RETURN BOUNDS NEED NOT LIMIT ADVANTAGES: PREDICTION REWARDS IN LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small coefficients and episodic return bounds are commonly used to limit the influence of auxiliary rewards. We show that neither necessarily constrains the weights used by a normalized policy update. We study a next-observation prediction reward whose consecutive score differences telescope, bounding its total episodic contribution by a coefficient λ. Nevertheless, the resulting prediction-reward conditions on ALFWorld exhibit a striking prediction–task divergence: across Qwen3-1.7B, 4B, and 8B, the rewarded prediction score approaches one while task success approaches zero. The episodic bound remains valid, but it need not survive estimator transformation. In the studied step-pooled Group Relative Policy Optimization (GRPO) implementation, interaction steps from multiple trajectories are standardized before prediction increments telescope. The implemented initialization $\Phi_0 = 0$ makes saturated predictions leave a sparse first-step boundary increment. When task returns tie, pooling this increment with later zeros produces a standard deviation proportional to λ; while this scale exceeds the numerical floor, standard-deviation division approximately cancels λ, yielding order-one update weights. The predicted late raw standard deviation (0.014035) matches the observed median (0.01404), and prediction accounts for 98.74% of late task-tied pre-normalization variance. A tenfold coefficient reduction does not prevent collapse, whereas a denominator intervention removing standard-deviation division restores baseline-range learning in a matched Qwen3-4B comparison. A matched four-arm estimator control further shows that the prediction-mode contrast is negative under step pooling but not when increments are aggregated within trajectories before normalization, yielding an 8.5-point interaction. The failure also recurs on WebShop, where mean-only normalization restores performance. These results show that auxiliary rewards should be evaluated after grouping, temporal aggregation, and normalization: return-level bounds alone do not guarantee update-level safety.

1 INTRODUCTION

Long-horizon large language model (LLM) agents receive sparse outcome feedback, complicating step-level credit assignment (Feng et al., 2025; Wang & Ammanabrolu, 2025; Wang et al., 2025a). Prediction-based auxiliary objectives have long provided denser feedback for representation learning and exploration (Jaderberg et al., 2017; Pathak et al., 2017; Voelcker et al., 2024), and recent LLM/VLM agent systems use next-observation or world-model prediction through auxiliary losses or rewards (Wang et al., 2025b; Yu et al., 2026; Lu et al., 2026). A natural safeguard is therefore to assign the auxiliary reward a small coefficient or to bound its total contribution to episodic return, echoing return-level guarantees in reward shaping and auxiliary-objective design (Ng et al., 1999; Forbes et al., 2024; Seth, 2026). These safeguards constrain the reward as written. Our central question is whether they also constrain the weights that reach the optimizer after credit assignment and normalization.

We examine this distinction using a next-observation prediction reward in ALFWorld (Shridhar et al., 2021). The reward is the change in a rule-based score over selected next-observation fields from one interaction step to the next. The increments telescope, so their episodic sum is bounded in magnitude by a coefficient λ, independently of episode length. Nevertheless, the resulting prediction-reward conditions exhibit the same failure mode with Qwen3-1.7B, 4B, and 8B (Yang

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.

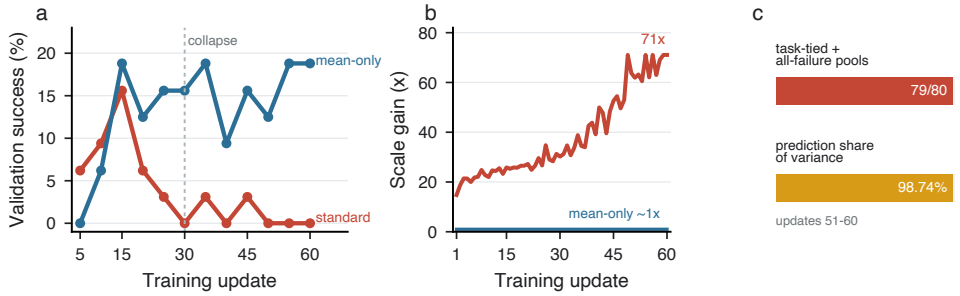


Figure 1: Matched Qwen3-1.7B standard and mean-only arms. **(a)** The standard-normalization arm collapses to zero success, whereas the mean-only arm remains nonzero. **(b)** The normalized-to-raw scale ratio reaches 71-fold under standardization and stays near one under mean-only. **(c)** Across updates 51–60, 79/80 standard-arm pools are task-tied and all-failure, and prediction supplies 98.74% of their pre-normalization variance. Curves show checkpoints or per-update medians.

et al., 2025): the rewarded prediction score approaches one while task success approaches zero. The policy becomes better at the auxiliary problem while becoming worse at the task the reward was intended to support. Figure 1 previews the failure and its rescue in a matched, instrumented 1.7B pair.

The episodic bound remains valid, but it need not constrain the weights after estimator transformation. In the implementation studied here, a terminal task score is assigned to every active interaction step, and steps from multiple trajectories are pooled and standardized, following Group Relative Policy Optimization (GRPO) and related agent reinforcement learning (RL) objectives (Shao et al., 2024; DeepSeek-AI, 2025; Feng et al., 2025). When task returns tie—as in an all-failure group—the task score cannot rank the samples. Because the prediction potential is initialized at zero, saturated scores leave one positive first-step boundary increment per trajectory. Step pooling mixes these increments with later zeros; if this sparse signal controls the denominator, its coefficient appears in both the numerator and standard deviation and approximately cancels while the scaled spread remains above the numerical floor. A reward that is small in return space can therefore become order one in update space. This connects to analyses of standard-deviation bias and low-variance amplification in group-relative objectives (Liu et al., 2025; Salmani-Zarchi et al., 2026); our setting adds a structured, telescoping difference reward in a multi-turn agent.

The distinction is structural, not merely scalar: the bounded object is a signed trajectory sum, whereas the optimizer consumes weights over pooled step samples. Grouping, centering, standardization, and temporal aggregation do not generally commute, so the estimator must be treated as part of the reward's effective specification. Consequently, a useful safety claim must constrain the transformed advantages, not only the reward sequence from which they are derived.

We test each component. A tenfold reduction in λ does not prevent collapse; a denominator intervention removing standard-deviation division restores baseline-range learning at 4B. In an instrumented 1.7B pair, the predicted late raw standard deviation (0.014035) matches the observed late-pool median (0.01404), and prediction supplies 98.74% of task-tied pre-normalization variance. A second matched 1.7B experiment changes temporal order: over updates 30–60, the prediction-mode contrast shifts from -7.6 points under step pooling to $+0.9$ after trajectory aggregation (an 8.5-point interaction). All-failure filtering also restores performance but changes the sample distribution. On WebShop (Yao et al., 2022), mean-only normalization prevents collapse in the small-product setting. Finally, an auxiliary-loss route avoids this reward-path failure at 4B, but content-free targets match or exceed correct targets; avoiding amplification therefore does not establish task-relevant semantic benefit.

Our contributions are:

1. **A counterexample to return-level safety.** An episodically bounded prediction-reward condition exhibits prediction–task divergence across three Qwen3 scales and ALFWorld/WebShop.
2. **An estimator-level mechanism.** Task ties and the implemented first-step boundary quantitatively explain the late denominator under step-pooled standardization.
3. **Matched mechanistic evidence.** Sweeps, telemetry, and matched normalizer and temporal-order interventions test the mechanism and its estimator dependence.

2 How a Bounded Reward Enters the Policy Update

We analyze the advantage weight that enters the policy objective. We first bound the raw episodic contribution. We then follow the same signal through step-pooled group normalization and identify when λ ceases to control its normalized weight.

2.1 Prediction reward and episodic bound

Our training setup follows the step-independent multi-turn rollout framework of Feng et al. (2025). At each environment step, the LLM receives the current observation together with a bounded summary of the interaction history. It then generates reasoning, an action, and selected features of the next observation in a structured `<predict>` block. After the action is executed, a rule-based scorer compares the prediction with the realized observation; Appendix D gives the template and a worked example.

Let $\Phi_{i,t} \in [0,1]$ denote the scorer-weighted accuracy over the rewarded prediction fields at step t of trajectory i. The implementation resets the prediction potential at each episode and sets $\Phi_{i,0} = 0$ before the first scored increment. We define

$$s_{i,t} = \Phi_{i,t} - \Phi_{i,t-1}, \qquad r_{i,t}^{\text{pred}} = \lambda s_{i,t}, \tag{1}$$

for coefficient $\lambda > 0$. Across the T_i scored steps of a trajectory,

$$\sum_{t=1}^{T_i} r_{i,t}^{\text{pred}} = \lambda\big(\Phi_{i,T_i} - \Phi_{i,0}\big), \qquad \left|\sum_{t=1}^{T_i} r_{i,t}^{\text{pred}}\right| \le \lambda. \tag{2}$$

This identity is exact: the raw episodic contribution shrinks with λ and cannot grow with the horizon. The question is whether this bound also controls the weights produced by the estimator.

2.2 Step-pooled group normalization

Each training group contains G trajectories sampled from the same task initialization. In ALFWorld, trajectory i receives a terminal task score $q_i \in \{0, 10\}$, while invalid actions may incur step-dependent penalties $p_{i,t} \ge 0$. The implementation assigns q_i to every active step of trajectory i and treats these steps as optimization samples. It pools samples across trajectories and time,

$$\mathcal{P}_g = \{(i,t) : i = 1, \dots, G;\ t = 1, \dots, T_i\}, \qquad n_g = \sum_{i=1}^{G} T_i,$$

and computes

$$x_{i,t} = q_i + \lambda s_{i,t} - p_{i,t}, \qquad \widehat{A}_{i,t} = \frac{x_{i,t} - \bar{x}_g}{\sigma_{x,g} + \epsilon}. \tag{3}$$

Here $\bar{x}_g$ and $\sigma_{x,g}$ are the mean and standard deviation over $\mathcal{P}_g$, and $\epsilon = 10^{-6}$ is the numerical floor used in our experiments. We refer to this operation as *step-pooled group normalization*. The resulting $\widehat{A}_{i,t}$ weights the response tokens generated at step t in the clipped GRPO objective. Clipping, Kullback–Leibler (KL) regularization, and token-level gradients also shape the update, so $\widehat{A}_{i,t}$ should be understood as an update weight rather than a complete gradient decomposition.

Equation 3 standardizes each increment, whereas the telescoping identity in Eq. 2 applies to the signed trajectory sum that the estimator never forms. The episodic bound and the policy update therefore act at different temporal levels. Step pooling also gives longer trajectories more samples in the optimization pool. We do not vary pooling independently, so we treat it as part of the analyzed implementation rather than an isolated cause.

2.3 When the coefficient cancels

Consider first an all-failure group, for which $q_i = 0$ for every sample. If the penalties are also constant within the pool, Eq. 3 becomes

$$\widehat{A}_{i,t} = \frac{\lambda(s_{i,t} - \bar{s}_g)}{\lambda\sigma_{s,g} + \epsilon} \approx \frac{s_{i,t} - \bar{s}_g}{\sigma_{s,g}} \quad \text{when } \lambda\sigma_{s,g} \gg \epsilon. \tag{4}$$

In this regime, λ scales both the numerator and denominator and approximately cancels, so lowering it need not proportionally shrink the normalized weights. This is a finite-scale statement, not invariance as $\lambda \to 0$: for fixed $\epsilon > 0$, the prediction component vanishes in that limit. In the measured late window, however, the median of $\sigma_{x,g}$ is 0.01404, about $1.4 \times 10^4 \epsilon$. The calculation is exact up to this floor under the stated task- and penalty-tie assumptions; it is not a claim about arbitrary training pools.

With varying penalties, the residual variance is

$$\sigma_{x,g}^2 = \lambda^2 \sigma_{s,g}^2 + \sigma_{p,g}^2 - 2\lambda \operatorname{Cov}_g(s, p). \tag{5}$$

Approximate cancellation follows only when prediction-related variation controls this denominator and the numerical floor is negligible; an all-failure group alone is not sufficient. We therefore log the per-group task, prediction, and penalty variance terms and their covariance in a telemetry-instrumented matched pair (Section 3.3). A coefficient sweep and a direct ablation of standard-deviation division provide complementary tests.

The implemented temporal boundary. The initialization $\Phi_{i,0} = 0$ matters even after later prediction scores saturate. If $\Phi_{i,t} = 1$ for all $t \geq 1$, each length-T trajectory has $s_i = (1, 0, \ldots, 0)$. Pooling G such trajectories therefore gives G values equal to λ and $G(T-1)$ zeros. Under the sample-standard-deviation convention used here,

$$\sigma_{\text{boundary}} = \lambda \sqrt{\frac{G(T-1)}{T(GT-1)}}. \tag{6}$$

For $G = 4$, $T = 50$, and $\lambda = 0.1$, this is 0.014035 and the positive boundary sample receives a z-score of approximately 6.98. The trajectory sum remains exactly λ; the amplification arises from representing that bounded sum as a sparse step-level signal before standardization. This limiting construction supplies a quantitative prediction for the late telemetry rather than an assumption that high mean accuracy implies vanishing increment variance.

The approximation is quantitative. For a task-tied pool with at most N samples, let $\eta_g = (\sigma_{p,g} + \epsilon)/(\lambda \sigma_{s,g})$ and $z_{i,t} = (s_{i,t} - \bar{s}_g)/\sigma_{s,g}$. Whenever $\eta_g < 1$, there is a pool-size constant K_N such that

$$\left| \widehat{A}_{i,t} - z_{i,t} \right| \leq \frac{2K_N \eta_g}{1 - \eta_g}. \tag{7}$$

Thus the update weight approaches a coefficient-free z-score as penalty variation and the numerical floor become small relative to $\lambda \sigma_{s,g}$. Appendix E proves the bound; the telemetry analysis evaluates its substantive premise rather than assuming it from task ties alone.

