FROM REPLAY TO NATURAL ARRIVAL: IDENTIFYING AGENT ACTION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Replay enables researchers to evaluate agent action repairs by reconstructing a trajectory and comparing a replacement action with the agent's own continuation. Whether this comparison captures the repair's benefit during natural execution, however, depends on what reconstruction preserves and what the experiment observes. We study when replay evidence identifies natural-arrival action effects. For an observation setting containing natural repair outcomes and replay repair/control outcomes, but no natural policy-control outcomes, we derive a sharp bias interval of width $1 - \alpha_N$ under an action–response factorization assumption, where α_N is the natural probability of selecting the repair. Additional replay samples cannot resolve this missing-control uncertainty. We introduce Replay Transport Audit (RTA), which adds independent live-control measurements to identify the bias under stated within-context sampling conditions, without requiring that factorization. In ALFWorld, complete replay agrees with natural-arrival effects within an absolute panel-level tolerance of 0.10. In WebShop, removing a backend option shifts the repair-versus-alternative effect by -0.165 on a $[0, 1]$ score scale despite identical visible inputs. In a separate calibrated-pool training study, live-audited corrections improve success over RL-only by 3.47 and 14.35 percentage points on Gemma4-31B and Qwen3.6-27B, respectively. These findings clarify which action-effect claims replay supports and which require additional live observations or assumptions about reconstruction.

1 INTRODUCTION

To test whether an action repair helps a language agent, a researcher can reconstruct a recorded trajectory, replace one action, and compare subsequent outcomes with those obtained by letting the policy act. Executing these alternatives in the environment makes their consequences observable (Shah, 2026; Bonagiri et al., 2026; Ma et al., 2026). Such comparisons inform agent evaluation, counterfactual repair and action-level credit. Their interpretation depends on whether the reconstructed experiment represents the agent's natural execution.

A repair's benefit can change when reconstruction changes the context. At *natural arrival*, the policy reaches the candidate during an uninterrupted rollout. *Replay* instead creates a context by executing a recorded prefix in a fresh environment. The visible observation may match while policy history, hidden state or execution conditions differ. Figure 1 illustrates a recorded WebShop manipulation: a selected product option is removed from the backend while the page and policy input remain unchanged. A purchase then acts on different underlying state.

The quantity that matters is the repair's benefit relative to the policy's own action. Reconstruction can change either response, so observing a matching page or a matching forced-repair outcome does not establish agreement in their difference. A replay-based credit estimate may accurately describe the reconstructed experiment while leaving the benefit to a naturally arriving agent unresolved.

We characterize the information needed to interpret a replay contrast as a natural-arrival effect. Using established statistical tools, we study an explicit three-arm setting: natural forced-repair and replay forced/control outcomes, without natural policy control. The missing response concerns the natural policy's alternatives. More replay cannot supply that information; live control identifies bias under stated within-context conditions.

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.

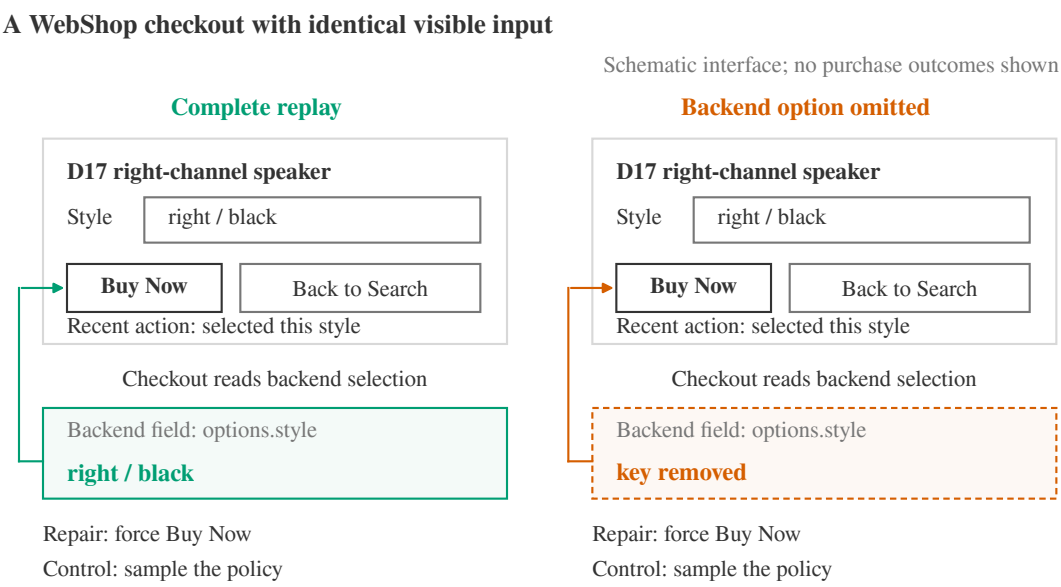


Figure 1: Identical visible input can conceal a different checkout state. The schematic abbreviates the recorded WebShop product page and prior style selection (goal 152). The controlled omission removes `options.style` while preserving the observation, action menu, prompt and tokens. Arrows show the backend selection read by `Buy Now`. Both repair and policy control are evaluated in each context. No purchase outcomes are depicted; Figure 3 reports the measured effects.

