; consistency follows only along $r \to \infty$. □

D.3 Lemma 1: Pairwise ranking threshold

Proof. Subtracting the two penalized advantages gives

$$A_i - A_\ell = (A_i^{\mathrm{DR}} - A_\ell^{\mathrm{DR}}) - \lambda(p_i - p_\ell).$$

When $p_i > p_\ell$, solving $A_i - A_\ell \leq 0$ gives the stated threshold. When $p_i = p_\ell$, the positive initial difference is unchanged; when $p_i < p_\ell$, every nonnegative λ can only increase it. The argument holds for any fixed pair and penalty vector and does not use the deployed rule for λ. □

D.4 Proof of Proposition 3

Proof. The residual score change is $(w - w')^\top \hat{\gamma}_i$. Cauchy–Schwarz gives $|(w - w')^\top \hat{\gamma}_i| \leq \|w - w'\|_2 \|\hat{\gamma}_i\|_2 = \sqrt{c}\|w - w'\|_2 \sqrt{p_i}$. The uniform-draw statement follows from $\mathbb{E}_j|\hat{\gamma}_{ij}| \leq (\mathbb{E}_j \hat{\gamma}_{ij}^2)^{1/2} = \sqrt{p_i}$. □

D.5 Premises for conditional structural extensions

Partial observation. With missing cells, ordinary observed row and column means need not estimate their full-layout counterparts without bias. Disconnected observation graphs also allow component-wise shifts between row and column effects that preserve every observed entry. A suitable extension must therefore specify the sampling design and weighting, resolve additive identifiability, and estimate main effects jointly with the interaction. Low rank alone does not supply these properties. Matrix-completion guarantees can only be invoked after matching the estimator and observation model to a specific result's incoherence, coverage, sample-size, and noise assumptions (Candès & Recht, 2009; Keshavan et al., 2010). Assumption 2(A5) records these requirements; no joint estimator or standalone rate is established here.

Sparse demixing. In the raw model, the additive matrix

$$L = \mu \mathbf{1}\mathbf{1}^\top + \boldsymbol{\alpha}\mathbf{1}^\top + \mathbf{1}\boldsymbol{\beta}^\top$$

has low rank. This motivates considering the existing convex program $\min_{L,E} \|L\|_* + \eta\|E\|_1$ subject to $R = L + E$. Its recovery conditions concern both the geometry of the low-rank spaces and the arrangement of sparse support, together with the tuning of η (Chandrasekaran et al., 2011). A

minority of nonzeros in each row is insufficient, since column concentration and alignment with the low-rank component also matter. Section C.4 reports the failure of the stated sufficient condition in our empirical regime. Consequently, this literature does not certify the archived sparse attribution or the implemented dense penalty.

E Numerical Properties and Implementation Boundaries

The examples below illustrate estimator accuracy, pairwise advantage changes, and the numerical limits of normalization. They complement the formal results without proving that the assumptions hold in the policy experiments.

E.1 Numerical illustrations

Row-effect estimation. At $r = 8$, the maximum error is $\max|\hat{\alpha} - \alpha| = 0.028$. Theorem 1 characterizes how estimator variance depends on replication.

Interaction estimation. With deterministic evaluators and $r = 1$, the maximum error is $\max|\hat{\gamma} - \gamma| = 2.8 \times 10^{-16}$, consistent with floating-point precision. With stochastic evaluators and $\sigma = 0.30$, RMSE is 0.277, 0.075, and 0.046 at $r = 1$, 8, and 32, respectively.

Pairwise ranking. For one externally labeled pair and a fixed penalty vector, the computed threshold produces a tie; a larger weight gives a penalized advantage difference of -0.100. This illustrates the fixed-pair algebra of Lemma 1, without establishing that the deployed normalization selects that threshold.

Undeployed sparse-demixing example. A separate 12×6 synthetic layout with five planted sparse entries gives support match $5/5$ after 50 alternations. This example concerns the optional sparse extension, not the dense penalty used in training. Its scope is governed by the conditions discussed in Section C.4.

E.2 Conditions for nonzero interaction

A nonzero interaction requires $K \geq 2$ and $c \geq 2$; either dimension equal to one makes double centering vanish. These conditions permit, but do not guarantee, useful discrimination across trajectories. Deterministic full crossing yields an exact ANOVA decomposition. Under stochastic scoring, $r \geq 2$ permits estimation of within-cell noise variance; finite-r unbiasedness and consistency as $r \to \infty$ require the conditions in Appendix C.

The experiments use full crossing. Partial-observation cost reductions require additional sampling and recovery assumptions, while sparse-support claims require separate structural conditions (Section C.4). Support stability and honest-row leakage are relevant to sparse attribution; they are not support-recovery requirements for the deployed dense penalty.

E.3 Numerical implementation and normalization behavior

Exact score arithmetic. In the primary implementation, finite decimal rewards retain their original representations and are treated as exact rational values for replication averaging and double centering. For a block-specific common denominator L, write $\bar{R}_{ij} = X_{ij}/L$ and define

$$r_i = \sum_j X_{ij}, \qquad h_j = \sum_i X_{ij}, \qquad t = \sum_{i,j} X_{ij}, \qquad N_{ij} = KcX_{ij} - Kr_i - ch_j + t.$$

Then

$$\hat{\gamma}_{ij} = \frac{N_{ij}}{LKc}, \qquad \hat{\alpha}_i = \frac{Kr_i - t}{LKc}, \qquad p_i = \frac{\sum_j N_{ij}^2}{c(LKc)^2}.$$

Arbitrary-precision integer and rational intermediates preserve exact zeros of the input decimal scores. Square roots, λ, and advantages are computed in float64. The block-dependent denominator L does not impose a common scoring grid. Finite-conversion checks are applied to every

block; the positive-energy median, λ, maximum scaled penalty and absolute advantage, and gradient norms before and after clipping are monitored during training. Global gradient clipping at 1.0 limits the gradient passed to the optimizer.

Small positive medians can amplify finite penalties. For an algebraic example with $K = 14, c = 4$, take $v = (1, -1, 1, -1)$ and let $a_1 = 1/4$, $a_2 = -1/4$, with the remaining twelve a_i consisting of six $+\epsilon$ and six $-\epsilon$, where $0 < \epsilon < 1/4$. Set

$$\bar{R}_{ij} = \tfrac{1}{2} + a_i v_j.$$

All rewards lie in $[1/4, 3/4]$. Row and column means are $1/2$, so $\hat{\alpha}_i = 0$, $\hat{\gamma}_{ij} = a_i v_j$, and $p_i = a_i^2$. Twelve positive energies equal ϵ^2, giving $q_{0.5}(P_+) = \epsilon^2$ and

$$\lambda p_1 = \frac{0.15}{16\epsilon} \longrightarrow \infty \quad \text{as } \epsilon \downarrow 0.$$

Thus bounded rewards and exact zero handling do not provide a uniform penalty bound. At $\epsilon = 0$ the twelve zero energies are excluded, illustrating the role of the strictly positive-energy quantile. This algebraic stress case complements the observed training behavior: all 60 primary runs completed without non-finite loss, gradients, or parameters (Appendix F.4). The construction describes an extreme input regime, whereas the completed runs describe the realized regime. Appendix J.6 trains the bounded floor variant and shows that it lowers the realized advantage and gradient tails without harming discrepancy, reward, or success.

F Experimental Protocol and Reproducibility

The primary comparison uses a shared optimizer and sampling schedule, method-specific advantages, fixed evaluator panels, and seed-level evaluation. This appendix specifies each component and the task-selection procedure used for the held-out extension. The alternative-rule studies retain the settings stated in Section I.7.

F.1 Training configuration and advantage rules

The five crossed methods share a training schedule and valid-score budget and differ in the advantage rules defined below. GRPO uses a single evaluator and a smaller grading budget. Section I.1.1 reports the corresponding primary outcomes.

Training setup. The six arms use Qwen3-8B, three epochs, $K = 14$, learning rate 10^{-6}, four accumulation microbatches, and 240 optimizer updates per run. The ten training seeds are 42, 49, 56, 63, 70, 77, 84, 91, 98, and 105, giving 60 completed training runs. Each run uses four A800-80GB GPUs for distributed policy training. The task set and task order are shared across methods, while each policy generates its own trajectories. The heterogeneous frozen-policy pilot retains its separate evaluator setup in Section I.4.1.

Baseline advantages. Write $u_i = \bar{R}_{i\cdot}$ and $m = \bar{R}_{\cdot\cdot}$ for the current group. All score-derived quantities are detached during differentiation. The actual advantages are

$$A_i^{\mathrm{DR}} = u_i - m, \qquad A_i^{\mathrm{MMSD}} = u_i - m - 0.50 s_i, \tag{17}$$

$$A_i^{\min} = \min_j \bar{R}_{ij} - m, \qquad s_i = \sqrt{\frac{1}{c-1}\sum_j (\bar{R}_{ij} - u_i)^2}, \tag{18}$$

$$A_i^{\mathrm{DR+}} = v_i - \frac{1}{K}\sum_\ell v_\ell, \qquad v_i = \sum_j w_j \bar{R}_{ij}. \tag{19}$$

ISA uses $A_i = \hat{\alpha}_i - \lambda p_i$, with $b = 0$, $\kappa = 0.15$, and $\theta = 0.5$. Its positive-energy median is the middle sorted value, or the arithmetic mean of the two middle values for an even count. After subtracting the penalty, advantages are neither re-centered, divided by their SD, nor clipped; thus their group mean is $-\lambda K^{-1}\sum_i p_i$. For the single-evaluator GRPO reference, $x_i = \bar{R}_{i1}$ and

$$A_i^{\mathrm{GRPO}} = \frac{x_i - \bar{x}}{\sqrt{K^{-1}\sum_\ell (x_\ell - \bar{x})^2 + 10^{-6}}}, \qquad \bar{x} = K^{-1}\sum_i x_i.$$

DR+ maintains running counts, means, and sums of squared deviations $M_{2,j}$ of historical cell means across tasks, separately for each run. Its first group uses $w_j = 1/c$. With N historical trajectories, subsequent groups use

$$V_j = \frac{M_{2,j}}{N-1}, \qquad w_j = \frac{(V_j + 10^{-4})^{-1}}{\sum_\ell (V_\ell + 10^{-4})^{-1}}.$$

DR+ fixes the weights before processing the current group and adds that group's cell means to its history only after a successful update. Replications are averaged within cells, rather than counted as additional historical trajectories. The weights therefore reflect marginal reward variance, not an estimate of optimal noise weights.

F.2 Policy optimization and trajectory generation

Let $\mathcal{T}_i$ contain the policy-generated token positions in trajectory i, and let $L_i = |\mathcal{T}_i|$. Policy-generated tool calls and EOS tokens are included; system/user prompts and environment or tool returns are excluded. We minimize

$$\begin{aligned}\mathcal{L}(\phi) = &-\frac{1}{K}\sum_i \frac{1}{L_i}\sum_{t\in\mathcal{T}_i} A_i \log \pi_\phi(a_{it} \mid h_{it}) \\ &+ \frac{0.01}{K}\sum_i \frac{1}{L_i}\sum_{t\in\mathcal{T}_i} D_{\mathrm{KL}}(\pi_\phi(\cdot \mid h_{it}) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid h_{it}))\,. \end{aligned} \tag{20}$$

The reference policy is the frozen initialization. We compute the full categorical KL over the vocabulary in chunks. Rewards, baselines, historical weights, λ, and advantages are detached from the gradient. Losses are averaged over policy tokens within each trajectory and then equally across trajectories. Each group receives one optimizer update, without PPO ratio clipping, an entropy bonus, or additional optimization passes.

Each epoch visits every task twice in a seed-specific fixed random permutation, giving $3 \times 40 \times 2 = 240$ fresh groups. Advantages are computed from the complete $14 \times c$ score block before dividing the trajectories into microbatches of 4, 4, 3, and 3. Their mean losses are weighted by $4/14$, $4/14$, $3/14$, and $3/14$, respectively. Parameters remain unchanged during accumulation and are updated once afterward. Four-GPU sharding preserves these global weights and the original group size.

Optimization and generation. The instruction-capable Qwen3-8B policy uses its corresponding tokenizer and chat template, with thinking disabled and all parameters fine-tuned. AdamW uses constant learning rate 10^{-6}, betas $(0.9, 0.999)$, epsilon 10^{-8}, and zero weight decay, without warmup or decay. Global gradient L2 norm is clipped to 1.0 after full-group accumulation and unscaling, where applicable. Appendix I.9 reports the pre-clipping norm and the resulting factor $\min(1, 1/(G + 10^{-6}))$ for the Raw/DC comparison. Weights and main forward computation use bfloat16; probability calculations, loss reductions, gradient norms, and optimizer states use float32. Activation checkpointing and parameter/optimizer-state sharding distribute training across the four GPUs. Dropout is disabled during sampling, gradient computation, and reference forwards.

Sampling uses temperature 1.0 and top-p 1.0, without top-k or other probability truncation, best-of-n, rejection sampling, or reward-based filtering. Generation and loss probabilities use the same full-vocabulary policy distribution. Each trajectory permits at most 8,192 policy-generated tokens and 32 assistant decision turns. A policy call has at most 32,768 input-plus-output tokens. With H current input tokens and G previously generated policy tokens, its remaining output budget is $\min(8192 - G, 32768 - H)$. A nonpositive budget terminates the trajectory; old history is not truncated to continue generation. Valid incomplete trajectories and model-generated tool failures remain in scoring. An EOS-only output has at least one policy token; an infrastructure failure before any policy token is generated is retried after restoring the task and random state.

Evaluation uses only the policy after update 240. Checkpoints at updates 80 and 160 are used for recovery, not selection by held-out performance. Recovery restores the optimizer state, random streams, task cursor, and DR+ history together. Matching seeds pairs training runs without requiring different policies to generate identical trajectories.

F.3 EVALUATOR PANELS AND REWARD CONSTRUCTION

A1–A6, D1–D2, and B/C are auxiliary evaluators of a common immutable evidence bundle. Their scores lie in $[0, 1]$ and are labeled as continuous component scores or semantic completion probabilities. These are measurement choices, not official leaderboard metrics. N0 applies the official binary rule with two distinct roles: $N0_{\mathrm{cal}}$ fits external-baseline maps on calibration evidence, while $N0_{\mathrm{confirm}}$ evaluates the task-success, reference-target, and Raw/DC studies. Current-block ISA does not use either set of labels in training. Table 6 distinguishes the evaluators individually.

Table 6: Evaluator roles and score construction. D means deterministic under fixed inputs and snapshots; S means stochastic. Auxiliary scores share a range but need not share calibration or semantics.

ID	Permitted use	Construction	Type
A1	Training; GRPO reference	Fixed weighted aggregation of auditable benchmark subchecks	D
A2	Crossed training	Independent parser and component recomputation; discrepancies with A1 retained	D
A3	Crossed training	Prespecified representation normalization, then A1 components and weights	D
A4	Crossed training	Frozen semantic verifier returning a completion probability	S
D1	Development only	Component, missing-component, and aggregation-rule development	D
D2	Development only	Separate semantic verifier for prompt/parser and coefficient selection	S
A5	Confirmation only	Minimum of required component scores; strict evidence validation	D
A6	Confirmation only	Separate model, prompt, and parser with the A4 evidence schema	S
B	External evaluation	Independently implemented component scoring and harmonic-mean aggregation	D
C	External evaluation	Container replay of recorded actions/patch; auxiliary completion fractions	D
N0	Calibration / confirmation	Official binary success; calibration labels fit only external maps; confirmation evaluates task-success, reference-target, and Raw/DC studies	D

Auxiliary components and official success. For tau-bench, A1 combines the fraction of required state fields satisfied, s, with the fraction of required outputs satisfied, o: $0.6s + 0.4o$ when outputs are required, and s otherwise. For SWE-bench/Lite, it uses $0.5f + 0.5p$, where f and p are the passed fractions in FAIL_TO_PASS and PASS_TO_PASS; weights are renormalized over nonempty groups under the fixed configuration. For Toolathlon, it uses the fraction of public atomic checks passed, reducing to a binary score where only a total pass/fail result is available. Denominators, empty groups, and missing components are validated explicitly. A5 applies the minimum to required components; B applies their harmonic mean, with score zero when a required component is zero. These auxiliary aggregations can assign partial credit even when official success is binary.

Table 7: Official binary endpoint N0, computed directly from benchmark evidence. Auxiliary component scores are never thresholded to infer these outcomes.

Benchmark	Evidence	N0 success criterion
tau-bench	Initial/terminal database state and agent-to-user outputs	Official adapter matches the target database and all required outputs
SWE-bench/Lite	Instance ID and generated patch on the frozen base commit	Official resolved: FAIL_TO_PASS all pass and PASS_TO_PASS remain acceptable
Toolathlon	Final workspace/application state	Official task-specific evaluator returns pass

Tau-bench evaluates tool-agent-user interactions (Yao et al., 2024); SWE-bench evaluates issue resolution by patches (Jimenez et al., 2023); and Toolathlon evaluates long-horizon tool use (Li et al., 2025). The benchmark definitions determine N0. Neither a semantic verifier's completion probability nor a continuous component average replaces the official binary criterion. A4/A6 completion probabilities are auxiliary scoring signals; probability calibration is not evaluated.

Common evidence and representation changes. Stateful tau-bench and Toolathlon inputs include the initial state, ordered tool calls with arguments and results, terminal snapshot, and environment version. SWE-bench evaluators inspect the same generated patch and associated test evidence, not a newly generated solution. A3 applies prespecified Unicode, line-ending, and representation normalization while retaining the original task evidence. Removal of nonsemantic metadata is allowlisted; it does not justify changing task requirements. Semantic evaluators receive the task requirements and evidence without hidden test answers. A4/A6 use temperature 0.7, top-p 1.0, and at most 1,024 output tokens; D2 uses temperature 0.3 and at most 768 output tokens. Model snapshots, prompts, and parsers are fixed within a protocol. Native main-study results use aggregation P, native ablation results use Q, and the sensitivity analysis applies Q to both frozen ISA policies. Applying P versus Q changes neither the trajectory nor its training-panel reward; their effect cannot be inferred from a cross-cohort mean difference.

Independent implementations and replay coverage. B does not reuse the Family-A scorer implementation. Its implementation is independent, but the benchmark specification and evidence remain shared. C restores the initial environment in a separate container and replays recorded actions or applies the same patch; it never resamples policy or user-model outputs. Stateful replay requires matching terminal-state hashes after allowlisted removal of nonsemantic fields. SWE replay requires the same patch hash and a successfully started harness. Tasks requiring fresh simulated-user responses or unfrozen dynamic data are marked not replayable. Family-C results are conditional on the common replayable subset. Replay coverage is not quantified in the reported summaries, limiting extrapolation to the full task pool. C's binary consistency field measures agreement under replay; N0 separately measures official task success.

Replication, validation, and retries. Each auxiliary evaluator is invoked four times per cell. For deterministic evaluators these are service-integrity checks and must agree exactly; they are not four independent semantic measurements. Stochastic evaluators use four independently seeded replications and their arithmetic mean defines the raw cell score; the held-out aggregation rule is applied separately as specified in Appendix I.1.1. Each replication allows at most three attempts, accepting its first valid return. Retries address invalid structure, mismatched identities, non-finite values, or infrastructure failure before scoring. A valid low score, test failure, patch-application failure, or maximum-turn termination is a task outcome and must not trigger a retry. The same evidence is retained across attempts.