Step pooling creates a related temporal mismatch. Suppose successive prediction scores change by $+a$ and $-a$; their episodic prediction reward is zero. In a task- and penalty-tied pool where these increments control the variance, Eq. 3 can nevertheless assign substantial positive and negative weights. Over the finite-scale regime $\lambda \sigma_{s,g} \gg \epsilon$, an episodic $O(\lambda)$ bound need not make normalized weights shrink proportionally with λ; this does not make the mismatch necessary for the observed behavioral failure.

The temporal-order control forms the quantity protected by the return-level argument before normalizing it. Because the implementation broadcasts the terminal task score q_i to every active row, we count it once and define

$$X_i = q_i + \sum_{t=1}^{T_i} (x_{i,t} - q_i) = q_i + \lambda(\Phi_{i,T_i} - \Phi_{i,0}) - \sum_t p_{i,t}, \qquad \widehat{A}_i^{\text{traj}} = \frac{X_i - \bar{X}_g}{\sigma_{X,g} + \epsilon}. \tag{8}$$

The trajectory weight is then broadcast back to the same response rows. This keeps the rollouts, row count, loss mask, and optimizer unchanged while moving the telescoping sum before group normalization. It does not make trajectory standardization universally safe: if endpoint differences dominate $\sigma_{X,g}$, λ can still cancel at the trajectory level. It is a targeted intervention on the temporal mismatch in Eq. 3.

2.4 Behavioral signature and targeted interventions

The estimator-level account does not require a behavioral feedback loop. The prediction potential and its first-step boundary depend on generated forecasts, while the resulting update also weights the action tokens produced in the same response. Repetitive behavior and reduced state visitation can therefore accompany collapse without establishing whether they precede or follow it. We treat those measurements as a behavioral signature, not as evidence for a separate causal pathway.

This analysis motivates four tests. First, decreasing λ need not reduce the normalized advantage scale while prediction variation controls the denominator. Second, replacing Eq. 3 with the mean-only weight $\widehat{A}^{\text{mean}}_{i,t} = x_{i,t} - \bar{x}_g$ restores explicit proportionality of the prediction component to λ. Because this intervention changes only the standard-deviation division, the paired result shows that the matched 4B collapse depends on this division. Third, excluding all-failure groups removes updates with no task discrimination, although it also changes the sample distribution and is therefore a less isolated intervention. Fourth, Eq. 8 predicts that aggregating within trajectories before normalization should remove prediction-signal damage caused by stepwise treatment of otherwise telescoping increments. Section 3 reports these tests. We use training telemetry and replay to measure the estimator-level signal and characterize the collapsed policy, not to infer a causal direction between exploration and task failure.

3 Experiments

We test how the bounded prediction reward behaves after the transformation in Section 2. ALFWorld establishes the failure and supports matched interventions on the normalizer and training groups; WebShop tests whether the same pattern extends beyond binary task returns.

3.1 Experimental design

ALFWorld. ALFWorld (Shridhar et al., 2021) is our main testbed. Its terminal task score is 10 for success and 0 for failure, with a horizon of 50 interaction steps. We train Qwen3-1.7B, 4B, and 8B using the step-independent multi-turn agent stack in Section 2.1, implemented in verl-agent (Feng et al., 2025). Unless an ablation changes the named factor, each update uses 32 trajectories (batch 8, group size 4); training lasts 150 updates with temperature 0.4. The telemetry pair uses 60 updates, spanning the standard arm's healthy, transition, and collapse window. A separate matched four-arm 1.7B control also uses 60 updates and crosses prediction reward with step-pooled versus trajectory-first normalization. Appendix C reports the complete configuration.

Evaluation. Training-time validation uses 32 sampled episodes to describe dynamics, including the 60-update telemetry diagnostic. Final-checkpoint battery comparisons use a common 140-game evaluation; Table 14 reports Wilson intervals and the unseen split. We base our claims on the large collapse and matched interventions, not on small endpoint differences. Appendix H reports all training conditions and evaluation protocols.

WebShop. WebShop (Yao et al., 2022) provides a second environment with continuous terminal matching scores and a shorter horizon of 15.

3.2 A bounded reward collapses task learning across scales

Figure 2a shows the central phenomenon. At each model scale, the rewarded-field prediction score rises toward one while task success falls to a sustained floor; the terminal training-time prediction score is 1.000 at each model scale. The unified final-checkpoint evaluation makes the contrast explicit (Table 1): the prediction-reward conditions reach at most 0.7% success, whereas no-prediction baselines range from 17.9% to 45.0%. Because the conditions also differ in whether a prediction block is elicited, this comparison establishes the phenotype rather than isolating reward addition.

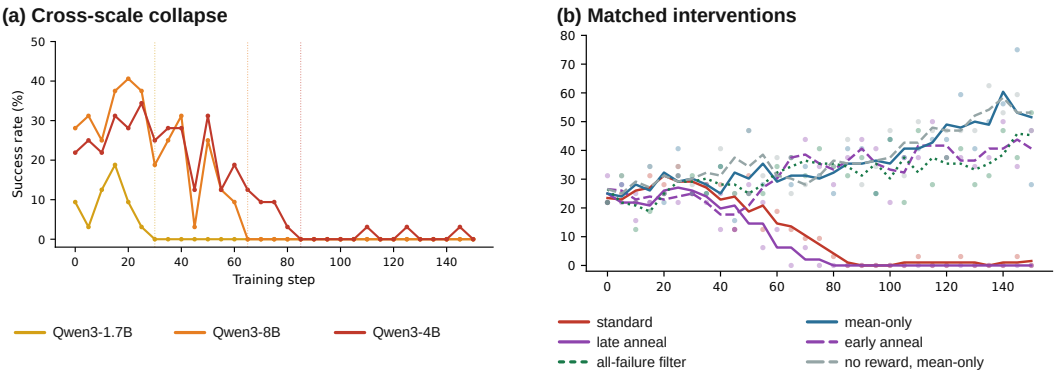


Figure 2: Failure and targeted interventions. **(a)** Qwen3-1.7B, 4B, and 8B prediction-reward conditions all enter a near-zero-success regime; dotted guides mark the first zero-success checkpoint. **(b)** In the matched Qwen3-4B comparison, mean-only normalization, all-failure filtering, and early annealing avoid the zero-success window; standard normalization and later annealing do not.

Table 1: ALFWorld success (%) under the 140-game final-checkpoint evaluation. Wilson intervals are in Table 14.

model	GRPO baseline	+ prediction reward	difference
Qwen3-1.7B	17.9	0.0	−17.9
Qwen3-4B	42.9	0.7	−42.2
Qwen3-8B	45.0	0.0	−45.0

Table 2: Final success for matched Qwen3-4B interventions under the 140-game protocol. Each Δ uses the no-signal reference with the same normalizer; filtering also changes the training data.

condition (Qwen3-4B)	normalization	success (%)	matched Δ
no prediction reward	standard	42.9	reference
prediction reward	standard	0.7	−42.2
no prediction reward	mean-only	57.9	reference
prediction reward	mean-only	52.9	−5.0
prediction reward, all-failure groups dropped	standard	40.7	−2.2

3.3 Standard-deviation division amplifies the tie-breaker

The most direct intervention removes only within-group standard-deviation division. Final-checkpoint success rises from 0.7% to 52.9% (Table 2). The mean-only no-signal control reaches 57.9%, and its interval overlaps the reward arm. The paired result shows that the matched 4B collapse depends on standard-deviation division, but it does not show a benefit from the prediction reward once that division is removed.

Figure 2b shows the intervention dynamics. Standard normalization collapses, whereas the matched mean-only arms continue to improve. All-failure filtering and early annealing also avoid collapse; late annealing does not. Because filtering changes the samples, we treat annealing only as prevention evidence.

Direct measurement of the denominator. A 60-update Qwen3-1.7B matched pair measures the normalization terms; the arms differ only in standard-deviation division (Figure 1). From updates 30–60, 32-episode validation success averages 0.9% under standardization and 15.6% under mean-only. In the standardized arm's final ten updates, 79/80 pools are task-tied and all-failure, and prediction accounts for 98.74% of the pre-normalization variance that sets their denominator. The raw standard deviation falls from 0.0463 early to 0.01404 late. Equation 6 predicts 0.014035, matching this median; no arm deletes only the boundary. The normalized root-mean-square (RMS) weight stays near one, as expected under standardization. Appendix E reports the full decomposition and integrity checks.

Table 3: Telemetry from the matched Qwen3-1.7B pair. Variance shares aggregate the 79 task-tied, all-failure standard-arm pools in updates 51–60.

quantity	standard, 1–10	standard, 51–60	mean-only, 51–60
raw group std	0.04634	0.01404	0.04629
update-weight RMS	0.99748	0.99743	0.04618
task / prediction variance share	—	0.00% / 98.74%	—
penalty / interaction share	—	1.21% / 0.05%	—

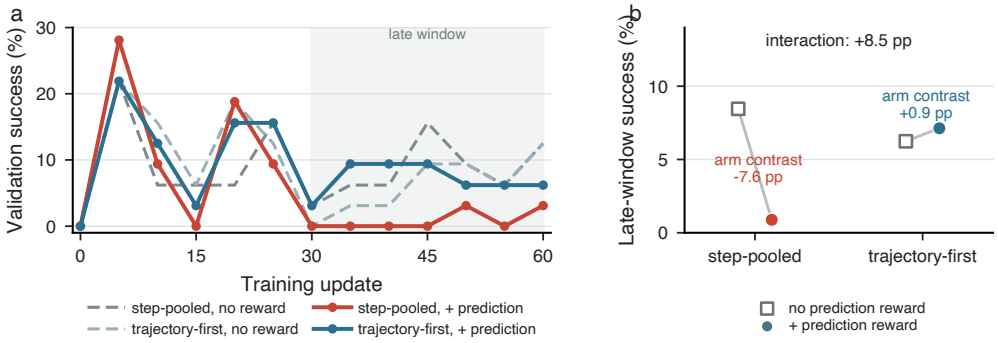


Figure 3: Matched temporal-order control at Qwen3-1.7B. **(a)** Under standard step pooling, the prediction arm enters a near-zero regime; after trajectory-first aggregation, its late curve tracks the no-signal control. **(b)** Mean success over updates 30–60. The arm contrast changes from -7.6 to $+0.9$ percentage points, an 8.5-point interaction. Each checkpoint samples 32 episodes, so the interaction is a checkpoint-level diagnostic rather than a final-evaluation confidence interval.

At update 1, the two arms contain the same 32 trajectories and 1,574 active step samples; raw rewards and group statistics match, and only the normalized weights differ. The late standard arm therefore combines shrinking raw spread with unit-scale weights, whereas mean-only preserves the raw scale. These measurements identify a denominator intervention, not an update-scale-matched optimizer comparison, and do not decompose the post-clipping policy gradient.

Aggregating before normalization tests temporal order. We next compare two complete estimators in a matched factorial. The four Qwen3-1.7B arms cross prediction/no signal with the step-pooled estimator in Eq. 3 versus trajectory-first Eq. 8, using the same 60-update budget and 32-episode validation. Over updates 30–60, the prediction-mode contrast is -7.6 points with step pooling but $+0.9$ points after trajectory aggregation, an 8.5-point interaction (Figure 3).

The interaction shows that prediction-mode-specific damage depends on which of the two complete estimators is used over this window. The trajectory-first arm also changes the normalization population, counts the task score once, and broadcasts one trajectory weight to its rows; it is therefore a targeted test of the temporal mismatch rather than a pure one-factor decomposition. It does not establish universal improvement: all late means remain modest, and a 140-game endpoint remains future work.

Filtering all-failure groups raises final success from 0.7% to 40.7%, near the 42.9% baseline, but changes sample count and composition. We therefore treat it as supporting evidence and the matched mean-only comparison as the factor-matched intervention.

3.4 Magnitude alone does not explain collapse

The 4B coefficient sweep gives a complementary scale test (Appendix Figure 5, left). All $\lambda \in \{0.01, 0.03, 0.1\}$ conditions collapse, and reducing λ tenfold does not contract the all-failure advantage range (approximately ± 7.3, ± 9.1, and ± 7.2). Unlike the instrumented pair, the sweep alone does not identify component shares.

A matched-marginal-variance random-potential control causes a substantial performance reduction (23.4% last-6 success versus 49.5% for baseline) but no sustained zero-success window. Across the

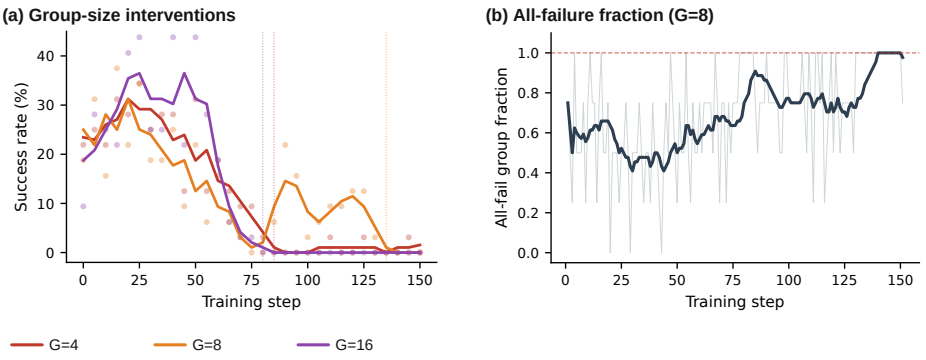


Figure 4: Group-composition diagnostics. All tested group sizes enter zero-success windows with nonmonotonic onset (left); the all-failure fraction approaches one near collapse for group size 8 (right).

signal-handling schemes in Figure 5 (right), mean-only exposes a larger raw magnitude than the decoupled arm, while standardization expands prediction and noise to larger scales with different outcomes. Magnitude alone therefore does not order performance. Differences in temporal structure and action dependence remain potential explanations, and the noise control does not match the temporal correlations of the prediction reward.

