SAME GROUPS, DIFFERENT VALUES: DECOMPOSING STEP-LEVEL CREDIT ASSIGNMENT FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Step-level group-relative estimators for language-model agents are often characterized by the states they place in the same group. We show that this description is incomplete: the rollout partition determines where statistics are shared, whereas the value functional determines what signal is aggregated. We introduce a component-resolved audit that varies these choices separately on matched rollouts, and complement it with a support-aware variance decomposition for measuring history-associated value variation after the task and current observation are fixed. The framework combines an unbalanced random-effects estimator with task-level resampling, explicit accounting of variance-bearing groups, and replay-based validation of the collected trajectories. Across replay-verified policy snapshots, two text environments, and an independently emitted matched-row batch, the audit reveals that apparently different grouping methods can remain closely aligned when they retain the same base operator, while added value machinery can dominate their differences. It also shows why sparse terminal feedback can leave most formally eligible groups uninformative even when a denser progress signal varies within them. The academic value is a reusable attribution protocol: it separates algorithmic structure, measurable support, inferential units, and data fidelity before a snapshot difference is assigned a mechanism.

1 INTRODUCTION

Group-relative policy optimization replaces a learned critic with comparisons among samples that share a prompt or state (Shao et al., 2024). In multi-step LLM agents, however, the same text observation can follow different histories, latent states, and remaining plans (Feng et al., 2025; He et al., 2026; Wang et al., 2026). Recent methods respond with flat state groups, history-conditioned groups, or state-transition graphs. This flat/history/graph taxonomy is useful, but it can make the grouping rule appear to be the estimator itself.

That conflation obscures mechanism attribution. A *partition* specifies which rollout steps share statistics; a *value functional* specifies what is evaluated and how multiple structures are aggregated. Two methods can share a partition yet produce very different advantages, while a hierarchical method can remain close to a flat estimator if its aggregation retains substantial base-level weight. Figure 1 separates these two axes.

We use this factorization to ask two controlled questions. First, when released estimators consume the same rollout rows, which differences come from grouping and which come from the value functional? Second, after fixing the task and current anchor, is there enough history-associated value variation to measure reliably? The first question separates GiGPO, faithful HGPO, a grouping-only G2PO control, and full G2PO. The second motivates a variance-component analysis of recent-history labels. Both questions concern computed advantage vectors and recorded policy distributions, not exact policy-gradient geometry.

For each task-and-anchor group, we treat recent histories as labels and estimate their associated share of value variance with a pooled unbalanced ICC(1), denoted $\hat{\rho}$. The estimand is deliberately narrow: it quantifies structured variation within the recorded policy distribution. It does not identify

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.

a hidden state, prove that conditioning on history improves learning, or imply generalization to unseen environments.

Contributions.

- **A factorized estimator audit.** We separate rollout partition from value functional and use matched rows to distinguish exact flat-group equivalence, hierarchical refinement, and additional value machinery.
- **Stress-tested measurement.** We specify a pooled unbalanced ICC(1), whole-task bootstrap, explicit support conditions, and fixed-data subsampling and label-shuffle diagnostics, while stating why the latter do not constitute a cluster-valid null test.
- **An effective-support result.** Replay-verified snapshots show that sparse terminal reward can make informative groups far rarer than nominally eligible groups; the available matched policy levels do not support an increasing $\hat{\rho}$ trend.
- **Independent scope checks.** A separately emitted matched-row batch tests producer sensitivity, and a formula-audited Yang progress-fallback implementation tests whether the partition/value distinction transfers to a related operator family. Neither check uses task outcomes as evidence.

Data-validity failures and null results are part of the answer: they determine which comparisons are identified by the available artifacts. Our contribution is therefore a measurement and attribution framework, not a new optimization algorithm. All conclusions concern fixed policy snapshots or matched-row operator calculations; we make no claim about downstream task improvement or learning dynamics.

Relationship to a concurrent anonymous submission. A concurrent anonymous submission, *DSA: Progress Inside the Group Standardizer in Sparse-Reward Agent Learning*, proposes and evaluates that training operator. This paper neither defines nor evaluates DSA; it shares only the problem setting and fixed-row audit interface, and no claim here uses DSA training outcomes or controls.

2 Related work

Credit assignment and policy-gradient estimation. Classical RL separates the target used for temporal credit from the update operator: temporal-difference and Q-learning bootstrap value targets (Sutton, 1988; Watkins & Dayan, 1992), whereas REINFORCE and the policy-gradient theorem connect sampled returns to action probabilities (Williams, 1992; Sutton et al., 1999). Actor–critic baselines, asynchronous actors, and GAE make bias–variance control explicit, while PPO changes how the resulting advantage enters an update (Konda & Tsitsiklis, 1999; Greensmith et al., 2004; Mnih et al., 2016; Schulman et al., 2016; 2017). Distributed and off-policy corrections further separate behavior data from targets and updates (Espeholt et al., 2018; Munos et al., 2016; Gu et al., 2017; Sutton & Barto, 2018). Our fixed-row audit asks a narrower attribution question and proposes neither a new return estimator nor an optimizer.

Group-relative estimation for LLM agents. GRPO compares completions of one prompt (Shao et al., 2024). GiGPO adds groups over repeated environment states (Feng et al., 2025); HGPO aggregates nested history groups (He et al., 2026); and G2PO adds graph values and edge-centric advantages (Wang et al., 2026). The resulting flat/history/graph taxonomy describes structure but not the entire estimator. We instead factorize each executable operator into the partition that shares statistics and the value functional applied to them. This supports component attribution, not a re-evaluation of the papers' task-performance claims.

History, sparsity, and text environments. A current observation need not be sufficient in a partially observable process (Kaelbling et al., 1998), but this does not imply that every repeated text observation is aliased or that memory improves learning. Sparse-feedback agent methods motivate step-level grouping (Feng et al., 2025; He et al., 2026; Wang et al., 2026); progress-conditioned variants add denser transition signals or an all-failed-group branch (Fan & Liu, 2026; Yang et al., 2026). ALFWorld and ScienceWorld provide terminal- and dense-reward measurement regimes

here (Shridhar et al., 2021; Wang et al., 2022). We use history only as a within-anchor label, treat the environments as contrasts rather than controlled effects, and use the Yang fallback only as a matched-row operator comparison rather than a training ranking.

Off-policy evaluation and inferential scope. Off-policy traces and policy gradients distinguish a target policy from the behavior data (Precup et al., 2000; Degris et al., 2012). Logged-data policy-value estimators use doubly robust, marginalized-ratio, or minimax constructions (Jiang & Li, 2016; Thomas & Brunskill, 2016; Liu et al., 2018; Nachum et al., 2019; Kallus & Uehara, 2020; Uehara et al., 2020); high-confidence methods and empirical benchmarks stress coverage (Thomas et al., 2015; Fu et al., 2021). Our matched-row cosines and snapshot ICC are not off-policy value estimates, so we report replay eligibility, effective support, and the independent inferential unit instead of extrapolating to an unobserved policy.

Reliability, calibration, and variance components. Deep-RL conclusions can depend on implementation, replication, and evaluation choices (Henderson et al., 2018; Agarwal et al., 2021; Engstrom et al., 2020; Ilyas et al., 2020; Andrychowicz et al., 2021). Predictive calibration, model calibration, ensembles, and distributional values address uncertainty objects different from ours (Kuleshov et al., 2018; Lakshminarayanan et al., 2017; Bellemare et al., 2017; Dabney et al., 2018). We instead stress-test the finite-sample behavior of a pooled one-way ICC(1) (Shrout & Fleiss, 1979; McGraw & Wong, 1996; Koo & Li, 2016), using unbalanced method-of-moments variance components (Searle et al., 1992; Gelman, 2005) and a whole-task bootstrap (Efron & Tibshirani, 1993). This adapts the estimand and dependence structure to repeated trajectories; it is neither a new variance estimator nor a general ICC theorem.

3 Problem formulation and measurement

We use two analyses with different counterfactuals. A matched-rollout audit holds the data rows and return target fixed while changing estimator components; it asks which operator causes an advantage-vector difference. A variance-component analysis holds the task and current anchor fixed; it asks how much variation in a selected value channel is associated with recent-history labels. Figure 1 shows both paths. Keeping them separate prevents a structural difference between estimators from being mistaken for a training effect.

3.1 Grouping unit and value channels

For task q and normalized current observation o, the flat anchor group is $G = (q, o)$. A row $i \in G$ carries recent-history label $h_i^{(k)}$, discounted return-to-go R_i, and—when available—an oracle diagnostic V_i^{oracle}. The Monte-Carlo return depends on the evaluated policy's trajectory. The oracle diagnostic is negative planner distance to goal: its value at a state does not depend on the policy, although the distribution of visited states does. Because the source random-policy collection has limited replay fidelity, we use that channel only for explicitly marked construct checks. Primary conclusions use the return channel and replay-verified datasets.