All four valid replications are required for a complete cell. Exhausting one replication's retry budget marks the cell incomplete; averaging only the remaining replications would violate the balanced design. Deterministic disagreement is an integrity failure. Missing scores are not imputed and unsuccessful valid trajectories are retained. Paired analyses use complete method pairs and their actual degrees of freedom. The first valid response is accepted independently of its score magnitude.

Evaluator roles and provenance. A1–A4 supply training rewards; D1–D2 supply development feedback; A5–A6 supply same-family confirmation; B/C supply external auxiliary measurements. Separate continuous human reference scores fit aggregation Q, whereas $N0_{\mathrm{cal}}$ supplies initial-policy binary labels used only for the frozen external-baseline maps; neither replaces the other. $N0_{\mathrm{confirm}}$ supplies the task-success, reference-target, and Raw/DC evaluation outcomes. A5–A6/B/C and all confirmation outcomes are excluded from coefficient, evaluator/prompt/parser, and checkpoint selection. Current-block ISA uses no N0 labels; historical and shrinkage offsets also depend only on scores. Configurations are frozen before confirmation, and a scoring-rule change receives a distinct protocol identity. This role separation is different from task-set disjointness: coefficient confirmation shares task IDs with training and calibration, and the three N090 cohorts share tasks but not rollout evidence.

Each accepted cell mean is associated with its replications, trajectory evidence, benchmark revision, task set, model snapshot, prompt/parser version, and execution environment. Invalid attempts are retained only for integrity and retry accounting; they do not contribute to cell averages.

F.4 Scoring cost and run completion

A crossed group requires $14\times4\times4 = 224$ valid grading replications, and GRPO requires $14\times1\times4 = 56$. Across 240 groups, the six-arm, ten-seed cohort uses 2,822,400 valid training replications. Invalid service returns are retried under the fixed policy above; all required replications and all 60 planned runs completed, with ten runs per arm, no replacement seeds, and no terminations due to non-finite loss, gradients, or parameters. The primary endpoint uses ten complete pairs, with $40 \times 8 = 320$ fresh evaluation trajectories per run and 19,200 across the six methods. Unsuccessful valid trajectories are retained. Four A800 GPUs support each policy-training run; benchmark execution and evaluator services require additional resources, so grading counts should not be interpreted as wall-clock cost. Calibration and confirmation sample sizes for the follow-up studies are reported with their corresponding analyses.

F.5 Task selection and evaluation separation

The task relationships are explicit: `T_MAIN` and `T_ABLATION` are disjoint 40-task sets with identical quotas, and `U180` excludes both. Candidate pools for each study were fixed before its policy training and outcome comparison. The coefficient-confirmation protocol allocates ten training-cohort tasks to development and 30 to confirmation. This establishes separation from coefficient selection, not disjointness of all evaluator-development and policy-training tasks. Eligibility concerned resettable environments, available tools, readable task requirements, and applicability of native evaluation, without using policy rewards, success rates, difficulty rankings, or interaction statistics. Deduplication used both benchmark-family task identifiers and exact normalized task content; SWE-bench-lite and SWE-bench shared a canonical SWE-family task key. Natural-language content used Unicode NFC and CRLF-to-LF normalization, retaining other whitespace, tool arguments, numeric values, and goal conditions; code retained its original bytes. Software tasks also checked the base repository state and task description.

Candidates were sorted by fixed protocol identifiers. For the primary study, task-selection seed 2026091801 sampled the 15/10/15 `T_MAIN` quotas without replacement. The residual-target ablation uses a separate `T_ABLATION` task set with the same quotas; its sampling seed was not retained, which prevents exact reconstruction of that draw. From the remaining pools, 60 tasks per family formed `U180` after excluding both training sets and exact normalized-content duplicates. The union of the two training sets and `U180` contains 260 unique task IDs; all three N090 cohorts use the same fixed 90-task subset of `U180`. The SWE extension draws from the SWE-bench family, beyond the lite training subset. Tasks are fixed across training seeds, and model failures do not trigger replacement. The zero-overlap result concerns identifiers and exact normalized content, not semantic near-duplicates or absence from model pretraining.

The 180 tasks, A5–A6, and final Family-B/C evaluations do not enter policy updates, checkpoint or coefficient selection, or evaluator development. The extension comprises three methods, ten seeds, and eight trajectories per task: 1,440 evaluation trajectories per method-seed and 43,200 overall, separate from the primary cohort's 19,200. Its bootstrap resamples paired seed indices and task indices within benchmark strata, using the same indices across methods. This accounts for seed and task variation and differs from the primary fixed-task paired-t analysis.

F.6 Consistency of computed advantages and endpoints

Author-side consistency checks covered all 14,400 training blocks (240 in each of 60 runs) and matched the specified score-to-advantage rules in every block; all blocks contain $K = 14$ rows. In the 2,400 ISA blocks, the penalty is $c^{-1}\sum_j \hat{\gamma}_{ij}^2$ and its weight is $\kappa/\sqrt{q_{0.5}(P_+)}$. Threshold and mask quantities used by alternative-rule diagnostics do not enter this dense penalty. Reconstructing an advantage additionally requires its baseline and, for DR+, the preceding variance state.

Author-side aggregation checks reproduce the signed and absolute primary endpoints from their underlying evaluation values to within 10^{-12}. These checks establish internal arithmetic consistency,

not independent reconstruction from the PDF. Statistical inference continues to use the training seed as its independent unit.

G Ranking Geometry and Fixed-Block Diagnostics

Lemma 2 (Main-effect contribution to row-wise raw energy)**.** *For a fully crossed observed block, let* $Z_i^{\mathrm{raw}} = (\bar{R}_{ij} - \bar{R}_{i\cdot})_{j=1}^{c}$, $B = c^{-1}\|\hat{\beta}\|_2^2$, *and* $p_i = c^{-1}\|\hat{\gamma}_i\|_2^2$. *Then*

$$Z_i^{\mathrm{raw}} = \hat{\beta} + \hat{\gamma}_i, \qquad p_i^{\mathrm{raw}} = B + p_i + x_i, \qquad x_i = \frac{2}{c}\langle \hat{\beta}, \hat{\gamma}_i \rangle.$$

Moreover, $K^{-1}\sum_i x_i = 0$ *and* $\overline{p^{\mathrm{raw}}} = B + \bar{p}$. *For two trajectories,*

$$p_i^{\mathrm{raw}} - p_\ell^{\mathrm{raw}} = (p_i - p_\ell) + \frac{2}{c}\langle \hat{\beta}, \hat{\gamma}_i - \hat{\gamma}_\ell \rangle.$$

Proof. Substitute the double-centering definition to obtain $Z_i^{\mathrm{raw}} = \hat{\beta} + \hat{\gamma}_i$, and expand its squared norm. Since each residual column sums to zero, $\sum_i \hat{\gamma}_i = 0$, so the averaged cross term vanishes. Subtracting the row identities gives the pairwise formula. □

Sum-to-zero constraints on $\hat{\beta}$ and individual residual rows do not imply $\langle \hat{\beta}, \hat{\gamma}_i \rangle = 0$. The ANOVA components are orthogonal over the complete block, while this row-wise inner product can be positive or negative. Consequently, subtracting B alone removes the block-average main-effect energy but does not remove the row-specific contamination. Double centering removes both. Raw sample SD equals $\sqrt{c/(c-1)}\sqrt{p_i^{\mathrm{raw}}}$ and preserves raw-energy ordering. Comparisons of complete advantages additionally depend on row rewards, coefficients, and normalization.

Contamination and rank diagnostics. Define $C_t = B_t/\bar{p}_t$ for task blocks with positive mean interaction energy. Then the ratio of mean raw energy to mean DC energy is exactly $1+C_t$. The ratio is undefined when $\bar{p}_t = 0$. Cauchy–Schwarz gives $|x_i| \le 2\sqrt{Bp_i}$; hence a vanishing main effect forces the cross term to vanish at fixed interaction energy. Large C_t need not imply a particular rank-reversal rate, because reversals also depend on row-energy gaps and the orientation of the residual vectors.

For each task, we compare the signs of $p_i^{\mathrm{raw}} - p_\ell^{\mathrm{raw}}$ and $p_i - p_\ell$ over all unordered trajectory pairs, excluding a pair if either difference is tied. Forty $K = 15$ blocks supply 4,200 candidate pairs before tie removal. Relationships with downstream discrepancy are descriptive: a task's geometric penalty reduction is distinct from its improvement in trained-policy D^{rollout}.

Detailed pilot geometry results. Recomputing the decomposition from all 40 15×4 cell-mean matrices gives a median contamination ratio of 7.54 (interquartile range $[6.75, 7.90]$), with all 40 ratios exceeding one. The range is $[6.24, 998.62]$; the upper extreme is driven by a small interaction-energy denominator ($B \simeq 0.0020753$, $\bar{p} \simeq 2.0782 \times 10^{-6}$). We therefore report the median and interquartile range rather than the outlier-sensitive mean. The observed cell means satisfy the row and block-average identities in Lemma 2.

Raw and double-centered energies reverse the penalty ordering in $1,763/4,200$ pairs (41.98%), with no ties. The median within-task reversal rate is 40.48%, ranging from 31.43% to 55.24% (Figure 3). These fixed-block contamination and rank-reversal statistics use the complete 15-trajectory blocks, including the planted controls. We do not report control-excluded estimates for these two diagnostics.

The task-level distributions are shown in Figure 3.

Aligning each task's contamination ratio with its ten-seed mean rollout-discrepancy improvement, $D_{\mathrm{MMSD}}^{\mathrm{rollout}} - D_{\mathrm{ISA}}^{\mathrm{rollout}}$, yields Pearson $r = 0.199$ and Spearman $\rho = 0.109$ over 40 tasks. These weak descriptive associations do not establish that greater contamination predicts larger performance gains. Moreover, ISA–MMSD changes more than residualization, so this association cannot replace the matched Raw-Energy/DC-Energy training contrast.

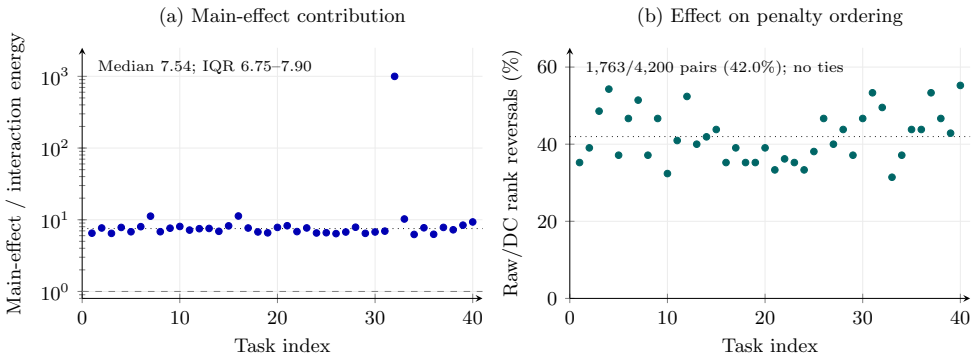


Figure 3: Shared evaluator shifts materially affect raw penalty signals in 40 frozen-policy blocks ($K = 15, c = 4$). (a) Main-effect energy divided by mean interaction energy, shown on a logarithmic scale; the dotted line marks the median and the dashed line marks one. (b) Fraction of unordered trajectory pairs whose raw/DC energy ordering reverses; the dotted line marks the pooled fraction. Each point represents one task; both panels describe penalty geometry in fixed scored blocks.

H Bounded Shifts, Composition Stability, and Reference Targets

This appendix reports three analyses with explicit inference units and role separation. The bounded-shift and companion-composition analyses use tasks as the inference unit; the reference-target comparison uses paired training seeds. The analyses were specified before their reported outcome comparisons, but the PDF does not provide timestamped preregistration evidence.

Bounded-shift feasibility. For each frozen score block, this audit evaluates Eq. (5) over the exact no-clipping box $[L, U]$. It reports the task-aggregated prevalence of pairs for which both raw orderings are feasible, the minimum ℓ_∞ reversal radius, and fixed-radius stress curves. Pairwise quantities are aggregated within task before a 10,000-draw task bootstrap. Author-side checks using rounded values reproduce the box extrema and tie-boundary radii to 4.6×10^{-8}; verification at the protocol tolerance of 10^{-10} requires full-precision cell scores that are not included in the PDF. Algebraically, all shifts in the box preserve the score range and double-centered energies. This is an existence and prevalence diagnostic, not a probability model for naturally occurring evaluator drift.

Companion composition. The composition audit and reference-target comparison share a calibration bank containing 64 independently sampled initial-policy trajectories for each task in `T_MAIN`, scored by evaluators A1–A4 with four valid replications per cell. A hash rule fixes eight anchors per task; each of 100 draws combines those anchors with six trajectories sampled without replacement from the remaining pool. Anchor penalties, ranks, top-quartile membership, and advantage signs are compared with the 64-trajectory full-pool reference. Absolute-agreement single-measure ICC, median Spearman correlation, top-quartile overlap, sign consistency, and anchor coefficient of variation are first computed per task, and uncertainty is then bootstrapped across tasks. These measures describe block-relative stability and carry no confirmatory efficacy threshold.

Reference-target comparison. Using the same frozen calibration bank, this study compares four targets at identical training and scoring budgets. Current-block and historical offsets are

$$\hat{\beta}_j^C = \bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot}, \qquad \hat{\beta}_j^H = \mathbb{E}_{\text{cal}}[R_{ij} - \bar{R}_{i\cdot}].$$

The external target fits and freezes evaluator-specific affine maps $Z_{ij} = g_j(\bar{R}_{ij})$ to $U_{\text{cal}} \in \{0, 1\}$, the official success labels on initial-policy calibration evidence, using task-grouped folds and ridge penalty $0.001b_j^2$ with an unpenalized intercept. Appendix H.1 gives all fitted coefficients and the five-fold membership rule. It subtracts a frozen calibration offset, $\hat{\beta}_j^E = \mathbb{E}_{\text{cal}}[Z_{ij} - \bar{Z}_{i\cdot}]$, giving $\hat{\gamma}_{ij}^E = Z_{ij} - \bar{Z}_{i\cdot} - \hat{\beta}_j^E$. This is reference residualization of calibrated scores, not current-block double centering of Z. Shrinkage uses $\hat{\beta}_j^S = \rho\hat{\beta}_j^C + (1-\rho)\hat{\beta}_j^H$, with $\rho \in \{0, 0.1, \ldots, 1\}$ selected

Table 8: Reference-target comparison on 40 tasks and 20 paired seeds. Values are mean±sample SD. All four arms use the same training grading budget and energy-penalty rule. The primary contrast is current ISA versus shrinkage.

Residual target	$D^{\mathrm{rollout}} \downarrow$	$D^{\mathrm{task}} \downarrow$	$R_{\mathrm{out}} \uparrow$
Historical	0.07359 ± 0.00242	0.06936 ± 0.00272	0.66806 ± 0.00338
External-anchor	0.06943 ± 0.00218	0.06541 ± 0.00325	0.67083 ± 0.00359
Shrinkage	0.06581 ± 0.00206	0.06243 ± 0.00302	0.67182 ± 0.00293
Current ISA	$\mathbf{0.06158 \pm 0.00223}$	$\mathbf{0.05797 \pm 0.00295}$	$\mathbf{0.67452 \pm 0.00338}$

only by task-grouped offset error without U_{cal}, giving $\rho = 0.4$. Twenty paired seeds compare historical, external, shrinkage, and current-block targets; the primary contrast is current-block minus shrinkage on fresh-rollout D^{rollout}. A policy-level superiority claim requires both a paired two-sided 95% interval with upper endpoint below zero and a one-sided 95% lower bound of at least -0.01 for the current-block-minus-shrinkage A5–A6 held-out-reward difference. $N0_{\mathrm{confirm}}$ and the other residual targets are secondary; the $N0_{\mathrm{confirm}}$ analysis resamples paired seed IDs and tasks within benchmark strata, and secondary confirmatory tests use Holm correction. All 80 method–seed outcomes satisfy $D^{\mathrm{global}} \le D^{\mathrm{task}} \le D^{\mathrm{rollout}}$. The intervals reported below are marginal intervals; secondary contrasts are presented descriptively without claiming completion of the protocol-wide Holm family, which also includes endpoints not reported in the paper.

Full reference-target comparison. The four arms share 20 paired seeds, 40 training tasks, $K = 14$, $c = 4$, $r = 4$, 240 updates, and the same energy normalization ($\kappa = 0.15$, $\theta = 0.5$). Historical offsets, affine maps, and the shrinkage weight $\rho = 0.4$ are obtained from the shared initial-policy calibration bank on `T_MAIN`; A5–A6 are reserved for evaluation. Each arm receives 53,760 valid training grades per seed, with 40,960 calibration grades and 2,560 $N0_{\mathrm{cal}}$ labels shared across the composition audit and reference-target comparison.

Current ISA reduces rollout discrepancy by 0.00423 relative to shrinkage (95% paired-t CI $[-0.00488, -0.00358]$), a 6.43% relative reduction. Held-out reward increases by 0.00270, with one-sided 95% lower bound $+0.00211$, above the -0.01 safeguard. The joint discrepancy and reward criterion is therefore met by the reported seed-level estimates. Historical and external-anchor targets also have higher mean discrepancies. On the distinct-evidence $N0_{\mathrm{confirm}}$ evaluation using the shared 90-task `N090` task set, current ISA succeeds on $9482/14400$ rollouts versus $9397/14400$ for shrinkage, a $+0.590$ percentage-point difference (reported paired-seed, benchmark-stratified task-bootstrap 95% CI $[+0.354, +0.840]$). This secondary result belongs to the reference-target comparison and is separate from the ten-seed DC-SD success comparison. The evidence favors current-block estimation among these fitted targets; it does not identify how that preference changes with calibration quality, block size, or drift.

Residual construction and budget accounting. For $M \in \{C, H, S\}$, $\hat{\gamma}_{ij}^{M} = \bar{R}_{ij} - \bar{R}_{i\cdot} - \hat{\beta}_{j}^{M}$; the external residual is defined above. Each method uses $p_i^M = c^{-1} \sum_j (\hat{\gamma}_{ij}^{M})^2$ and its own positive-energy median to scale the penalty, with the ISA zero-energy branch. The row-reward term always uses the original scores, including for the external-anchor arm. Thus affine calibration changes the penalty input rather than the task-reward objective. The four arms account for $4 \times 20 \times 53{,}760 = 4{,}300{,}800$ valid training grades, plus a study-shared calibration bank of $40 \times 64 \times 4 \times 4 = 40{,}960$ grades. Each method–seed has 320 fresh auxiliary-evaluation rollouts and 720 N0 binary outcomes. These totals count valid grades and rollouts; they do not establish retry costs or equality of wall-clock compute.