The experiments distinguish an identification gap from an observed mismatch. Our *Replay Transport Audit* (RTA) finds close panel-average agreement under complete ALFWorld replay despite wide three-arm identification intervals. A 12-cell controlled WebShop study isolates omissions that change action effects, including under identical visible inputs. Downstream correction training also improves over RL-only on two model families, with controls leaving the incremental value of live-effect weighting unresolved. The paper establishes three findings:

- **Replay precision cannot resolve missing-control uncertainty.** Under marginal factorization, the sharp interval has width $1 - \alpha_N$; even agreement in the action distribution and forced-repair response leaves alternative-action responses unresolved (Sections 2.2 and 2.3).
- **Live control makes replay bias directly measurable.** Four directly observed arm means identify the replay-minus-natural effect under within-context common laws, without marginal factorization. The audit makes this comparison operational (Section 2.4).
- **Reconstruction choices can change relative action benefit.** Complete reconstruction yields small measured average biases on the studied panels, while specified omissions shift WebShop effects and point repair decisions. Visible-input agreement alone misses some of these shifts (Sections 3.3 to 3.5).

2 From Replay Effects to Natural-Arrival Effects

An action effect compares interventions within one execution context. The transport question asks whether that difference is preserved in another. Answering it requires information about the responses in both contexts.

2.1 The target is a policy-replacement effect at natural arrival

The target is the benefit of a fixed repair at a specified candidate under two execution laws. Let $h_t = (o_0, a_0, \ldots, o_t)$ be the interaction history and let x identify the candidate, including the required

observable signature and policy history. Natural arrival occurs at its first exact match, $T_x < \infty$, during a fresh policy rollout. Law N describes the complete context at that live arrival; law R describes the context produced by executing the frozen discovery prefix in a fresh environment. Effects are conditional on arrival, so a candidate without sufficient natural support is reported as unsupported, not assigned an effect of zero.

Within either context, the forced arm executes a prespecified action $a^+(x)$, and policy control executes a draw from the agent's declared policy. The repair is the next executable action in a verified expert continuation; the continuation verifies the repair but is not shown to the policy. Let F_z and C_z be the corresponding population mean outcomes for context $z \in \{N, R\}$. With binary success in ALFWorld, or a bounded task score $Y \in [0, 1]$ where specified, define

$$\tau_z = F_z - C_z, \qquad b = \tau_R - \tau_N. \tag{1}$$

Thus τ_N is the benefit of replacing the naturally arriving policy's action, τ_R is the replay-local benefit, and positive transport bias b means that replay overstates that benefit.

The three claims use different conditions. The three-arm sharp set requires marginal factorization (Assumption A.1) and the specified observation pattern. Four-arm identification instead requires within-context common laws, consistency, noninterference and stable protocols (Assumption A.2; Proposition A.3). Forced and control arms must represent the same complete-context distribution *within each* z; natural and replay distributions need not coincide. Sufficient reconstruction of all relevant state and action-response mechanisms *across* N and R establishes zero bias (Proposition A.7). Matching visible signatures and serialized history does not establish that sufficiency (Proposition A.5). Complete context includes hidden environment and sampling-relevant policy state (Appendix A.1).

2.2 Three arms leave the natural control response unidentified

The three-arm setting separates what replay reveals from what natural execution reveals. Suppose we observe the repair's natural outcome F_N and both replay means F_R, C_R, together with natural action samples. For example, querying the policy but always executing the repair at live arrival records its action probability without observing its unmodified continuation. This analytical setting isolates the missing information; it is not a characterization of every replay workflow.

Existing logs can supply C_N when their unmodified continuations represent the same policy, arrival rule, complete-context law and continuation protocol. Logs selected by success, by the action taken, or from reconstructed continuations generally target a different mean unless an appropriate adjustment identifies the required law.

The missing quantity is the outcome of the actions the natural policy chooses instead of the repair. Write $p_z(a)$ for its action probability, $q_z(a) = \mathbb{E}[Y \mid z, \mathrm{do}(A = a)]$ for a forced-action response, and $\alpha_z = p_z(a^+)$. To connect these marginal quantities to policy control, Assumption A.1 fixes the policy input, separates the action draw from latent context and continuation noise, and aligns the within-context arm laws. This yields *marginal action–response factorization*:

$$C_z = \sum_a p_z(a) q_z(a) = \alpha_z F_z + (1 - \alpha_z) M_z,$$

where M_z is the average response after a non-repair action. Stochastic decoding alone does not establish this assumption.

Proposition 2.1 (Missing-control identification interval)**.** *Under the factorization assumption, suppose* $(\alpha_N, F_N, F_R, C_R)$ *are known and natural responses after non-repair actions are unrestricted in* $[0, 1]$*. The sharp identified set for* b *is*

$$[F_R - C_R - (1 - \alpha_N) F_N, \quad F_R - C_R + (1 - \alpha_N)(1 - F_N)]. \tag{2}$$

Its width is $1 - \alpha_N$*.*

The width follows from letting the unobserved mean M_N range from zero to one. Each value is compatible with the observed moments; the endpoints are attainable, so the interval cannot be tightened without more information (Proposition A.4). This is an identification limit even with infinite samples, not a confidence interval for estimated moments. When the repair is rarely chosen

naturally, uncertainty about the alternatives dominates. More replay samples improve estimates of F_R and C_R but do not observe M_N. The corresponding set for natural benefit contains zero, so the three arms alone do not certify a strictly positive τ_N.