3.2 Pooled unbalanced ICC(1)

Fix a value channel Y. In anchor group g, let $j = 1, \dots, J_g$ index history labels, n_{gj} their sizes, $n_g = \sum_j n_{gj}$, and Y_{gji} a step value. Define

$$\mathrm{SSB}_g = \sum_j n_{gj}(\bar{Y}_{gj} - \bar{Y}_g)^2, \qquad d_{B,g} = J_g - 1, \tag{1}$$

$$\mathrm{SSW}_g = \sum_j \sum_i (Y_{gji} - \bar{Y}_{gj})^2, \qquad d_{W,g} = n_g - J_g, \tag{2}$$

$$n_{0,g} = \frac{n_g - \sum_j n_{gj}^2/n_g}{J_g - 1}. \tag{3}$$

We pool sums of squares and degrees of freedom over eligible anchor groups, $\mathrm{MSB} = \sum_g \mathrm{SSB}_g / \sum_g d_{B,g}$ and $\mathrm{MSW} = \sum_g \mathrm{SSW}_g / \sum_g d_{W,g}$, with $\bar{n}_0 = \sum_g d_{B,g} n_{0,g} / \sum_g d_{B,g}$. The

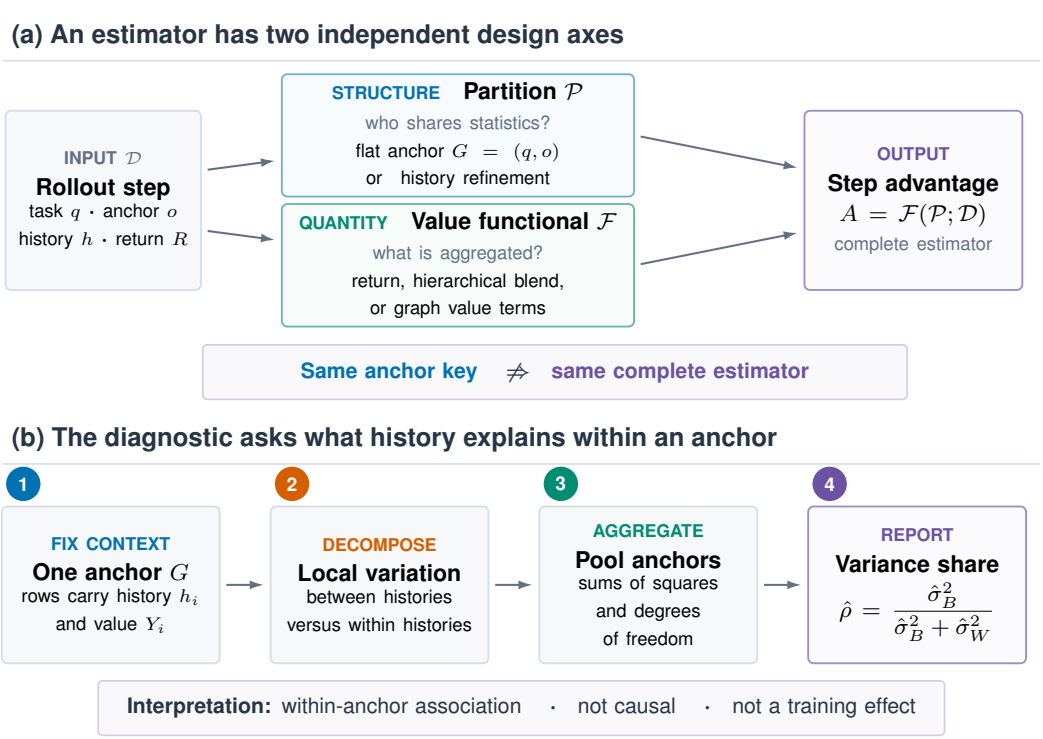

Figure 1: Two questions, two controlled comparisons. (a) Every step estimator chooses both who shares statistics (the partition) and what is aggregated (the value functional); a shared anchor key does not identify the complete estimator. (b) The pooled ICC fixes an anchor, separates between-history from within-history variation, pools eligible anchors, and reports association rather than a causal or training effect.

method-of-moments component estimates and reported diagnostic are

$$\hat{\sigma}_B^2 = \max\left(0, \frac{\mathrm{MSB} - \mathrm{MSW}}{\bar{n}_0}\right), \quad \hat{\sigma}_W^2 = \mathrm{MSW}, \quad \hat{\rho} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}. \tag{4}$$

This is the pooled unbalanced ICC(1) method-of-moments estimator (Shrout & Fleiss, 1979; Searle et al., 1992), not an average of unstable per-group ICCs. A group is estimable only if $J_g \geq 2$ and $n_g - J_g \geq 1$; the reported default additionally requires $n_g \geq 4$. We count every group in coverage diagnostics but pool only these "preferred" groups. Thus $\hat{\rho}$ summarizes the modeled within-anchor variance share associated with the history label under the random-effects decomposition—not a causal effect or a universal property of the environment.

Proposition 1 (Pooled unbalanced moment identity)**.** *For each eligible anchor g, suppose*

$$Y_{gji} = \mu_g + B_{gj} + \varepsilon_{gji}, \qquad \mathbb{E}B_{gj} = \mathbb{E}\varepsilon_{gji} = 0,$$

where the label effects and residuals are mutually independent, with common variances $\mathrm{Var}(B_{gj}) = \sigma_B^2$ and $\mathrm{Var}(\varepsilon_{gji}) = \sigma_W^2$, and condition on fixed, non-informative label sizes. Then

$$\mathbb{E}[\mathrm{MSW}] = \sigma_W^2, \qquad \mathbb{E}[\mathrm{MSB}] = \sigma_W^2 + \bar{n}_0 \sigma_B^2.$$

Consequently, the untruncated estimate $\widetilde{\sigma}_B^2 = (\mathrm{MSB} - \mathrm{MSW})/\bar{n}_0$ is unbiased for σ_B^2 at finite sample size.

The result explains why the pooled estimator retains the complete unbalanced label-size pattern rather than substituting a nominal average group size. Under additional Gaussian and independent-anchor assumptions, its conditional variance depends on that same pattern: limited degrees of freedom and concentrated label sizes can reduce precision even when the total row count is unchanged. Appendix C gives the proof and exact quadratic-form variance.

The guarantee is deliberately narrow. Proposition 1 makes the *untruncated* components unbiased only for fixed, non-informative label sizes and common independent random effects. Truncation, the nonlinear ICC ratio, and trajectory dependence can break that guarantee. We therefore use whole-task resampling for uncertainty and treat within-anchor label shuffles only as descriptive references: independent shuffles do not preserve joint trajectory/task dependence and are not calibrated permutation tests.

Rows enter the variance sums but are not experimental replicates: the bootstrap resamples tasks. A fixed-data subsampling ladder exposes strong sample-size sensitivity and motivates a dataset-specific reporting heuristic, not a universal minimum sample size (Section 4.3).

3.3 Estimator audit on matched rollouts

All implementations consume the same rollout batch and return one per-step advantage vector.

- **GiGPO** standardizes return within each flat G and combines the result with an episode-level term.
- **HGPO** includes the current-observation group at depth 1, constructs deeper history-keyed subgroups, drops singletons, and renormalizes a length-weighted blend over surviving depths. Here K is the maximum history depth and α is the released aggregation-weight parameter. HGPO augments rather than replaces the flat partition.
- **G2PO** uses the same task-and-anchor partition in its grouping stage and then adds a next-group score, invalid-action penalty, and score-gain machinery. Our grouping-only version is a diagnostic control, not the full published estimator.

The HGPO implementation is derived from the released cluster construction and validated against upstream `hgpo_advantage_estimate`: maximum per-step error is 2.2×10^{-7} for $K \in \{2, 4\}$ and $\alpha \in \{0, 1\}$. Appendix G documents why a pure-depth ablation is not a faithful HGPO implementation.

For vectors A and A' on identical rows, we report cosine similarity, sign agreement where both entries are nonzero, and the reference's squared-norm share on materially disagreeing rows. These diagnose advantage vectors, not policy-gradient geometry, which also depends on per-token score functions and on whether estimator transformations preserve the state-baseline identity (Tucker et al., 2018); exact gradients were not stored.

Environments, models, schema, and collection are specified in Appendix A, including the fixed analysis parameters in Table 3; finite-sample stability details appear in Appendix D.

4 Results

We organize the evidence as a sequence of controls. Matched rows isolate estimator components; effective-support and sample-size checks determine what the return channel can identify; replay determines which datasets are eligible; and independent matched-row checks test whether the operator decomposition survives a new producer and a related method family. The primary evidence comes from the 1.5B and 7B Qwen2.5 policy snapshots; smaller auxiliary snapshots are used only for sensitivity checks. Exact success rates and sample sizes are reported where they are needed to interpret a comparison.