The median bounded-shift radius is 0.04524 (interquartile range $[0.02189, 0.07193]$). Observed raw/DC disagreement is 40.66% (task-bootstrap 95% CI $[38.21, 43.08]\%$). The degenerate full-box bootstrap interval reflects that every observed task has feasibility rate one; it is not proof of universal feasibility on unseen tasks. The $K = 14$ audit must not be pooled with the separate $K = 15$ pilot containing planted controls.

The mean aggregate discrepancies are 0.01003, 0.00874, 0.00845, and 0.00712 for historical, external-anchor, shrinkage, and current ISA, respectively. The current-minus-shrinkage task-

Table 9: Bounded-shift stress curve on 40 $K = 14$ blocks. Rates concern separately optimized pairwise shifts; intervals bootstrap tasks. The radius reaches a tie, with strict reversal feasible beyond the boundary.

ℓ_∞ budget	Pair fraction (%)	95% task-bootstrap CI (%)
0.010	10.69	[9.40, 12.12]
0.025	28.46	[25.91, 31.37]
0.050	54.81	[51.26, 58.60]
0.075	77.75	[75.08, 80.44]
0.100	93.90	[92.45, 95.14]
Full score-feasible box	100.00	[100.00, 100.00]

Table 10: Companion-composition diagnostics. Aggregation is across 40 tasks; top-quartile membership means the top two of eight anchors, not the top quarter of all 14 block trajectories.

Task-aggregated diagnostic	Estimate	95% task-bootstrap CI
Median absolute-agreement ICC	0.9235	[0.9147, 0.9444]
Median Spearman agreement	0.9286	[0.8929, 0.9524]
Mean top-two-anchor overlap	0.8084	[0.7556, 0.8595]
Mean advantage-sign consistency	0.9188	[0.8946, 0.9409]
Median task-level anchor CV	0.2200	[0.1986, 0.2390]

discrepancy difference is -0.00446 (95% paired-t CI $[-0.00517, -0.00375]$). These finer-grained endpoints avoid interpreting a small aggregate discrepancy as uniformly small rollout gaps.

Analysis and reproducibility scope. The PDF reports bounded-shift summary statistics, task-aggregated composition diagnostics, seed-level reference-target summaries, calibration definitions and fitted parameters, task relationships, and the stated confidence intervals. It does not include the full-precision bounded-shift cell scores, auxiliary rollout-level scores, bootstrap resampling indices, executable training or evaluation code, or checkpoint-to-code mappings. Consequently, the reported formulas and tables support method reimplementation and selected arithmetic checks, but not exact end-to-end replay. Cross-family B/C outcomes for the reference-target comparison are not reported; the cross-family ISA–MMSD results use a different cohort.

H.1 Calibration utility and frozen reference parameters

Official utility on initial-policy evidence. For each valid calibration trajectory, $U_{\mathrm{cal}} \in \{0, 1\}$ is the fixed-version official binary label: native success for tau-bench, official harness resolution for SWE-bench-lite, and the native success predicate for Toolathlon. Zero denotes a valid execution that fails the success criterion; infrastructure failures, missing evidence, and verifier faults are invalid records, not zeros. The composition audit and reference-target comparison share $40 \times 64 = 2{,}560$ initial-policy trajectories on `T_MAIN`, including 1,314 positive labels. This count describes the calibration sample, not a trained-policy success result. Each trajectory has four A1–A4 cell means, each averaging four valid grades: 10,240 cell means represent 40,960 grading replications. Each trajectory supplies one utility label, not four independent utility observations.

Table 11: Seed-paired reference-target contrasts: current ISA minus each alternative target. Discrepancy intervals are marginal two-sided 95% paired-t intervals; the reward column is a one-sided 95% lower bound. Shrinkage is the primary comparator.

Comparator	$\Delta D^{\mathrm{rollout}}$ [95% CI]	ΔR_{out}	Reward lower bound
Historical	-0.01201 $[-0.01277, -0.01125]$	$+0.00645$	$+0.00559$
External-anchor	-0.00785 $[-0.00862, -0.00708]$	$+0.00369$	$+0.00290$
Shrinkage	-0.00423 $[-0.00488, -0.00358]$	$+0.00270$	$+0.00211$

Table 12: Secondary task-success endpoint for the reference-target comparison: 90 tasks, 20 paired seeds, and eight rollouts per task and seed (57,600 binary outcomes). Difference intervals use 10,000 paired-seed and within-benchmark task-bootstrap draws.

Target	Successes / 14,400	Success rate	Current ISA minus target (pp)
Historical	9,290	64.51%	$+1.333$ [$+1.028, +1.667$]
External-anchor	9,380	65.14%	$+0.708$ [$+0.451, +1.007$]
Shrinkage	9,397	65.26%	$+0.590$ [$+0.354, +0.840$]
Current ISA	9,482	65.85%	—

Affine fitting and residuals. For each evaluator j, the external baseline fits $g_j(R) = a_j + b_j R$ using

$$L_j(a,b) = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}(U_{\mathrm{cal},i} - a - b\bar{R}_{ij})^2 + 0.001b^2.$$

The intercept is unpenalized. Five task-grouped folds each hold out eight tasks (tau/SWE/Tool: $3/2/3$), train on the other 32, and keep all evidence for a task together. Within each benchmark, task number n is assigned to fold $(n-1) \bmod 5$ using the ordering defined for this study. After cross-validation, each map is refitted on the full bank and frozen before reference-target training. The following full-bank parameters use nine decimal places; $\hat{\beta}_j^H$ and $\hat{\beta}_j^E$ are the historical and external offsets defined in Appendix H.

Table 13: Frozen reference-target calibration parameters. External offsets round to zero at nine decimal places; their unrounded magnitudes are of order 10^{-17}.

Evaluator	a_j	b_j	$\hat{\beta}_j^H$	$\hat{\beta}_j^E$
A1	0.168224209	0.813466051	-0.075221056	0.000000000
A2	0.115770191	0.784149983	0.007530135	0.000000000
A3	0.119237089	0.786816277	0.001406050	0.000000000
A4	0.067714008	0.787656617	0.066284870	0.000000000

Writing $Z_{ij} = g_j(\bar{R}_{ij})$, stationarity of the free intercept gives $\mathbb{E}_{\mathrm{cal}}[Z_{ij}] = \mathbb{E}_{\mathrm{cal}}[U_{\mathrm{cal}}]$ for every j. The common labels and equal bank weights therefore imply $\hat{\beta}_j^E = \mathbb{E}_{\mathrm{cal}}[Z_{ij} - \bar{Z}_{i\cdot}] = 0$ in exact arithmetic. Near-zero fitted offsets are an algebraic consequence, not a missing correction. The external residual is $Z_{ij} - \bar{Z}_{i\cdot} - \hat{\beta}_j^E$, not current-block double centering of Z; the row-reward term still uses the original scores. Transformed bank scores span $[0.182667, 0.853789]$. Over the entire input domain $R \in [0,1]$, these fitted maps span $[0.067714, 0.981690]$ (rounded endpoints), so no output clipping is applied.

Shrinkage selection and label separation. The weight in $\hat{\beta}_j^S = \rho\hat{\beta}_j^C + (1-\rho)\hat{\beta}_j^H$ is selected from $\{0, 0.1, \ldots, 1\}$ by mean offset squared error on calibration pseudo-blocks, without using U_{cal}. Each task's 64 trajectories are separated into a current block of 14 and a reference block of the remaining 50; historical offsets come from tasks in the other folds. The objective compares the combined current/historical offset with the 50-trajectory reference, averaging over tasks, draws, and evaluators. Selection uses unrounded errors, with exact ties resolved toward smaller ρ, yielding $\rho = 0.4$. Current-block ISA, historical offsets, and shrinkage use no utility labels. The external baseline uses U_{cal} indirectly in training through its frozen affine maps. In contrast, $N0_{\mathrm{confirm}}$ outcomes are evaluation-only and do not select parameters, evaluators, or checkpoints.

I Matched Training, Transfer, and Mechanistic Results

The empirical question is whether penalizing trajectory-specific evaluator dependence improves policy transfer beyond penalizing raw score dispersion. The primary ISA–MMSD comparison estimates the overall training effect, and finer-grained endpoints test its sensitivity to cancellation. A separate ablation examines residual choice, followed by studies of baseline coefficients, task transfer, and evaluator-family transfer. Controlled scope checks clarify the mechanism and its interpretation.

Statistical conventions. Unless stated otherwise, tables report mean±sample SD across training seeds, and contrasts use matched-seed differences with two-sided 95% paired-t intervals. Evaluator replications and trajectories are aggregated within each seed before inference. The primary cohort has ten paired seeds (df$= 9$), with ISA minus MMSD as its primary contrast and -0.01 as the held-out-reward non-inferiority margin. DR, DR+, and Pess-min form a separate Bonferroni-corrected family of three secondary comparisons; GRPO is a lower-budget reference. The primary choices were fixed before outcome inspection. Extension-specific endpoints and resampling methods are given with their results.

I.1 Primary comparison and aggregation sensitivity

The primary cohort compares six methods using Qwen3-8B on 40 training tasks, with 240 updates per run and ten seeds. Five methods use the same four-evaluator panel and valid-score budget; GRPO uses one evaluator. Two held-out settings from the same family score fresh trajectories on those tasks, measuring evaluator transfer with the task set fixed. The shared loss, method-specific advantages, and scoring procedures are specified in Appendix F.

I.1.1 Aggregate discrepancy and held-out reward

Evaluation and aggregation. Training uses A1–A4 for crossed methods and A1 for GRPO; all methods are evaluated with A5–A6 on the same generated trajectory. Let $\mathcal{E}_v$ denote held-out aggregation rule v. For arm a, seed s, task t, rollout m, and evaluator j, write

$$q^{\mathrm{in}}_{astmj} = \frac{1}{4}\sum_{k=1}^{4} R_{astmjk}, \qquad q^{\mathrm{out},(v)}_{astmj} = [\mathcal{E}_v(\mathcal{B})]_{astmj}\,.$$

Here $\mathcal{B}$ denotes the held-out cell means and frozen parameters required by the aggregation rule for one evaluation record. This notation makes aggregation-rule dependence explicit without identifying P with Q or treating their difference as a universal additive constant. Panel aggregation is

$$\bar{q}^{\mathrm{in}}_{astm} = \frac{1}{|J^{\mathrm{in}}_a|}\sum_{j\in J^{\mathrm{in}}_a} q^{\mathrm{in}}_{astmj}, \qquad \bar{q}^{\mathrm{out},(v)}_{astm} = \frac{1}{|J^{\mathrm{out}}|}\sum_{j\in J^{\mathrm{out}}} q^{\mathrm{out},(v)}_{astmj}, \tag{21}$$

$$\Delta^{(v)}_{ast} = \frac{1}{M_t}\sum_{m}\left(\bar{q}^{\mathrm{in}}_{astm} - \bar{q}^{\mathrm{out},(v)}_{astm}\right), \qquad D^{\mathrm{task},(v)}_{as} = \frac{1}{T}\sum_{t}|\Delta^{(v)}_{ast}|, \tag{22}$$

$$\Delta^{\mathrm{global},(v)}_{as} = \frac{1}{N_s}\sum_{t,m}\left(\bar{q}^{\mathrm{in}}_{astm} - \bar{q}^{\mathrm{out},(v)}_{astm}\right), \quad D^{\mathrm{global},(v)}_{as} = |\Delta^{\mathrm{global},(v)}_{as}|, \tag{23}$$

$$R^{\mathrm{in}}_{as} = \frac{1}{N_s}\sum_{t,m}\bar{q}^{\mathrm{in}}_{astm}, \qquad R^{\mathrm{out},(v)}_{as} = \frac{1}{N_s}\sum_{t,m}\bar{q}^{\mathrm{out},(v)}_{astm}, \qquad N_s = \sum_t M_t.$$

Here $|J^{\mathrm{in}}_a| = 1$ for GRPO and 4 for crossed methods, and $|J^{\mathrm{out}}| = 2$. Each primary seed evaluates the 40 `T_MAIN` tasks with $M_t = 8$, hence $N_s = 320$; suite weights are $15/40$, $10/40$, and $15/40$. Native primary tables use P and native ablation tables use Q; their superscripts are suppressed only within a fixed-protocol comparison. The P/Q sensitivity analysis keeps checkpoints and evaluator scores fixed when comparing aggregation rules and separately evaluates frozen policies on both task sets under Q (Appendix L). $\mathcal{E}_P$ is the equal-weight mean of the raw held-out evaluator cell means, whereas $\mathcal{E}_Q$ is the reliability-weighted mean of the same cell means after frozen evaluator-specific affine calibration and clipping to $[0, 1]$. They use identical held-out evidence and differ only in aggregation. Appendix L.2 gives the formulas, functional pseudocode, and an illustrative arithmetic check; the paper does not provide the record-level inputs needed to regenerate every reported endpoint. Absolute values are taken after overall averaging for D^{global}, after within-task averaging for D^{task}, and before either averaging for D^{rollout}.

For $S = 10$ seeds, the primary comparison uses matched-seed differences $C_s = D^{\mathrm{global}}_{\mathrm{ISA},s} - D^{\mathrm{global}}_{\mathrm{MMSD},s}$, with a two-sided Student-t interval $\bar{C} \pm t_{0.975,S-1}\,\mathrm{sd}(C)/\sqrt{S}$ and the corresponding paired two-sided p value. Held-out non-inferiority uses $H_s = R^{\mathrm{out}}_{\mathrm{ISA},s} - R^{\mathrm{out}}_{\mathrm{MMSD},s}$ with margin -0.01. The three secondary D^{global} contrasts against DR, DR+, and Pess-min use Bonferroni-corrected intervals and p values for a family of three. Evaluator scores and task–rollout observations define each seed

endpoint, but only the ten seed values enter these tests. For extensions that specified task-level discrepancy, subsequent aggregate recomputations are treated as robustness analyses.

Table 14: Primary cohort on T_MAIN under native aggregation P (Qwen3-8B, ten seeds; mean±sample SD). The five crossed arms use $c = 4$; GRPO uses $c = 1$. Lower D^{global} and higher held-out reward are better. Signed gaps retain direction. MMSD is the primary comparator; bold marks the best discrepancy and held-out reward.

Method	$D^{\text{global}} \downarrow$	Signed Δ^{global}	Held-out reward ↑
GRPO ($c = 1$)	0.1617 ± 0.0042	$+0.1617 \pm 0.0042$	0.5724 ± 0.0057
DR	0.1329 ± 0.0052	$+0.1329 \pm 0.0052$	0.6208 ± 0.0077
DR+	0.0923 ± 0.0033	$+0.0923 \pm 0.0033$	0.6456 ± 0.0056
Pess-min	0.0841 ± 0.0035	$+0.0841 \pm 0.0035$	0.6409 ± 0.0079
MMSD 0.50	0.0833 ± 0.0034	$+0.0833 \pm 0.0034$	0.6517 ± 0.0059
ISA	$\mathbf{0.0604 \pm 0.0016}$	$+0.0604 \pm 0.0016$	$\mathbf{0.6742 \pm 0.0042}$

ISA improves same-family evaluator transfer. ISA reduces aggregate discrepancy from 0.0833 ± 0.0034 to 0.0604 ± 0.0016 (Table 14). The paired difference is -0.02292 (95% CI $[-0.02579, -0.02005]$, $t = -18.090$, df$= 9$, two-sided $p = 2.196 \times 10^{-8}$), with all ten seeds favoring ISA. The three secondary comparisons also favor ISA after Bonferroni correction (Table 15).

Table 15: Secondary primary-cohort contrasts on D^{global}. Differences are ISA minus comparator; intervals and p values are Bonferroni-corrected for three comparisons.

Comparator	Difference	Corrected 95% CI	Corrected p
DR	-0.07249	$[-0.07782, -0.06716]$	5.794×10^{-11}
DR+	-0.03194	$[-0.03534, -0.02854]$	1.585×10^{-9}
Pess-min	-0.02369	$[-0.02718, -0.02020]$	2.867×10^{-8}

The smaller gap accompanies a held-out gain. Held-out reward is 0.6742 ± 0.0042 for ISA and 0.6517 ± 0.0059 for MMSD. The paired gain is $+0.02247$ (95% CI $[+0.01935, +0.02559]$, $t = 16.274$, df$= 9$, two-sided $p = 5.547 \times 10^{-8}$). Its one-sided 95% lower bound is $+0.01994$ ($p = 1.083 \times 10^{-9}$ for the -0.01 safeguard), satisfying non-inferiority for any nonpositive margin. The chosen margin represents one percentage point on the normalized reward scale, rather than a utility-calibrated minimum effect. In-panel reward changes little: -0.00045 (95% CI $[-0.00414, +0.00324]$, $p = 0.7891$). Thus the primary gap contraction is accompanied by improved held-out scoring rather than a detectable decrease in in-panel reward.

Sensitivity to the seed-level test. The primary paired effect size is $d_z = -5.72$. All ten discrepancy differences have the same sign, giving an exact two-sided Wilcoxon signed-rank value $p = 2/2^{10} = 1.95 \times 10^{-3}$. Both tests describe training-seed variability on the fixed suite; the unseen-task analysis in Section I.3.1 additionally propagates task variation.

Table 43 provides all primary seed-level outcomes.

Single-evaluator reference. At $c = 1$, the interaction term vanishes algebraically. GRPO additionally uses group SD normalization (Appendix F), so this identity does not equate its training updates with those of ISA. The evaluator-count sweep in Section I.5 provides a separate empirical comparison.

I.1.2 Task- and rollout-level discrepancy

An aggregate gap can decrease when positive and negative discrepancies cancel. We therefore compute task- and rollout-level discrepancies from the same primary records (Table 16), taking absolute values before aggregation at each level. Equal task and rollout weights imply, by the triangle inequality,

$$D^{\text{global}} \leq D^{\text{task}} \leq D^{\text{rollout}} \tag{24}$$

for each seed and arm. The primary endpoint remains D^{global}; the finer summaries describe the discrepancy before cancellation at successive aggregation levels.

Table 16: Finer-grained primary-cohort discrepancies. Arm summaries are mean±sample SD across ten seeds; differences are ISA minus MMSD with two-sided 95% paired-t intervals (df$=9$). Task and rollout contrasts are rounded to four decimal places.

Endpoint	ISA	MMSD 0.50	Difference	95% CI
D^{global}	0.0604 ± 0.0016	0.0833 ± 0.0034	-0.02292	$[-0.02579, -0.02005]$
D^{task}	0.0694 ± 0.0018	0.0973 ± 0.0040	-0.0279	$[-0.0313, -0.0246]$
D^{rollout}	0.0754 ± 0.0019	0.1053 ± 0.0044	-0.0299	$[-0.0334, -0.0264]$

ISA reduces discrepancy at all three aggregation levels. Its task- and rollout-level differences are more negative than the aggregate difference, indicating that the improvement is not attributable solely to cancellation in the global endpoint. These intervals describe individual comparisons within the primary cohort and are not jointly adjusted for multiplicity.

Table 17: Benchmark-family breakdown of the 40-task primary comparison. Entries are approximate ISA-minus-MMSD mean differences, rounded to three decimals. All three suites show lower task-level discrepancy and higher held-out reward; the breakdown is descriptive.