The same argument extends to stochastic repairs. A repair distribution g reveals $g^\top q_N$, while natural control requires $p_N^\top q_N$; optimizing the latter over compatible bounded responses gives sharp endpoints (Proposition A.9). These use established bounded-outcome and linear-programming tools to isolate the information missing from the replay experiment. Finite-sample outer regions are given in Proposition A.10.

2.3 Matching selected diagnostics does not identify transport

Agreement on actions and on the forced repair still leaves the alternatives' responses unconstrained. Under the same factorization, even $p_N = p_R$ and $q_N(a^+) = q_R(a^+)$ permit

$$b \in [-(1-\alpha_N),\, 1-\alpha_N]. \tag{3}$$

Both endpoints arise by changing only non-repair responses (Proposition A.5). Even zero action TV leaves this ambiguity; a small action-TV diagnostic alone cannot certify a tighter bound without response restrictions.

Proposition A.8 supplies such a restriction, while retaining marginal factorization. For its finite action set and binary outcome, $\mathrm{TV}(p_N, p_R) \le \epsilon_p$ together with $\max_a |q_N(a) - q_R(a)| \le \epsilon_q$ gives $|b| \le \epsilon_p + 2\epsilon_q$. Applying this result requires a bound on population action-law discrepancy and a uniform response bound over *all* actions, including unmeasured natural responses to alternatives. Visible-input checks and the forced-repair response do not supply that bound. Exact sufficient reconstruction can instead establish $b = 0$ under Proposition A.7.

2.4 Live policy control completes the comparison

Natural policy control observes the response that the three-arm design lacks. Figure 2 shows how the four-arm audit obtains F_N, C_N, F_R, C_R and computes

$$b = (F_R - C_R) - (F_N - C_N). \tag{4}$$

Under Assumption A.2, this identifies bias even when control action probabilities depend on latent context: C_N is measured directly, rather than reconstructed from marginal probabilities and responses.

Live measurement preserves the reference context by intervening in the environment the policy has already reached. Candidate matching rules and repairs are fixed before intervention outcomes, and each arm uses separate attempts allocated before rollout. ALFWorld uses deterministic allocation, so equality of the two conditional arm laws remains an assumption (Appendix B.1). Table 5 distinguishes the implemented checks and potential design measures from the assumptions they cannot establish.

2.5 Estimation targets the supported population

The audit estimates an average over a declared supported population. For checkpoints $\mathcal{K}_p$ in panel p and retained cells $\mathcal{X}_k$ in checkpoint k, the primary ALFWorld weights are $w_x = (|\mathcal{K}_p||\mathcal{X}_k|)^{-1}$. We estimate each cell's four arm means and aggregate $\sum_x w_x \widehat{b}_x$. Population selection precedes forced/control intervention outcomes and requires a failed natural trajectory as well as sufficient arrival support.

Sampling uncertainty is assessed separately from identification uncertainty. The fixed-cell bootstrap resamples outcomes within arms while retaining cell identities and weights. A weighted-Hoeffding analysis supplies bounded-outcome intervals under independent usable draws and stable arm laws, conditional on observed arrival counts (Proposition A.11). Preregistered confidence families and post-hoc joint analyses are identified separately in the results. The declared $[-0.10, 0.10]$ equivalence margin is an absolute tolerance whose practical meaning depends on the effect scale.