4.1 MATCHED ESTIMATOR COMPARISON

Table 1: Matched-rollout estimator comparison. "cos" is cosine between per-step advantage vectors, not actual policy-gradient alignment. Grouping-only G2PO holds the value functional fixed; faithful HGPO includes the flat group plus history refinements.

quantity	1.5B policy	7B policy	ScienceWorld
G2PO grouping ≡ flat anchor grouping?	true	true	true
cos(GiGPO, grouping-only G2PO)	**1.000**	**1.000**	**1.000**
cos(GiGPO, faithful HGPO)	0.965	0.944	0.937
GiGPO signal zeroed by HGPO	0.0%	0.0%	0.0%
cos(GiGPO, full G2PO)	0.187	0.272	0.183

Same partition, identical controlled estimator. For every row in all three evaluated datasets, G2PO's grouping-stage identifier equals GiGPO's task-and-normalized-observation identifier. When we also hold the return functional fixed, the two advantage vectors are elementwise equal (cosine 1.000; Table 1). This establishes an implementation-level identity for the controlled operators, not equivalence of the full published estimators.

History refinement need not imply a large vector change. HGPO's depth-1 term is the flat anchor, while deeper terms use history-conditioned subgroups. Under the released settings, singleton filtering and weight renormalization often return weight to that base term. The resulting cosine with GiGPO is 0.937–0.965, and HGPO zeros no nonzero GiGPO entry. The partitions differ, but their aggregated advantages remain close on these rollouts; this is not a training comparison.

Value machinery explains the controlled G2PO gap. Restoring G2PO's next-group score, invalid-action penalty, and score-gain term lowers the cosine to 0.183–0.272. Because the grouping-only and full controls share their partition, this drop isolates the added value machinery on the matched rows. Thus "same state group" is insufficient for mechanism attribution. Appendix F localizes the squared-norm disagreement; Appendix G gives the faithful HGPO construction and retains our earlier port retraction.

4.2 SPARSE REWARD COLLAPSES EFFECTIVE SUPPORT

A support bound. Zero reward on failed trajectories makes every all-failure anchor return-constant. Thus nonzero variance requires a success-containing group: mixed groups suffice, while discounted all-success groups may vary with terminal distance. Success therefore bounds return-channel information.

The bottleneck is tight in both audited cold-start snapshots: every success-containing preferred group carries return variance. For the 1.5B policy, however, fewer than one in twenty preferred groups is variance-bearing, so nominal eligibility substantially overstates observed support. The appendix reports the exact support count for this primary 1.5B snapshot alongside a dense-reward contrast. One successful trajectory can inform multiple anchors, and the observed equality is not a theorem for every terminal-reward batch.

The silent return channel does not imply identical latent states. The preceding calculation uses the preferred-ICC subset, whereas the activation comparison uses all flat anchor groups. In that broader denominator, progress varies in substantially more groups than terminal return in both snapshots. History-associated structure is therefore measurable even where the optimized return is silent. This is an availability result, not evidence that using the structure improves a policy. ScienceWorld contributes only a dense-reward measurability contrast (Appendix I); its $\hat{\rho}$ magnitude is withdrawn.

An injected-aliasing check makes observations coarser and recomputes both keys. Label-shuffle-centered $\hat{\rho}$ rises from 0.20 to 0.44 over six of seven levels. Because these levels are nested transforms of one dataset rather than independent replicates, the pattern is descriptive. Appendix Figure 8 reports the rank summary and its exact enumeration. The replay-flagged random-policy dataset makes this check *indicative*, not primary evidence.

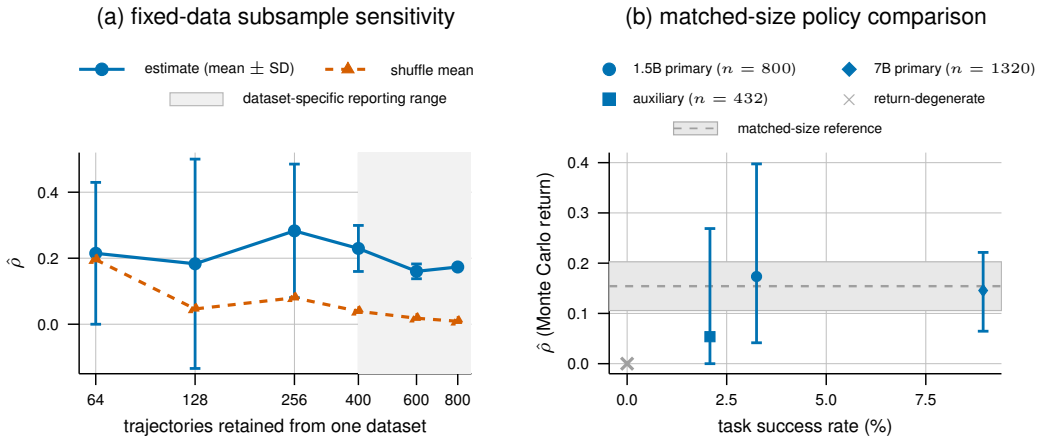

Figure 2: Finite-sample stability and matched-policy sensitivity. (a) Fixed-data subsampling reveals substantial estimator sensitivity and motivates a dataset-specific reporting threshold. (b) A matched-size comparison does not support a larger diagnostic at the higher-success policy. Ranges summarize repeated subsampling of recorded data rather than uncertainty across independently trained policies; withdrawn dial results are omitted.

The attempted ϵ-oracle success dial is removed from all results. Observation replay fails at every retained dial level, and terminal returns inherit the unverified success labels. Accordingly, dial-derived success, informative-group, $\hat{\rho}$, cosine, and headroom values support no claim. Appendix K retains the failure as a methodological warning.

4.3 Finite-sample stability stress test

Raw ICC(1) is highly sample-size sensitive in this sparse, unbalanced regime. The fixed-data ladder and the auxiliary snapshot both show that changing the retained trajectory count can materially move the estimate. We therefore match sample sizes where possible and use $n \geq 400$ only as a dataset-specific reporting heuristic. For unequal sizes we show the label-shuffle-centered statistic $\hat{\rho} - \mathbb{E}_{\text{shuffle}}[\hat{\rho}]$, not a generally bias-corrected estimator. Because the ladder resamples trajectories rather than complete task blocks, its draws can alter task and group composition; within-anchor shuffles also break joint trajectory dependence. These diagnostics expose sensitivity but neither estimate frequentist coverage nor establish a universal sample requirement.

4.4 Implications at fixed policy snapshots

On the verified 1.5B-policy Monte-Carlo return, define the matched-row difference $b_i = A_i^{\text{GiGPO}} - A_i^{\text{HGPO}}$ and the normalized aggregate $\text{mean}_i\, |b_i| / \text{sd}_i(A_i^{\text{GiGPO}})$. Its value is 0.072; the strict sign-reversal rate is 0.09%. Together with the GiGPO/HGPO cosine, these diagnostics bound the observed snapshot-level footprint under the selected diagnostics: the vector changes, but rarely reverses sign. They neither measure a policy update nor imply that history is unimportant during training.

At matched $n = 800$, the higher-success 7B policy does not have a larger $\hat{\rho}$ than the 1.5B policy; their fixed-data resampling comparison has descriptive tail area 0.40. The subsamples quantify sensitivity within recorded data, not variation across independently trained policies. The point estimates and resampling ranges are shown in Figure 2; the appendix documents the resampling protocol. ScienceWorld is excluded from this magnitude comparison. The separate history-depth pattern is non-monotone and limited by subgroup fragmentation and data density. On verified data, structural-bias sign agreement remains near its descriptive reference in both datasets; pair dependence precludes a cluster-valid comparison with chance. Appendix J reports the full negative-result ledger.

Table 2: Dataset eligibility after observation-level replay. Successful trajectories are checked exhaustively for the two primary policy datasets; failures use a stratified sample. "Trusted" is scoped to the audited claims, not a blanket endorsement of every field.

dataset	policy	episodes reproducing exactly	verdict
1.5B primary	model, 3.25% success	successes **26/26**; failures **240/240**	trusted
7B primary	model, 8.94% success	successes **118/118**; failures **260/260**	trusted
auxiliary snapshots	model	all sampled trajectories reproduced	limited supporting replay checks
random oracle	random admissible actions	**10/32**	indicative only
$\epsilon = 0.2$ dial	model + planner injection	**0/32**	withdrawn
ScienceWorld	model, 7B	separate environment instances	structural/alignment diagnostics; ICC magnitude excluded

4.5 Data validity and claim eligibility

The fidelity audit changes which measurements are eligible. Our original offline ALFWorld collector used multiple slots in one shared world, so one slot's action could alter another slot's state; the upstream training path instead creates one instance per rollout. This is a defect in our collection, not in the studied methods, and implies nothing about published results. We replay each recorded action sequence in isolation and compare every observation before assigning an evidence role.