3.5 Task-tied groups and behavioral collapse

With a fixed rollout budget, group sizes 4, 8, and 16 enter zero-success windows at updates 85, 135, and 80, respectively (Figure 4); for sizes 8 and 16, the all-failure fraction approaches one near collapse. Because zero success itself creates all-failure pools, this describes the collapsed regime rather than its causal direction.

A matched replay of 12 baseline and 12 collapsed 1.7B episodes provides a behavioral check. It still emits valid actions, but state-visitation entropy falls from 0.91 to 0.09, visited locations from 4.6 to 1.25, and action repetition rises from 0.21 to 0.93. This establishes low diversity, not whether it causes or follows task failure; Appendix J gives details.

3.6 Collapse depends on the estimator

The tested alternatives are Group-in-Group Policy Optimization (GiGPO) (Feng et al., 2025), Proximal Policy Optimization (PPO) with Generalized Advantage Estimation (GAE) (Schulman et al., 2017; 2016), REINFORCE Leave-One-Out (RLOO) (Ahmadian et al., 2024), and REINFORCE++ (Hu et al., 2025). Only GRPO and GiGPO, which retain same-task within-group standard-deviation division, sustain collapse (Table 4). Because the algorithms are not factor-matched, this comparison is descriptive; the matched mean-only intervention remains the main localization evidence.

3.7 WebShop reproduces the collapse and mean-only rescue

The same failure under standard normalization appears on WebShop (Table 4, right). In the small-product regime, standard normalization reduces final-checkpoint mean task score from 60.0% to 0.5%, while the mean-only reward arm reaches 57.5%. In the full regime, the standard-normalized reward arm ends at 0.5%, compared with 28.5% for the baseline, after initially tracking it (Appendix Figure 8). WebShop uses continuous task returns and does not reproduce ALFWorld's strict all-failure pools. WebShop therefore reproduces the collapse and, in the small regime, the mean-only rescue without establishing the same temporal-boundary decomposition.

Beyond the reward channel. These results do not imply that prediction supervision is intrinsically harmful. A related auxiliary-loss route bypasses reward standardization, but at 4B shuffled and random-vocabulary targets match or exceed correct targets in matched comparisons. It therefore avoids the identified reward-path failure without attributing its gain to correct semantics; full controls and unstable 8B results are in Appendix L.

Table 4: Scope tests under unified final-checkpoint evaluation. **Left:** ALFWorld estimator comparison. **Right:** WebShop transfer. Estimator families are not factor-matched.

ALFWorld estimator	no reward	+ prediction
GRPO, step-pooled std	42.9	0.7
GiGPO, within-group std	—	0.0
PPO/GAE, trajectory advantage	—	36.4
RLOO, leave-one-out mean	45.0	42.1
REINFORCE++, batch normalization	55.0	46.4

WebShop	base	standard	mean-only
small	60.0	0.5	57.5
full	28.5	0.5	—

Scope. The mechanism is localized to one agent implementation and model family, and the trajectory-first test uses sampled validation. We therefore do not claim a universal pathology or normalizer; further scope tests appear in Appendices A and L.

Practical implications. Auxiliary-reward safety should be audited after estimator transformation. Such audits should compare raw and normalized component variances and covariances under task ties and intervene on the denominator and temporal order. Mean-only normalization and trajectory-first aggregation serve those diagnostic roles here, not as universal replacements. Delivery and content require separate tests: auxiliary losses bypass reward-path amplification, whereas placebo targets test whether gains use the intended predictive information.

4 Related Work

Group-relative normalization. GRPO standardizes returns within sampled groups (Shao et al., 2024; DeepSeek-AI, 2025). The Dr.GRPO correction identifies standard-deviation and length biases (Liu et al., 2025). MDP-GRPO reports low-dispersion amplification (Salmani-Zarchi et al., 2026); Noise-corrected GRPO studies corrupted rewards (El Mansouri et al., 2025), and RAGEN-2 links low reward variance to collapse (Wang et al., 2026b). GDPO decouples reward channels (Liu et al., 2026), while Constrained GRPO identifies shared-denominator coupling from combining rewards before normalization (Girgis et al., 2026). Our bound instead appears only after trajectory aggregation.

Reward shaping and advantage safety. Classical policy-invariance constrains shaped returns (Ng et al., 1999); potential-based intrinsic motivation (PBIM) covers complex intrinsic rewards (Forbes et al., 2024), while Sign-Safe Advantage Composition constrains auxiliary advantages (Seth, 2026). PGPO derives adjacent-state potential differences from anchor-state groups (Zheng et al., 2026). We instead ask whether telescoping control survives grouping, aggregation, and rescaling.

Prediction supervision for agents. Predictive objectives support representations, exploration, and agent training (Jaderberg et al., 2017; Voelcker et al., 2024; Wang et al., 2025b; Yu et al., 2026; Lu et al., 2026). We study their optimization pathway, not general value: the tested loss avoids the reward-path failure, but placebo targets preclude a semantic interpretation. Appendix I gives broader comparisons.

5 Conclusion

A return-level guarantee is not automatically preserved by the estimator that converts rewards into policy-update weights. Here, a valid episodic bound coexists with high prediction scores and task collapse across model scales and on WebShop. With $\Phi_0 = 0$, saturated prediction leaves a sparse first-step boundary. In task-tied pools, standard-deviation division approximately cancels its coefficient at the measured scale, before the increments telescope. The boundary calculation predicts the late raw group standard deviation (0.014035 versus 0.01404 observed), and matched normalizer and temporal-order controls remove prediction-mode damage within their evaluated windows.

This conditional mechanism is not an argument against prediction rewards or standardization generally. Mean-only normalization and trajectory-first aggregation are diagnostic controls, not universal replacements. Guarantees based on coefficients, bounded returns, or telescoping should therefore be audited after grouping, aggregation, and rescaling: the relevant safety object is the estimator-transformed signal that drives the policy update.

Reproducibility Statement

Appendix C specifies the aligned training configuration, reward implementation, decoding settings, hardware, and evaluation protocol. Training curves use the fixed validation protocol described in §3.1; ALFWorld final-checkpoint battery comparisons use the unified evaluations in Table 14. Appendix H reports every training arm and its evaluation protocol. The implementation builds on the open-source verl-agent stack. Appendix E gives the exact trajectory-first construction and the four-arm window summary. For double-blind review, project-specific repository and registration links are omitted. The appendix provides the reward and loss definitions, environment harness details, training configurations, experiment inventory, and an author-reported prediction ledger for auditing the reported comparisons.

AI Use Statement

Generative AI tools assisted with literature discovery, manuscript organization, formulation and checking of mathematical claims, code review, experiment orchestration, critique of hypotheses and methodology, interpretation, and language editing. The authors specified the experimental questions, reviewed the implementation, configurations, logs, and raw telemetry, verified all reported values and claims, and take responsibility for the complete content of this work.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024.

Manuel Baltieri and Christopher L. Buckley. The dark room problem in predictive processing and active inference, a legacy of cognitivism? In *Artificial Life Conference Proceedings*, 2019.

Yong Yi Bay and Kathleen A. Yearick. GRPO, Dr.GRPO, and DAPO are three operations on one number: The group-standard-deviation identity. *arXiv preprint arXiv:2607.00152*, 2026.

Paniz Behboudian, Yash Satsangi, Matthew E. Taylor, Anna Harutyunyan, and Michael Bowling. Policy invariant explicit shaping: an efficient alternative to reward shaping. *Neural Computing and Applications*, 2021.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: GRPO induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *ICLR*, 2019.

Théophile Champion, Marek Grześ, Lisa Bonheme, and Howard Bowman. Deconstructing deep active inference: a contrarian information gatherer. *Neural Computation*, 2024.

Ganqu Cui et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025a.

Ganqu Cui et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025b.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Sam Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. In *AAMAS*, 2012.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Hao Dou. CIGPO: Contextual information-gain policy optimization for multi-turn evidence-reading LLM agents. *arXiv preprint arXiv:2607.16244*, 2026.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected GRPO: From noisy rewards to unbiased gradients. *arXiv preprint arXiv:2510.18924*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *NeurIPS*, 2025.

Grant C. Forbes, Leonardo Villalobos-Arias, Jianxun Wang, Arnav Jhala, and David L. Roberts. Potential-based intrinsic motivation: Preserving optimality with complex, non-Markovian shaping rewards. *arXiv preprint arXiv:2410.12197*, 2024.

Karl Friston, Christopher Thornton, and Andy Clark. Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3:130, 2012.

Cheng Ge et al. Why GRPO needs normalization: A local-curvature perspective on adaptive gradients. *arXiv preprint arXiv:2601.23135*, 2026.

Roger Girgis, Rodrigue de Schaetzen, Luke Rowe, Azalée Robitaille, Christopher Pal, and Liam Paull. Constrained group relative policy optimization. *arXiv preprint arXiv:2602.05863*, 2026.

Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. Expressing arbitrary reward functions as potential-based advice. In *AAAI*, 2015.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2602.22817.

Xixiang He, Qiyao Sun, Ao Cheng, et al. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *arXiv preprint arXiv:2605.21125*, 2026b.

Zhibo Hou et al. Beyond noisy-TVs: Noise-robust exploration via learning progress monitoring. *arXiv preprint arXiv:2509.25438*, 2025.

Jian Hu et al. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Yuki Ichihara et al. MO-GRPO: Mitigating reward hacking of group relative policy optimization on multi-objective problems. *arXiv preprint arXiv:2509.22047*, 2025.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *ICLR*, 2017.

Junguang Jiang, Baixu Chen, Junwei Pan, Ximei Wang, Liu Dapeng, Jie Jiang, and Mingsheng Long. ForkMerge: Mitigating negative transfer in auxiliary-task learning. In *Advances in Neural Information Processing Systems*, 2023.

Shih-Yang Liu et al. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Ning Lu et al. Policy and world modeling co-training for language agents. *arXiv preprint arXiv:2606.02388*, 2026.

Augustine N. Mavor-Parker, Kimberly A. Young, Caswell Barry, and Lewis D. Griffin. How to stay curious while avoiding noisy TVs using aleatoric uncertainty estimation. In *ICML*, 2022.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021.

Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Gongrui Nan, Siye Chen, Jing Huang, et al. NGRPO: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*, 2025.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, 1999.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *ICLR*, 2022.

Jaesung Park et al. Clip-low increases entropy and clip-high decreases entropy in reinforcement learning of large language models. *arXiv preprint arXiv:2509.26114*, 2025.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

Kai Ruan, Jinghao Lin, Qianshan Wei, Ziqi Zhou, and Zihe Huang. SPO++: Stream-aligned policy optimization for asynchronous agentic RL. *arXiv preprint arXiv:2608.24870*, 2026.

Mohammad Mahdi Salmani-Zarchi, Zahra Rahimi, Heshaam Faili, and Mohammad Javad Dousti. MDP-GRPO: Stabilized group relative policy optimization for multi-constraint instruction following. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42784–42797, 2026. doi: 10.18653/v1/2026.acl-long.1982. URL https://aclanthology.org/2026.acl-long.1982/.

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

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

Aryan Seth. Sign-safe advantage composition for multi-objective RL. Master's thesis, University of California, Los Angeles, 2026. URL https://escholarship.org/uc/item/666912nq.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *ICLR*, 2025.

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Vaishnavi Shrivastava, Piero Kauffmann, Ahmed Awadallah, and Dimitris Papailiopoulos. ECHO: Terminal agents learn world models for free. *arXiv preprint arXiv:2605.24517*, 2026.

Claas Voelcker, Tyler Kastner, Igor Gilitschenski, and Amir-massoud Farahmand. When does self-prediction help? Understanding auxiliary tasks in reinforcement learning. *arXiv preprint arXiv:2406.17718*, 2024.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn LLM agents. *arXiv preprint arXiv:2510.14967*, 2025a.

Kangrui Wang et al. VAGEN: Reinforcing world model reasoning for multi-turn VLM agents. In *Advances in Neural Information Processing Systems*, 2025b. arXiv:2510.16907.

Ruiyi Wang and Prithviraj Ammanabrolu. A practitioner's guide to multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2510.01132*, 2025.

Xiaohua Wang et al. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026a.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-GRPO: Pairwise preference reward-based GRPO for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025c.

Zihan Wang, Chi Gui, Xing Jin, Qineng Wang, Licheng Liu, Kangrui Wang, Shiqi Chen, Linjie Li, Zhengyuan Yang, Pingyue Zhang, Yiping Lu, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN-2: Reasoning collapse in agentic RL. In *ICML*, 2026b. Oral.

Zihan Wang et al. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025d.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alexander Alemi, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

Junkang Wu et al. Quantile advantage estimation: Stabilizing RLVR for LLM reasoning. *arXiv preprint arXiv:2509.22611*, 2025.

Zhiheng Xi et al. AgentPRM: Process reward models for LLM agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference*, 2026.

Zhongwen Xu and Zihan Ding. Single-stream policy optimization. *arXiv preprint arXiv:2509.13232*, 2025.

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

Kaibing Yang, Guangfeng Cai, Shengtian Yang, Shuo He, Yu Li, Mengyi Liu, Pengwei Chen, Jun Xu, and Lei Feng. Progress-conditioned group policy optimization for long-horizon agentic tasks. *arXiv preprint arXiv:2607.22724*, 2026.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xiao Yu et al. Reinforcement world model learning for LLM-based agents. *arXiv preprint arXiv:2602.05842*, 2026.