Benchmark	Tasks	D^{task} difference	Held-out reward difference
tau-bench	15	-0.029	$+0.027$
SWE-bench-lite	10	-0.028	$+0.030$
Toolathlon	15	-0.027	$+0.013$

Each suite-level contrast is first computed within a seed. Weighting its unrounded mean by $15/40$, $10/40$, or $15/40$ recovers the overall task-level discrepancy and held-out-reward contrasts. Rounding accounts for small differences between the weighted table entries and the overall means. Overall confidence intervals use paired seed differences. Within this fixed task selection, all three benchmarks show the same direction of change.

I.2 Contribution of residual target and penalty geometry

The ISA–MMSD comparison changes both the residual target and penalty geometry. A separate five-arm study examines these choices by crossing raw and double-centered residuals with SD and adaptive energy penalties, alongside MMSD 0.50. Each arm uses ten seeds and the same `T_ABLATION` task set ($15/10/15$), which is disjoint from `T_MAIN`; its native held-out aggregation rule is Q. The principal residual-target contrast compares DC-Energy/ISA with Raw-Energy on D^{rollout}, with held-out reward as a safeguard. For $d_{astm} = \bar{q}^{\mathrm{in}}_{astm} - \bar{q}^{\mathrm{out}}_{astm}$, rollout discrepancy is $N_s^{-1} \sum_{t,m} |d_{astm}|$. All four endpoints in Table 1 are computed from the same rollout scores within this cohort.

Residuals, geometries, and coefficients. For $Z^{\mathrm{raw}}_{ij} = \bar{R}_{ij} - \bar{R}_{i\cdot}$ and $Z^{\mathrm{dc}}_{ij} = \hat{\gamma}_{ij}$, define

$$e_i(Z) = \frac{1}{c} \sum_j Z_{ij}^2, \qquad s_i(Z) = \sqrt{\frac{1}{c-1} \sum_j Z_{ij}^2}.$$

The four factorial arms use

$$A_i^{\mathrm{SD}}(Z) = \hat{\alpha}_i - 0.15\sqrt{e_i(Z)} = \hat{\alpha}_i - \eta_{\mathrm{SD}} s_i(Z), \qquad \eta_{\mathrm{SD}} = 0.15\sqrt{\frac{c-1}{c}}, \tag{25}$$

$$A_i^{\mathrm{Energy}}(Z) = \hat{\alpha}_i - w(Z) e_i(Z), \qquad w(Z) = \frac{0.15}{\sqrt{q_{0.5}(\{e_\ell(Z) : e_\ell(Z) > 0\})}}. \tag{26}$$

Set $w = 0$ if all energies vanish. Each arm uses additional baseline zero and no subsequent advantage centering. At $c = 4$, $\eta_{\mathrm{SD}} \simeq 0.129904$, so Raw-SD is distinct from MMSD 0.50. DC-Energy

uses the deployed ISA rule. The coefficients specify the tested geometry comparison and do not establish superiority over all tuned SD penalties.

Effect of double centering. DC-Energy/ISA improves both rollout- and task-level discrepancy relative to Raw-Energy within the same adaptive energy family (Table 1). The paired rollout-level difference is -0.0241 (95% paired-t CI $[-0.0272, -0.0210]$), and the task-level difference is -0.0251 ($[-0.0285, -0.0218]$). The aggregate difference is smaller and inconclusive: -0.0031 ($[-0.0079, 0.0017]$). Thus the finer-grained endpoints provide the evidence for the residual-target effect, while the aggregate endpoint permits comparison with the primary study's estimand. Held-out reward increases by $+0.0197$, with a one-sided 95% lower bound of $+0.0187$, above the pre-specified -0.01 safeguard.

The SD branch shows a similar residual-target effect: DC-SD lowers rollout discrepancy relative to Raw-SD by -0.0187 (95% CI $[-0.0227, -0.0147]$). DC-Energy/ISA also improves on MMSD 0.50, with a rollout discrepancy difference of -0.0198 ($[-0.0230, -0.0166]$) and a held-out reward gain of $+0.0207$. Together, these comparisons support double centering for the tested fine-grained discrepancy endpoints and adaptive energy rule. The subsequent official-success contrast for the frozen Raw/DC checkpoints is reported separately in Appendix I.8.

Sensitivity bounds from aggregate summaries. The DC-Energy arm also has lower rollout discrepancy and higher held-out reward than DC-SD. The aggregate SDs in Table 1 give a conservative bound without specifying the covariance between paired outcomes. Since $s_d \le s_E + s_S$, an ordinary two-sided 95% paired-t interval on ten pairs is contained in

$$\left[\bar{d} - t_{0.975,9}\frac{s_E + s_S}{\sqrt{10}}, \quad \bar{d} + t_{0.975,9}\frac{s_E + s_S}{\sqrt{10}}\right].$$

Allowing each displayed four-decimal mean and SD to vary by 0.00005 for rounding and rounding the final bounds outward, Table 1 implies an envelope of $[-0.0146, -0.0058]$ for DC-Energy minus DC-SD on D^{rollout}, and $[0.0084, 0.0122]$ on held-out reward. These conservative bounds account for unknown paired covariance and rounding; they are not directly estimated paired intervals and are not adjusted for multiplicity. They favor energy at the tested coefficients.

Relation to the primary cohort. Both cohorts use $15/10/15$ task quotas, $K = 14$, $c = 4$, 240 updates, learning rate 10^{-6}, ten runs per arm, and eight evaluation rollouts per task. However, `T_MAIN` and `T_ABLATION` have no task IDs in common, and their native held-out aggregation rules are P and Q, respectively. Table 18 preserves their native summaries, not a randomized cross-cohort method contrast. The sensitivity analysis separately holds evidence fixed when comparing P with Q and evaluates the frozen checkpoints on both task sets under Q (Appendix L).

Table 18: Descriptive alignment of the primary and ablation ISA endpoints. Values are mean±sample SD over ten training seeds; the last row is the ratio of the reported task- and aggregate-discrepancy means.

Endpoint	Main ISA: `T_MAIN`/P	Ablation ISA: `T_ABLATION`/Q
D^{global}	0.0604 ± 0.0016	0.0076 ± 0.0044
D^{task}	0.0694 ± 0.0018	0.0590 ± 0.0035
D^{rollout}	0.0754 ± 0.0019	0.0605 ± 0.0030
R_{out}	0.6742 ± 0.0042	0.6799 ± 0.0011
$\overline{D^{\text{task}}}/\overline{D^{\text{global}}}$	1.15	7.76

The native task-to-aggregate ratios are 1.15 and 7.76, but cancellation alone is not a source attribution. On the same main-study evaluator scores, switching the held-out aggregation rule from P to Q changes mean D^{global} from 0.06037 to 0.00897, a paired difference of 0.05140 (95% CI $[0.05054, 0.05226]$). Under unified Q, both frozen policies score similarly within each task set, whereas in-panel reward remains higher on `T_MAIN` than on `T_ABLATION`. The crossed-mean identity in Appendix L separates aggregation-rule, task-set, and within-task-set policy contrasts descriptively; it is not causal mediation. All primary and ablation method effects retain their original within-cohort comparisons.

I.3 Robustness across coefficients, tasks, and evaluator families

I.3.1 Baseline coefficients and transfer to unseen tasks

The next studies examine whether the result depends on the fixed MMSD coefficient or on the 40 training tasks. Both use their specified task-level discrepancy endpoint; aggregate recomputations provide a separate sensitivity analysis.

Coefficient sensitivity. The primary comparison is a six-arm, ten-seed cohort (60 runs) using native P and primary D^{global}. The coefficient-sensitivity cohort is separately trained: for each of ten seeds it contains one ISA arm and five MMSD arms (60 additional runs), uses aggregation P and the same frozen 40 task identities, but distinct run IDs, training seeds, checkpoints, generated trajectories, and evaluation records. Its prespecified endpoint is D^{task}; D^{global} is post-hoc. At the protocol-specified $\kappa_{\mathrm{SD}} = 0.50$ anchor, the frozen D^{task} difference is -0.01195 (95% CI $[-0.01424, -0.00966]$), and the one-sided held-out-reward lower bound is 0.00165, above -0.01. All five coefficients satisfy the reward safeguard. The four exploratory non-anchor task-level contrasts remain negative after Holm correction (largest adjusted $p = 1.36 \times 10^{-6}$). However, the post-hoc D^{global} contrast at coefficient 0.25 is inconclusive (95% CI $[-0.01123, 0.00285]$). This is a baseline-robustness analysis, not the DEV10 selection study in Appendix J.1; its effects are neither interchangeable with nor additive to the primary or confirmation contrasts.

Table 19: Coefficient sensitivity: ISA-minus-MMSD mean differences across ten paired seeds. The frozen endpoint is D^{task}; D^{global} is post-hoc. The last column reports whether the paired 95% D^{global} interval excludes zero.

κ_{SD}	D^{task} difference	D^{global} difference	Excludes zero
0	-0.02089	-0.0118	Yes
0.25	-0.01438	-0.00419	No
0.50 (anchor)	-0.01195	-0.0081	Yes
0.75	-0.01075	-0.0082	Yes
1.00	-0.00747	-0.0060	Yes

Unseen 180-task extension. Three arms, ISA, MMSD(0.50), and DR, are trained on the fixed 40-task training suite and evaluated at ten paired seeds on 180 unseen tasks, with 60 each from tau-bench, SWE-bench, and Toolathlon and eight rollouts per task. The held-out set has no task-identifier or exact normalized-content overlap with training (Appendix F.5). Under frozen D^{task}, a fixed 10,000-draw paired-seed by suite-stratified-task percentile bootstrap gives ISA $-$MMSD $= -0.01276$ (95% CI $[-0.01556, -0.00980]$) and ISA $-$DR $= -0.03666$ ($[-0.03979, -0.03362]$). Their held-out reward differences are $+0.00361$ and $+0.02486$, with one-sided 95% lower bounds 0.00055 and 0.02181, satisfying the specified discrepancy and reward criteria jointly. Reusing the same frozen draws for post-hoc D^{global} gives -0.00862 ($[-0.01289, -0.00395]$) versus MMSD and -0.02351 ($[-0.02933, -0.01775]$) versus DR. The 180 tasks broaden coverage across the three benchmark suites. Both aggregation orders use the same paired seed and suite-stratified task resamples.

I.3.2 Transfer to independent evaluator implementations

This study tests whether the advantage extends beyond variations of the training routine. ISA improves both discrepancy and reward on Families B and C (Table 20). Appendix F.3 specifies their scoring procedures and the interpretation of the normalized rewards.

Policies trained with Family-A variants are evaluated on its held-out settings and on the independently implemented Families B and C. Table 20 reports mean$\pm$sample SD over ten paired seeds. Task and rollout scores are aggregated within each seed before forming ISA–MMSD contrasts.

The cross-family gaps are larger than the same-family A-holdout gaps, as expected for independent evaluator families, but the ordering remains favorable to ISA. Relative to MMSD 0.50, ISA reduces $D^{\mathrm{global}}_{A\to B}$ by 0.0230 and $D^{\mathrm{global}}_{A\to C}$ by 0.0272, while the corresponding task-level endpoints move in the same direction (-0.0187 and -0.0210). Independent-family reward is higher for ISA on both

Table 20: Dense-rule cross-family evaluation, mean±sample SD over ten paired seeds. Lower D^{global} is better; rewards are higher-is-better.

Arm	R_{in}	$D^{\text{global}}_{A\to A_{\text{holdout}}}$	$D^{\text{global}}_{A\to B}$
DR	0.7536 ± 0.0068	0.1324 ± 0.0059	0.1826 ± 0.0103
DR+	0.7403 ± 0.0058	0.0921 ± 0.0042	0.1537 ± 0.0091
Pess-min	0.7281 ± 0.0077	0.0842 ± 0.0041	0.1420 ± 0.0098
MMSD 0.50	0.7354 ± 0.0013	0.0830 ± 0.0038	0.1280 ± 0.0072
ISA dense	0.7330 ± 0.0022	$\mathbf{0.0600 \pm 0.0035}$	$\mathbf{0.1050 \pm 0.0055}$
Arm	$D^{\text{global}}_{A\to C}$	Reward B	Reward C
DR	0.2007 ± 0.0118	0.5710 ± 0.0108	0.5529 ± 0.0127
DR+	0.1708 ± 0.0109	0.5866 ± 0.0097	0.5695 ± 0.0115
Pess-min	0.1576 ± 0.0119	0.5861 ± 0.0107	0.5705 ± 0.0118
MMSD 0.50	0.1450 ± 0.0077	0.6074 ± 0.0069	0.5904 ± 0.0076
ISA dense	$\mathbf{0.1178 \pm 0.0068}$	$\mathbf{0.6280 \pm 0.0048}$	$\mathbf{0.6152 \pm 0.0061}$

Table 21: Paired seed-level contrasts for the dense-rule cross-family evaluation. Differences are ISA dense minus MMSD 0.50.

Endpoint	Mean diff	SD(diff)	95% paired-t CI	Seed-level check
R_{in}	-0.0024	0.0018	$[-0.0037, -0.0011]$	10/10 did not cross the -0.01 collapse threshold
$D^{\text{global}}_{A\to B}$	-0.0230	0.0026	$[-0.0249, -0.0211]$	10/10 favored ISA
Reward B	$+0.0206$	0.0030	$[+0.0185, +0.0227]$	10/10 favored ISA
$D^{\text{task}}_{A\to B}$	-0.0187	0.0026	$[-0.0205, -0.0169]$	10/10 favored ISA
$D^{\text{global}}_{A\to C}$	-0.0272	0.0029	$[-0.0293, -0.0251]$	10/10 favored ISA
Reward C	$+0.0248$	0.0032	$[+0.0225, +0.0271]$	10/10 favored ISA
$D^{\text{task}}_{A\to C}$	-0.0210	0.0032	$[-0.0233, -0.0187]$	10/10 favored ISA

families; the one-sided 95% lower bounds for Reward B and Reward C are $+0.0189$ and $+0.0230$, both above the -0.0100 non-inferiority margin. The in-panel reward difference is small (-0.0024) and remains above the pre-set -0.01 collapse threshold for all ten seeds. The agreement across B and C supports transfer beyond the Family-A settings, within the two evaluated implementations. External utility remains a separate measurement question.

I.4 Mechanistic analyses of the interaction penalty

I.4.1 Interaction signal in frozen-policy trajectories

The frozen-policy pilot asks whether trajectory–evaluator interaction is large enough relative to noise and row-effect differences to motivate a training penalty. Its specified criterion requires both conditions on at least 30% of tasks; the criterion concerns interaction magnitude rather than gaming labels.

Setup. Qwen3-8B (frozen, temperature 0.9) generates $K = 15$ trajectories per task across 40 agent tasks from Toolathlon (15 tasks), tau-bench (15), and SWE-bench-lite (10). Each trajectory is graded under $c = 4$ semantics-preserving harness variants: two independently-implemented deterministic checkers (different parsing/normalization), a tool-schema rephrase variant, and one stochastic LLM-judge variant (Qwen-Plus, $r = 4$ replications). One planted exploiter per task is injected as a positive control. Measured cost: 1×A800, 1h27m rollout, 22m grading, 2,400 judge calls, $0.31. The pilot uses $K = 15$ frozen rollouts per task, including the planted exploiter; the main experiment uses training groups of $K = 14$.

Interaction relative to noise and row effects. From the $K \times c$ cell means we compute

$$\rho_{\text{int}} = \frac{\widehat{\mathbb{E}}[\gamma^2]}{\widehat{\text{Var}}(\hat{\alpha})}, \quad \text{SNR}_{\text{int}} = \frac{\widehat{\mathbb{E}}[\gamma^2]}{(\hat{\sigma}^2/r)(1-1/K)(1-1/c)}, \quad \text{ratio}_{\text{mat}} = \frac{\max_{ij} |\hat{\gamma}_{ij}|/c}{\text{median}_i \, |\Delta\hat{\alpha}_{(i)}|}. \tag{27}$$

Criterion A is $\mathrm{SNR}_{\mathrm{int}} \geq 1$, placing interaction above its estimated noise scale (Theorem 2); criterion B is $\mathrm{ratio}_{\mathrm{mat}} \geq 1$, placing the interaction proxy above the ANOVA row-effect gap. A task meets the pilot criterion when both hold. For a single-cell raw increment, the corresponding interaction cell is scaled by $(1 - 1/c)(1 - 1/K) = 0.70$.

Interaction is substantial on most tasks. Both criteria hold on 29 of 40 tasks (72.5%), exceeding the specified 30% threshold. The interaction-to-row-effect ratio has median 2.35, mean 3.67, and range 0.04–12.60. Thus interaction energy is substantial relative to ANOVA row-effect variation on this suite.

The pattern persists without planted exploiters. Removing the planted trajectory leaves 29/40 tasks meeting both criteria, with 90% meeting criterion A and 82.5% meeting criterion B. The median interaction-to-row-effect ratio changes from 2.35 to 2.11. The signal therefore extends beyond the planted controls, although its remaining sources can include both endogenous exploitation and legitimate evaluator dependence.

Table 22: Pilot results on 40 agent tasks (frozen Qwen3-8B, $K = 15$, $c = 4$, $r = 4$).

Benchmark	Tasks	Both	Both (%)	ρ_{int} median	Criteria A / B
tau-bench	15	12	80%	2.26	93% / 87%
Toolathlon	15	11	73%	4.25	87% / 87%
SWE-bench-lite	10	6	60%	1.28	90% / 70%
Overall	**40**	**29**	**72.5%**	**2.35**	**90% / 82.5%**

Positive-control check. The planted exploiter coincides with the largest $|\hat{\gamma}|$ cell in 25/40 tasks (62.5%, chance $1/K = 6.7\%$); the largest cell lies elsewhere in the remaining 37.5%. This measures localization of the planted control. Section I.6.1 evaluates association with separately assigned gaming labels.

Multiplicative non-additivity check. Tukey's test isolates the multiplicative component $\hat{\alpha}_i \hat{\beta}_j$: median 1.3% of interaction SS (max 12.8%), and 1/40 tasks rejects additivity (median $p = 0.47$). This result argues against the tested multiplicative component as the main source of interaction.

Concentration of interaction energy. The largest cell carries a median 59.9% of a row's interaction energy (uniform baseline 25%), and 77.3% of the 600 rows put over half in one cell. These are concentration statistics; the sufficient condition for certified sparse recovery is not met (Section C.4). The deployed method retains the full dense energy.

Evaluator main-effect diagnostics. Per-variant $\hat{\beta}_j$ means are $(-0.000, -0.003, +0.003, -0.000)$, with SD 0.033–0.047 and maximum $|\hat{\beta}_j| = 0.103$; only 2/40 tasks exceed 0.10. These statistics characterize average column shifts; trajectory-dependent scoring changes are assessed through the interaction residual. Section I.6.3 summarizes the perturbation diagnostics.

Tasks with weaker interaction. Among the 11 tasks that do not meet both criteria, 4 have signal below noise at $r = 4$ and 7 have interaction below the row-effect-gap threshold; none fail both. All 40 tasks enter training, so the primary comparison is not restricted to tasks selected for strong interaction.