All successful trajectories in the two primary datasets reproduce exactly. The failure samples also reproduce exactly, with Wilson 95% contamination upper bounds of 1.6% and 1.5%. No group is silently removed. The residual uncertainty is sampling uncertainty in the failure stratum, not license to reuse the failed random-oracle or dial magnitudes. The operational lesson is that non-interference between rollout groups is a collection-design property, not an estimator guarantee; rollout replay therefore belongs in the validation protocol for group statistics.

4.6 Independent matched-row scope checks

Two additional controls ask whether the factorization is an artifact of the original collector or of the three primary estimator implementations. Both checks freeze a newly emitted row batch and compare executable operators on exactly those rows. They do not update a policy, compare task returns after optimization, or use checkpoints as independent evidence.

The independent producer reproduces exact key equality and the grouping-only vector identity while leaving the full operators separated. This matters because it decouples the finding from the historical offline collector. In the second check, the Yang et al. (Yang et al., 2026) implementation inherits the same grouping rule but changes the value assigned to eligible all-failure anchors. Its activation rate therefore illustrates the same partition/value distinction in a related family. Neither result is evidence that one operator trains a better policy; it is evidence that a group label alone does not specify the computation performed on a fixed batch. Protocol details and the admission boundary appear in Appendix B.

5 Discussion

A method name should resolve to an executable operator. GiGPO and G2PO's grouping stages induce the same flat equivalence relation, yet full G2PO differs substantially because its value functional changes. HGPO refines the partition but retains the flat term and stays close to GiGPO. A grouping claim is therefore under-specified unless the value channel and aggregation are also fixed.

Effective support is an identification constraint. Terminal reward makes every all-failure group return-constant. In the lower-success primary snapshot, only a small fraction of otherwise eligible groups contains observed return variance. The estimator remains defined in the silent groups, but its return channel cannot identify history-associated variation there. Support, dependence-preserving uncertainty, and stability diagnostics are therefore part of the result. The dense-reward contrast shows channel dependence, not cross-environment equality of $\hat{\rho}$.

Negative results distinguish absence of evidence from absence of structure. Planner progress varies inside many return-silent groups, so the data do contain within-anchor structure. Yet faithful HGPO changes vectors only modestly, and the matched snapshots do not show a larger $\hat{\rho}$ estimate at the higher observed success rate; the history-depth pattern is nonmonotone and limited by subgroup fragmentation and data density. ScienceWorld's raw $\hat{\rho}$ magnitude is excluded after its no-op robustness failure, and the unfaithful ϵ-dial is withdrawn. The independent-producer and Yang checks broaden operator scope without introducing post-update outcomes as evidence. History may matter; nominal grouping alone neither attributes an estimator difference nor guarantees a measurable signal.

A minimal reporting standard follows. Report the partition, value functional, aggregation and normalization, variance-bearing support, independent inferential unit, and replay or provenance checks. This exposes two confounds: naming only the group while changing the operator, and treating repeated rows or checkpoints as independent evidence.

6 LIMITATIONS

Fixed-policy scope. The decomposition is fixed-policy: its vectors and $\hat{\rho}$ characterize sampled policies, not the feedback loop after an update. The independent-producer and related-operator checks also compare only fixed matched rows. This paper deliberately contains no training-arm comparison or post-update task outcome, so it cannot establish task gain, wall-clock efficiency, or optimization behavior. We likewise do not equate advantage cosine with policy-gradient alignment.

Narrow empirical regime. The primary regime is ALFWorld at 2–9% success, Qwen2.5-Instruct at 1.5B and 7B, group size eight, and symbolic interaction; $n \geq 400$ is only a reporting heuristic for this dataset. ScienceWorld supports dense-reward measurability, but its $\hat{\rho}$ magnitude fails the no-op check. One independently emitted batch and one related progress fallback extend operator coverage, not model- or environment-level generalization.

Collection and implementation validity. The offline ALFWorld collector shared a world across slots, unlike the training path. Observation replay supports both primary policy datasets, but only a minority of sampled historical random-oracle episodes replay exactly, so that channel remains indicative. The faithful HGPO port matches its upstream reference on the equivalence batch, not every training path. Appendix L details the remaining depth, parser, and coverage limits.

7 CONCLUSION

Step-level credit estimators are defined by both who shares statistics and what is aggregated. On matched rollouts, this distinction resolves the apparent grouping disagreement: faithful history refinement changes the resulting advantage vectors only modestly, whereas G2PO's additional value machinery changes them substantially. Under sparse terminal reward, effective support—not nominal group count—then limits what history-sensitive measurement can resolve. Matched-row checks from an independent producer and a related fallback operator support the decomposition without making a downstream learning claim. Separating operator, support, inferential unit, and data fidelity is the paper's durable result. Accordingly, future credit-assignment comparisons should establish variance-bearing support before interpreting estimator differences and quantify uncertainty at the task, independently initialized run, or environment unit appropriate to the claim. This decomposition also makes null results informative by locating their boundary in support, aggregation, value construction, or evidence quality rather than in a method name.

AI USE STATEMENT

The authors used AI assistance only for manuscript drafting and English-language polishing. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The reported snapshot and matched-row results are produced through a checked analysis pipeline linking admitted rollout records, derived statistics, tables, and plotted values. The analysis refits the pooled estimator under whole-task resampling, recomputes the fixed-row operator comparisons, and excludes evidence blocks that fail the stated replay-admission rules. These checks establish arithmetic and provenance consistency; they do not turn descriptive within-anchor label shuffles into cluster-valid inference. No training-arm result or post-update task outcome is imported into this manuscript.

Data provenance and the reproduction test. Section 4.5 reports, per dataset, the fraction of recorded episodes that reproduce when their own actions are replayed in isolation. The test needs no ground truth: it re-executes a trajectory's recorded actions in a fresh environment and compares every observation with what was recorded. Complete success censuses and stratified failure samples for both primary datasets reproduce exactly, placing the Wilson upper bounds on contamination below $2\,\%$. Disabling an unused planner query makes this verification practical at scale. We release the checker because the defect it detects is a property of data-collection code rather than of any algorithm, and other rollout dumps may have it.

Estimator implementations. All three estimators are re-implemented against a common interface so that their advantage vectors are computed on identical inputs. The faithful hierarchical re-implementation is additionally validated against its released reference: its per-step advantages agree with the upstream implementation to 2.2×10^{-7} across the hyper-parameter settings we test. We report this because an earlier re-implementation of ours was wrong and produced a result we retracted (Appendix G); the validation is the safeguard we added in response.

Release. Code, analysis scripts, the verification harness, the reproduction checker, and the machine-readable records behind the reported measurements will be released publicly on publication. No repository is linked here, to preserve anonymity during review.

ETHICS STATEMENT

This work involves no human subjects, no personally identifying data, and no data collected from people. All experiments run in simulated text environments (household tasks and a science laboratory) using publicly available benchmark data and open-weight language models.

The one ethical dimension we think worth stating is a limit on what we claim. This is a fixed-policy measurement study, not evidence of task improvement after optimization. We also exclude a failed-replay success dial and retract an earlier result produced by an unfaithful hierarchical implementation. Both decisions reduce apparent claim strength, and we regard stating them plainly as the minimum standard for a measurement paper.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems 34*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nIAxjsniDzg.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11791. URL https://ojs.aaai.org/index.php/AAAI/article/view/11791.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 179–186, 2012. URL https://icml.cc/2012/papers/268.pdf.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993. ISBN 978-0-412-04231-7. doi: 10.1007/978-1-4899-4541-9. URL https://doi.org/10.1007/978-1-4899-4541-9.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. *arXiv preprint arXiv:2005.12729*, 2020. URL https://arxiv.org/abs/2005.12729.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1407–1416. PMLR, 2018. URL https://proceedings.mlr.press/v80/espeholt18a.html.

Mingxuan Fan and Peiyang Liu. Progress- and reliability-oriented group policy optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.04242*, 2026. URL https://arxiv.org/abs/2607.04242.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://arxiv.org/abs/2505.10978.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Thomas Paine. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kWSeGEeHvF8.

Andrew Gelman. Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005. doi: 10.1214/009053604000001048. URL https://doi.org/10.1214/009053604000001048.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJ3rcZcxl.

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*, 2026. URL https://openreview.net/forum?id=T8Dev99qnz.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxdEkHtPS.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661. PMLR, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient and robust off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5078–5088. PMLR, 2020. URL https://proceedings.mlr.press/v119/kallus20b.html.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems 12*, pp. 1008–1014. MIT Press, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012. URL https://doi.org/10.1016/j.jcm.2016.02.012.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/dda04f9d634145a9c68d5dfe53b21272-Abstract.html.

Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996. doi: 10.1037/1082-989X.1.1.30. URL https://doi.org/10.1037/1082-989X.1.1.30.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937. PMLR, 2016. URL https://proceedings.mlr.press/v48/mniha16.html.