Xichen Zhang, Sitong Wu, Yinghao Zhu, et al. Scaf-grpo: Scaffolded group relative policy optimization for enhancing llm reasoning. *arXiv preprint arXiv:2510.19807*, 2025.

Zhi Zhang et al. Train less, learn more: Adaptive efficient rollout optimization for group-based reinforcement learning. *arXiv preprint arXiv:2602.14338*, 2026.

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.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. *arXiv preprint arXiv:2402.19446*, 2024.

Mingkang Zhu et al. Stratified GRPO: Handling structural heterogeneity in reinforcement learning of LLM search agents. *arXiv preprint arXiv:2510.06214*, 2025.

A Scope and Limitations

Scope of the mechanism. The matched normalizer results and telemetry localize the observed collapses to within-group standard-deviation division over their evaluated windows. The boundary construction predicts the observed late scale, but we do not have an arm that deletes only the first-step boundary; the match is quantitative mechanistic evidence, not a necessity test of that boundary. Cancellation is also finite-scale: fixed $\epsilon > 0$ restores shrinkage as $\lambda \to 0$, while the measured late standard deviation is about $1.4 \times 10^4 \epsilon$. Mean-only removes the divisor but consequently changes update magnitude, so it is a denominator intervention rather than an update-scale-matched optimizer. The four-arm control tests estimator order, but prediction/no-signal arms also differ in their prediction-generating interface. Its 60-update interaction therefore identifies estimator dependence of prediction-mode damage, not a pure reward-only effect or universal superiority of trajectory-first normalization. All-failure filtering additionally changes pool composition and update count.

Practical interpretation. The result is not an argument against prediction rewards or standardization in general. Standardization is useful when within-group spread reflects task quality. The risk arises when task outcomes offer little discrimination while a sparse boundary or other auxiliary component supplies the remaining variation. In that regime, the effective scale is set after grouping and normalization, not by the reward coefficient alone. A useful audit should therefore report centered component variances and covariances, compare raw and normalized scales, and include a matched intervention on the normalizer. Mean-only normalization and trajectory-first aggregation serve complementary diagnostic roles here; we do not claim either is a universal replacement for variance normalization.

Behavior and empirical scope. Telemetry establishes late prediction-dominated normalization, while replay shows that the collapsed policy is repetitive and visits fewer states. These measurements do not determine whether reduced exploration precedes or follows task collapse. The trajectory-first control lacks the 140-game final-checkpoint battery used for the main 4B comparisons. Our empirical scope is one model family and one agent framework, with the failure reproduced in ALFWorld and WebShop. The WebShop result uses continuous returns, showing that the empirical failure is not restricted to binary all-failure groups. We anchor the main claim on the large matched gaps and factor-matched normalizer interventions, rather than on small differences among endpoints.

B Experiment Prediction Ledger (Author-Reported)

The predictions below are author-reported as having been recorded before the corresponding experiments completed and are reproduced with their outcomes. The authors retained full 64-character SHA256 digests of the original plaintext records; the anonymous source package does not include the complete registry, so this ledger should not be read as independently verifiable preregistration evidence.

Table 5: Prospective predictions and outcomes (excerpt of the most informative entries; the retained 25-entry registry also covers the estimator, WebShop, ScienceWorld, and 8B-family arms). Unless explicitly labeled as final-checkpoint, scalar outcomes are training-time last-6 values.

prediction	outcome	note
Dose–response: weaker collapses earlier, 8B latest	falsified	observed val-zero ordering is 1.7B@30 $<$ 8B@65 $<$ 4B@85; three points do not define a law
1.7B collapse at $\sim$step 30	hit	
Anneal arm: no recovery after $\lambda \to 0$	hit	irreversibility evidence #2; the turn precedes the anneal's tail ($\lambda \approx 0.068$ at the turn)
Rescue arm: no collapse without std	hit	shows that the paired collapse depends on std division
Arm D: upper-bound Φ hard to learn	hit	base-rate plateau 0.90–0.93

prediction	outcome	note
Arm D: prediction converts to success	not testable from training-time validation; revised to partial hit by final-checkpoint evaluation	prediction did not exceed base rate in training; the 140-game endpoint is 24.3% vs. 13.6% for the matched mean-only reference, with overlapping intervals; the probe shows a $+0.16$ descriptive difference (App. D)
Self-report: confidence saturates day one	hit	retrospective variance check
Self-report: harmless under std (variance $\rightarrow$0)	hit in training-time validation; revised by final-checkpoint evaluation	training-time last-6 49.5% $=$ baseline; the 140-game endpoint revises the arm to 38.6%, below baseline, the table's largest downward revision (Table 14)
Anchor-QA: harmless, no gain (revised pre-launch)	premise falsified; final-checkpoint endpoint level with baseline	recall did not saturate (0.65$\rightarrow$0.95); the final-checkpoint endpoint (44.3%) lands level with the 44.1% baseline mean, and drag appears only in training-time validation
Δacc progress reward: safe under std	premise falsified; outcome consistent with trajectory form	accuracy plateaued at 0.92, variance persisted, chronic drag (28.1%); endpoint form falsified prospectively a second time
Group-size ablation (n=8, constant budget): collapse delayed but not prevented	hit	literal zero at step 135 vs. 85; all-fail fraction climbs from an early $\sim$0.5–0.6 (rolling) to a terminal lock at 1.00, measuring the co-occurrence expected when task ties expose auxiliary variation
Group-size ablation (n=16): collapse further delayed	half-hit	collapsed (✓) at step 80, *earlier* than n=4; the two-point delay trend was falsified by the third point
λ-sweep ($\lambda \in \{0.01, 0.03, 0.1\}$): std conditions behave similarly	consistent with cancellation	advantage ranges $\pm 7.3/\pm 9.1/\pm 7.2$, no trend over tenfold λ; component dominance was not logged in the sweep arms
All-fail-group filter: rescues without touching std	hit	final-checkpoint 0.7% $\rightarrow$ 40.7%; filtering also discards 23–77% of samples
Aux-loss gain under matched comparisons	mixed	matched differences $+19.8/+24.4/-1.6$ (vs. baseline differences $+22.6/+16.8/+3.3$); the magnitude varies and one comparison lands at its matched baseline
Random-vocabulary strong placebo: gains like gold/shuffled $\Rightarrow$ pure update/format effect; baseline band $\Rightarrow$ environment statistics are the active ingredient	partial	content-free targets match or exceed gold; correct semantic pairing is unnecessary, but the gain mechanism is unresolved
R46a (compute-matched replay, no aux target): no gain	hit	48.9% $\approx$ baseline; no detectable improvement under this condition
R46b (random-token placebo, format-stripped): between baseline and random-vocabulary placebo	hit	61.5%, inside the predicted band; supports the observed ladder without completing a mechanism
Rescue from inside the absorbing state (mean-only switch): no recovery	hit	11 checkpoints at literal zero over a 50-step window; measured advantage range falls; no recovery within the tested window
Fast anneal completed inside the early high-success window: avoids collapse	hit	40.1%; timed withdrawal is the schedule-side preventive route
Matched-variance noise reward: collapses, early high-success window absent	half-hit	no early peak (✓), severe drag -26pt (✓), but no zero-success lock; suggests policy controllability as a candidate factor

train batch × group size	8×4 (= 32 trajectories/step)
validation	32 episodes (ALFWorld/WebShop), 16 (HiddenRule-Gym at 1.7B; the 4B HRG capacity arm used 32); sampled decoding, temperature 0.4, every 5 steps
horizon / max steps	50 (ALFWorld), 30 (HiddenRule-Gym), 15 (WebShop)
optimizer / lr	AdamW, 10^{-6}, KL loss coef 0.01 (low-var estimator)
lengths	prompt 1536 / response 512 tokens
task reward	terminal; ALFWorld/HiddenRule-Gym binary success, WebShop continuous matching score in $[0, 1]$
shaping	$\lambda = 0.1$, difference-form (superficially PBRS), injected per step-sample
advantage normalization unit	GRPO mean/std pooled over all step-samples of a group (cross-step pooling, the framework default; pool = 4 trajectories × ≤50 steps), z-score broadcast to each sample's response tokens (§2.3); the trajectory-first control instead forms one score per trajectory using Eq. 8, standardizes within the same task group, and broadcasts that weight to the unchanged response rows
normalization floor	$\epsilon = 10^{-6}$ in both step-pooled and trajectory-first standardization
invalid-action penalty	0.1
auxiliary CE weight (loss-channel arms)	$\beta = 0.1$ (full weight; the 8B half-weight arm uses $\beta = 0.05$)
training length	150 steps (≈ 18–24 h per condition; the post-hoc rescue extends from the step-125 checkpoint to step 175; the telemetry pair and four temporal-order arms use 60 steps)
precision	FP32 master weights; BF16 compute (FSDP mixed precision, FP32 gradient reduce); BF16 rollout (vLLM). Critic (PPO/GAE arm) follows the actor scheme
hardware	8×96 GB GPUs per arm (4B/8B); 2×32 GB (original 1.7B, HRG); 2×NVIDIA H200 per telemetry or temporal-order arm

Table 6: Default configuration for matched comparison arms. Named ablations vary only the stated factor unless otherwise noted.

C Reproducibility Details

D The Predict Block: Template and Worked Example

Every step's prompt inserts the following instruction between the `<think>` reasoning and the `<action>` choice (ALFWorld schema protocol, quoted verbatim from the released code):

```
After your reasoning, predict the outcome of your action, enclosed
within <predict> </predict> tags in exactly this format:
<predict>next_location: [the location you will be at after this
action, e.g. 'cabinet 2', or 'none' if unchanged]; objects_visible:
[yes/no - will the next observation list at least one object
("you see ...")?]; visible_objects: [comma-separated objects you
expect to see next, or 'none']; task_done: [yes/no - will the task
be completed after this action?]</predict>
```

Worked scoring example. Suppose the agent, holding no task-relevant object, chooses `<action>go to fridge 1</action>` and emits `<predict>next_location: fridge 1; objects_visible: no; visible_objects: none; task_done: no</predict>`, and the environment returns "You arrive at fridge 1. The fridge 1 is closed." The rule parser extracts the four fields and compares each against this realized observation: `next_location` matches (score 1, weight 0.5); `objects_visible` matches, since no "you see" listing appears (score 1, weight 0.5); so $\Phi_t = 1.0$. The `visible_objects` F1 (here 1.0) and the `task_done` bit are logged but carry zero weight. The per-step reward is the difference form $r_{\text{pred}}(t) = \lambda(\Phi_t - \Phi_{t-1})$ of §2.1. In the loss-channel arms the same extractors are applied in reverse: the realized next observation is rendered into a gold predict string of the identical format (here exactly the block above), whose round trip through the parser scores 1.0 by construction, and the auxiliary CE teacher-forces that string. HiddenRule-Gym uses the same pattern with a device-state field added; WebShop replaces the fields with its own page-level schema.

E Analysis of Cross-Step Normalization

This section supplies the details behind the coefficient-cancellation argument in Section 2.3. Means and variances use the same convention as the standard deviation in Equation 3.

Exact and approximate scale cancellation. Suppose the task score is tied, $q_i = C_g$, for every sample in a finite pool and $\sigma_{s,g} > 0$. If $p_{i,t}$ is also constant within the pool and $\epsilon = 0$, centering removes both constants. Because $\sigma_{\lambda s,g} = \lambda \sigma_{s,g}$ for $\lambda > 0$, the normalized weight is exactly

$$\widehat{A}_{i,t} = \frac{s_{i,t} - \bar{s}_g}{\sigma_{s,g}}. \tag{9}$$

For a finite task-tied pool of at most N samples, define $\eta_g = (\sigma_{p,g} + \epsilon)/(\lambda \sigma_{s,g})$ and $z_{i,t} = (s_{i,t} - \bar{s}_g)/\sigma_{s,g}$. Whenever $\eta_g < 1$, there is a constant K_N depending only on the pool-size convention such that

$$\left| \widehat{A}_{i,t} - z_{i,t} \right| \leq \frac{2K_N \eta_g}{1 - \eta_g}.$$

Thus the normalized weight approaches the coefficient-free z-score uniformly as penalty variation and the numerical floor become small relative to $\lambda \sigma_{s,g}$. This remains a conditional statement: tied task scores alone do not establish its premise.

For the approximate case in Eq. 7, write $d = \sigma_{\lambda s - p,g} + \epsilon$. The reverse triangle inequality gives $|d - \lambda \sigma_{s,g}| \leq \sigma_{p,g} + \epsilon$. Under either the sample or population convention, $|z_{i,t}| \leq K_N$ and $|p_{i,t} - \bar{p}_g| \leq K_N \sigma_{p,g}$. Applying these two bounds to the centered numerator and using $\eta_g < 1$ yields Eq. 7. The bound is conditional: tied task scores alone do not establish that prediction variation dominates penalty variation or the numerical floor.