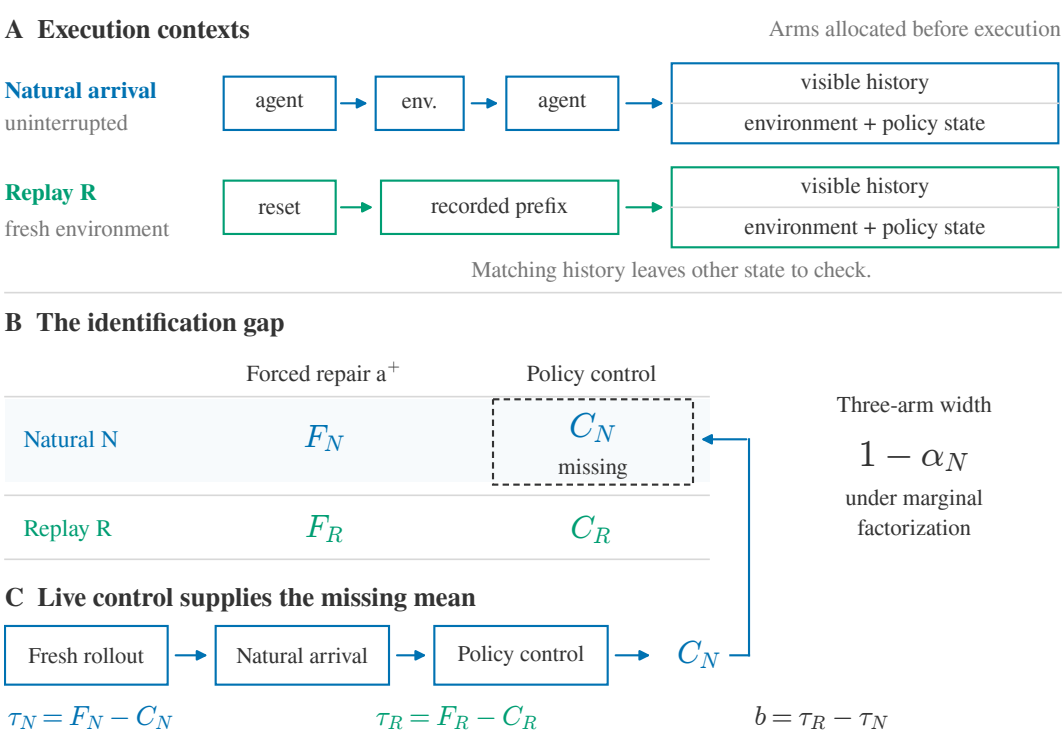


Figure 2: Execution contexts, missing information, and its measurement. (A) Natural arrival retains the live rollout; replay executes a recorded prefix in a fresh environment. (B) The three-arm setting lacks C_N; its displayed width requires marginal factorization. (C) Separately preallocated natural-control attempts supply that mean, continuing the live environment at intervention. No arrival is cloned or forked. Four-arm identification requires the conditions in Section 2.4; execution details are in Appendix B.1.

3 Empirical Diagnosis of Replay Effects

The information gap and the size of replay bias are different empirical questions. ALFWorld tests whether effects agree once the missing live control is observed; WebShop tests whether specific reconstruction omissions change that agreement. Together they show what a replay comparison can establish about its natural-arrival counterpart.

3.1 Natural support enables a direct live comparison

A live comparison requires candidates that the policy reaches often enough to observe both interventions. In ALFWorld (Shridhar et al., 2021), five of six step-8 Qwen3-8B checkpoints trained with GRPO or GiGPO provide 56 checkpoint–template cells. Qualification requires at least 16 exact-history arrivals, a failed arrival and a verified expert continuation, before forced/control outcomes are collected. The sixth checkpoint, GiGPO-201, has no qualifying population.

Qualification establishes where live comparison is supported. The 25,910 four-arm outcomes estimate effects conditional on reaching those candidates; fixed aggregation weights then define panel averages. Non-arrivals contribute cost but do not enter these means. Neither overall policy success nor a task-population average is estimated. The native execution kernel and allocation are documented in Appendix B.1.

Table 1: Primary ALFWorld panel effects: estimates with the same post-hoc six-estimand joint weighted-Hoeffding intervals as Figure 3c (at least 95% simultaneous coverage). The 33 GRPO and 23 GiGPO cells form fixed populations. The absolute equivalence tolerance for bias is ± 0.10.

Panel	Natural effect τ_N	Replay effect τ_R	Bias b
GRPO	0.1617	0.1672	$+0.0056$
	$[0.1257, 0.1976]$	$[0.1175, 0.2170]$	$[-0.0558, 0.0669]$
GiGPO	0.1426	0.1392	-0.0034
	$[0.0984, 0.1867]$	$[0.0794, 0.1990]$	$[-0.0777, 0.0709]$

3.2 The missing control leaves substantial ambiguity

The repair is rarely chosen naturally, leaving most of the natural control response unobserved in the three-arm setting. A post-hoc analysis places the lower bound on identification width $1 - \alpha_N$ above 0.80 in 53/56 cells, with at least 95% simultaneous coverage (Appendix C.1). Under factorization, the three observed arms therefore leave substantial uncertainty about natural benefit.

Low α_N makes the missing-control ambiguity consequential on this support-qualified population. It does not establish replay failure or extend the result to unsupported candidates; direct live control measures the bias.

3.3 Complete replay agrees with live effects at the panel level

Complete replay closely tracks the live effects in both ALFWorld panels (Table 1). The joint intervals show positive action effects and bias within the absolute ± 0.10 tolerance, consistent with the original preregistered analysis (Table 7). The wide three-arm sets therefore coexist with small measured average bias: live control supplies information needed to establish agreement.

Panel agreement allows local heterogeneity: effects vary across checkpoints, and 27/56 cells have zero observed successes in all four arms, which does not establish zero population effects (Table 8). Appendix C.1 consolidates the tolerance, relative-scale descriptions and tighter-margin sensitivity.

3.4 History and backend omissions shift WebShop action effects

The controlled WebShop study tests whether visible-input agreement suffices to preserve action effects. It varies policy history and hidden environment state while keeping the action comparison fixed. For 12 checkout-ready cells spanning 12 product categories, a Qwen3-4B policy is evaluated at live arrival, under complete replay, and under replay omitting history, a goal-relevant backend option, or both (Yao et al., 2022). The preregistered study contains 11,520 outcomes. Its primary outcome is bounded task score; exact success is secondary (Appendix B.2).

The action comparison asks whether `Buy Now` helps relative to the policy or to the forced alternative `Back to Search`. Writing these actions as g and h, the effects are $\tau_g = F^g - C$ and $\rho = F^g - F^h$. Complete replay has small estimated bias for both effects; its bootstrap bias intervals lie inside ± 0.05 within the families shown in Figure 3a,b. History omission instead increases the effects relative to live arrival by 0.5396 and 0.5198, respectively. An equal shift in all arm means would cancel in these contrasts.