Remi Munos, Thomas Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems 29*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cf9a242b70f45317ffd281241fa66502-Abstract.html.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann, 2000. URL http://www.cs.ualberta.ca/~papersdb/uploaded_files/467/paper_PSS-00.pdf.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

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*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. Wiley, New York, 1992. doi: 10.1002/9780470316856. URL https://doi.org/10.1002/9780470316856.

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. URL https://arxiv.org/abs/2402.03300.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420. URL https://doi.org/10.1037/0033-2909.86.2.420.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057–1063. MIT Press, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148. PMLR, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. doi: 10.1609/aaai.v29i1.9541. URL https://ojs.aaai.org/index.php/AAAI/article/view/9541.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard E. Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5015–5024. PMLR, 2018. URL https://proceedings.mlr.press/v80/tucker18a.html.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9659–9668. PMLR, 2020. URL https://proceedings.mlr.press/v119/uehara20a.html.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is Your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-graph policy optimization for long-horizon agentic reinforcement learning. *arXiv preprint arXiv:2606.22995*, 2026. URL https://arxiv.org/abs/2606.22995.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

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. URL https://arxiv.org/abs/2607.22724.

A Environments, models, and collection

Environments. ALFWorld (Shridhar et al., 2021) (text) with terminal reward 10 upon success, and ScienceWorld with a graded 0–100 score (Δscore/100 per step) — chosen specifically to contrast sparse against dense reward (Section 4.2). Both are text-only symbolic interactive environments with discrete action spaces, but they use distinct simulators.

Models. Qwen2.5-Instruct (Qwen Team, 2024) at 1.5 B and 7 B, served with vLLM (Kwon et al., 2023) at temperature 0.4, 512 max new tokens, using each environment's released structured <think>/<action> prompt.

Grouping structure. Each task instance (an ALFWorld game file, or a ScienceWorld task variation) contributes $m = 8$ rollouts, with m fixed throughout. Within a task, steps are first partitioned into the anchor groups $G = (q, o)$ defined in Section 3.1; the ICC sums are then pooled over eligible anchors. Episodes run to at most 50 steps. Rollouts are dumped to a fixed 26-column schema (observation, normalized observation, observation hash, history hashes at $k = 0, \ldots, 4$, action, per-step reward, discounted and undiscounted return-to-go, success, episode length, validity, and the task/rollout group ids), so every diagnostic is a pure function of the dumped data and is recomputable without re-collection.

Table 3: Fixed parameters for the reported Measurement analyses. Resampling counts refer to refits of the complete estimator, not to independent policies. History-depth analyses are stratified sensitivity checks and do not define a monotone trend test.

component	fixed analysis choice
anchor / group	exact (task ID, normalized current observation); rollout group size 8
primary value	discounted return-to-go, $\gamma = 0.95$; dense progress is labelled diagnostic
history label	one-step history hash ($k = 1$) for the primary decomposition; $k = 0, \ldots, 4$ reported only as depth-stratified sensitivity
eligible anchor	at least two history labels, positive within-label degrees of freedom, and at least four rows
estimator / diagnostic	pooled unbalanced ICC(1), with the between component truncated at zero
uncertainty / reference	whole-task bootstrap; fixed-count within-anchor label shuffle (descriptive, not cluster-valid)
sample-size stress test	$n \in \{64, 128, 256, 400, 600, 800\}$ trajectories; four fixed subsamples per size; 100 permutations per subsample
matched-policy check	$n = 800$ for each policy; 24 fixed-data subsample draws from the larger policy (descriptive, not policy replicates); one-sided empirical comparison
data trust	observation-by-observation replay; complete success census and stratified failure replay; failed fidelity blocks are ineligible
estimator comparison	identical rows and return target; grouping-only and full-value operators reported separately

Table 4: Fixed protocol for the two matched-row scope checks. Neither check admits a post-update outcome.

check	fixed construction	admitted inference
independent producer	discounted return-to-go with $\gamma = 0.95$, success reward 10, invalid-action penalty 0.1, complete matched groups	exact key and advantage-vector relations on one independently emitted batch
Yang fallback	exact-string first-visit progress, $\tau_r = 10^{-3}$, $\tau_p = 10^{-4}$, auxiliary coefficient 0.3, $\epsilon = 10^{-6}$, population-standard- deviation normalization	grouping inheritance and fallback activation on fixed rows

B INDEPENDENT PRODUCER AND RELATED-OPERATOR PROTOCOL

The two scope checks in Section 4.6 preserve the paper's matched-row counterfactual: every compared implementation consumes the same rows, and only the executable operator changes. Their unit is a fixed row batch, not an episode, checkpoint, or independently initialized training run. Consequently, these checks can test structural replication but cannot support a task-performance or learning claim.

Independent producer replication. The estimator implementations consume a newly emitted row stream from a producer separate from the historical offline collector. We first compare the complete task-and-normalized- observation key for every row, then compare the grouping-only and full-estimator advantage vectors. Exact grouping-only equality survives, whereas restoring each estimator's value machinery separates the full operators. Because the rows are fixed throughout, this result isolates producer sensitivity from operator design without measuring optimization.

Yang progress-fallback check. The project-side Yang et al. implementation was audited against the stated formula and the exact first-visit signal. Its inherited grouping component and deployed all-failed fallback were executed separately on matched rows. The former preserves the grouping identity; the latter activates roughly one third of eligible anchors whose terminal-return channel is

silent. This is evidence about where the fallback changes a computed advantage, not about the return of a policy trained with that fallback.

Admission boundary. Fan–Liu is excluded because its action-boundary hidden-state and normalization conventions are not independently resolved in the available implementation. Dense Signal Augmentation, training arms, policy-update trajectories, and post-update outcomes are likewise outside this paper; they belong to the concurrent algorithm submission described in the introduction. No generated training-result table or figure is imported into this manuscript.

C Proof and boundary of the finite-sample moment identity

Proposition 1 states the finite-sample component guarantee in the main text. The following calculation proves it and then records the narrower conditions needed for an exact variance expression. Neither result asserts that the reported ICC ratio is unbiased.

Proof of Proposition 1. Within label j, subtracting the label mean removes B_{gj}, so $\mathbb{E}[\mathrm{SSW}_g] = (n_g - J_g)\sigma_W^2$. Write $X_{gj} = B_{gj} + \bar{\varepsilon}_{gj}$. Independence gives $\mathrm{Var}(X_{gj}) = \sigma_B^2 + \sigma_W^2/n_{gj}$. Expanding the weighted sum of squares around its weighted grand mean yields

$$\begin{aligned}\mathbb{E}[\mathrm{SSB}_g] &= \sum_j n_{gj}\,\mathrm{Var}(X_{gj}) - n_g\,\mathrm{Var}\left(\sum_j \frac{n_{gj}}{n_g} X_{gj}\right) \\ &= (J_g - 1)\sigma_W^2 + \left(n_g - \frac{\sum_j n_{gj}^2}{n_g}\right)\sigma_B^2 \\ &= d_{B,g}\big(\sigma_W^2 + n_{0,g}\sigma_B^2\big).\end{aligned}$$

Summing these expectations over anchors and dividing by the pooled degrees of freedom gives the two displayed moment equations because $\bar{n}_0 = \sum_g d_{B,g} n_{0,g} / \sum_g d_{B,g}$. □

Under the additional assumption that the random effects are Gaussian and anchors are independent, the same calculation also gives an exact conditional variance for the untruncated component. Let $A_g = \mathrm{diag}(\boldsymbol{n}_g) - \boldsymbol{n}_g \boldsymbol{n}_g^\top / n_g$ and $\Sigma_g = \mathrm{diag}(\sigma_B^2 + \sigma_W^2/n_{g1}, \ldots, \sigma_B^2 + \sigma_W^2/n_{gJ_g})$. Gaussian quadratic-form identities and independence of label means from within-label residuals yield

$$\mathrm{Var}(\mathrm{MSB}) = \frac{2}{d_B^2}\sum_g \mathrm{tr}\left[(A_g\Sigma_g)^2\right], \qquad \mathrm{Var}(\mathrm{MSW}) = \frac{2\sigma_W^4}{d_W},$$

$$\mathrm{Var}(\widetilde{\sigma}_B^2) = \frac{\mathrm{Var}(\mathrm{MSB}) + \mathrm{Var}(\mathrm{MSW})}{\bar{n}_0^2}, \qquad \mathrm{Cov}(\mathrm{MSB}, \mathrm{MSW}) = 0,$$

where $d_B = \sum_g d_{B,g}$ and $d_W = \sum_g d_{W,g}$. The trace term retains the full unbalanced size pattern: limited degrees of freedom and concentrated label sizes can therefore increase uncertainty even when total row count is unchanged. A separately frozen deterministic simulation verifies these expressions. We do not use the formula as a confidence interval for the agent data because Gaussianity, anchor independence, and non-informative sizes are not established there.