Telemetry-instrumented matched pair. We reran the Qwen3-1.7B prediction-reward condition for 60 updates in two arms that differed only in within-group standard-deviation division. At update 1, all 32 complete trajectories and all 1,574 active step samples match across arms; rollout content, raw rewards, and group statistics are identical, and only the normalized advantages differ. Telemetry reconstruction checks pass for all 120 arm-updates, and all 960 group-updates contain four distinct exact-response/telemetry trajectory signatures. At update 60, standardized versus mean-only validation success is 0.0% versus 18.8%, and training-rollout prediction accuracy is 99.8% versus 41.7%. In updates 51–60 of the standardized arm, 79 of 80 pools are task-tied and all-failure. Summing the task, prediction, penalty, and interaction variance terms over those 79 pools and dividing each sum by the summed total pre-normalization variance gives shares of 0.00%, 98.74%, 1.21%, and 0.05%. Median raw group standard deviation declines from 0.04634 in updates 1–10 to 0.01404 in updates 51–60, while advantage RMS stays at 0.99748 and 0.99743, respectively, as expected under standardization. In the mean-only arm's final ten updates, raw group standard deviation is 0.04629 and advantage RMS is 0.04618, with no comparable rescaling. For $G = 4$, $T = 50$, and $\lambda = 0.1$, the implemented $\Phi_0 = 0$ boundary predicts a raw standard deviation of 0.014035 (Eq. 6), matching the observed late median. These measurements verify late prediction dominance, identify its limiting temporal scale, and document the paired outcome over the evaluated window. They do not establish temporal ordering or decompose the final gradient after clipping, KL regularization, and token weighting.

Matched temporal-order control. The aggregate-before-normalize implementation uses the same per-step reward rows as the standard arm. For trajectory i, it computes $X_i = q_i + \sum_t (x_{i,t} - q_i)$, counting the broadcast terminal task score once, standardizes the four trajectory scores from the same task, and broadcasts the resulting trajectory z-score to every original response row. The loss mask, number of rows, optimizer, rollout generation, and remaining hyperparameters are unchanged. We cross this switch with prediction on/off, yielding four 60-update Qwen3-1.7B arms. Mean validation success over checkpoints 30–60 is 0.89% (step/prediction), 8.46% (step/no signal), 7.13% (trajectory/prediction), and 6.24% (trajectory/no signal). The corresponding prediction-mode contrasts are -7.57 and $+0.89$ percentage points, for an 8.46-point interaction. These 32-episode checkpoint means are a matched checkpoint-level diagnostic, not final-evaluation uncertainty or a 140-game final evaluation.

claim	targeted ablation	verdict	where
the observed lock appears in prediction-reward arms and not in no-prediction baselines	baseline and reward arms	large contrast establishes the phenotype; the prediction prompt and sampled output also differ	§3.2
the matched 4B collapse depends on std division	single-flag mean-only intervention	final-checkpoint 0.7% → 52.9%	§3.3
prediction dominates late task-tied pre-normalization variance at the predicted boundary scale	instrumented 1.7B standard/mean-only pair	98.74% prediction share; predicted 0.014035 vs. observed 0.01404 std	§3.3
prediction-mode damage differs between step-pooled and trajectory-first estimators	2×2 estimator × prediction/no-signal control	contrast -7.6 vs. $+0.9$ points; 8.5-point interaction	§3.3, Fig. 3
dropped all-failure groups avoid collapse	all-failure filter, std kept	final-checkpoint endpoint 40.7%; sample dropping is confounded	§3.3
nominal coefficient does not control observed advantage range	λ sweep $\{0.01, 0.03, 0.1\}$	no range trend; all collapse	§3.3, Fig. 5
not specific to group size 4	constant-budget group sizes $\{4, 8, 16\}$	collapse at every size, non-monotone	§3.5, Fig. 4
policy controllability may matter	matched-marginal-variance noise	drags without locking; temporal structure is unmatched	Appendix K
endpoint accuracy is insufficient	alternate prediction-reward forms	different drag and saturation patterns	Appendix P
correct semantics are unnecessary for the 4B gain	paired shuffled and random-vocabulary targets	both placebo classes exceed gold across the matched comparisons	§L
auxiliary loss is not uniformly safe	8B gold-target conditions	endpoints span 0.0–85.0%	Appendix L
estimator pattern is consistent with a within-group-σ boundary	GiGPO / PPO / RLOO / REINFORCE++	among alternative estimators, only GiGPO collapses; algorithms are not factor-matched	Appendix M
semantic usefulness has further boundaries	exploratory HRG coverage sweep	one intermediate rung improves; curve is non-monotonic	Appendix M, Fig. 9

Table 7: Ablation index. The verdict column states the evidence at the strength supported by the corresponding control; referenced sections give protocols, evaluation protocols and remaining confounds.

E.1 Vanishing increment variance

The main text does not infer vanishing increment variance from a high mean prediction score. The following elementary bound records the relevant conditional statement. Suppose $n_g \le N$, $0 < \lambda \le L$, and $\epsilon > 0$ is fixed. For the prediction component

$$a_{i,t}^{\text{pred}} = \frac{\lambda(s_{i,t} - \bar{s}_g)}{\sigma_{x,g} + \epsilon},$$

$$\max_{(i,t)\in\mathcal{P}_g} |a_{i,t}^{\text{pred}}| \le \frac{L\sqrt{N}}{\epsilon}\,\sigma_{s,g} \longrightarrow 0 \quad \text{if } \sigma_{s,g} \longrightarrow 0. \tag{10}$$

The inequality follows because, under either sample or population standard deviation, $|s_{i,t} - \bar{s}_g| \le \sqrt{n_g}\,\sigma_{s,g}$, while $\sigma_{x,g} + \epsilon \ge \epsilon$. The premise concerns increment variance, not mean accuracy. Our implementation sets $\Phi_{i,0} = 0$, so a first-step boundary increment remains even when later prediction scores are constant. Equation 6 gives its exact scale under saturated, equal-length trajectories.

F Claim-by-Claim Ablation Index

G Additional Experimental Detail

Boundary conditions of the rescue. Std normalization is genuinely useful where it was designed: an adaptive gradient with provable convergence benefits when within-group reward variance is task-informative (Ge et al., 2026). Our analysis defines a candidate failure regime—a dense policy-dependent shaping term whose spread dominates within-group variance, with sparse terminal success and all-fail groups as the extreme case—and our original battery tests observable consequences. The telemetry-instrumented 1.7B pair directly measures late component dominance in the measured setting; it does not imply the same composition generally. evidence that std removal repairs GRPO in stochastic-outcome domains (Bereket & Leskovec, 2025) suggests the repair generalizes along the axis of "variance not aligned with task quality." A learned critic derives its scale from bootstrapped value estimates rather than the spread of a small same-prompt group, so the tied-pool cancellation calculation has no direct analogue there. At the unified final-checkpoint evaluation, PPO/GAE

reaches 36.4% without a matched PPO baseline, the RLOO reward arm reaches 42.1% against its 45.0% baseline, and the REINFORCE++ reward arm reaches 46.4% against its 55.0% baseline. None shows a sustained collapse window; the PPO/GAE prediction score peaks at 0.60 during training. These estimator arms are consistent with a within-group σ boundary, but the algorithms are not factor-matched (Appendix M).

Post-hoc belief probes (HRG). Privileged-latent probes on final checkpoints (leakage-audited questions; yes/no cells reported against majority-class baselines) give the belief-side view. State-memory probe accuracy orders the arms C=0.48 (0.61) $>$ task-weighted (0.46) $>$ pure GRPO (0.40) $>$ C=0.90 $\approx$ Arm D (0.35) $>$ generic-PS (0.31) against a 0.14 random floor. This ordering is consistent with the training-time F1 pattern, but these post-hoc probes do not establish representation loss or strengthening. Every arm except the mid-coverage and task-weighted ones scores at or below the no-signal 0.40, including the privileged-feature Arm D (0.35). Arm D's lower general-memory score coexists with a $+0.16$ descriptive difference on the specific latent it was trained to predict: it probes at 0.99 against a 0.83 majority-class baseline. This is a descriptive dissociation, not evidence of a breadth-specific cost or of internalized knowledge. The post-hoc yes/no probe is a different instrument from the training-time generative prediction accuracy, whose 0.90–0.93 plateau in Appendix M is scored against the environment, hence the different baselines. Task success converts only partially (24.3% vs. 13.6% for the matched mean-only reference at the 140-game endpoint, Table 12; the 16-game window had read it as floor). Probe scores and behavioral endpoints can therefore differ in these measurements, but causal interpretation requires a controlled probe study. No arm exceeds the majority baseline on rule-relevance probes, consistent with none cracking the rules.

H Complete experiment inventory

The current inventory contains 81 arms: 74 in the original freeze, one post-freeze HRG mean-only reference, two post-freeze telemetry diagnostics, and four post-freeze temporal-order controls. Design-factor columns are σ for within-group standard-deviation division, R for next-observation supervision in the reward, L for related but non-equivalent supervision through auxiliary CE, and $\varnothing$ for a content-free target or control. "last-6" is the training-time endpoint; "seen" and "unseen" are final-checkpoint evaluations where available (ALFWorld: 140 seen and 134 unique unseen games, with six duplicate loader outputs reported separately; WebShop: 200 goals; HiddenRule-Gym: 140 probe-family games). Arms use the 150-update budget unless the configuration states otherwise; the telemetry and temporal-order diagnostics use 60 updates, and the post-hoc rescue instead extends its parent from step 125 to 175. Timestamped outcome predictions were recorded before the corresponding experiments completed. Appendix B reproduces the most informative subset. Within σ✓R✓ rows, outcome differences trace to the modifier named in the Configuration column (group filtering, anneal schedule, decoupled or mean-only normalization): the σ column marks presence of the within-group divisor, not absence of every mitigation.

Table 8: ALFWorld, Qwen3-4B (main battery)

		signal			Metrics (%)			
Configuration	σ	R	L	$\varnothing$	last-6$^{(\uparrow)}$	seen$^{(\uparrow)}$	unseen$^{(\uparrow)}$	outcome
Baselines and controls								
GRPO baseline	✓				49.5	42.9	42.1	baseline
GRPO baseline	✓				39.1	40.7	37.1	baseline
GRPO baseline	✓				51.6	48.6	46.4	baseline
mean-only control					52.6	57.9	44.3	control
2-epoch replay (compute)	✓				48.9	48.6	39.3	parity
Reward channel, std normalization on								
PS prediction reward	✓	✓			1.0	0.7	0.0	**collapse**
PS prediction reward	✓	✓			0.0	0.0	0.0	**collapse**
PS, λ=0.01	✓	✓			1.0	—	—	**collapse**
PS, λ=0.03	✓	✓			0.0	—	—	**collapse**
PS, group n=8	✓	✓			2.6	—	—	**collapse**
PS, group n=16	✓	✓			0.0	—	—	**collapse**
PS, slow anneal $\lambda\to0$	✓	✓			0.0	—	—	**collapse**
PS + fast anneal	✓	✓			40.1	—	—	**rescue**
PS, GiGPO (step groups)	✓	✓			0.0	0.0	0.0	**collapse**
matched-variance noise	✓	✓		✓	23.4	—	—	severe drag

Configuration	σ	signal R	signal L	signal ∅	Metrics (%) last-6$^{(\uparrow)}$	Metrics (%) seen$^{(\uparrow)}$	Metrics (%) unseen$^{(\uparrow)}$	outcome
Δacc progress reward	✓	✓			28.1	32.1	29.3	chronic drag
Δacc progress reward	✓	✓			34.4	27.1	25.7	chronic drag
anchor-QA reward	✓	✓			44.3	44.3	33.6	training-only drag
self-report reward	✓	✓			49.5	38.6	39.3	parity/drag
Reward channel, normalization modified								
PS + mean-only		✓			51.6	52.9	40.7	**rescue**
PS + all-fail filter	✓	✓			39.6	40.7	35.7	partial rescue
PS + decoupled norm	✓	✓			31.2	27.9	30.0	drag
PS + decoupled norm	✓	✓			42.2	32.1	26.4	drag
mean-only@step125 (post hoc)		✓			0.0	—	—	no recovery
Loss channel (auxiliary CE)								
aux gold	✓		✓		69.3	68.6	65.0	gain
aux gold	✓		✓		63.5	57.9	45.0	gain
aux gold	✓		✓		50.0	52.1	39.3	parity
aux shuffled	✓		✓	✓	76.0	81.4	73.6	gain
aux shuffled	✓		✓	✓	60.4	67.9	66.4	gain
aux shuffled	✓		✓	✓	48.9	60.7	43.6	final-checkpoint gain
aux random-vocab	✓		✓	✓	73.4	**87.1**	71.4	gain
aux random-vocab	✓		✓	✓	77.6	75.0	51.4	gain
aux random-vocab	✓		✓	✓	60.9	67.1	48.6	gain
aux random-tokens	✓		✓	✓	61.5	67.9	57.9	gain
aux noise (R60a)	✓		✓	✓	43.2	47.9	38.6	no gain
aux half-gold (R60b)	✓		✓	✓	10.9	10.0	7.9	**catastr. drag**
Estimator family (same PS recipe)								
PPO/GAE + PS		✓			37.5	36.4	40.0	survives
RLOO baseline					44.8	45.0	40.7	baseline
RLOO + PS		✓			40.6	42.1	35.7	survives
R++ baseline (E11a)					55.2	55.0	46.4	baseline
R++ + PS (E11b)		✓			48.4	46.4	46.4	no collapse

Table 9: ALFWorld, Qwen3-8B (bistability family)

Configuration	σ	signal R	signal L	signal ∅	Metrics (%) last-6$^{(\uparrow)}$	Metrics (%) seen$^{(\uparrow)}$	Metrics (%) unseen$^{(\uparrow)}$	outcome
baseline	✓				32.8	45.0	39.3	baseline
PS prediction reward	✓	✓			0.0	0.0	0.0	**collapse**
aux gold	✓		✓		0.0	0.0	0.0	**absorbing**
aux gold	✓		✓		72.4	**85.0**	72.1	**best gain**
aux gold	✓		✓		0.0	0.0	0.0	**absorbing**
aux gold, β=0.05	✓		✓		65.1	77.1	75.0	escape+gain
aux shuffled	✓		✓	✓	78.6	79.3	66.4	gain
prompt-only	✓				47.4	52.1	44.3	level

Table 10: ALFWorld, Qwen3-1.7B / Qwen2.5-1.5B

Configuration	σ	signal R	signal L	signal ∅	Metrics (%) last-6$^{(\uparrow)}$	Metrics (%) seen$^{(\uparrow)}$	Metrics (%) unseen$^{(\uparrow)}$	outcome
1.7B baseline	✓				20.9	17.9	18.6	baseline
1.7B PS reward	✓	✓			0.0	0.0	0.0	**collapse**
1.7B PS, telemetry (60 updates)	✓	✓			1.0	—	—	diagnostic collapse
1.7B PS, mean-only telemetry (60 updates)		✓			15.6	—	—	diagnostic rescue
1.7B step-pooled, no signal (60 updates)	✓				9.4	—	—	reference
1.7B step-pooled PS (60 updates)	✓	✓			1.0	—	—	damage
1.7B trajectory-first, no signal (60 updates)	✓				7.3	—	—	reference
1.7B trajectory-first PS (60 updates)	✓	✓			7.8	—	—	no damage
1.5B full budget (anchor)	✓				67.2	—	—	anchor

For the 60-update telemetry and temporal-order rows, "last-6" averages the six validation checkpoints at updates 35–60. The main text reports the temporal control over the predeclared late window of updates 30–60; both summaries use 32 sampled episodes per checkpoint.