I.4.2 Sensitivity to positive reward rescaling

The normalization rule predicts how the penalty changes under positive reward rescaling. We examine this prediction using the primary cohort's double-centered penalties and a fitted curve relating penalty strength to discrepancy. This analysis evaluates the fitted response; it does not retrain policies under rescaled rewards.

Comparison with an alternative normalization. The comparison uses the non-deployed dimensionless heuristic

$$\lambda_{\text{old}} = c_1 \cdot \frac{\sigma^2_{\text{ref}}}{\text{median}_i(p_i)}, \tag{28}$$

with $\sigma^2_{\text{ref}} = 0.01$ and $c_1 = 1.5$, and $\lambda_{\text{old}} = 0$ when $\text{median}_i(p_i) < 10^{-8}$. Because p_i and σ^2_{ref} both carry reward-squared units, their ratio is dimensionless, so Eq. (28) cannot supply the inverse-reward units the advantage requires. With σ^2_{ref} held fixed, its failure mode is predictable: under $R \to sR$, $\lambda_{\text{old}} \to s^{-2}\lambda_{\text{old}}$, so $\lambda_{\text{old}} p_i$ is invariant while the signal term grows by s, diluting the penalty by exactly $1/s$. The response-model comparison below evaluates this predicted scaling behavior.

Response under rescaling. Across five seeds and $s \in \{0.1, 1, 10\}$, the old rule yields gaps of 0.0724, 0.0451, and 0.1172 (range 0.0721). The dense scale-equivariant rule yields 0.0436, 0.0431, and 0.0437 (range 0.0006). At $s = 1$, the new-minus-old difference has a 95% interval of $[-0.0053, 0.0013]$. At this scale, the interval permits either direction of change; a non-inferiority claim would additionally require a specified upper tolerance. The fitted responses are consistent with the predicted scaling behavior over these three positive scales, without establishing training invariance or robustness to affine, nonlinear, or evaluator changes.

I.4.3 Separating interaction from dispersion: the β sweep

This simulated-block diagnostic compares interaction energy with raw within-row dispersion as evaluator main effects vary.

Raw within-row dispersion combines the shared evaluator effect β_j with the trajectory-specific interaction γ_{ij}. The shared component contains no trajectory-specific information, but remains in the within-row SD; double centering removes it. We vary $\text{sd}(\beta) \in \{0.03, 0.10, 0.18\}$ on common simulated blocks and measure the within-block AUC for ranking a planted exploiter above honest trajectories.

Across $\text{sd}(\beta) = 0.03/0.10/0.18$, the block AUC of ISA is $0.9671/0.9701/0.9689$, a range of 0.003. Mean-minus-SD changes from 0.9856 to 0.8578 to 0.7054. At $\text{sd}(\beta) = 0.18$, the AUC difference is 0.2635 in favor of ISA; at 0.03, the absolute difference is 0.0185, with mean-minus-SD higher. Pess-min remains above chance but also declines ($0.8814/0.7409/0.6303$).

The common planted-exploiter structure isolates the effect of evaluator shifts on these score statistics. The sweep thus illustrates the residual geometry: double centering removes the controlled column effect retained by raw dispersion. It is distinct from the trained-policy residual-target comparison in Section I.2.

I.4.4 Penalty-assignment controls

The assignment study examines whether the mapping between trajectories and penalties matters in addition to the amount of penalty applied. It compares separately trained policies under correct and disrupted mappings.

Four arms are trained with 20 paired seeds: zero applies no penalty; constant applies common pressure; deranged retains heterogeneous penalty values but breaks their trajectory mapping; correct restores the intended mapping. Correct assignment yields lower frozen D^{task} than deranged assignment (Table 23). At the seed level, the contrasts satisfy total=targeting+residual and residual=common+heterogeneous exactly.

Table 23: Penalty-assignment contrasts over 20 paired seeds on frozen D^{task}. Positive differences favor the second arm in each contrast.

Component	Contrast	Difference	95% paired-t CI
Common pressure	zero−constant	0.00621	[0.00465, 0.00777]
Heterogeneous pressure	constant−deranged	0.00474	[0.00375, 0.00573]
Residual	zero−deranged	0.01095	[0.00947, 0.01244]
Targeting	deranged−correct	0.01123	[0.01009, 0.01237]
Total	zero−correct	0.02218	[0.02097, 0.02339]

The protocol-specified ratio-of-means paired bootstrap assigns 0.5062 of the frozen D^{task} total to targeting (95% CI $[0.4571, 0.5530]$) and 0.4938 to the residual ($[0.4470, 0.5429]$); all 10,000 draws are finite. Under post-hoc D^{global}, targeting remains positive at 0.0079 (95% paired t CI $[0.0037, 0.0120]$), residual is 0.0087 ($[0.0052, 0.0122]$), and total is 0.0166 ($[0.0132, 0.0199]$). Reusing the frozen paired-seed draws gives a targeting share of 0.474 (95% CI $[0.272, 0.659]$) and residual share 0.526 ($[0.341, 0.728]$), showing that the numerical split is aggregation-order sensitive and imprecise. Correct-minus-deranged held-out reward has a one-sided 95% lower bound of 0.00763, above the -0.01 margin. A fixed-input calculation on 640 blocks per arm (2,560 total) checks the decomposition of the contrasts and the score-to-advantage mapping. Both partitions compare separately trained policy arms whose sampling and optimization paths differ; neither percentage is a causal share.

Interpretation across studies. The residual-target ablation, simulated shift control, and assignment study examine complementary aspects of the penalty. The ablation supports double centering on rollout- and task-level discrepancy (Section I.2); the simulated blocks explain how removing column main effects changes the residual geometry; and the assignment study shows that preserving the trajectory-to-penalty mapping improves the reported discrepancy endpoints. Their agreement supports a trajectory-specific training effect under the tested rules. The assignment shares summarize contrasts between separately trained policies, and gaming attribution is assessed through the label study (Section I.6.1).

I.5 Evaluator count and grading cost

Increasing panel size changes both the information available for residual estimation and the cost of obtaining it. This sweep measures the resulting cost–discrepancy trade-off across evaluator counts.

Grading cost. A fully crossed block of K trajectories and c evaluators requires Kc trajectory–evaluator score evaluations when each cell is scored once, or Kcr with r replications per cell. With evaluator-specific replication counts r_j, the total is $K \sum_j r_j$. These are score-evaluation counts; API request counts also depend on which evaluators use an API and whether calls are batched.

The pilot used one A800, 1h27m of frozen rollouts plus 22m of grading, 2,400 judge calls, and \$0.31 of API spend. The separate alternative-rule 8B and 14B studies each used 76,000 calls, costing \$8.45 and \$9.01, respectively. These measurements concern different experimental settings and cannot be combined into the cost of the deployed dense rule.

We evaluate $c \in \{1, 2, 3, 4, 6, 8\}$ for ISA and DR with five seeds per setting. Table 24 reports ISA means. Although the lowest mean occurs at $c = 6$, the paired $c = 4$ minus $c = 6$ interval includes zero (95% CI $[-0.0043, 0.0155]$), leaving that comparison inconclusive. Performance is worse at $c = 8$ than at $c = 6$, with a positive difference and 95% CI $[0.0007, 0.0221]$. Across both methods and all settings, the sweep uses $1{,}545{,}600$ judge calls at an API cost of \$170.01.

Evaluator composition, grading cost, and the interaction estimate all vary with c in this model and task suite. The non-monotonic pattern is compatible with changes in both penalization and estimation quality; the design does not isolate their separate contributions.

Table 24: Evaluator-count sweep: mean reported ISA gap over five seeds. Evaluator composition and grading cost change with c, so this is an observed trade-off rather than an isolated effect of evaluator count.

Evaluator count c	1	2	3	4	6	8
Mean gap	0.19764	0.10639	0.07321	0.04323	0.03764	0.04902

I.6 CONSTRUCT-VALIDITY CONTROLS

I.6.1 INTERACTION ENERGY AND GAMING LABELS

The usefulness of a training regularizer is distinct from its behavior as a classifier of gaming labels. We evaluate the latter through within-block discrimination, incremental information beyond simple row statistics, and overlap with legitimate-specialization labels.

Design and labels. The study uses $K=8$ trajectories and $c=4$ evaluators per block, 40 blocks per seed, and five seeds, for $6{,}400$ trajectories. Two Qwen-Max annotators, blinded to all decomposition outputs, label trajectories as gaming, legitimate specialization, or honest. Using more than two trajectories avoids the degeneracy in which double centering assigns equal squared penalties to both rows. The labeled cohort retains the primary ordering across GRPO, DR, DR+, and ISA, with gaps of 0.2016, 0.1289, 0.0881, and 0.0466, respectively.

Within-block discrimination. We estimate $\Pr(p_i > p_j \mid Z_i = \text{gaming}, Z_j = \text{honest}, \text{same block})$, comparing trajectories on the same task under the same evaluators. The mean is 0.69286 (95% CI $[0.6375, 0.7482]$). The interval lies above chance, although its lower endpoint remains below 0.65. The 876 gaming–honest pairs share trajectories and blocks and are not independent experimental units; inference uses the five seed values 0.6802, 0.6890, 0.6936, 0.7626, and 0.6389. The result indicates a trajectory-level association with gaming labels beyond alignment with planted controls.

Incremental discrimination beyond row statistics. A logistic model on three simple row features (row mean, within-row SD, row minimum) reaches AUC 0.7475–0.8659 across seeds. Adding p_i increases AUC by $+0.00506$ on average (per-seed $+0.0040$ to $+0.0093$), indicating limited incremental discrimination beyond these row features. The training comparisons assess its separate role as a regularizer (Sections I.1.1, I.4.3, I.4.4).

Overlap with legitimate specialization. At 0.70 gaming recall, the share of trajectories labeled as legitimate specialization above the threshold is 0.184–0.273 across seeds. This thresholded diagnostic differs from the continuous training penalty, but the overlap supports interpreting interaction energy as a panel-dependence signal rather than a gaming-specific label.

Human-label comparison. A spot check of 300 trajectories assesses the model-produced labels. Human–human agreement is 0.792, compared with human–proxy agreement of 0.630. Against human labels, penalty AUC is 0.6517, versus 0.6838 against proxy labels, a decrease of 0.0321. Proxy precision/recall against human labels is $0.683/0.860$ for gaming, $0.856/0.873$ for honest behavior, and $0.729/0.520$ for legitimate specialization; 35 of 98 human-labeled specialization trajectories are classified as gaming. These discrepancies are consistent with a contribution from proxy-label noise. The comparison assesses agreement with human judgments; it does not partition the observed classification error between label noise and the penalty.

I.6.2 EVALUATOR-SHARED REWARD PATTERNS

This control separates reward increments common to every evaluator from increments that affect individual evaluators. The former are absorbed into the row effect, while the latter appear as interaction.

We plant two kinds of high-reward pattern. E_{var} affects a particular evaluator and therefore appears as interaction; E_{uni} is shared by all c evaluators and therefore appears mostly in the row mean. For E_{var}, ISA and DR win rates are 0.04408 and 0.3280. For E_{uni}, they are 0.67818 and 0.71212. Under ISA, the pattern's median penalty is 0.0745 for E_{var} and 0.01588 for E_{uni}, against a reference median of 0.0146: ratios of $5.10\times$ and $1.09\times$.

The result follows the decomposition: a component constant across evaluators enters $\hat{\alpha}_i$ and is absent from the interaction. ISA therefore targets evaluator-specific dependence, while Family-B/C outcomes in Section I.3.2 measure transfer across evaluator implementations.

I.6.3 Evaluator-perturbation diagnostics (summary)

Appendix K distinguishes fixed evaluator-column offsets in Monte Carlo calibration from modified scoring rules in the training study. The magnitude-threshold diagnostic is sensitive to large offsets, and its behavior changes when multiple columns shift. In training, removing flagged variants moves the ISA gap toward its uncorrupted level (0.057 versus 0.046), while the suppression measure is 0.40–0.60 relative to 1.00 in the uncorrupted condition. These studies characterize screening behavior and training sensitivity under the specified perturbations.

I.7 Alternative penalty rules

The following studies examine thresholded or sparse penalties with different normalization. They clarify which observations extend across these alternatives; their outcomes remain separate from evidence for the deployed dense rule.

I.7.1 Residual targets under matched penalty pressure

This factorial compares residual targets under arm-local matched pressure, using a weighting rule distinct from the deployed method.

Endpoint. This analysis tests whether removing the evaluator column main effect from the residual target shrinks the transfer gap while meeting a held-out reward non-inferiority margin after normalizing each arm to its own block-level raw-SD pressure reference.

Setup. The study uses Qwen3-8B, the same 40-task composition (15 tau-bench, 10 SWE-bench-lite, and 15 Toolathlon tasks), $K = 14$, $c = 4$, three epochs, learning rate 10^{-6}, eight evaluation rollouts, and ten seeds. The four factorial arms cross row-centered versus double-centered residual targets with SD versus threshold-energy geometry under the arm-local pressure normalization defined below. DR and the alternative sparse ISA rule are included only as descriptive controls.

Thresholded energy. For the double-centered target, let $t = \mathrm{median}_{k,j} |\hat{\gamma}_{kj}|$ over the block and $\hat{S}_{ij} = \hat{\gamma}_{ij} \mathbf{1}\{|\hat{\gamma}_{ij}| > t\}$. The threshold-energy penalty is $p_i^{\mathrm{thr}} = \sum_j \hat{S}_{ij}^2$, using a strict threshold. In the deployed dense rule, $\hat{S}$ is computed only as a diagnostic and does not enter $p_i = c^{-1} \sum_j \hat{\gamma}_{ij}^2$; the two values differ by a factor of roughly 3.2 over the evaluated blocks (Section F). The factorial below uses its own normalization to compare residual targets and penalty geometries.

Matched-pressure protocol. This follow-up separates the residual target from the penalty geometry under its own arm-local pressure normalization. Let $Z_{ij}^{\mathrm{raw}} = R_{ij} - \bar{R}_{i\cdot}$ be the row-centered residual and $Z_{ij}^{\mathrm{dc}} = \hat{\gamma}_{ij}$ the double-centered residual. For either target Z, the SD geometry (called *dense* in the original arm labels) is

$$q_i^{\mathrm{SD}}(Z) = \sqrt{\frac{1}{c-1} \sum_j Z_{ij}^2}, \tag{29}$$

whereas the threshold-energy geometry is

$$q_i^{\mathrm{thr}}(Z) = \sum_j \left[Z_{ij} \mathbf{1}\{|Z_{ij}| > \mathrm{median}_{k,\ell} |Z_{k\ell}|\} \right]^2 . \tag{30}$$

Each arm returns $\bar{R}_{i\cdot} - w q_i$. Within each arm and training block, the arm-local raw-SD reference is $B = \mathrm{median}_i \, q_i^{\mathrm{SD}}(Z^{\mathrm{raw}})$, and $w = B / \mathrm{median}_i \, q_i$ for non-degenerate arms. The raw–SD arm returns exactly row mean minus within-row sample standard deviation; its implementation uses $w = 1$ in non-degenerate blocks and $w = 0$ when both quantities vanish. This construction normalizes each arm's median subtraction to the raw-SD scale of its own current block while retaining the row mean. It does not impose one shared numerical B across separately trained arms or equate their trajectory-level penalty rankings. The factorial arms are row-centered/double-centered $\times$ SD/threshold-energy. This SD geometry differs from the mean-squared interaction energy used by the deployed ISA rule. DR and a non-deployed sparse ISA variant are descriptive controls and do not enter factorial main-effect estimates.

Table 25: Matched-pressure factorial follow-up on Qwen3-8B (10 seeds; mean±sample SD). Lower transfer gap is better. The first four rows form the pre-specified 2×2 factorial; DR and sparse ISA do not enter the main-effect estimates.

Condition	Target	Geometry	Gap Δ	Held-out reward
Raw–SD	Row-centered	SD	0.19922 ± 0.00933	0.49912 ± 0.00334
Raw–threshold	Row-centered	Threshold energy	0.18018 ± 0.01051	0.50753 ± 0.00232
DC–SD	Double-centered	SD	0.17523 ± 0.01082	0.49986 ± 0.00327
DC–threshold	Double-centered	Threshold energy	0.16308 ± 0.01099	0.50820 ± 0.00252
DR	N/A	N/A	0.20404 ± 0.00956	0.49395 ± 0.00596
Sparse ISA	Double-centered	Sparse energy	0.16179 ± 0.01118	0.51699 ± 0.00272

The independent unit is the training seed. All four factorial arms must be present for each of ten seeds, and arms are paired by seed. With Δ denoting transfer gap, the pre-specified centering effect is

$$d_s = \tfrac{1}{2}\left[(\Delta_{\mathrm{raw,SD},s} - \Delta_{\mathrm{dc,SD},s}) + (\Delta_{\mathrm{raw,thr},s} - \Delta_{\mathrm{dc,thr},s})\right]. \tag{31}$$

The held-out reward effect h_s replaces each gap contrast with double-centered minus raw held-out reward. We report two-sided paired Student-t 95% intervals over the ten seed values. The specified success criterion requires $\bar{d} \geq 0.015$, the lower 95% confidence limit for d_s to exceed zero, and the lower limit for h_s to be at least -0.01. Geometry and target×geometry interaction effects are descriptive and do not enter this decision.

Double centering under matched pressure. Averaged across the two geometries, double centering improves transfer gap over the row-centered target by 0.02055 (sample sd 0.00327, paired 95% CI $[0.01821, 0.02288]$), with positive d_s on all ten seeds. This exceeds the pre-specified minimum effect 0.015, and the interval excludes zero. The corresponding held-out reward effect is $+0.00070$ (sample sd 0.00123, 95% CI $[-0.00018, 0.00159]$); its lower confidence limit is well above the non-inferiority margin -0.01. The gap effect exceeds the specified 0.015 threshold, with held-out reward meeting the -0.01 non-inferiority criterion.

Descriptive factorial structure. Averaged across targets, SD minus threshold-energy transfer gap is 0.01560 (95% CI $[0.00881, 0.02238]$), favoring threshold energy. The target×geometry difference-in-differences is 0.00689 (95% CI $[0.00495, 0.00883]$), indicating that the double-centering benefit is larger under SD geometry in this fixed protocol. These two intervals describe the factorial structure beyond the specified residual-target comparison. Sparse ISA has the smallest descriptive gap in Table 25, but its alternative weighting rule is outside the matched-pressure factorial and cannot be used to estimate either main effect.

Interpretation. Under the alternative arm-local pressure normalization, double centering shrinks the reported gap relative to row-centered residuals and passes the held-out non-inferiority criterion. Approximately 96.6% of the mean gap change is accounted for by a decrease in the in-distribution side of the gap, whereas the mean held-out change is only $+0.00070$ with 95% CI $[-0.00018, 0.00159]$. The result supports gap contraction with held-out non-inferiority under this normalization. The held-out interval includes zero, so the study does not establish a capability gain.