The implementation reports $\hat{\sigma}_B^2 = \max(0, \widetilde{\sigma}_B^2)$ and then forms a ratio. Truncation creates positive boundary bias when $\sigma_B^2 = 0$, and a ratio of random quantities is not generally unbiased. Informative group sizes, heterogeneous components, repeated trajectory dependence, and selection into eligible anchors also fall outside the proposition. A deterministic simulation check reproduces the component moment identity and the truncation bias, but it is a code sanity check rather than agent data. These boundaries motivate the empirical sample-size ladder and whole-task bootstrap. The within-anchor label shuffle is retained only as a descriptive reference because it does not preserve the complete trajectory/task clustering.

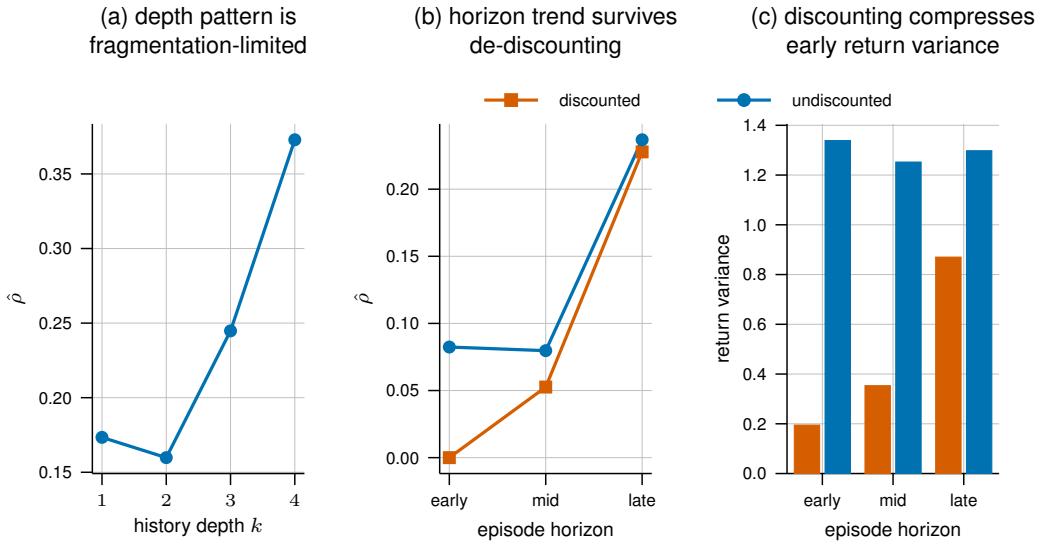

Figure 3: Where the history signal lives, and a control for discounting. (a) The history-depth pattern is nonmonotone and limited by subgroup fragmentation and data density; it does not support a monotonic depth conclusion. (b) The history-associated component becomes more visible later in an episode. Discounting makes the early bucket appear silent, whereas removing discounting reveals a small early component and preserves the late rise. The associated label shuffle is descriptive rather than cluster-valid inference. (c) Terminal-reward discounting suppresses early return variance far more strongly; without discounting, variance remains comparatively stable across the horizon.

D THE SAMPLE-SIZE STABILITY HEURISTIC (SETUP AND RESULT)

ICC(1) is sample-size sensitive in this regime. Along the fixed-data trajectory-subsampling ladder, the estimate decreases as more of the recorded dataset is retained, while both the label-shuffle reference and across-subsample spread contract. This exercise does not keep complete task blocks intact, has only four subsamples per size, and does not hold the realized task/group composition fixed. We therefore use $\mathbf{n \geq 400}$ **trajectories** only as a dataset-specific reporting heuristic, match sample sizes when possible, and call $\hat{\rho} - \mathbb{E}_{\text{shuffle}}[\hat{\rho}]$ *label-shuffle-centered*, not bias-corrected. Appendix H gives the full ladder and a separate-dataset check. Neither the ladder nor the non-cluster-preserving shuffle estimates coverage or defines a generally valid hypothesis test.

E HOW THE THREE METHODS DESCRIBE THEIR OWN GROUPING

Why this needed measuring. Each method foregrounds grouping, but the released definitions mix a shared anchor partition with additional structure. GiGPO gathers repeated anchor states; HGPO adds increasingly long historical contexts and a weighted aggregation; G2PO builds a graph over identical anchor observations. Reading the implementations reveals the narrower exact identity: GiGPO and G2PO share the flat partition, while HGPO retains it as a base level and adds refinements. What we do **not** claim is that anyone was misled about an experimental result: we did not re-run published training, and the primary sources may attribute gains to aggregation or optimization choices rather than to the flat partition alone.

F DISAGREEMENT MASS: FULL TABLE

Where the disagreement sits, not just how large it is. A cosine says how aligned two advantage vectors are; it does not identify where their entries differ. We therefore also report the share of GiGPO's **squared advantage norm** carried by rows where the other estimator materially disagrees

Table 5: Where the disagreement between estimators sits. The last column is the share of GiGPO's squared advantage norm carried by the steps on which the other estimator materially disagrees.

comparison	cos	steps disagreeing	**share of GiGPO's squared norm there**
GiGPO vs grouping-only G2PO	1.000	0 %	**0 %**
GiGPO vs real HGPO (1.5B)	0.965	0.35 %	**3.89 %**
GiGPO vs real HGPO (7B)	0.944	1.43 %	**9.49 %**
GiGPO vs real HGPO (ScienceWorld)	0.937	11.79 %	**11.08 %**
GiGPO vs full G2PO (1.5B)	0.187	84.7 %	**86.1 %**
GiGPO vs full G2PO (7B)	0.272	82.8 %	**79.5 %**
GiGPO vs full G2PO (ScienceWorld)	0.183	86.4 %	**80.9 %**

($|A_1 - A_2| > \frac{1}{2}|A_1|$). This remains an advantage-vector diagnostic, not actual policy-gradient geometry:

This table supports a controlled attribution, not a universal statement about the full methods. Grouping-only G2PO equals GiGPO on the matched rows, whereas disagreements with the full implementation cover rows carrying most of GiGPO's squared advantage norm. Because the two G2PO variants share their partition, the additional next-group score, invalid-action penalty, and score-gain machinery causes that within-implementation gap. The statistic does not identify actual gradient contribution or training impact.

G HGPO REDUCES TO THE FLAT GROUP: FULL ARGUMENT

HGPO contains the flat group and adds refinements. Its hierarchy starts with the current observation alone and then adds history-conditioned subgroups. Singleton subgroups are skipped and weights renormalize over surviving levels, so fragmentation shifts weight back toward the flat term. Across the two primary policy datasets and ScienceWorld, the full HGPO advantage remains highly aligned with GiGPO and never zeroes a nonzero GiGPO entry. This establishes snapshot alignment under released settings; it does not make the HGPO partition identical to GiGPO or prove equivalence during training.

Where the controlled G2PO variants differ, the partition is held fixed. Full G2PO has low cosine similarity with GiGPO across the audited datasets, whereas grouping-only G2PO is identical to it. The controlled implementation changes the value machinery while retaining the partition; it therefore localizes this observed gap without claiming that every difference among all three full methods is a value-only effect.

A result we retract, because it is instructive. An earlier version of this analysis incorrectly reported that history-conditioned grouping discards a substantial share of the gradient. That comparison used a *pure* depth-k grouping—history label only, with no flat term—which no published method uses; faithful HGPO discards none of the nonzero GiGPO signal on these rows. The error had two causes in our code: an index-origin mismatch that dropped HGPO's flat term and an equivalence test that covered only a helper. Both are fixed and now guarded by an end-to-end equivalence test (Section 3.3). We retain the pure-depth curves in Figure 4 as an explicitly labelled extreme design point, because they show how a genuinely history-only method would differ from faithful HGPO.

Takeaway. A shared flat anchor partition can coexist with different history refinements and value functionals. Any performance attribution must isolate these components with training evidence; the present snapshot audit does not identify why one full method trains better.