A backend omission changes the comparison even when the policy sees the same input. Removing the selected option preserves the observation, action menu, prompt and token IDs in all 12 cells and passes the surface-plus-action-TV rule in 11/12 (Appendix B.2). Nevertheless, it shifts the repair-versus-alternative effect by -0.1651. This controlled intervention demonstrates that visible-input agreement can coexist with a change in relative action benefit. Appendix C.2 reports every reconstruction condition.

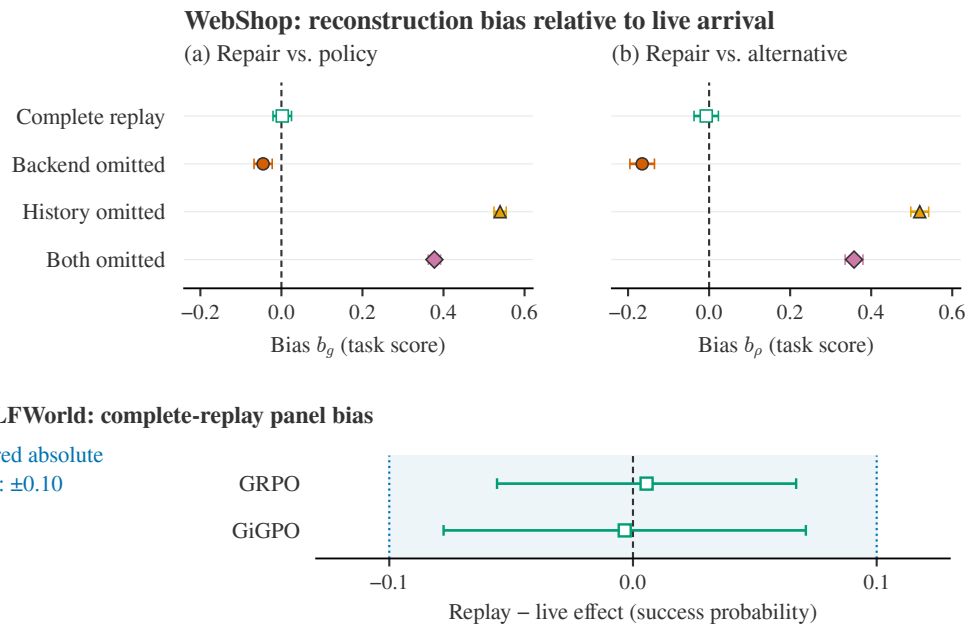


Figure 3: Complete replay and reconstruction omissions yield different action effects. (a,b) WebShop task-score bias relative to live arrival on 12 fixed checkout-ready cells: `Buy Now` versus policy control (b_g) or the forced `Back to Search` alternative (b_ρ). Zero denotes equal effects. Bars are 95% simultaneous fixed-cell bootstrap intervals within separate eight-estimand families: components for complete replay (open teal squares), recovery biases for omissions (filled markers). These families do not form one joint 95% statement. (c) ALFWorld panel-average bias with the post-hoc six-estimand, at-least-95% joint weighted-Hoeffding intervals. The blue region is the absolute ± 0.10 tolerance around the live reference. This panel uses a different scale and inference family. Numerical results and the original preregistered intervals are in Tables 1, 7 and 11.

3.5 Omissions change point decisions; the primary certificate abstains

The effect shifts change which repairs a point-estimate rule would select. At the fixed WebShop threshold 0.05, history-omitted replay adopts the repair in 8/12 cells whose live-effect estimates are below threshold; complete replay does so in none. History omission also changes the repair-versus-alternative threshold decision in 6/12 cells and reverses strict effect signs in 2/12 (Table 12).

The primary certificate abstains on all 12 cells under its declared rule. It uses weighted-Hoeffding intervals with Bonferroni allocation: total error 0.025 for the effect family and 0.025 for the bias family. Even complete-replay bias intervals extend outside the ± 0.10 tolerance, and aggregate direct-live intervals cross the 0.05 benefit threshold. These conditions, rather than cell count alone, prevent certification (Appendix C.2 and Proposition A.13). The bootstrap families in Figure 3 are distinct; point-decision disagreements remain uncertified as errors.

3.6 Average agreement extends across the measured trained policies

Complete replay also agrees with live effects at the panel level across the three trained-policy groups on common-task support. For Qwen3-8B policies trained with GRPO, GiGPO or Agent-DAPO, all six ALFWorld/WebShop bias intervals lie inside ± 0.10 in a separate post-hoc analysis (Table 13). Each method has three training seeds; the single fixed checkpoint without additional task RL has wider intervals that do not establish agreement at this tolerance.

Local omission sensitivity persists despite this average agreement. History-minus-full-replay shifts are detected in 11/236 ALFWorld and 21/262 WebShop eligible cells in the separate 504-contrast cellwise family (Appendix B.3), while aggregate omission intervals include zero. The average pattern therefore does not describe every candidate. Nor does it isolate an optimizer effect: common

Table 2: Final success (%) after eight calibrated-pool RL updates. Each entry aggregates 432 episodes: 144 per seed across three training seeds. All five arms are retained; differences use unrounded rates. AULC results and paired-comparison intervals are in Tables 18 and 19.