I.7.2 Baseline comparisons with a sparse penalty

This comparison evaluates a sparse penalty with a dimensionless weight, separate from the deployed dense scale-equivariant rule. Hyperparameters are selected on a fixed development split and evaluated on a distinct held-out split. The fixed methods are sparse ISA ($c_1 = 1.5$), mean-minus-sample-SD ($\kappa = 1$), lower-tail CVaR ($q = 0.25$), minimum, and DR+. Across five seeds, sparse ISA attains a gap of 0.04498, compared with 0.06880 for mean-minus-sample-SD. The paired MMSD-minus-sparse-ISA difference is 0.02382 (95% t interval $[0.01419, 0.03345]$). Held-out reward is 0.65150 versus 0.62846, a paired gain of 0.02304 (95% interval $[0.01715, 0.02893]$). These results support the alternative sparse rule over the selected aggregator on this evaluation split.

I.7.3 A SECOND MODEL CONFIGURATION

The second-configuration study evaluates the same non-deployed sparse penalty and dimensionless weight on Qwen3-14B. It uses the 40-task suite, the same five seeds, $K = 14$, $c = 4$, and 240 steps, with different adaptation and optimizer settings. Transfer gaps retain the ordering ISA $<$ DR+ $<$ DR $<$ GRPO in both the 8B and 14B configurations; the 14B sparse-ISA gap is 0.041. Because model size, adaptation, and optimization change together, the comparison is descriptive and does not isolate a model-scaling effect.

I.7.4 CROSS-FAMILY EVALUATION OF THE SPARSE RULE

The sparse-rule cross-family study trains with the parameterized LLM-grader Family A and evaluates with an independent LLM judge (Family B) and an execution verifier (Family C). Across five seed summaries, sparse ISA has positive margins over DR+ and a fixed mean-minus-sample-SD baseline on both families. The positive margins suggest that cross-family gains also occur with this sparse formulation. Table 20 and Section I.3.2 report the separate comparison for the deployed dense rule.

I.8 FOLLOW-UP OFFICIAL SUCCESS FOR THE RAW/DC TARGET INTERVENTION

The Raw/DC follow-up evaluates ten paired Raw-Energy and DC-Energy/ISA checkpoints frozen at update 240 from the residual-target ablation. It is a subsequent study, not retrospectively part of the initial preregistration. Both arms use the fixed `N090` task set: 90 tasks ($30/30/30$ across tau-bench, SWE-bench-lite, and Toolathlon), with eight rollouts per task and checkpoint, or 7,200 outcomes per arm. The task set is the same as in the task-success and reference-target studies, but the policies and rollout evidence differ.

Table 26: Raw/DC follow-up under official binary success. The contrast is DC-Energy/ISA minus Raw-Energy; only the contrast has a reported confidence interval.

Arm or contrast	Successes / outcomes	Rate or difference
Raw-Energy	$4133/7200$	57.40%
DC-Energy / ISA	$4395/7200$	61.04%
DC $-$ Raw	—	$+3.639$ pp
95% CI for DC $-$ Raw	—	$[+1.027, +6.292]$ pp

The reported interval uses 10,000 bootstrap draws accounting for seeds and benchmark-stratified tasks while retaining method pairing. Exact replay of the interval is not possible from the PDF because the resampling indices are not included. This contrast supports a residual-target effect within the tested adaptive energy rule. It is not added to the $+4.36$ pp adaptive-rule contrast (Appendix J.2) or the $+0.590$ pp current-versus-shrinkage contrast (Appendix H). Shared tasks do not make these distinct evidence sets a single comparison or independent task-set replications.

I.9 GRADIENT COVERAGE AND CLIPPING IN THE RAW/DC COHORT

The analysis covers all $10 \times 240 = 2{,}400$ updates per method, with 240 update indices per seed. Raw-Energy clips 347 updates (14.46%; between-seed sample SD 2.36 pp), whereas DC-Energy/ISA clips 149 (6.21%; SD 1.60 pp). The measured quantity is the global gradient norm G after accumulation and unscaling, before the optimizer step. With threshold 1 and $\epsilon = 10^{-6}$, the clipping factor is

$$f_{\text{clip}} = \min\left(1, \frac{1}{G + \epsilon}\right), \qquad G_{\text{post}} = f_{\text{clip}} G.$$

This is gradient clipping, not PPO ratio clipping. Neither an additional normalization floor nor post-hoc advantage recentering changes the deployed rule.

Table 27: Global gradient norms over 2,400 updates per method. All quantiles are pre-clipping; post-clipping maxima round to 1.0000.

Method	p50	p95	p99	p99.9	Pre max	Post max
Raw-Energy	0.5566	1.3590	1.8685	2.4895	4.1205	1.0000
DC-Energy / ISA	0.4905	1.0417	1.3817	2.0367	2.3260	1.0000

No non-finite loss, gradient, or parameter event occurs in either arm, and the score-block analysis covers all 2,400 updates per method. These observations characterize the Raw/DC cohort, not every training configuration. They neither exclude effects of rare large updates nor remove the small-positive-median amplification described in Appendix E.3; the paired floor-versus-original training in Appendix J.6 directly measures and reduces that update tail.

J Adaptive-Rule Comparison and Official Task Success

This appendix reports two targeted confirmations of the main interpretation. The coefficient-confirmation study compares ISA with a tuned double-centered SD penalty under a shared development-grid selection rule. The task-success study asks whether the ISA–DC-SD comparison also favors ISA under benchmark-native binary task completion; both arms already use double-centered residuals. These experiments are separate from the primary 40-task MMSD comparison and the residual-target ablation in Section 4, so each table states its cohort, endpoint, and selection role explicitly.

J.1 Grid-selected DC-SD versus ISA

Both methods are tuned using D1–D2 on the same ten-task development split and the same four-point coefficient grid $\{0.075, 0.15, 0.30, 0.60\}$. The selection stage uses three development seeds for each method–coefficient combination ($2 \times 4 \times 3 = 24$ runs). DC-SD uses

$$A_i = \hat{\alpha}_i - a\sqrt{p_i},$$

while ISA uses Eq. (4). For each method, coefficients within 0.01 of the best development-panel reward remain eligible; the selected coefficient is the eligible value with the lowest development-set D^{rollout}, with ties broken by higher reward and then by the smaller coefficient. This selects $a = 0.15$ for DC-SD and $\kappa = 0.15$ for ISA. The frozen rules are then retrained with ten new paired confirmation seeds (20 runs) and evaluated under aggregation P with A5–A6 on the 30-task confirmation subset. Table 28 lists the split. These confirmation tasks are the remaining `T_MAIN` tasks after `DEV10`. They are held out from coefficient selection, not from policy training or reference-target calibration. They share 30 task IDs but no rollout evidence with the calibration bank; unseen-task transfer is assessed separately on `U180`.

Table 28: Development/confirmation split for coefficient selection. Numbers are local positions in the 15/10/15 training-task lists, not public benchmark identifiers. This split does not apply to the primary 40-task evaluation.

Benchmark	Development positions	Development N	Confirmation N
tau-bench	3, 7, 11, 14	4	11
SWE-bench-lite	2, 5, 9	3	7
Toolathlon	1, 6, 12	3	12
Total	D1–D2 / A5–A6 panels	10	30

The study uses ten training-cohort tasks for development with D1–D2 and the remaining thirty for confirmation with A5–A6. Development scores determine coefficient eligibility and tie-breaking before confirmation. The confirmation tasks are held out from coefficient selection but remain part of the policy-training task set. D1–D2 also support evaluator development within this training cohort.

The rollout-discrepancy reduction is 18.4% relative to DC-SD. The one-sided 95% lower bound for the reward difference is $+0.0043$, so the reward safeguard is satisfied. Both methods operate

Table 29: Coefficient-selection rule and frozen outcome. The same development split and coefficient grid are used for both methods. Development-set rewards and rollout discrepancies determine eligibility and tie-breaking before any confirmation-set result is inspected.

Method	Grid	Eligibility and tie-breaker	Frozen coefficient
DC-SD	0.075, 0.15, 0.30, 0.60	R_{out} within 0.01 of method best; then lowest D^{rollout}, higher reward, smaller coefficient	$a = 0.15$
ISA	0.075, 0.15, 0.30, 0.60	Same development split, window, and tie-breakers	$\kappa = 0.15$

Table 30: Paired confirmation results after grid selection. Values are mean±sample SD over ten paired seeds. Confidence intervals are paired-t intervals for ISA minus DC-SD.

Endpoint	DC-SD	ISA	ISA−DC-SD
$D^{\text{rollout}} \downarrow$	0.0760 ± 0.0040	$\mathbf{0.0620 \pm 0.0035}$	$-0.0140\ [-0.0172, -0.0108]$
$R_{\text{out}} \uparrow$	0.6650 ± 0.0045	$\mathbf{0.6710 \pm 0.0040}$	$+0.0060\ [+0.0039, +0.0081]$

on double-centered residuals and pass through the same coefficient-selection rule. The contrast therefore tests ISA's adaptive energy rule against DC-SD with the residual target fixed; it does not separately identify the effects of squared energy and block-adaptive normalization.

J.2 Separately measured official task success

Here $N0_{\text{confirm}}$ scores each rollout under the official binary completion rule. Confirmation outcomes are excluded from training, evaluator development, coefficient selection, and checkpoint selection. This exclusion does not extend to $N0_{\text{cal}}$, which fits the reference-target study's external baseline on separate initial-policy evidence. The task-success study uses the fixed `N090` subset of `U180`, with 30 tasks each from tau-bench, SWE-bench-lite, and Toolathlon. Eight rollouts per task and ten paired seeds give 7,200 outcomes per method. Method pairing is retained within seeds and benchmark blocks. The reference-target comparison and Raw/DC follow-up reuse these task IDs but generate distinct policy evidence; endpoint separation alone does not imply independent task samples.

Table 31: Suite mapping and binary success criterion. The endpoint uses benchmark-native task completion rather than an additional manual rubric.

Suite	Benchmark	Tasks	Binary success criterion	Rollouts per method
suite-1	tau-bench	30	Official adapter: target database state and all required outputs	$30\times10\times8 = 2400$
suite-2	SWE-bench-lite	30	Official resolved: all FAIL_TO_PASS pass and PASS_TO_PASS remain acceptable	$30\times10\times8 = 2400$
suite-3	Toolathlon	30	Official task-specific pass/fail evaluator	$30\times10\times8 = 2400$

Table 32: Binary task-success results by benchmark block. Rates are successes divided by rollouts for that method and block.

Benchmark block	Tasks	DC-SD	ISA	Difference
tau-bench	30	1332/2400 (55.50%)	1433/2400 (59.71%)	+4.21 pp
SWE-bench-lite	30	1306/2400 (54.42%)	1412/2400 (58.83%)	+4.42 pp
Toolathlon	30	1282/2400 (53.42%)	1389/2400 (57.88%)	+4.46 pp
Overall	90	3920/7200 (54.44%)	4234/7200 (58.81%)	+4.36 pp

A fixed 10,000-draw bootstrap resamples paired seeds and tasks within benchmark blocks while keeping the two methods paired. The resulting 95% percentile interval for the overall success-rate

difference is $[+2.35, +6.36]$ percentage points. The block-level signs are consistent across all three benchmarks, so the task-success result is not driven by a single benchmark block.

J.3 Prevalence scope for shared patterns

The planted-prevalence sweep clarifies what interaction energy measures. It uses $K = 14$ and $c = 4$ with a panel-specific high-score pattern $E = (0.9, 0.5, 0.5, 0.5)$ and a reference pattern $H = (0.5, 0.5, 0.5, 0.5)$, holding external utility fixed. When the high-score pattern is rare, its residual energy is larger than the reference rows' energy, so ISA can reverse its row-mean advantage. As the pattern becomes common, the panel reference shifts and the comparison becomes a population-composition effect rather than an isolated row interaction.

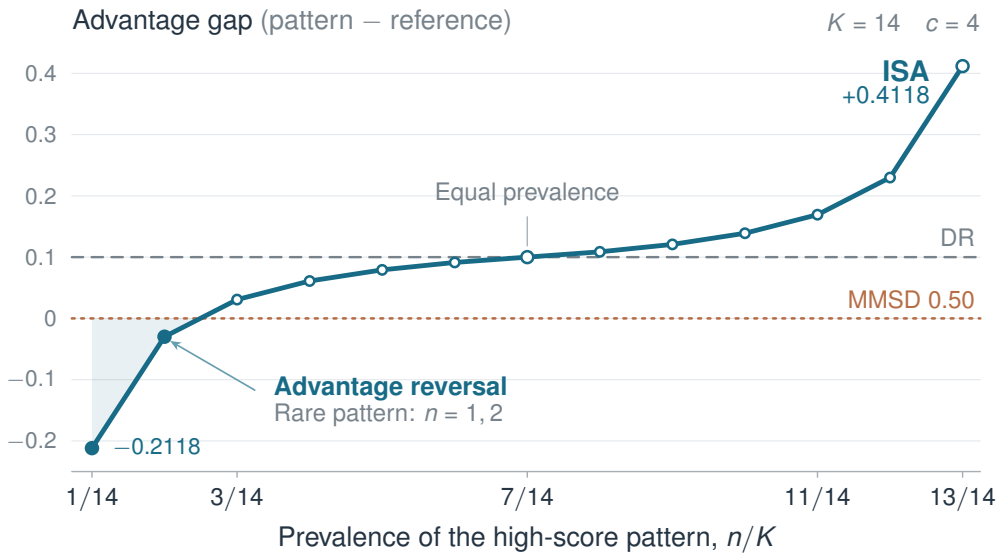


Figure 4: **ISA responds to pattern prevalence through the panel reference.** In the analytic $K = 14, c = 4$ sweep, ISA reverses the pattern–reference advantage at $n = 1, 2$ (shaded area). The gap rises as the pattern becomes more common and equals the DR gap at equal prevalence ($n = 7$). DR remains at $+0.10$; MMSD 0.50 remains at zero. Markers show the 13 interior compositions; lines connect adjacent values.

The first two interior points show reversal ($n = 1$ gives an ISA advantage gap of -0.2118, and $n = 2$ gives -0.0299 for pattern minus reference). From $n = 3$ onward the high-score pattern retains a positive ISA advantage, reaching $+0.1000$ at the balanced point and $+0.4118$ at $n = 13$. The sweep helps interpret the penalty as a panel-relative interaction regularizer: it distinguishes isolated evaluator-specific dependence from row effects that are common in the current block. The $+0.4118$ figure is the analytic boundary of the $E_{\text{count}} = 13$ composition (one evaluator preferring 13 of 14 trajectories), which exceeds DR's constant $+0.10$; it is a conditional amplification that matters only if that composition is actually common during training. The next subsection tests that prevalence empirically.

J.4 Training-period companion-composition dynamics

The analytic prevalence sweep and initial-policy composition audit are both static. To check whether the $E_{\text{count}} = 13$ dominant mode emerges or worsens during on-policy training, we audit frozen checkpoints at updates $0/60/120/180/240$. For each seed, update, and task we draw a *contemporaneous* 64-trajectory pool (no cross-checkpoint borrowing) and reuse the composition-audit design: eight hash-fixed anchors plus 100 without-replacement six-companion draws from the remaining 56, scored by evaluators A1–A4 with four replications. The full design is 10 seeds $\times$ 5 checkpoints $\times$ 40 tasks $\times$ 64 trajectories $\times$ 4 evaluators $\times$ 4 replications $=$ 2,048,000 valid grading replications. Update 0 retains ten independent sampling seeds on the shared initial checkpoint so that seed vari-

ation is not a tenfold copy of one calibration bank. Interaction signatures (E/H) are frozen before scores are viewed. This uses a fresh cohort and an update-wise mean-over-seed-and-task estimand; it is explicitly *not directly comparable* to the initial-policy audit's median estimand.

We separate $E_{\text{count}} = 13$ from $E_{\text{count}} = 1$, 14, and 0: the exact-13 event merges only the first two, and the fully homogeneous $14/14$ and $0/14$ blocks are excluded from the singleton-condition denominator. Table 33 reports the crossed seed×benchmark-task bootstrap (10,000 draws).

Table 33: Training-period composition dynamics across frozen checkpoints (fresh cohort; update-wise estimand). The estimated exact-13 rate is the prevalence under the preregistered 64-pool resampling design, not a natural training-block frequency; it stays below 1% and does not increase with training.

Update	0	60	120	180	240
Anchor Spearman (mean)	0.8906	0.8893	0.8841	0.8927	0.8885
Advantage-sign consistency	0.9171	0.9217	0.9239	0.9201	0.9247
Estimated exact-13 rate	0.333%	0.320%	0.295%	0.323%	0.310%

The smallest anchor-Spearman lower bound over the five checkpoints is 0.8695 and the smallest advantage-sign lower bound is 0.9033. The largest exact-13 upper confidence bound is 0.723%, and the update-240-minus-0 change is -0.023 percentage points (95% CI $[-0.443, +0.408]$), so the dominant single-evaluator composition is rare and does not grow during training. Across all draws there are 593 blocks with $E_{\text{count}} = 13$ and 39 with $E_{\text{count}} = 1$; excluding the 31 pure-$14/14$ and 2 pure-$0/14$ blocks, the singleton receives the maximum interaction penalty in $474/632 = 75.0\%$ of the conditioned blocks. The conditional amplification therefore does occur when a single evaluator dominates, but that regime is empirically uncommon and non-worsening; we retain it as a boundary rather than a resolved case.

J.5 Normalization behavior near small positive medians

The deployed ISA rule uses $\lambda = \kappa/\sqrt{q_{0.5}(P_+)}$ with zero energies excluded from the positive-energy quantile. Appendix E.3 characterizes the boundary case in which bounded rewards can still produce large scaled penalties when the positive-energy median is very small. The algebraic diagnostic evaluates a candidate floor,

$$\lambda_{\text{floor}} = \frac{\kappa}{\sqrt{\max(q_{0.5}(P_+), \delta\,\bar{p})}}, \qquad \delta = 0.05,$$

on three algebraic blocks. It is a numerical check of the normalization rule; the deployed experiments use the rule in Eq. (4).

Table 34: Normalization behavior under a candidate floor. The floor reduces finite scaled penalties in small-positive-median algebraic blocks. The deployed training rule remains Eq. (4).

ϵ	Positive-energy median	Original max λp	Floor-rule max λp
10^{-1}	10^{-2}	0.0938	0.0938
10^{-3}	10^{-6}	9.3750	0.4437
10^{-6}	10^{-12}	9375.0000	0.4437

The primary training results complement this algebraic check: all 60 planned runs completed without non-finite loss, gradients, or parameters (Appendix F.4). Together, the exact-arithmetic description, the algebraic stress case, and the completed runs separate the normalization rule's mathematical boundary from the realized numerical regime. Absence of non-finite events alone does not establish a controlled update tail, so the next subsection trains the floor rule and compares its realized tails, clipping, and outcomes against the un-floored rule.

J.6 Empirical floor-versus-original paired training

The algebraic case bounds the worst case but not the behavior at realized medians. To measure that behavior, we train the floor rule $\lambda_{\text{floor}} = \kappa/\sqrt{\max(q_{0.5}(P_+), \delta\bar{p})}$ ($\delta = 0.05$, chosen a priori

and not re-searched) against the original rule in a paired study. Twenty new paired seeds run 240 updates with $K = 14$, $c = 4$, $r = 4$ ($\approx$ 2,150,400 valid training grading replications total). Within a pair, the two arms share the initial model, training seed, task set and order, sampling seed, evaluator configuration, and evaluation samples; they differ only in whether the floor is enabled and in the resulting checkpoints. These runs are independent of the reference-target training cohort.