H FINITE-SAMPLE STABILITY STRESS TEST FOR THE ICC ESTIMATOR

Finite-sample behavior materially affects the raw diagnostic in the studied regime. Along the fixed-data ladder, retaining more trajectories lowers the estimate, contracts the label-shuffle reference, and reduces across-subsample dispersion; an auxiliary snapshot shows the same qualitative instability (Figure 2). We therefore avoid raw unequal-size comparisons and use $\geq$**400** only as a reporting

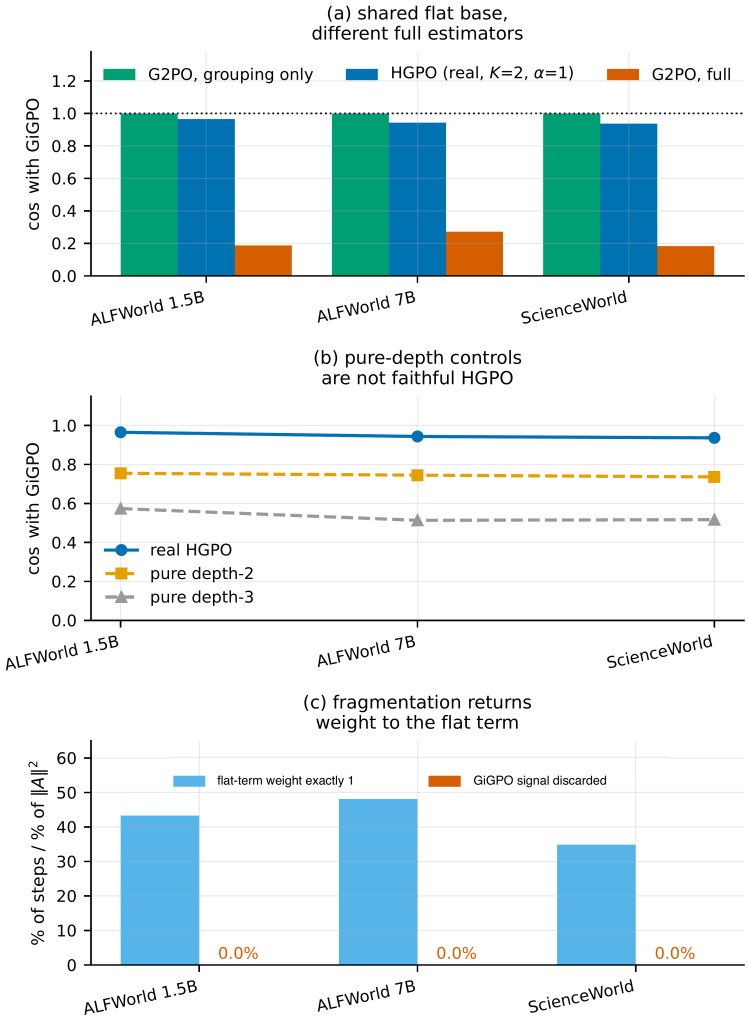

Figure 4: What the estimators actually differ in. Cosines compare per-step advantage vectors, not policy gradients; without the score-function Gram matrix, no gradient-cosine approximation is asserted. (a) G2PO's grouping-only variant is identical to GiGPO, faithful HGPO remains closely aligned, and full G2PO changes sharply because of its value machinery. (b) Pure depth-k grouping is an extreme design point, not a published-method comparison; faithful HGPO discards none of the nonzero GiGPO signal. (c) HGPO skips singleton clusters and renormalises its surviving weights, so fragmentation returns weight to the flat term.

heuristic for these data. Because the ladder contains few subsamples per size and samples trajectories rather than intact task blocks, it neither estimates coverage nor establishes a universal cutoff. Likewise, the within-anchor shuffle breaks joint dependence; its centered statistic and tail areas are descriptive sensitivity summaries, not a calibrated permutation test.

I SCIENCEWORLD RHO MAGNITUDE: WHY IT IS NOT PAPER-USABLE

ScienceWorld's *ρ-magnitude* is not paper-usable because invalid actions produce a dominant stereotyped no-op observation, and removing those steps collapses the estimate. A prespecified prompt repair did not improve parsing (Section 6). The retained claim is narrower: dense reward leaves many groups informative even after every no-op step is removed. The table below reports the exact support and robustness values.

Table 6: Sparse versus dense reward; plotted in Figure 5. The informative-group fraction measures effective return-channel support, not a statistical effective sample size. The $\hat{\rho}$-magnitude row is struck through because it fails the robustness check described in this appendix.

quantity	ALFWorld (terminal)	ScienceWorld (graded)
non-zero returns	1.31 %	**30.97 %**
informative groups	4.58 %	**72.57 %**
informative = successful?	90/90 exact	decoupled (46 succ, 418 var)
~~$\hat{\rho}$ magnitude~~	~~0.173~~	~~0.154 → 0.026 w/o no-ops~~

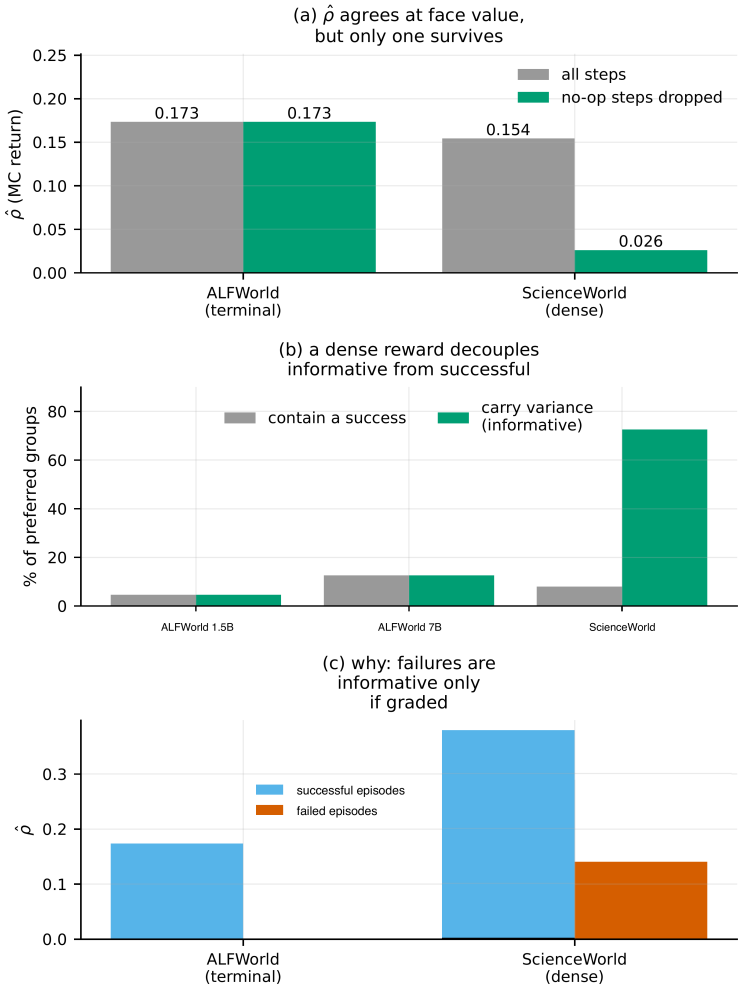

Figure 5: Robustness and support diagnostics under terminal and graded reward. Bars reproduce the aggregate quantities in Table 6; the ScienceWorld magnitude comparison is withdrawn.

Figure 5 separates the failed magnitude comparison from the retained measurability result. Panel (a) shows that the apparent ScienceWorld $\hat{\rho}$ does not survive no-op removal. Panel (b) shows instead that dense reward preserves many informative groups, while panel (c) explains the contrast: failed episodes can remain return-informative under graded reward but are return-constant under terminal reward. Thus the figure supports a statement about available signal, not cross-environment magnitude replication.

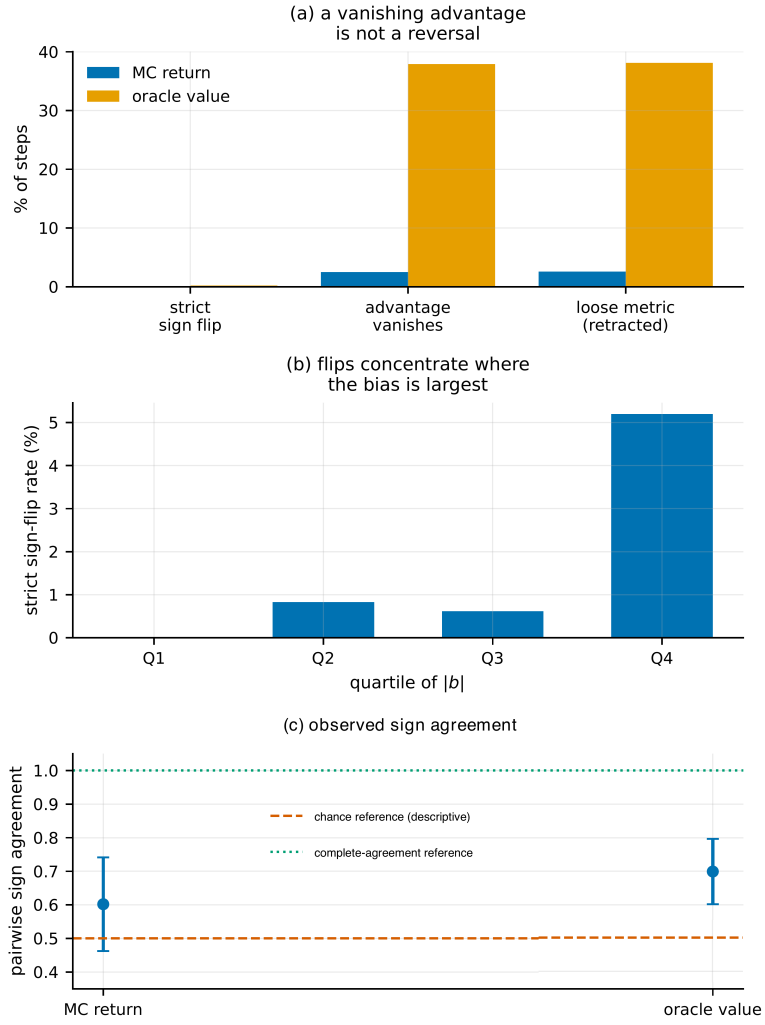

Figure 6: The practical footprint of ignoring history. Here $b_i = A_i^{\mathrm{GiGPO}} - A_i^{\mathrm{HGPO}}$; $|b|$ denotes the mean absolute matched-row difference, and $\sigma(A)$ the GiGPO-vector standard deviation. (a) Strict sign reversals are rare; the larger "loose" rate also counts an advantage vanishing. (b) Changes concentrate among rows with the largest $|b_i|$. (c) Bars are nominal binomial intervals over elementary pairs; dependence makes them descriptive rather than calibrated confidence intervals. Oracle-channel marks use the replay-flagged dataset and are indicative only.