Training arm	Gemma4-31B	Qwen3.6-27B
RL-only	39.35	38.66
Uniform	40.28	47.45
Conservative replay	40.97	40.74
Raw replay	43.06	57.41
Live RTA	42.82	53.01

tasks retain policy-specific states and repairs, interaction budgets differ, and every method-difference interval includes zero. This larger study examines other policies and supported states; it does not provide broader replication of the 12-cell controlled probe. Only three WebShop own-support cells qualify for backend and joint omissions here (Appendices B.3 and C.3).

3.7 Downstream training and control comparisons

Live-audited corrections improve final success over RL-only in the measured calibrated-pool setting (Table 2). The gains are 3.47 percentage points for Gemma4-31B and 14.35 for Qwen3.6-27B. Relative to uniform weighting, the gains are 2.55 and 5.56 points; relative to raw replay, the differences are -0.23 and -4.40. Thus the correction pipeline yields useful improvements over RL-only, while simpler correction schemes remain competitive.

The controls do not establish an added benefit from live-effect weighting. Both live-minus-uniform 95% intervals include zero. These paired Student-t intervals use three training seeds (two degrees of freedom), assume independent normal paired effects, and are unadjusted for multiple comparisons (Table 19). Raw replay differs in weight mass and coverage, so its higher point estimates also do not isolate weighting shape.

A separate broad-data SFT experiment shows gains from uniform imitation. Half-strength LoRA improves ALFWorld unseen-frame success by 8.56 and 15.74 points for Gemma and Qwen (Table 20). Its data and training conditions differ from the calibrated-pool RL study; the two are not a matched-budget comparison or evidence for an RTA-weighting advantage. The task frames had already informed development, so fresh training seeds do not make them untouched. WebShop transfer is inconsistent: corrected Qwen checkpoints have zero successes, with a post-hoc diagnostic showing an action-format failure (Table 21 and Appendix C.4.1). Retail and the other downstream controls are reported in Appendix C.4.

4 Related Work

Executed credit assignment and repair methods evaluate consequences within their declared intervention contexts (Bonagiri et al., 2026; Ma et al., 2026). Causal Agent Replay estimates outcome-distribution changes after reconstructing a prefix and resampling downstream decisions under the same stochastic policy (Shah, 2026). C3 estimates decision-level advantages in restored, fixed-history groups under a frozen continuation distribution (Chen et al., 2026); a fixed distribution does not mean a fixed realized trajectory. Our additional question is whether an action comparison in that reconstructed context identifies its counterpart conditional on natural arrival.

Decoder reconstruction addresses one mechanism that can change a comparison. Re-feeding measures counterfactual token-credit discrepancies between fresh prefill and decode-state resume, using a repeated-resume noise floor and a batch-invariant confirmation (Matteson, 2026). Its reference is decoder state; ours is an arrival-conditional agent action effect, which also depends on environment and external-session state. Model-switching evaluation changes the policy (Gonuguntla, 2026); recovery and snapshot branching address permitted continuation and diverse user responses (Zhang & Liu, 2026; Nakash et al., 2026). With the policy fixed, sufficiently restoring all relevant state and mechanisms establishes zero bias under Proposition A.7. Live control is therefore not universally required when justified transport assumptions already suffice.

Off-policy evaluation targets policy value from data collected by another policy; overlap or response restrictions supply information on unobserved outcomes (Oberst et al., 2020; Khan et al., 2024). Here the policy can be unchanged while reconstruction changes the context and action-response law. Transportability studies the invariances needed to move effects across contexts (Bareinboim & Pearl, 2013); our partial-identification result isolates the missing natural-control response in a specified observation pattern. The bounded-outcome and interval-decision tools are established (Manski, 2011; Stoye, 2012; Aradillas Fernández et al., 2025); their application here separates replay precision from natural-effect identification.

5 Discussion

An identification gap and a measured replay mismatch call for different conclusions. The gap describes what the observations leave unresolved; its width is not evidence that reconstruction introduces a large bias. More replay improves precision for observed replay quantities, but cannot recover the missing natural-control mean. Under the three-arm factorization assumption, low α_N leaves a wide population identified set unless further response restrictions apply. Finite-sample inference must also account for uncertainty in the observed moments (Proposition A.10).

These intervals support decisions at an explicit benefit threshold. For a chosen benefit threshold δ, adopt only if the entire valid natural-effect interval lies above δ, or reject if it lies below. An interval that contains the threshold certifies neither decision; live control or justified response restrictions may be needed (Proposition A.12). Once all four arms are observed under the within-context conditions, the same-data identity $\widehat{\tau}_R - \widehat{b} = \widehat{\tau}_N$ means that direct inference for τ_N avoids the unnecessary conservatism of subtracting dependent intervals. This identity provides no efficiency advantage; the cost and sequential analyses likewise demonstrate none (Table 10 and Appendix C.5).

The audit directs effort toward the information needed for the intended claim. Independent candidates broaden coverage; repetitions improve within-candidate precision. Estimating how often replay mismatches occur additionally requires a defined sampling frame of actual replay workflows.