Stability, tail, and clipping. The floor fires on 2.854% of updates. Its primary stability endpoint, the paired ratio of the per-seed q99 of block-max λp, is $q99_{\text{floor}}/q99_{\text{original}} = 0.5723$ (95% CI $[0.5577, 0.5869]$). Across all $240 \times 14 = 3{,}360$ trajectory-update observations per pair-arm (Table 35), the floor lowers the realized absolute-advantage and pre-clipping gradient-norm tails, with the largest effect at the maximum: absolute-advantage max falls from 0.52505 to 0.34542 (paired -0.179635, 95% CI $[-0.200127, -0.159143]$) and pre-clip gradient-norm max from 1.22593 to 1.10626 (paired -0.11967, 95% CI $[-0.19128, -0.04806]$). The gradient-clipping rate drops by 1.000 percentage point (95% CI $[-1.643, -0.357]$). All three non-finite event classes remain 0 in both arms.

Table 35: Realized floor-versus-original tails over all 3,360 trajectory-update observations per pair-arm (20 paired seeds; $(n-1)q$ linear-interpolation quantiles). Differences are floor minus original with 95% paired CIs. These tail metrics are descriptive secondary endpoints rather than prespecified primary outcomes.

Quantity	Level	Original	Floor	Floor − original [95% CI]
|advantage|	p95	0.15832	0.15810	-0.000221 $[-0.000373, -0.000068]$
	p99	0.21127	0.20912	-0.002148 $[-0.002964, -0.001333]$
	max	0.52505	0.34542	-0.179635 $[-0.200127, -0.159143]$
Pre-clip grad norm	p95	0.88717	0.86366	-0.02351 $[-0.03822, -0.00880]$
	p99	1.05914	1.00273	-0.05641 $[-0.09574, -0.01709]$
	max	1.22593	1.10626	-0.11967 $[-0.19128, -0.04806]$

Performance and success are preserved. The floor does not degrade the learning outcome. The floor-minus-original rollout-discrepancy difference is -0.000186 (95% CI $[-0.000297, -0.000076]$), the held-out-reward difference has one-sided 95% lower bound $+0.000443$ (above the -0.01 safeguard), and the N0 official-success difference is $+0.757$ percentage points (95% CI $[-0.247, +1.761]$). N0 does not enter the primary gate family and carries no completed Holm correction, so we report "no detected decrease" and do not claim non-inferiority, equivalence, or improvement on it. The floor is an empirically supported stabilizer: it cuts the update tail with negligible change to discrepancy, reward, and success relative to the original rule, and the main-text results still use the un-floored rule in Eq. (4).

K Evaluator-Perturbation Checks

Evaluator perturbations test two separate properties: detection of fixed column offsets and sensitivity of training to changes in scoring rules. We first calibrate a screening statistic under controlled offsets, then examine training with altered evaluators. These checks complement the Family-B/C transfer evaluation by characterizing how evaluator construction changes the measured interaction signal.

K.1 Calibration setup

The diagnostic flags evaluator-column effects estimated from a frozen reference policy with r replications per cell. In Monte Carlo calibration, a fixed additive bias Δ_{corr} is applied to all trajectory scores from one or two variants. We use 2×10^4 trials per cell and pilot-derived noise $\sigma_\varepsilon = 0.0604$ and legitimate variant spread $\sigma_\beta = 0.0399$.

K.2 Calibration performance

Testing a zero main effect. Studentizing $\hat{\beta}_j$ and testing $H_0 : \beta_j = 0$ at $\alpha = 0.05$ yields 97.9% recall for $\Delta = 0.20$ with $r = 4$. The clean-variant flag rate is 21.3% and rises to 71.8% at $r = 64$,

showing that increased replication makes the test sensitive to legitimate evaluator differences as well as corruption. Semantic equivalence does not require a zero main effect.

Magnitude-threshold diagnostic. Thresholding $|\hat{\beta}_j|$ against the legitimate spread keeps the clean flag rate low in the tested large-offset settings. At $r = 16$ and threshold 0.10, recall is 91.3% for $\Delta = 0.20$ and 100% for $\Delta = 0.30$, with a 0.7% clean flag rate. Threshold 0.08 gives 97.1% recall and 3.0% clean flags. Recall is 25.1%, 4.6%, and 1.3% for $\Delta = 0.10$, 0.05, and 0.02, respectively, matching the smaller effect sizes.

Half-shifted panel. Shifting two of four variants by $\Delta = 0.30$ produces 91.0% recall and a 91.1% clean flag rate. The shift increases the grand mean by 0.15, moving the centered clean-column effects toward $\hat{\beta}_j \approx -0.15$ and beyond the magnitude threshold. This setting illustrates how column-effect screening depends on panel composition.

K.3 Training response to scoring-rule changes

The training intervention modifies evaluator scoring rules, rather than adding the fixed column offsets used for calibration. With all four evaluators valid, the reported gap is 0.046 ± 0.005 for ISA and 0.153 ± 0.006 for DR. In the reported one-corrupted setting ($\Delta = 0.2$), the ISA gap is 0.087 ± 0.008; with two corrupted evaluators ($\Delta = 0.3$), it is 0.126 ± 0.009, compared with 0.190 ± 0.004 for DR. Relative to the uncorrupted values, the latter setting increases the gap by factors of approximately 2.7 for ISA and 1.25 for DR. These are relative increases from different starting gaps.

Removing flagged evaluators. Discarding flagged variants moves the ISA gap toward the uncorrupted value of 0.046: the reported gaps are 0.057 ± 0.003 and 0.057 ± 0.006 after one and two corrupted evaluators, respectively. The reported suppression measure remains at 0.40 and 0.60, compared with 1.00 in the uncorrupted condition. This intervention was not separately applied to the all-valid condition, and these reported values do not by themselves establish full recovery of the gap.

K.4 Interpretation

The calibration and training interventions answer different questions: offset detection characterizes a screening statistic, while the training intervention measures policy response under changed scoring rules. Removing flagged evaluators also changes c and the panel defining the interaction. The single-shifted and half-shifted settings therefore serve as controlled sensitivity analyses for evaluator construction, complementing the Family-B/C transfer results.

L Cohort Structure, Calibration, and Statistical Scope

This section consolidates cohort definitions, aggregation-rule sensitivity, calibration parameters, task overlap, and evidence separation. It distinguishes within-cohort method contrasts from descriptive cross-cohort comparisons and states what can and cannot be checked from the submitted PDF.

L.1 Protocol overview

The aggregation sensitivity analysis compares frozen ISA checkpoints from the primary and residual-target ablation cohorts while retaining each cohort's seed identities. Both used ten seeds and identical $15/10/15$ task quotas, but `T_MAIN` and `T_ABLATION` are disjoint. Their native held-out aggregation rules are P and Q, respectively. The main-study tables retain their native P estimand; the unified-Q results below are diagnostics, not replacements for the primary ISA–MMSD contrast.

Same-evidence aggregation contrast. Applying P and Q to the same main-study evidence leaves its training-panel scores unchanged. Mean D^{global} is 0.06037 under P and 0.00897 under Q. The paired P-minus-Q difference across ten seeds is 0.05140 (95% paired-t CI $[0.05054, 0.05226]$). The original main-study in-panel pairing is retained, with $R_{\mathrm{in}} = 0.7346 \pm 0.0046$. This identifies

Table 36: Protocol overview for reported cohorts. Each row defines the comparison unit, evaluator role, endpoint, and valid comparison scope; method effects are interpreted only within a fixed cohort unless a paired extension is stated.

Cohort	Task and methods	Evaluator role	Endpoint and selection role	Direct comparison scope
Primary matched study	`T_MAIN`: 15/10/15 tasks; six arms, ten seeds (60 runs)	A1–A4 training; A5–A6 held out; aggregation P	Primary D^{global} and held-out reward	Six arms within the matched 40-task cohort
Residual-target ablation	`T_ABLATION`: disjoint 15/10/15 tasks; five arms, ten seeds	Family-A evaluators; aggregation Q	D^{rollout} and held-out reward; mechanism check	Residual-target arms within the ablation cohort
MMSD sensitivity	Same frozen 40 task IDs; one ISA plus five MMSD arms, ten seeds (60 separate runs)	Same Family-A protocol; aggregation P	Primary D^{task}; post-hoc D^{global}	ISA and MMSD at stated coefficients
Unseen-task transfer	`U180`: 180 tasks disjoint from both training sets; ISA, MMSD, DR, ten seeds	A5–A6 held-out settings	D^{task} bootstrap; held-out extension	Three methods on the unseen-task cohort
Cross-family transfer	Same evaluation trajectories; ISA versus MMSD, ten pairs	Family B and Family C	Family-B/C discrepancy and reward; cross-family extension	ISA–MMSD pairs under each independent family
Coefficient confirmation	`T_MAIN`: 10-task development (24 runs) / 30-task confirmation (20 runs); DC-SD and ISA	D1–D2 select; A5–A6 confirm under P	Primary D^{rollout}, R_{out}; coefficients frozen	Tuned DC-SD versus tuned ISA only
Task-success confirmation	Shared `N090`, 30 tasks per benchmark; DC-SD and ISA, ten pairs	$N0_{\mathrm{confirm}}$ binary success	Binary success-rate difference	Paired contrast within each benchmark block and overall
Shared-pattern scope	Analytic $K = 14, c = 4$ prevalence sweep; DR, MMSD, ISA scores	Constructed evaluator panel	Pattern–reference gap	Analytic score rules within the planted-prevalence block
Normalization diagnostic	Algebraic blocks and completed training runs; ISA normalization rule	Fixed scored blocks	Max scaled penalty and finite training behavior	Normalization variants in the stated algebraic blocks
Bounded-shift audit	40 frozen-policy blocks, $K = 14$; 3,640 unordered trajectory pairs	A1–A4 original $[0, 1]$ cell means	Feasible-box extrema and minimum ℓ_∞ reversal radius; task bootstrap	Fixed-block diagnostic only; distinct from observed raw/DC disagreement
Companion-composition audit	`T_MAIN`; 64 initial-policy trajectories per task; eight anchors and 100 companion draws	A1–A4, four valid replications per cell	ICC, rank, top-quartile, sign, and CV stability; task bootstrap	Current-block estimates versus frozen full-pool references
Reference-target comparison	`T_MAIN`; four targets, 20 pairs; shared `N090`	A1–A4 training; external uses $N0_{\mathrm{cal}}$; $N0_{\mathrm{confirm}}$ evaluates	D^{rollout}, held-out reward, and N0; paired comparison	Current-block versus historical, external-anchor, and shrinkage targets at matched budgets
P/Q sensitivity	Two frozen ISA cohorts on both 40-task sets; ten independent seeds per cohort	Native P/Q and unified Q	Same-evidence contrast; independent-seed equivalence	Descriptive aggrega-tion/task/policy decomposition
Raw/DC success	Ablation checkpoints; ten pairs on shared `N090`	$N0_{\mathrm{confirm}}$ only	Official binary success	Raw-Energy versus DC-Energy under one adaptive rule

an aggregation-rule contribution for this cohort and does not imply the same shift for other policies. Appendix L.2 states both functions and gives an illustrative arithmetic check. Applying both rules uniformly to all six T_MAIN and five T_ABLATION arms preserves the prespecified D^{rollout} method-mean ordering (Spearman 1.00 in both cohorts); for D^{global}, the ordering is unchanged in T_MAIN and has one adjacent inversion in T_ABLATION.

Frozen-policy evaluation on both task sets. Each policy is evaluated on both task sets using eight rollouts per task and aggregation Q. Existing same-task evaluations are reused; the two cross-task evaluations generate new trajectories. Table 37 reports the four combinations. For either policy, R_{in} is approximately 0.734 on T_MAIN and 0.688 on T_ABLATION; the native in-panel difference is therefore predominantly associated with task composition in this crossed comparison, not a large same-task policy difference.

Table 37: Frozen ISA policies evaluated on both task sets under Q. Values are mean±sample SD over ten seeds per policy cohort.

Policy cohort	Task set	D^{global}	D^{task}	D^{rollout}	R_{in}	R_{out}
Main	T_MAIN	0.0090 ± 0.0020	0.0694 ± 0.0018	0.0694 ± 0.0018	0.7346 ± 0.0046	0.7256 ± 0.0043
Ablation	T_ABLATION	0.0076 ± 0.0044	0.0590 ± 0.0035	0.0605 ± 0.0030	0.6875 ± 0.0055	0.6799 ± 0.0011
Main	T_ABLATION	0.0077 ± 0.0043	0.0592 ± 0.0034	0.0608 ± 0.0030	0.6880 ± 0.0054	0.6803 ± 0.0011
Ablation	T_MAIN	0.0088 ± 0.0018	0.0691 ± 0.0018	0.0692 ± 0.0018	0.7340 ± 0.0045	0.7252 ± 0.0040

Within-task-set equivalence assessment. The specified absolute margins are 0.005 for D^{global}, R_{in}, and R_{out}, and 0.003 for D^{task} and D^{rollout}. For each task set, 10,000 bootstrap draws resample the two independently trained seed groups independently; equal seed labels across cohorts are not assumed to create matched pairs. Table 38 reports the resulting 90% intervals. All ten intervals lie inside their corresponding margins. This supports equivalence only for the two fixed task sets, these endpoints, the specified tolerances, and this bootstrap analysis; it is not inferred merely from a nonsignificant difference, nor generalized to all tasks. The submitted PDF does not include the resampling indices or timestamped evidence that the margins were registered before this analysis, so the intervals cannot be replayed exactly and the timing claim is not independently verifiable from the paper.

Table 38: Main-policy minus ablation-policy differences under Q, comparing independent seed groups on the same task set. Intervals and differences are rounded to five decimals.

Task set	Endpoint	Difference	90% CI	Margin
T_MAIN	D^{global}	$+0.00017$	$[-0.00119, +0.00148]$	± 0.005
	D^{task}	$+0.00030$	$[-0.00096, +0.00159]$	± 0.003
	D^{rollout}	$+0.00020$	$[-0.00106, +0.00149]$	± 0.003
	R_{in}	$+0.00055$	$[-0.00265, +0.00376]$	± 0.005
	R_{out}	$+0.00038$	$[-0.00257, +0.00332]$	± 0.005
T_ABLATION	D^{global}	$+0.00010$	$[-0.00296, +0.00311]$	± 0.005
	D^{task}	$+0.00020$	$[-0.00223, +0.00259]$	± 0.003
	D^{rollout}	$+0.00030$	$[-0.00181, +0.00238]$	± 0.003
	R_{in}	$+0.00050$	$[-0.00334, +0.00427]$	± 0.005
	R_{out}	$+0.00040$	$[-0.00037, +0.00116]$	± 0.005

Descriptive decomposition of native differences. For any reported endpoint, let Q_{XY} be its seed mean for policy cohort $X \in \{M, A\}$ evaluated on task set $Y \in \{M, A\}$ under Q, and let P_{MM} be

the native main-study mean. Then

$$P_{MM} - Q_{AA} = \underbrace{(P_{MM} - Q_{MM})}_{\text{aggregation rule}} + \underbrace{\tfrac{1}{2}[(Q_{MM} - Q_{MA}) + (Q_{AM} - Q_{AA})]}_{\text{average task-set contrast}} + \underbrace{\tfrac{1}{2}[(Q_{MM} - Q_{AM}) + (Q_{MA} - Q_{AA})]}_{\text{average within-task-set policy contrast}}. \tag{32}$$

This additive identity holds for the endpoint means; it does not express causal mediation or a causal percentage attributable to scoring or tasks. Table 39 reports six decimals to retain the small residual components. The unrounded components recover the native difference; displayed sums can differ in the final decimal through rounding.

Table 39: Descriptive decomposition of native main-P/T_MAIN minus ablation-Q/T_ABLATION. Task and policy terms are averaged across the two crossed comparisons.

Endpoint	Native difference	Aggregation rule	Task-set term	Policy term
D^{global}	$+0.052770$	$+0.051400$	$+0.001235$	$+0.000135$
D^{task}	$+0.010400$	$+0.000000$	$+0.010150$	$+0.000250$
D^{rollout}	$+0.014900$	$+0.006000$	$+0.008650$	$+0.000250$
R_{in}	$+0.047050$	$+0.000000$	$+0.046525$	$+0.000525$
R_{out}	-0.005720	-0.051400	$+0.045290$	$+0.000390$

The aggregation rule accounts for most of the native aggregate-discrepancy difference, whereas task composition dominates the in-panel-reward difference. The smaller within-task-set policy terms accord with the equivalence intervals. These results explain why native cohort scales should not be compared as method effects; the primary ISA–MMSD and ablation Raw/DC contrasts remain within their respective cohorts and protocols.

L.2 P/Q AGGREGATION DEFINITIONS AND SENSITIVITY ANALYSIS

P and Q are two frozen aggregations of the *same* held-out evaluator scores, not independent measurements. For one evaluation record, let $\bar{R}^{\text{out}}_{\cdot j}$ be evaluator j's held-out cell mean over valid replications. The two rules are

$$\mathcal{E}_P = \frac{1}{|J^{\text{out}}|} \sum_{j \in J^{\text{out}}} \bar{R}^{\text{out}}_{\cdot j}, \qquad \mathcal{E}_Q = \frac{\sum_j w_j \,\text{clip}_{[0,1]}\big(s_j\, \bar{R}^{\text{out}}_{\cdot j} + o_j\big)}{\sum_j w_j},$$

where (s_j, o_j, w_j) are frozen per-evaluator affine scales, offsets, and reliability weights. $\mathcal{E}_P$ is the equal-weight raw mean; $\mathcal{E}_Q$ applies the frozen affine calibration, clips to $[0, 1]$, and takes a reliability-weighted mean. Both fail closed if a required value is missing or non-finite and use a fixed evaluator order. The signed discrepancy is $R_{\text{in}} - \mathcal{E}_v$. Algorithm 3 specifies these aggregation functions; the submitted PDF does not include the record-level inputs needed to regenerate every reported endpoint.