J Practical impact and negative results, in full

Ignoring history has a small footprint under the reported snapshot diagnostics. On the 1.5B-policy Monte-Carlo return, the normalized mean absolute matched-row difference is **0.072**, and the strict sign-reversal rate is **0.09%**. Together with the GiGPO/HGPO advantage cosine, this describes per-step advantage vectors at the verified snapshot; it is not a measured policy update.

The higher-success matched snapshot has no larger $\hat{\rho}$ estimate. After matching trajectory counts, the higher-success policy has a slightly smaller diagnostic. Its repeated-subsample mean and standard deviation, together with the descriptive comparison tail area, are shown in Figure 2; they summarize sensitivity within recorded data rather than variation across trained policies. ScienceWorld's $\hat{\rho}$ magnitude is withdrawn and is not part of this comparison.

History-associated structure is visible but its inferential resolution is weak. Episode-stratified comparisons show a larger diagnostic in one return stratum, while the history-depth pattern remains

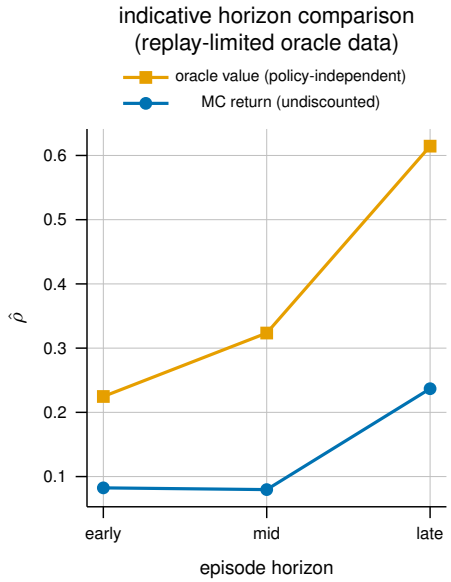

Figure 7: Indicative horizon sanity check on two value channels. The oracle value is planner-derived and policy-independent; the Monte-Carlo return comes from the evaluated policy. Both channels rise from early to late at comparable multiplicative rates, but the oracle dataset later failed replay trust, so the comparison is supporting context rather than primary evidence.

nonmonotone and fragmentation-limited. Structural-bias signs are only modestly aligned with the Monte-Carlo return, and dependence among elementary pairs precludes a definitive sign test. The stronger oracle-channel alignment comes from the replay-flagged random-policy dataset. These observations do not establish that a bias accumulates across updates or rises above chance under cluster-valid inference.

Summary of the negative results. HGPO's added history levels make only a modest advantage-vector difference at the verified snapshots; the higher-success snapshot does not have a larger matched-n ρ estimate; and dependent structural-bias pairs provide no cluster-valid comparison with chance. These observations do not make history irrelevant in principle and do not predict training behaviour.

K THE DIAL CONFOUND, IN FULL

A confound intrinsic to the withdrawn dial. The ϵ-oracle dial changes success labels, episode length, and state visitation together, and its replay is not faithful. Consequently every dial-derived success, informative-group, $\hat{\rho}$, cosine, and headroom result is withdrawn. The historical design remains useful only as a warning that any success sweep must control data density and episode length; it supports no current measurement conclusion.

L ADDITIONAL LIMITATIONS, IN FULL

Two text environments. ALFWorld and ScienceWorld are distinct text-only symbolic simulators with discrete actions and short episodes. Observation aliasing — the phenomenon ρ measures — is likely to behave differently where observations are images or long documents. Construct validity (Section 4.2) was established by *injected* aliasing, which does not establish that naturally occurring aliasing has the same character.

ScienceWorld agent quality. Invalid actions are much more common in ScienceWorld than in ALFWorld and make a single stereotyped no-op dominate the recorded steps. Consequently, the $\hat{\rho}$-*magnitude* replication is withdrawn. The prespecified prompt-engineering attempt did not improve

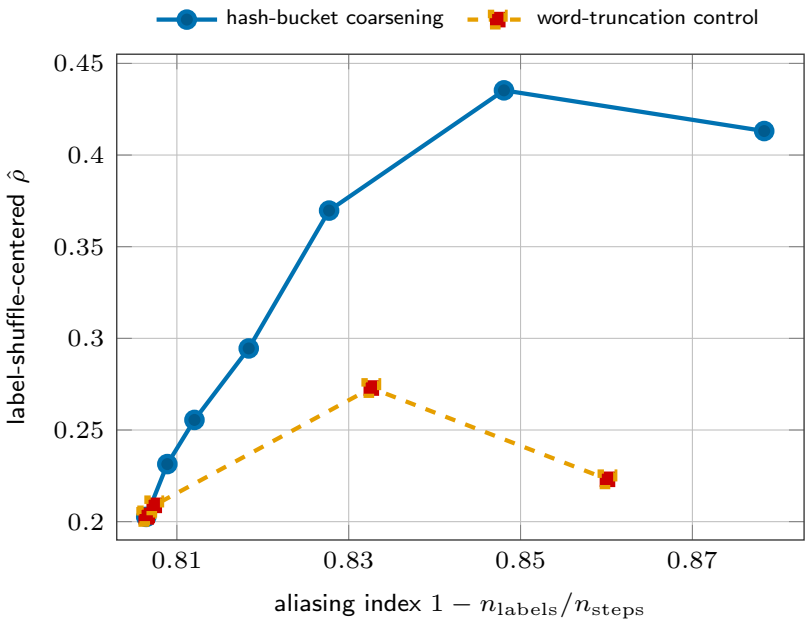

Figure 8: Injected-aliasing sanity check. Coarsening recomputes both observation and history keys; the label-shuffle-centered diagnostic rises over most levels and exhibits strong rank monotonicity. Because the levels are nested deterministic transforms of one dataset, the exact rank enumeration is descriptive rather than calibrated inference. The most extreme level turns over when both keys collide. This replay-flagged oracle dataset supports the construct only indicatively.

the parse rate, and the exclusion rule was applied without adjustment. The ScienceWorld results we retain—dense-reward measurability, estimator alignment, and bias cancellation—were each re-verified after removing every no-op step.

$\hat{\rho}$ sample-size sensitivity. Raw ICC(1) varies strongly along the fixed-data ladder. The $n \geq 400$ rule is only a reporting heuristic from one ALFWorld dataset at $m = 8$ rollouts per task, not a calibrated validity threshold.

Advantage cosine $\neq$ gradient cosine. Let $s_i = \nabla_\theta \log \pi_\theta(a_i \mid h_i)$ and let $\mathbf{K}_{ij} = s_i^\top s_j$ be the score-function Gram matrix. For advantage vectors A, A', the corresponding gradient cosine is $A^\top \mathbf{K} A' / \sqrt{(A^\top \mathbf{K} A)(A'^\top \mathbf{K} A')}$. It reduces to the ordinary vector cosine only when $\mathbf{K}$ acts like a common scalar multiple of the identity on the relevant subspace—a condition that requires both negligible cross-row products and comparable score norms. Baseline manipulations need not preserve gradient properties automatically (Tucker et al., 2018). Because per-token scores were not stored, the gradient cosine is unmeasured and no near-orthogonality approximation is asserted.

Depth trend is fragmentation-limited. $\hat{\rho}$ appears to rise with history depth on one dataset but not the others. At the deepest evaluated grouping, a large majority of steps are singletons in two datasets, so we cannot separate "deep history stops mattering" from "we ran out of data." We therefore report fragmentation rather than claiming a depth trend.

Implication for follow-up studies. A useful evaluation order is to verify variance-bearing support within the intended partition, compare value constructions on the same recorded rows, and only then test learning effects across independently initialized runs. This sequence prevents a silent return channel from being mistaken for evidence that history is irrelevant, while keeping measurement, mechanism, and downstream-performance claims separately falsifiable.