5.1 Scope

The empirical conclusions concern fixed supported populations and the execution laws studied here. The 12-cell WebShop probe establishes sensitivity to specified omissions under one policy, not their prevalence in ordinary replay. The separate cross-method study broadens policy coverage but does not replicate that controlled probe. The arrival requirement limits coverage of rarely reached candidates; panel-level agreement at an absolute tolerance does not establish uniformly small state-level error.

Common-law validity and independent usable draws remain assumptions about the experiment, conditional on its support and observed arrival counts. The checks reported here do not establish them. Broader candidate coverage, response measurements for additional actions, and downstream comparisons that isolate the value of live information require further evidence.

6 Conclusion

The specified three-arm observation pattern leaves a quantifiable gap between replay evidence and natural-arrival action effects: under marginal action–response factorization, the sharp bias interval has width $1 - \alpha_N$. Direct natural policy control identifies transport bias under the stated within-context conditions, without requiring that factorization. Our experiments show panel-level agreement under complete reconstruction and action-effect shifts under controlled omissions. Together, these results guide which live observations or justified reconstruction assumptions are needed to support an action-effect claim in agent evaluation.

Ethics statement

The study evaluates action effects in controlled agent environments and collects no human-subject data. Its intended use is to make the scope and uncertainty of replay-based evidence explicit. Deployment decisions require evaluation under their own operating conditions.

Reproducibility statement

The project retains 20,884 ALFWorld reconstruction-control outcomes, 29,056 ALFWorld live-arrival records (26,496 primary attempts, including 25,910 usable outcomes, plus 2,560 history-stress outcomes), and 11,520 WebShop omission-study outcomes. Locked support pools, seed registries, preregistrations and analysis scripts specify the frozen analyses. Appendices B and A give the protocols and proofs; Appendix C.4 records every retained downstream arm and seed, including the native `ms-swift` configurations. The manuscript builds independently from saved figures and tables. The figure package provides editable SVGs, vector PDFs, a plotting script, exact plotted data and source provenance.

Use of Large Language Models

Generative AI tools assisted with manuscript revision and figure preparation. GPT-5.6-Sol was also used to evaluate language assertions in the Retail experiment; Appendix C.4 describes the judging protocol and its limitations. The authors are responsible for the final manuscript, experimental results, and conclusions.

References

Andrés Aradillas Fernández, José Luis Montiel Olea, Chen Qiu, Jörg Stoye, and Serdil Tinda. Robust Bayes treatment choice with partial identification. *arXiv preprint arXiv:2408.11621*, 2025. URL https://arxiv.org/abs/2408.11621.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013. doi: 10.1515/jci-2012-0004. arXiv:1312.7485.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures. *arXiv preprint arXiv:2605.25338*, 2026.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026. Also titled Contextual Counterfactual Credit Assignment (C3).

Ashritha Gonuguntla. The replay gap: Static evaluation of model switching in LLM agents scores the wrong world. *arXiv preprint arXiv:2608.08239*, 2026.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757, 2024. URL https://proceedings.mlr.press/v235/khan24b.html.

Ming Ma, Jue Zhang, Fangkai Yang, Yu Kang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. DoVer: Intervention-driven auto debugging for LLM multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2512.06749.

Charles F. Manski. Choosing treatment policies under ambiguity. *Annual Review of Economics*, 3: 25–49, 2011. doi: 10.1146/annurev-economics-061109-080359.

Nils Matteson. Re-feeding is not replaying: Measuring replay noise in counterfactual token-credit estimation. *arXiv preprint arXiv:2606.15621*, 2026.

Itay Nakash, George Kour, and Ateret Anaby Tavor. Efficient agent evaluation via diversity-guided user simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1627–1648. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.112. URL https://aclanthology.org/2026.acl-industry.112/.

Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 788–798, 2020. URL https://proceedings.mlr.press/v108/oberst20a.html.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

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.

Jörg Stoye. Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156, 2012. doi: 10.1016/j.jeconom.2011.06.012.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

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

Zhihui Zhang and Wei Liu. Recoverability as a system primitive for long-horizon AI agents. *arXiv preprint arXiv:2609.13672*, 2026. URL https://arxiv.org/abs/2609.13672.

A Complete Proofs and Essential Extensions

This appendix gives the assumptions and complete proofs used in the paper. Bounded concentration, linear programming and interval propagation are standard tools; the identification results specify what they imply for the declared replay-to-natural observation problem.

A.1 Context laws and the two assumption sets

For a fixed candidate x, let $U_z \sim P_{z,x}$ contain the environment, policy memory and sampling-relevant execution state under context law z. Natural law N is conditional on first exact arrival in a fresh policy rollout; replay law R is induced by reconstructing the frozen prefix. With Y^{F_z} and Y^{C_z} the outcomes of the complete forced and policy-control protocols, define

$$F_z(x) = \int \mathbb{E}[Y^{F_z} \mid U_z = u] P_{z,x}(du), \quad C_z(x) = \int \mathbb{E}[Y^{C_z} \mid U_z = u] P_{z,x}(du). \tag{5}$$

The control propensity may depend on the actual policy input within u. Marginal action probabilities and marginal forced responses factorize only under the following additional condition.