Algorithm 3 Functional pseudocode for the frozen P/Q aggregation rules

Require: Training-panel reward R_{in}; frozen evaluator order J^{out}
Require: Held-out cell means $\{\bar{R}^{\text{out}}_{\cdot j}\}_{j \in J^{\text{out}}}$; frozen tuples $\{(s_j, o_j, w_j)\}_{j \in J^{\text{out}}}$
if a required value is missing or non-finite, or $\sum_j w_j \le 0$ **then**
  return INVALID ▷ fail closed; do not average a subset
end if
$\mathcal{E}_P \leftarrow |J^{\text{out}}|^{-1} \sum_{j \in J^{\text{out}}} \bar{R}^{\text{out}}_{\cdot j}$
for all $j \in J^{\text{out}}$ in the frozen order **do**
  $z_j \leftarrow \min\big(1, \max\big(0, s_j \bar{R}^{\text{out}}_{\cdot j} + o_j\big)\big)$
end for
$\mathcal{E}_Q \leftarrow \dfrac{\sum_j w_j z_j}{\sum_j w_j}$
$d_P \leftarrow R_{\text{in}} - \mathcal{E}_P$; $d_Q \leftarrow R_{\text{in}} - \mathcal{E}_Q$
return $(\mathcal{E}_P, \mathcal{E}_Q, d_P, d_Q)$

Frozen deployed parameters. The two-held-out-evaluator layout uses x_5 and x_6, each the mean of four valid A5 or A6 replications. Its deployed rules are

$$z_5 = \text{clip}_{[0,1]}(1.25x_5 - 0.125), \quad z_6 = \text{clip}_{[0,1]}(0.80x_6 + 0.10),$$
$$\mathcal{E}_P = \frac{x_5 + x_6}{2}, \qquad \mathcal{E}_Q = 0.80z_5 + 0.20z_6.$$

The affine maps were fit independently by least squares on four frozen calibration tasks (F01–F04) against the mean continuous score of two human raters, without fit-stage clipping or regularization. Reliability weights were fixed on a disjoint four-task set (W01–W04) using inverse post-calibration MSE: 0.0001 for A5 and 0.0004 for A6, corresponding to raw weights $10{,}000$ and $2{,}500$ and normalized weights 0.80 and 0.20. The continuous calibration target is used only to fit Q and is distinct from the binary $N0_{\text{cal}}$ labels used for the external-baseline maps. Evaluator order and all calibration parameters were frozen before policy evaluation.

All-method sensitivity under both aggregation rules. Table 40 applies both fixed rules to every arm using the same underlying evaluator scores. The prespecified D^{rollout} method-mean ordering is identical under P and Q (Spearman $= 1.00$ in both cohorts), and all recomputed seed endpoints satisfy $D^{\text{global}} \le D^{\text{task}} \le D^{\text{rollout}}$. For the displayed D^{global} means, the ordering is unchanged in `T_MAIN` but Raw-Energy and MMSD exchange adjacent positions in `T_ABLATION`; the best and worst arms remain unchanged in both cohorts. Each primary within-table comparison uses one fixed native aggregation rule. The sensitivity results therefore show substantial scale dependence with mostly stable ordering in these cohorts, not a universal shared scale factor or invariance to the aggregation rule.

Table 40: All-method sensitivity of D^{global} to P versus Q (ten seeds; mean±sample SD), computed from identical underlying evaluator scores. Ordering is unchanged in `T_MAIN`; one adjacent inversion occurs in `T_ABLATION`.

Cohort	Method	D^{global} under P	D^{global} under Q
`T_MAIN`	GRPO ($c = 1$)	0.16170 ± 0.00420	0.03400 ± 0.00300
	DR	0.13290 ± 0.00520	0.02800 ± 0.00280
	DR+	0.09230 ± 0.00330	0.01950 ± 0.00250
	Pess-min	0.08410 ± 0.00350	0.01750 ± 0.00230
	MMSD 0.50	0.08329 ± 0.00340	0.01477 ± 0.00250
	ISA	$\mathbf{0.06037 \pm 0.00160}$	$\mathbf{0.00897 \pm 0.00200}$
`T_ABLATION`	Raw-SD	0.05000 ± 0.00450	0.01560 ± 0.00710
	Raw-Energy	0.04400 ± 0.00420	0.01070 ± 0.00500
	MMSD 0.50	0.04100 ± 0.00410	0.01250 ± 0.00620
	DC-SD	0.03800 ± 0.00400	0.00810 ± 0.00680
	DC-Energy-ISA	$\mathbf{0.03400 \pm 0.00380}$	$\mathbf{0.00760 \pm 0.00440}$

L.3 Task intersections, evidence reuse, and accounting

The task union contains 260 unique IDs. T_MAIN and T_ABLATION each contain 40 tasks (15/10/15) and are disjoint; U180 contains 180 tasks (60/60/60) disjoint from both. The coefficient study's DEV10 (4/3/3) and CONFIRM30 (11/7/12) partition T_MAIN. Thus confirmation is isolated from coefficient selection, not from all project training or calibration (Appendix J.1). The task-success, reference-target, and Raw/DC studies use the same fixed 90-task subset of U180, not three disjoint test sets.

Table 41: Task-ID intersection counts; diagonal entries are set sizes. M: T_MAIN; A: T_ABLATION; D: coefficient development; C: coefficient confirmation; U: U180; S, T, R: task-success, reference-target, and Raw/DC N090, respectively.

	M	A	D	C	U	S	T	R
M	40	0	10	30	0	0	0	0
A	0	40	0	0	0	0	0	0
D	10	0	10	0	0	0	0	0
C	30	0	0	30	0	0	0	0
U	0	0	0	0	180	90	90	90
S	0	0	0	0	90	90	90	90
T	0	0	0	0	90	90	90	90
R	0	0	0	0	90	90	90	90

Table 42: Evidence reuse is distinct from task overlap. Each pair among the three N090 studies shares tasks but no rollout evidence.

Evidence-set pair	Shared tasks	Shared evidence
Composition / reference-target calibration	40	2,560
Calibration / coefficient confirmation	30	0
Calibration / any N090	0	0
Task-success / reference-target / Raw–DC, pairwise	90	0
U180 / any N090	90	0

Identical task IDs do not imply reused rollout evidence: the three N090 studies use distinct policy trajectories, while the composition audit and reference-target comparison intentionally share the same calibration cohort. The P/Q comparison applies two aggregation rules to the same held-out scores; the cross-task comparison instead evaluates frozen policies on newly generated trajectories.

Reproducibility scope. The submitted PDF provides task-set relationships, calibration definitions and fitted parameters, inference units, confidence intervals, and per-seed primary outcomes. It does not provide full-precision cell- or rollout-level scores, bootstrap resampling indices, executable training or evaluation code, complete random-number metadata, or checkpoint-to-code mappings. The formulas, pseudocode, and reported summaries therefore support method reimplementation and selected arithmetic checks, but not exact end-to-end replay or independent reconstruction of every table.

M Per-seed Experiment Values

Table 43 reports per-seed outcomes for the primary dense-rule comparison. Section I.7.1 reports aggregate contrasts for the alternative-rule factorial study.

Table 43: Per-seed values underlying Table 14 for the six-arm Qwen3-8B primary cohort. The independent analysis unit is the training seed ($n = 10$). All recorded signed gaps are positive, so $D = |\Delta| = \Delta$ in this cohort; together with held-out reward, the rows also reconstruct $R^{\text{in}} = R^{\text{out}} + \Delta$ and all paired effects. The identity $\Delta_{as} = R^{\text{in}}_{as} - R^{\text{out}}_{as}$ holds within 10^{-12} for every run.

Metric	Condition	s42	s49	s56	s63	s70
$D = \|\Delta\|$	GRPO	0.155812	0.167866	0.157797	0.164366	0.162825
	DR	0.131512	0.144998	0.132674	0.130976	0.134739
	DR+	0.093327	0.091096	0.092106	0.099153	0.092178
	Pess-min	0.087089	0.083471	0.088076	0.089474	0.085680
	Mean-minus-SD	0.083249	0.083014	0.080014	0.089973	0.080068
	ISA	0.059845	0.059868	0.060218	0.061011	0.058847
Held-out	GRPO	0.571386	0.562367	0.582460	0.569645	0.575417
	DR	0.625299	0.607492	0.630817	0.610975	0.624430
	DR+	0.642651	0.640481	0.655630	0.636443	0.648935
	Pess-min	0.650149	0.625783	0.653469	0.633401	0.643446
	Mean-minus-SD	0.654017	0.653626	0.661209	0.637844	0.653717
	ISA	0.671994	0.669521	0.680096	0.667049	0.675816

Metric	Condition	s77	s84	s91	s98	s105
$D = \|\Delta\|$	GRPO	0.168177	0.157438	0.162042	0.160943	0.160133
	DR	0.135636	0.128975	0.125389	0.132958	0.130742
	DR+	0.088600	0.087450	0.091180	0.094859	0.093151
	Pess-min	0.082097	0.082036	0.083955	0.079877	0.078844
	Mean-minus-SD	0.078366	0.086127	0.083127	0.084303	0.084659
	ISA	0.062813	0.058289	0.063288	0.059039	0.060482
Held-out	GRPO	0.576767	0.573713	0.570541	0.566320	0.575185
	DR	0.621754	0.623991	0.622699	0.612903	0.627941
	DR+	0.649029	0.651155	0.644554	0.642631	0.644691
	Pess-min	0.643614	0.642636	0.638689	0.636811	0.640701
	Mean-minus-SD	0.654139	0.651432	0.651614	0.651780	0.647723
	ISA	0.680008	0.675545	0.673221	0.672672	0.675879

N Additional Background on Evaluator-Dependent Rewards

In agent reinforcement learning, a *harness* comprises the evaluator, prompt scaffold, tool interface, and control flow used to produce a scalar reward (Yao et al., 2023; Qin et al., 2023; Jimenez et al., 2023; Luo et al., 2025). Evaluators intended to assess the same task can assign different rewards to an identical trajectory (Mehta et al., 2026). Such disagreement establishes evaluator dependence, but attributing it to reward gaming requires evidence that the trajectory exploits the evaluator rather than satisfies the task.

A fully crossed design scores every trajectory with the same evaluator panel. Consistent differences in evaluator strictness appear as column main effects; trajectory-specific departures from the additive row-and-column model appear as interactions. Double centering separates these departures from shared shifts, providing the panel-relative signal used by ISA. Section 3 defines the construction, and Appendix C establishes its estimation properties under explicit assumptions.

The source of an interaction remains ambiguous: it may reflect exploitation, legitimate specialization, semantic differences, or scoring noise. Conversely, an error shared by every evaluator can enter the row effect without contributing to the interaction. We therefore evaluate interaction energy as a training regularizer through held-out evaluator outcomes and assess its relation to gaming through separate label-based analyses (Section I.6.1). Appendix A locates these complementary sources of evidence.

O Additional Related Work

Measurement validity and evaluator correction. Reward misspecification and overoptimization arise when measured reward diverges from intended task success (Amodei et al., 2016; Pan et al., 2022; Gao et al., 2023); agent generalization studies examine the related problem of benchmark transfer (Mehta et al., 2026), and harness-level audits quantify how scaffolding and evaluation-harness choices move measured performance for a fixed policy (Kapoor et al., 2025), consistent with our premise that cross-evaluator disagreement carries a harness-attributable component. With LLM judges, evaluator selection and prompting are themselves part of the measurement procedure (Zheng et al., 2023; Liu et al., 2023). Length-controlled evaluation and process-reward length debiasing address particular scoring biases (Dubois et al., 2024; Zheng et al., 2025). ISA instead takes an evaluator panel as given and uses its crossed score structure to define a penalty. Panel validity is a separate measurement question: specialization or scoring drift can produce interaction, and shared errors can remain in the row mean. The proxy-label and human spot-check analyses examine this measurement distinction (Section I.6.1).

Aggregation objectives and direct comparators. Domain randomization motivates varying training conditions (Tobin et al., 2017); our DR baseline implements uniform evaluator averaging. Setting the ISA penalty weight to zero removes its interaction term, although equality of complete training procedures also requires matching baselines and optimizer settings. MMSD is the primary comparator; DR, DR+, and Pess-min provide secondary comparisons. DR+ sets its weights using past marginal reward variance, which combines grading noise with variation across trajectories and is not an estimate of optimal noise weights. Pess-min and MMSD operate on each trajectory's score vector, whereas ISA uses the full block to estimate shared evaluator shifts. Group-robust optimization targets specified group risks (Sagawa et al., 2020), a different objective from minimizing over evaluator scores within a trajectory. Lemma 1 characterizes pairwise changes under the row-mean-plus-penalty rule. The residual-target ablation supplies the direct training comparison (Section I.2).

Decomposition and combination of reward signals. Information-theoretic and multiplicative decompositions study reward structure in RLHF and tool-using agents (Mao et al., 2025; Liu, 2026). Multi-objective RL addresses preferences over several objectives (Yang et al., 2019), while reward shaping can combine correlated signals (Brys et al., 2014). ISA decomposes scores along a different pair of axes: sampled trajectories and evaluator variants intended to measure the same task objective. Its scalar advantage combines the centered row mean with a penalty on double-centered residual energy. Its distinguishing feature is the use of a crossed measurement layout to define the regularization target.

Turn-level advantage in agentic RL. A recent line of work refines credit assignment along the *temporal* axis of multi-turn rollouts. GiGPO forms group-relative advantages at both episode and per-turn level (Feng et al., 2025); SPO estimates segment-level advantages between fixed cut points (Guo et al., 2025); TCPO and Turn-PPO attach turn-level value or advantage estimates to PPO-style updates (Liao et al., 2026; Li et al., 2026); IAPO shapes credit from influence-aware estimates and PGPO from potential functions (Ren et al., 2026; Zheng et al., 2026). These methods decompose *when* within a trajectory reward is assigned. ISA instead decomposes *which evaluator* drives a reward: it operates on the trajectory$\times$evaluator block and suppresses cross-evaluator interaction, leaving the temporal credit-assignment axis untouched. The two axes are orthogonal and composable—a turn-level advantage can be computed on rewards that ISA has already double-centered—so we treat these methods as complementary rather than as baselines.

Statistical lineage and conditional extensions. Two-way interaction contrasts and thresholding are classical statistical tools. The optional partial-observation and sparse-demixing extensions draw on matrix completion (Candès & Recht, 2009; Keshavan et al., 2010) and robust PCA (Candès et al., 2011; Chandrasekaran et al., 2011; Chen et al., 2011). These extensions are not used in dense ISA. Their assumptions about sampling, rank, and sparsity are additional to the complete-block residualization used in the experiments (Section C.4); the cited recovery results do not supply a sparse-support guarantee for the dense penalty.

Relation to environment invariance. Domain-adversarial learning and invariant risk minimization seek stability across environments (Ganin et al., 2016; Arjovsky et al., 2019). Evaluator variants provide a measurement-level notion of environment variation. The ISA row mean is panel-dependent and includes reward increments shared by all evaluators. Proposition 3 instead bounds changes in residual scores when the same panel is reweighted. Transfer to new evaluator implementations is an empirical question addressed by the experiments.

P Consolidated Limitations and Scope

Measurement and attribution. Interaction energy measures trajectory-specific dependence on the selected evaluator panel. It is designed as a training regularizer for evaluator-dependent scoring rather than as a semantic classifier of trajectory intent. Reward increments shared by all evaluators enter the row effect, while panel-specific deviations enter the residual energy. The prevalence sweep makes this scope concrete: sparse single-evaluator exploiters can be suppressed because they carry larger residual energy, but when the exploiter pattern becomes common within the block, the panel reference shifts and the exploiter–honest ISA advantage gap turns positive (Appendix J.3). The proxy- and human-label studies are therefore used as construct-validity diagnostics: adding energy to row mean, SD, and minimum improves AUC by $+0.00506$, and the 300-trajectory human comparison quantifies how the proxy labels align with human judgments (Section I.6.1).

Evaluator and task coverage. The evaluator variants are treated as part of the measurement object: A1–A4 define the crossed training panel, A5–A6 provide same-family held-out transfer, and B/C provide independent implementation checks. Fixed-offset calibration is most informative for large shifts, and the scoring-rule intervention shows that removing flagged variants moves discrepancy toward the uncorrupted reference level (Appendix K). Dense-rule experiments cover Qwen3-8B, two disjoint 40-task training sets, an additional 180-task pool, and two independent evaluator implementations. The task-success, reference-target, and Raw/DC studies reuse the same 90-task subset of that pool with distinct policy evidence; coefficient confirmation is isolated only from coefficient selection. The 14B study is reported as a separate sparse-rule configuration, so the dense-ISA claims are scoped to the model, task distributions, and evaluator configurations specified in the cohort table.

Block and reference dependence. The bounded-shift audit quantifies pairwise ranking sensitivity inside score-feasible boxes; its separately optimized shifts do not model naturally occurring drift or establish one offset vector that reverses every counted pair. The companion-composition audit measures sensitivity at $K = 14$ using eight fixed anchors: high rank agreement coexists with nonzero magnitude variation and sign changes. The reference-target comparison favors current ISA over the fitted historical, external-anchor, and shrinkage targets at a common training grading budget. These findings do not determine the preferred estimator across block sizes, calibration quality, or drift regimes. Appendix H reports the estimates, inference units, and scope restrictions.

Experimental interpretation. Discrepancy endpoints retain different information: D^{global} summarizes the panel average, whereas D^{task} and D^{rollout} preserve finer variation. The residual-target ablation supports double centering on the finer endpoints and held-out reward, and its subsequent Raw/DC evaluation supports official success under the same adaptive rule. The inconclusive aggregate contrast remains a separate view. The P/Q equivalence assessment is limited to the two fixed task sets, aggregation Q, the specified margins, and the reported bootstrap; its additive decomposition is descriptive rather than causal. In the coefficient sweep, the frozen task-level contrasts favor ISA across the grid, and the post-hoc aggregate recomputation is reported for endpoint transparency. These comparisons concern the tested coefficients; the planted-exploiter simulation identifies a regime with small evaluator main effects where MMSD ranks the planted rows slightly better (Section I.4.3).

The fitted rescaling model and simulated shift blocks characterize controlled estimator behavior. The penalty-assignment study compares policies with different sampling and optimization paths, so its targeting and residual shares are interpreted as descriptive contrasts. Fixed-input calculations establish arithmetic allocation. The binary task-success endpoint strengthens the interpretation that lower discrepancy can coincide with better task completion in the tested blocks, while external

utility and trajectory-level intent remain separate constructs. The training comparisons therefore support interaction energy as a regularizer for evaluator-dependent scoring, and the label diagnostics describe how that signal aligns with proxy and human annotations (Sections I.4.2, I.4.3, I.4.4, J).

Normalization and computation. The coefficient κ and quantile θ are fixed design choices. Positive proportional score equivariance requires scaling the baseline and concerns the advantage rule, rather than training invariance or guaranteed pairwise reversal. The normalization analysis characterizes the small-positive-median regime in which bounded rewards can induce large scaled penalties (Appendix E.3); the appendix floor calculation shows a bounded variant for that algebraic regime while keeping the deployed experiments tied to Eq. (4) (Appendix J.5). Across the observed Raw/DC runs, no non-finite events occurred and DC had a lower gradient-clipping frequency, but neither observation guarantees stability in other configurations or eliminates rare-update effects. Full crossing costs Kc evaluations per block before replication, and evaluator count changes panel composition and estimation quality together with cost (Section I.5). The alternative thresholded penalties are reported as exploratory sparse-rule variants, while the main dense-ISA result uses the support-stable penalty defined in Section B.

Ethics Statement

The study uses public agent benchmarks, blinded model-proxy labels, and a blinded human-label comparison. That comparison assesses construct validity within its sampled trajectories. Because interaction can reflect legitimate specialization or scoring drift, evaluator selection shapes which behaviors receive the regularization signal. We therefore pair the penalty analysis with task-success and safety-oriented evaluation.

LLM Usage Disclosure

LLM tools assisted methodology development, experimental design, implementation and scaffolding code, statistical analysis and numerical checking, literature search, and manuscript/LaTeX drafting and editing. The authors remain responsible for the accuracy and integrity of the manuscript.