Table 11: WebShop, Qwen3-4B

Configuration	σ	signal R	signal L	signal ∅	Metrics (%) last-6$^{(\uparrow)}$	Metrics (%) seen$^{(\uparrow)}$	Metrics (%) unseen$^{(\uparrow)}$	outcome
base, small	✓				60.4	60.0	—	baseline
PS reward, small	✓	✓			1.0	0.5	—	**instant collapse**
base, full	✓				29.2	28.5	—	baseline
PS reward, full	✓	✓			0.5	0.5	—	**collapse**
aux gold, small	✓		✓		71.4	63.0	—	gain→tie
aux gold, small	✓		✓		50.5	46.0	—	below base
aux shuffled, small	✓		✓	✓	70.3	**67.0**	—	gain
PS + mean-only, small		✓			59.4	57.5	—	**rescue**

Table 12: HiddenRule-Gym, Qwen3-1.7B/4B

Configuration	σ	signal R	signal L	signal ∅	Metrics (%) last-6$^{(\uparrow)}$	Metrics (%) seen$^{(\uparrow)}$	Metrics (%) unseen$^{(\uparrow)}$	outcome
pure GRPO, 1.7B	✓				8.3	10.7	—	floor
mean-only no-signal (post-freeze)					10.4	13.6	—	matched ref.
PS generic Φ	✓	✓			10.4	14.3	—	floor; F1 collapse
PS task-weighted Φ	✓	✓			5.2	8.6	—	floor; F1 held
arm D, privileged Φ		✓			10.4	24.3	—	partial conversion
coverage C=0.23		✓			12.5	26.4	—	gain
coverage C=0.48		✓			24.0	**34.3**	—	best HRG
coverage C=0.66		✓			14.6	13.6	—	mid dip
coverage C=0.86		✓			22.9	22.1	—	gain
coverage C=0.90		✓			3.1	10.7	—	benefit consumed
aux gold @C=0.48	✓		✓		14.6	17.1	—	marginal
aux shuffled @C=0.48	✓		✓	✓	14.6	15.0	—	tied w/ gold
pure GRPO, 4B	✓				26.6	—	—	off floor

Table 13: ScienceWorld, Qwen3-4B

Configuration	σ	signal R	signal L	signal ∅	Metrics (%) last-6$^{(\uparrow)}$	Metrics (%) seen$^{(\uparrow)}$	Metrics (%) unseen$^{(\uparrow)}$	outcome
base (boil)	✓				0.0	—	—	env. floor
PS reward (boil)	✓	✓			0.0	—	—	floor; n.d.

I Extended related work

Group-normalized RL and its advantage pathologies. Group-relative policy optimization (Shao et al., 2024; DeepSeek-AI, 2025) normalizes each group's returns by their mean and standard deviation. Dr.GRPO (Liu et al., 2025) identified this estimator's std and length biases. We use its mean-only variant as a controlled intervention: removing within-group std division prevents collapse in our tested setup. Bereket & Leskovec (2025) independently show that removing group-std normalization fixes GRPO-induced overconfidence for stochastic outcomes, while Ge et al. (2026) defend std normalization as an adaptive gradient on reasoning benchmarks. Our results suggest a complementary regime in which a sparse temporal boundary or other auxiliary differences can be amplified when task scores have little spread. In the telemetry-instrumented 1.7B pair, late task-tied pools are prediction-dominated; other arms lack this component-level logging. Related diagnoses include heterogeneous-group bias (Zhu et al., 2025), local-normalization bias (Hu et al., 2025), baseline-induced entropy hazards (Wu et al., 2025), mean-only versus mean-plus-variance calibration theory (Mroueh, 2025), and GDPO's decoupled multi-reward normalization (Liu et al., 2026), which our decoupled arm tests and finds insufficient under sustained dense pressure. Constrained GRPO identifies shared-denominator coupling when rewards are scalarized before normalization (Girgis et al., 2026); SPO++ separately shows that trajectory centering need not center the token-weighted actor quantity (Ruan et al., 2026). A distinct line directly recovers signal from homogeneous groups through calibrated or virtual samples (Nan et al., 2025; He et al., 2026b), while other methods prune, replace, or reweight difficult groups (Xu & Ding, 2025; Zhang et al., 2026; 2025).

With an auxiliary reward, such groups can instead retain a temporal boundary and other prediction or penalty variation, which standardization may amplify (Eq. 4). Filtering all-failure groups raises the final-checkpoint endpoint to within 2.2 points of baseline, but also removes samples and changes the data distribution. On agent tasks, ProGPO supplies progress credit only in all-failed groups (Yang et al., 2026), whereas CIGPO injects separately normalized per-turn information gain to avoid zero-advantage lock-in (Dou, 2026). Recent work formalizes the group-std term as the single quantity on which GRPO, Dr.GRPO, and DAPO differ (Bay & Yearick, 2026), consistent with our intervention. Qualitative precedents of the amplification exist ("illusory advantages" in text-to-image GRPO, Wang et al., 2025c; scale-equalizing per-component normalization, Ichihara et al., 2025); we apply the scale-cancellation observation to step-pooled auxiliary rewards in sparse-success multi-turn training. Our infrastructure follows the step-independent multi-turn line (Feng et al., 2025; Zhou et al., 2024); successors refine step-level advantage estimation (He et al., 2026a) while retaining the group-std normalizer, and our results concern that shared normalization machinery, not any particular grouping scheme.

Dense signals, shaping, and reward hacking in agent RL. Prediction-*error* seeking (Pathak et al., 2017) is famously vulnerable to the noisy-TV trap (Burda et al., 2019; Mavor-Parker et al., 2022); our endpoint behavior is consistent with the mirror image: concentration on predictable interactions (Friston et al., 2012; Baltieri & Buckley, 2019), a pathology one deep active-inference agent already exhibited by mastering a single repeated action (Champion et al., 2024). On the shaping side, Ng-style invariance (Ng et al., 1999) assumes a policy-independent potential; dynamic and learned potentials break it in known ways (Devlin & Kudenko, 2012; Harutyunyan et al., 2015; Behboudian et al., 2021; Forbes et al., 2024), but that line does not consider the interaction with group normalization. Dense process rewards are known to be hacking-prone (Cui et al., 2025b; Pan et al., 2022; Wang et al., 2026a); the standard mitigation is the progress principle, rewarding change rather than level (Setlur et al., 2025; Hou et al., 2025; Wang et al., 2025a; Xi et al., 2026). Our original difference-form reward isolates a separate point: return-level boundedness does not control step-level normalized weights. RAGEN's "Echo Trap" (Wang et al., 2025d) and stability reports for dense turn-level rewards (Wang & Ammanabrolu, 2025) are near neighbors observed without a controlled dense-signal culprit.

Auxiliary objectives and world modeling for agents. From UNREAL (Jaderberg et al., 2017) to modern LLM agents, world-model signals delivered through the loss channel have a consistent track record (Shrivastava et al., 2026; Lu et al., 2026; Voelcker et al., 2024). Nor is the reward channel uniformly hostile: RWML (Yu et al., 2026) succeeds with an embedding-space gap reward on ALFWorld while warning that token-level next-state prediction "can lead to model collapse," a one-line anticipation of the failure we characterize, and VAGEN (Wang et al., 2025b) rewards world-model reasoning under a redesigned estimator. These successes do not test the same reward–normalization pathway studied here. Our reward and loss routes concern next-observation prediction but are not signal-matched: they differ in supervised fields, granularity, and credit assignment. We use their contrast as a boundary rather than a channel-only intervention. Relative to ECHO (Shrivastava et al., 2026), our placebo ladder shows that correct content–context pairing is unnecessary for the observed 4B gain; the mechanism of that gain remains unresolved.

Training instability across scales. Instability in language-model fine-tuning is well documented, including divergence partway through training (Dodge et al., 2020), with vanishing gradients as one driver (Mosbach et al., 2021). Our 8B gold-target endpoints span persistent zero and high-success outcomes. This pattern is consistent with basin-like behavior, but the available evidence cannot distinguish it from unusually large training variability. Scale-dependent instabilities invisible at smaller scales are also documented for pretraining (Wortsman et al., 2023). Finally, the auxiliary-update framing connects to negative transfer in auxiliary-task learning (Jiang et al., 2023): the positive mean effect at 4B does not transfer uniformly to 8B.

J Collapse mechanism: measurement detail

Budget vs. published ALFWorld numbers. Published full-budget GRPO baselines on this stack reach $\sim$73–74% (1.5B) and $\sim$79–83% (7B) on ALFWorld (reported as the GRPO baseline in Feng et al., 2025; He et al., 2026a). Our main battery deliberately uses one quarter of that budget (32

trajectories/update, 150 updates) so that a broad matched battery fits a fixed compute envelope. The 60-update telemetry pair is a separate matched diagnostic spanning the standard arm's healthy, transition, and collapse window. The core collapse and matched-intervention claims use equal-budget arm-to-arm comparisons and do not depend on absolute success rates. A separate full-budget sanity anchor (Qwen2.5-1.5B, 128 trajectories/step) reaches 67.2% in our hands, on the published baseline's convergence trajectory. We do not use cross-paper leaderboard comparisons because training budgets, checkpoints, and evaluation protocols differ substantially.

A replay of a collapsed 1.7B checkpoint against a healthy same-scale policy (12 episodes each, identical harness) documents the endpoint behavior: state-visitation entropy drops tenfold (0.91→0.09; 1.25 locations per episode vs. 4.6), consecutive-action repetition rises 0.21→0.93, episode length pins at the horizon, and the collapsed policy is more compliant on two format and action-validity metrics (invalid-action rate 0.005 vs. 0.173, prediction-parse validity 1.00), even as its behavioral diversity and task success deteriorate. Constant-budget group-size ablations (4×8 and 2×16) test whether the pathology is an artifact of unstable small-group $\hat{\sigma}$ at $n{=}4$: collapse occurs at every group size with *non-monotone* timing (literal zero at steps 85/135/80), the $n{=}8$ delay a two-point comparison would read as a trend being overturned by the third point. These conditions also log the all-fail group fraction: it climbs and locks at 1.00. The task-score component then has zero within-pool variation, while residual variation can still come from prediction and penalties. The lock coincides with the literal-zero window in both measured conditions (Figure 4); a calibrated retrodiction of the original collapse condition shows the same ramp ($\sim0.65 \rightarrow \geq0.9$), while the rescued and baseline arms' fractions instead decay to ~0.25 as success accumulates. Finally, a tenfold coefficient reduction leaves the observed total advantage range nearly unchanged (±7.2 at $\lambda{=}0.1$ vs. ±7.3 at 0.01) and the arm collapses on schedule (last-6 1.0%). This is consistent with Eq. 4, but the logs cannot apportion prediction and penalty contributions.

Exploratory early-warning diagnostic. The joint precursor is entropy monotone decline $\wedge$ prediction saturation $\wedge$ length pinning, leading the turn by 15–30 steps. Entropy alone false-alarms three ways (victory sharpening: the rescued arm dips to 0.095 while winning; early sharpening: the decoupled arm dips to 0.094 with prediction *declining*; pre-breakthrough consolidation: the auxiliary-loss arm reaches 0.04 right before its 28→59% phase transition), and it also declines for reward-independent reasons under clipped policy-gradient training (Park et al., 2025), so the conjunction is required. We froze the rule as entropy-has-declined $\wedge$ prediction ≥0.95 $\wedge$ length $\geq0.95\,H$, evaluated over history since the components arrive sequentially. Its thresholds were calibrated only on the definition-set collapse conditions (the three-scale collapse arms and the slow-anneal arm); the conjunction requirement itself was motivated both by those retrodictions and by the entropy-only dips already observed in healthy conditions. Evaluated once, at a fixed date, on the fourteen conditions then available outside the calibration set,[1] it detects all three evaluated *reward-channel* collapses in that frozen roster, 73/53/13 steps ahead of the collapse point, defined as the first validation opening a sustained dead band ($\leq6.5\%$ with no recovery; the first *exactly*-zero checkpoint can lag it, e.g. step 135 vs. 130 for $n{=}8$, and the 15–30-step precursor lead above is measured to the still-earlier *turn*: three distinct events). Its scope is that failure class: the 8B auxiliary-loss collapses bypass the length-pinning component altogether (Appendix M), structurally outside the rule's coverage rather than a miss. Among the eleven non-collapsing arms it raises one alarm, on the self-report arm, whose final-checkpoint endpoint later revises below baseline (38.6%, Table 14) but never enters the defined collapse regime. Under the declared target, the record is therefore three of three evaluated collapses detected and one of eleven non-collapsing arms alarmed. Because the thresholds were

[1]Three collapses: group-size $n{=}8$ ($+73$), the $\lambda{=}0.01$ arm ($+53$), and group-size $n{=}16$ ($+13$). Eleven training-healthy arms: the two baseline conditions, mean-only PS and its no-signal control, anchor-QA, Δacc, the all-fail filter, the gold, shuffled-gold, and random-vocabulary auxiliary arms, and self-report. "Healthy" means no sustained zero-success window in training-time validation, so drag arms count as healthy. Two of the healthy arms (mean-only PS, aux gold) supplied the pre-freeze entropy-dip observations that motivated the conjunction, so their clearing is confirmatory rather than independent; excluding them leaves nine independent healthy arms with the same single alarm. The decoupled arm, though completed by the evaluation date, is absent from the recorded roster, an omission we flag rather than justify; at its entropy dip its prediction accuracy was declining, below the rule's 0.95 gate, so the conjunction would not have alarmed on it, and its inclusion could only have enlarged the healthy denominator. Arms completing after the evaluation date, whether collapsing ($\lambda{=}0.03$, GiGPO, an additional 4B PS arm, both WebShop arms) or healthy (matched-variance noise, fast anneal), were never scored against the rule; three-of-three is a claim over the evaluated set, not over all later collapses.

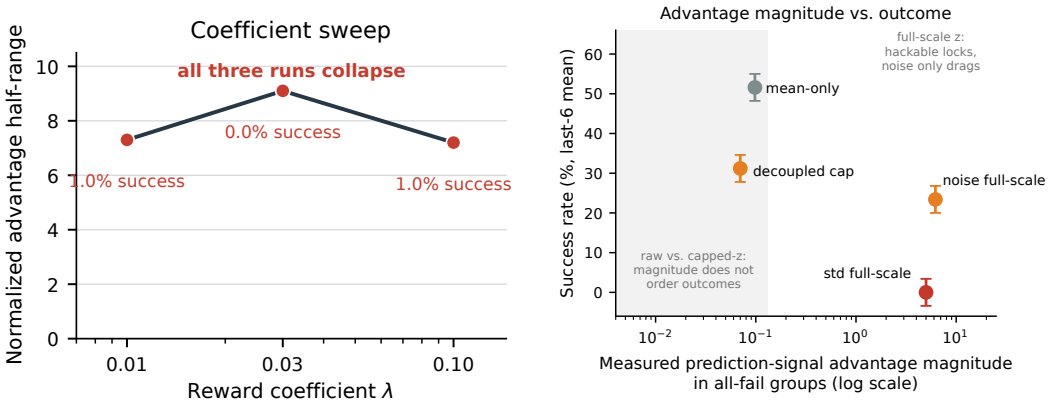


Figure 5: Magnitude controls. **Left:** a tenfold coefficient sweep leaves normalized advantage ranges at comparable scale, and every tested condition collapse. **Right:** measured auxiliary magnitude does not order task outcomes across signal-handling controls.

calibrated on four earlier collapses and two healthy cases informed the conjunction, this remains an exploratory diagnostic rather than a prospective performance estimate.

K Fix hierarchy and signal-form detail

No recovery in the tested post-collapse windows. Constant pressure: 120 post-collapse steps at fixed λ (1.7B), zero recovery. Annealing: cosine $\lambda \to 0$ still collapses, the turn arriving at $\lambda \approx 0.068$ (step $\sim$57 of the schedule; literal zero follows by $\sim$85), and its final $\sim$30 steps at $\lambda < 0.01$ (effectively pure GRPO) show zero recovery; entropy remelts to 0.439, matching the un-annealed arm's 0.44 (Fig. 2b). Post-collapse entropy rebounds at all scales (1.7B 0.39, 4B 0.44, 8B 0.52; ordered by reference-model entropy, consistent with KL-anchored dynamics; Cui et al., 2025a). Entropy alone therefore does not explain the persistent zero-success behavior. The late-rescue arm recorded in Appendix B completes the intervention set: resuming the collapsed arm from its step-125 checkpoint, 70 steps past that arm's first zero-success evaluation, and switching to mean-only normalization (the intervention that prevents collapse from the start) recovers *nothing* over a 50-step window, 11 validation checkpoints at literal 0%. The measured advantage range indicates that the amplifier is gone: magnitudes fall from ± 5–9 to ≈ 0.1, the raw per-step scale that λ and the invalid-action penalty share. The prediction score is saturated at 1.0, but the retained logs do not identify the shaping component's within-group spread. The remaining ≈ 0.1 excursions are on the raw reward scale and are consistent with rare invalid steps, but cannot be decomposed. No recovery occurs within this 50-step window. The tested late scheduling and post-hoc structural interventions also fail; three earlier interventions avoid or weaken collapse: structural (mean-only from the start), sample-level (dropping all-fail groups, partial rescue at 39.6%, §3.3), and *timed withdrawal* (a fast-anneal arm recorded in the same ledger completes the cosine schedule early and avoids collapse, 40.1%, Fig. 2b). These observations establish persistence over the measured windows, not mathematical irreversibility.

Matched-variance noise probes policy controllability. A control replaces the prediction score with random potentials matched in per-step marginal variance. The arm drags severely (23.4%, -26 vs. baseline) but never enters a sustained zero-success window over 150 steps. Variance alone therefore does not reproduce the observed lock in this control, and policy controllability is one candidate explanation. The temporal distribution and correlations of the random signal are not fully matched, so the experiment does not isolate controllability by itself.

Decoupling is not enough. Per-channel advantage decoupling (GDPO-style; prediction channel independently normalized, contribution capped at λ) does not collapse; the cap holds empirically (0.066–0.077). But it ends at **31.2%** last-6, well below the mean-normalized arms; an additional matched comparison lands level with its baseline in training-time validation (42.2 vs. 39.1) but reproduces the deficit at the final-checkpoint evaluation (27.9/32.1 against matched baselines 42.9/40.7;

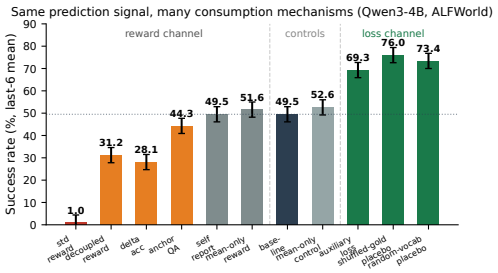


Figure 6: Overview of the tested reward and auxiliary-loss arms. The routes concern related prediction supervision but are not signal-matched, so the comparison is descriptive rather than channel-only.

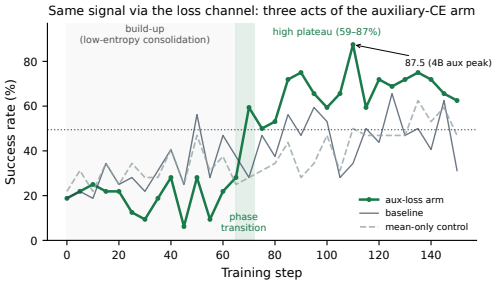


Figure 7: Training dynamics of the 4B gold auxiliary-loss arm. The interference probe detects no task-batch log-probability shift at its reporting resolution; it does not establish gradient orthogonality.

descriptive comparison). The measured scales carry the lesson (Fig. 5): the healthy mean-only arm's raw shaping advantage (± 0.098) is *larger* than the dragging decoupled arm's capped one (0.066–0.077), so outcomes are not monotonic in the reported nominal magnitude. The per-channel z-score re-stretches vanishing raw differences to a fixed λ-scale pressure regardless of the signal's actual spread, a capped copy of the same amplifier, whereas the mean-only contribution is the raw $\lambda(r - \bar{r})$, tied to that spread. Across the tested arms, the observed pattern ranges from an unamplified, healthy arm to capped drag and full-scale collapse; a dedicated factorial comparison would be needed to attribute that ordering to amplification structure alone.

L ATTRIBUTION LADDER DETAIL

The central attribution result uses matched comparisons. Across them, final-checkpoint success for correct targets is 68.6, 57.9, and 52.1%; shuffled targets reach 81.4, 67.9, and 60.7%, while random-vocabulary targets reach 87.1, 75.0, and 67.1%. Correct content–context pairing is therefore unnecessary for the observed 4B gain. These three pairs do not show that content is generally irrelevant, nor do they identify the mechanism behind the content-free effect.

A compute-matched replay control, which doubles PPO epochs without an auxiliary target, reaches 48.6% final-checkpoint success, and a structure-matched gradient-noise control reaches 47.9%; both lie in the baseline range. These controls show no detectable benefit from extra PPO epochs alone or from the tested zero-mean perturbation. A random-token target without the structured schema reaches 67.9%. Conversely, a freshly resampled half-corrupt target performs poorly (10.0%). Together these observations implicate a learnable, structured CE target distribution, but do not establish a specific regularization mechanism. They motivate matched placebo controls whenever gains are attributed to world-model semantics.

M WEBSHOP AND HIDDENRULE-GYM DETAIL

The same failure pattern appears in the tested WebShop condition, but its training dynamics differ by regime. In full WebShop, the reward arm follows its baseline for roughly 60 updates and then reaches a sustained floor from about step 105. In the small regime it remains near the floor from the beginning. The final-checkpoint task score is 0.5% for both reward arms, compared with 28.5% and 60.0% for their baselines. The small-regime mean-only arm reaches 57.5%. Because task returns are continuous and component variances were not logged, we do not infer the same strict all-failure mechanism as in ALFWorld. The experiment reproduces the collapse and, in the small regime, the mean-only rescue.

The auxiliary-loss evidence is less stable. Across the WebShop comparisons, final-checkpoint task-score differences relative to the 60.0% baseline are $+3.0$ and -14.0 points. A shuffled-target arm reaches 67.0%. On HiddenRule-Gym, correct and shuffled auxiliary targets reach 17.1% and 15.0%, respectively, both overlapping the 10.7% no-signal result. No environment in this study shows

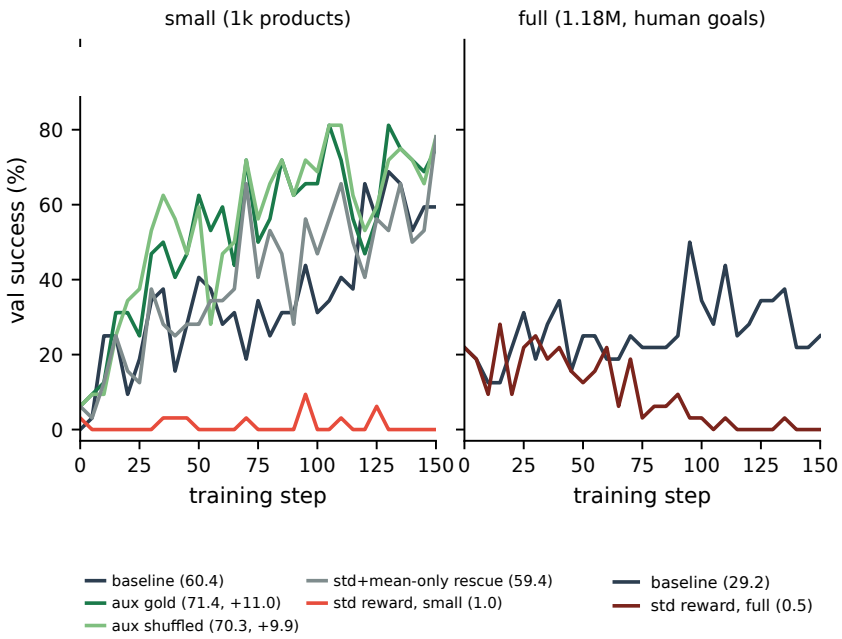


Figure 8: WebShop validation curves. In the small regime, the standard-normalized reward arm remains near the floor and the mean-only arm tracks the baseline. In the full regime, the standard-normalized reward arm follows the baseline initially and then falls to a sustained floor. The auxiliary-loss curve in the small-regime panel is included only as a boundary comparison.

a clear advantage of correct targets over shuffled targets. The main content-attribution statement therefore relies on the matched ALFWorld comparisons.

M.1 Exploratory coverage and capacity evidence

HiddenRule-Gym makes feature coverage exactly computable. Under mean-only normalization, the five measured coverage rungs reach final-checkpoint success rates of 26.4, 34.3, 13.6, 22.1, and 10.7%, against a 13.6% matched-normalization no-signal reference (Figure 9). Only the 0.48-coverage rung is clearly separated from that reference. The non-monotonic curve precludes a general claim that increasing coverage improves task success. A privileged-feature arm reaches 24.3% versus 13.6%, with overlapping intervals, so partial conversion is descriptive.

A 4B no-signal condition reaches 26.6% in the training-time protocol compared with 8.3% at 1.7B, suggesting that capacity contributes to escaping the smaller model's floor. This is one cross-scale pair and the 4B arm lacks a unified final-checkpoint endpoint, so it establishes neither necessity nor sufficiency. We use HRG only to illustrate three distinct gates—target information, optimization, and conversion by the learner—that ALFWorld does not isolate.

The 8B auxiliary-loss split. The full-weight gold auxiliary loss gives final-checkpoint endpoints spanning 0.0–85.0%. Low-success trajectories remain persistent, while the high-performing trajectory enters an early dip and then escapes. These observations are consistent with basin-like behavior but do not distinguish that interpretation from unusually large training variability. They do establish that moving supervision outside reward normalization is not a uniform safety guarantee.

The associated controls remain healthy: shuffled targets reach 79.3%, a half-weight gold loss reaches 77.1%, and the prompt-only control reaches 52.1%. The matched half-weight comparison supports a role for loss scale in one failure, while the other cells suggest that full-weight correct targets may contribute to instability. They do not localize a general cause or estimate a failure probability.

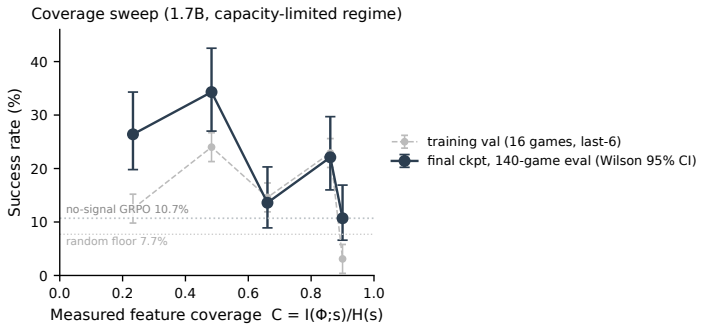


Figure 9: HiddenRule-Gym final-checkpoint success across five measured feature-coverage levels. The response is non-monotonic; the sweep is exploratory rather than a calibrated coverage law.

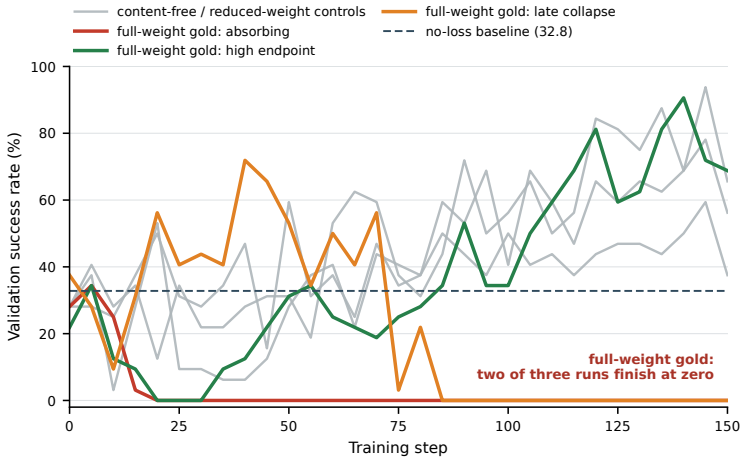


Figure 10: Qwen3-8B auxiliary-loss trajectories. Full-weight gold targets produce both zero and high-success endpoints; shuffled, prompt-only, and half-weight controls remain healthy.

N UNIFIED FINAL-CHECKPOINT EVALUATION AND CALIBRATION

Protocol. Last-6 summaries use 32 sampled validation episodes (16 for 1.7B HiddenRule-Gym) and describe training dynamics rather than training-instability uncertainty. We re-evaluate the final checkpoint of each arm in Table 14, each HiddenRule-Gym arm with a final-checkpoint endpoint, and the WebShop arms under a common protocol. ALFWorld uses all 140 `valid_seen` games and 140 loader outputs from the `valid_unseen` split. The latter contains 134 unique games and six duplicates; we therefore treat unseen point estimates as descriptive. Decoding is sampled at temperature 0.4. HiddenRule-Gym uses 140 games from held-out rule families, and WebShop uses 200 fixed test goals.

The unified final-checkpoint evaluation preserves the large qualitative contrasts. The 1.7B, 4B, and 8B reward-channel endpoints remain between 0.0 and 0.7%. The matched mean-only pair remains close (52.9 vs. 57.9%), and all-failure filtering reaches 40.7% against a 42.9% baseline. The 4B condition means are 44.1% for the baseline, 59.5% for correct auxiliary targets, 70.0% for shuffled targets, and 76.4% for random-vocabulary targets. These are descriptive condition-level summaries; row-level Wilson intervals do not capture training variability.

The unified evaluation also corrects a misleading training-time value: the 8B baseline rises from a 32-game last-6 estimate of 32.8% to 45.0% over 140 games. Smaller differences, including the self-report and compute controls, remain sensitive to the evaluation set and training variability. We therefore base the main paper on collapse-scale gaps and matched interventions, not fine-grained ranking.

arm	channel / role	seen (140)	[95% CI]	unseen†	last-6	Δ (seen)
1.7B						
baseline	—	17.9	[12.4, 25.1]	18.6	20.9	(ref)
prediction reward, std-normalized	reward	0.0	[0.0, 2.7]	0.0	0.0	−17.9
4B: reference arms						
baseline	—	42.9	[35.0, 51.2]	42.1	49.5	(ref)
baseline	—	40.7	[32.9, 49.0]	37.1	39.1	(ref)
baseline	—	48.6	[40.5, 56.8]	46.4	51.6	(ref)
no-signal control, mean-only	control	57.9	[49.6, 65.8]	44.3	52.6	(ref, mean-only)
compute-matched replay (2× epochs)	control	48.6	[40.5, 56.8]	39.3	48.9	+5.7
4B: reward channel						
prediction reward, std-normalized	reward	0.7	[0.1, 3.9]	0.0	1.0	−42.2
prediction reward, mean-only (rescue)	reward	52.9	[44.7, 61.0]	40.7	51.6	−5.0
prediction reward, all-fail filter	reward	40.7	[32.9, 49.0]	35.7	39.6	−2.2
prediction reward, decoupled cap	reward	27.9	[21.1, 35.8]	30.0	31.2	−15.0
accuracy-progress reward (Δacc)	reward	32.1	[24.9, 40.2]	29.3	28.1	−10.8
recall reward (anchor-QA)	reward	44.3	[36.3, 52.6]	33.6	44.3	+1.4
self-report confidence reward	reward	38.6	[30.9, 46.9]	39.3	49.5	−4.3
4B: loss channel						
auxiliary CE, true gold	loss	68.6	[60.5, 75.7]	65.0	69.3	+25.7
auxiliary CE, true gold	loss	57.9	[49.6, 65.8]	45.0	63.5	+17.2
auxiliary CE, true gold	loss	52.1	[43.9, 60.2]	39.3	50.0	+3.5
auxiliary CE, shuffled gold	loss placebo	**81.4**	[74.1, 87.0]	**73.6**	76.0	+38.5
auxiliary CE, shuffled gold	loss placebo	67.9	[59.8, 75.1]	66.4	60.4	+27.2
auxiliary CE, shuffled gold	loss placebo	60.7	[52.4, 68.4]	43.6	48.9	+12.1
auxiliary CE, random vocabulary	loss placebo	**87.1**	[80.5, 91.7]	71.4	73.4	+44.2
auxiliary CE, random vocabulary	loss placebo	75.0	[67.2, 81.4]	51.4	77.6	+34.3
auxiliary CE, random vocabulary	loss placebo	67.1	[59.0, 74.3]	48.6	60.9	+18.5
auxiliary CE, random tokens	loss placebo	67.9	[59.8, 75.1]	57.9	61.5	+25.0
gradient-noise falsifier (aux channel)	loss control	47.9	[39.8, 56.1]	38.6	43.2	+5.0
auxiliary CE, half-corrupted gold	loss boundary	10.0	[6.1, 16.1]	7.9	10.9	**−32.9**
4B: estimator family (post-freeze final-checkpoint round)						
GiGPO + prediction reward	reward	0.0	[0.0, 2.7]	0.0	0.0	—
PPO/GAE + prediction reward	reward	36.4	[28.9, 44.6]	40.0	37.5	—
RLOO baseline	—	45.0	[37.0, 53.3]	40.7	44.8	(ref)
RLOO + prediction reward	reward	42.1	[34.2, 50.4]	35.7	40.6	−2.9
REINFORCE++ baseline	—	55.0	[46.7, 63.0]	46.4	55.2	(ref)
REINFORCE++ + prediction reward	reward	46.4	[38.3, 54.6]	46.4	48.4	−8.6
8B						
baseline	—	45.0	[37.0, 53.3]	39.3	32.8	(ref)
prediction reward, std-normalized	reward	0.0	[0.0, 2.7]	0.0	0.0	−45.0
prompt-only control	—	52.1	[43.9, 60.2]	44.3	47.4	+7.1
auxiliary CE, true gold	loss	0.0	[0.0, 2.7]	0.0	0.0	−45.0
auxiliary CE, true gold	loss	**85.0**	[78.2, 90.0]	72.1	72.4	+40.0
auxiliary CE, true gold	loss	0.0	[0.0, 2.7]	0.0	0.0	−45.0
auxiliary CE, half-weight (β=0.05)	loss	77.1	[69.5, 83.3]	75.0	65.1	+32.1
auxiliary CE, shuffled gold	loss	79.3	[71.9, 85.2]	66.4	78.6	+34.3
4B: additional reward-channel conditions						
prediction reward, std-normalized	reward	0.0	[0.0, 2.7]	0.0	0.0	−40.7
prediction reward, decoupled cap	reward	32.1	[24.9, 40.2]	26.4	42.2	−8.6
accuracy-progress reward, Δacc	reward	27.1	[20.4, 35.0]	25.7	34.4	−13.6

Table 14: Unified final-checkpoint evaluation. Seen-split Wilson 95% intervals condition on a fixed checkpoint and do not quantify training variability. Δ uses the matched-normalization reference where available; it is descriptive. †Unseen values are computed from 140 loader outputs containing 134 unique games and six duplicates, so no independence-based interval is attached. All listed seen and unseen evaluations are complete. Anchor-QA's two 44.3 values coincide by arithmetic.

On the unseen split, every near-zero reward-channel endpoint remains zero, and most large qualitative patterns persist with smaller magnitudes. The loader evaluates 140 episodes drawn from 134 unique unseen games; six tasks are duplicated, so we report those point estimates descriptively rather than attach independence-based confidence intervals.

Additional scope checks. Additional group sizes 8 and 16 address the possibility that the result is only an unstable small-group estimate; both also collapse, although the three onset times do not support a scaling law. WebShop addresses the single-environment concern for the failure phenotype and gives one cross-environment mean-only rescue. The matched trajectory-first control directly tests normalization order at 1.7B and removes prediction-mode damage over updates 30–60. Component telemetry outside the instrumented 1.7B pair remains unavailable.

O EXTENDED PRACTICAL GUIDANCE

1. Inspect an auxiliary reward after every transformation that precedes the policy gradient. Raw scale and episodic bounds do not describe normalized stepwise weights.
2. Log task, auxiliary, and penalty variance and covariance for each pool, together with the fraction of tied or low-spread task groups. Our 1.7B diagnostic shows that this direct decomposition is more informative than a coefficient sweep alone; it should be routine rather than limited to one matched pair.
3. Treat mean-only normalization as a diagnostic and a candidate design choice, not a universal cure. When a reward's safety argument relies on telescoping, include a trajectory-first aggregation control; ours localizes the timing failure over the evaluated window.
4. When using auxiliary CE, include content-free targets that match length, schema, masking, and optimization budget. Our 4B results rule out attributing the gain solely to correct next-observation semantics, while the 8B and WebShop results show that the route is not uniformly beneficial or stable.
5. Monitor prediction accuracy, task-score spread, behavior diversity, and episode length jointly. The frozen early-warning roster in this study is an exploratory diagnostic, not a prospective estimate of detector performance.

Statistical interpretation. The reported endpoints support the large qualitative statements in the main text but not precise effect sizes, failure probabilities, or fine-grained rankings. In particular, WebShop auxiliary-loss effects change sign relative to baseline, and 8B gold-target endpoints span 0.0–85.0%. All rows in Table 14 now have complete seen and unseen endpoints; mechanism-only arms omitted from that table retain training-time endpoints.

Scope. ALFWorld, WebShop, and HiddenRule-Gym use one base-model family and one agent framework. ScienceWorld contributes two completed but floor-bound arms and no interpretable comparison. The cancellation algebra is architecture-agnostic; the empirical failure and behavioral signatures remain specific to the tested implementation. The estimator comparison includes algorithms that differ in multiple ways and should be read as supporting scope evidence, not a factorial estimator study.

P CRITERION: COMPATIBILITY AND IDENTIFIABILITY DETAIL

Three quantities should be kept distinct when assessing a dense reward. Task scores determine how much discrimination the task objective supplies; residual component variance determines what standardization can amplify; and policy controllability determines whether the agent can change that residual through its behavior. Equation 4 addresses only the second quantity under explicit limiting conditions. The matched 4B result shows that the observed collapse depends on standard-deviation division, but it does not establish a necessary-and-sufficient behavioral criterion.

Signal form matters because saturation of a mean score does not imply that its stepwise increments have zero variance. The anchor-QA score climbs slowly and the self-report score becomes nearly constant; neither condition reaches the strict zero-success lock. The Δacc control uses $r_{\Delta\text{acc}} = \lambda \,\text{clip}(\Phi_t - \Phi_{t-1}, 0, 1)$ and is not episodically telescoping after positive clipping. It ends below baseline at point estimate without locking. These arms show that endpoint accuracy alone does not predict training behavior; they do not yield a complete taxonomy.

The matched-marginal-variance noise arm drags but does not lock, suggesting that policy controllability may matter. Its temporal correlations are not fully matched, so the experiment is not a clean isolation. The invalid-action penalty also varies inside task-tied groups. The original battery retains only batch means, but the instrumented 1.7B pair resolves this ambiguity: prediction accounts for 98.74% of late task-tied pre-normalization variance, compared with 1.21% from penalties and 0.05% from their interaction. This measures late prediction dominance in that setting, not its generality, temporal precedence, or the final post-clipping gradient.

Published successes do not contradict this scoped result: RWML uses a level-scored embedding reward and VAGEN changes the estimator, so neither tests the same normalization pathway.