Assumption A.1 (Marginal action–response factorization)**.** For a fixed candidate and z, the declared policy input is constant across reconstructed latent contexts. The policy action draw is randomized independently of latent context and suffix noise, and forced and control arms sample the same context law. Thus p_z is the experimental action distribution and $q_z(a)$ is the mean outcome under its

randomized action intervention. The empirical protocol operationalizes this condition with distinct deterministic action and suffix seed transforms; sharing replicate indices across arms is a variance-reduction coupling, not a proof of the assumption.

Assumption A.2 (Direct-arm common-law validity)**.** For every fixed cell x and context label $z \in \{N, R\}$, forced and control attempts are assigned before their outcomes are observed and sample the same declared law $P_{z,x}$ of complete pre-intervention contexts. The observed outcome is consistent with the executed arm protocol, attempts do not interfere with one another, and the forced and control continuation kernels remain the declared kernels throughout collection. The forced protocol executes the frozen $a^+(x)$; the control protocol samples from the declared policy using the input actually constructed from its context. Deterministic preallocation is permitted, but then equality of the two arm-specific context laws is an assumption rather than a consequence of randomized assignment.

Proposition A.3 (Direct-arm completion without marginal factorization)**.** *Under Assumption A.2, define F_z and C_z as the population means of the complete forced and control protocols. Then the context-specific policy-replacement effect is $\tau_z = F_z - C_z$, and four arm means point-identify*

$$b = \tau_R - \tau_N = (F_R - C_R) - (F_N - C_N). \tag{6}$$

This conclusion does not require the control action to be independent of U_z, nor $C_z = \sum_a p_z(a)q_z(a)$ for marginal p_z and q_z.

Proof. Common-law arm samples and consistency identify, for each z, the two means under the forced and policy-control interventions on the same population of contexts. Their difference is τ_z by definition; subtracting the two context-specific effects gives the displayed identity. If policy action propensities vary with U_z, the control mean is instead the latent-context average $\mathbb{E}_{U_z} \sum_a \pi(a \mid U_z)q_z(a, U_z)$, which is directly observed by the control arm even though the product of marginal propensity and marginal response need not equal it. □

A.2 The missing control and the limits of fidelity proxies

Proposition A.4 (Sharp missing-control frontier)**.** *Under Assumption A.1, let $F_z = q_z(a^+)$, $C_z = \sum_a p_z(a)q_z(a)$, and $\alpha_N = p_N(a^+)$. Suppose the population moments $(\alpha_N, F_N, F_R, C_R)$ are observed, $Y \in [0, 1]$, and no restriction is placed on natural outcomes after non-repair actions. Then the sharp identified set for $b = \tau_R - \tau_N$ is Equation (2); its width is $1 - \alpha_N$. Observing C_N point-identifies $b = (F_R - F_N) + (C_N - C_R)$.*

Proof. Write the natural control mean as $C_N = \alpha_N F_N + (1 - \alpha_N)m_N$, where $m_N \in [0, 1]$ is the mean response under the natural policy conditional on selecting a non-repair action (arbitrary when $\alpha_N = 1$). Substitution into $b = (F_R - C_R) - (F_N - C_N)$ gives Equation (2). Its endpoints are attained by setting all non-repair action responses to zero or one; any interior value is attained by interpolation. Hence the set is sharp. Once C_N is observed, direct substitution gives the stated identity. □

Proposition A.5 (Sharp non-identification from proxy fidelity)**.** *Under Assumption A.1, fix a finite action set and the repair action a^+. Suppose $p_N = p_R = p$, $q_N(a^+) = q_R(a^+)$, and $p(a^+) = \alpha$. Over all response functions $q_N, q_R : \mathcal{A} \to [0, 1]$ satisfying these restrictions, the identified set for the replay bias $b = \tau_R - \tau_N$ is exactly*

$$b \in [-(1 - \alpha),\ 1 - \alpha]. \tag{7}$$

Thus zero action TV and zero forced-action suffix gap do not identify the sign of transport bias unless $\alpha = 1$.

Proof. The restrictions cancel the repair-action term and give $b = \sum_{a \neq a^+} p(a)[q_N(a) - q_R(a)]$. Since every bracketed difference lies in $[-1, 1]$, $|b| \leq \sum_{a \neq a^+} p(a) = 1 - \alpha$. Both endpoints are attained: set $(q_N(a), q_R(a)) = (1, 0)$ for every $a \neq a^+$ for the upper endpoint and reverse the pair for the lower endpoint. Any interior value follows by continuously interpolating one response function, so the interval is sharp. □

Proposition A.6 (Collapsed transport identity)**.** *Under Assumption A.1, for $z \in \{N, R\}$, let $\alpha_z = p_z(a^+)$, $F_z = q_z(a^+)$, and $M_z = \mathbb{E}_{A\sim p_z}[q_z(A) \mid A \neq a^+]$, with any fixed convention when $\alpha_z = 1$. Then*

$$C_z = \alpha_z F_z + (1 - \alpha_z) M_z, \tag{8}$$

$$\tau_z = (1 - \alpha_z)(F_z - M_z), \tag{9}$$

$$b = \tau_R - \tau_N = (\alpha_N - \alpha_R)(F_R - M_R) + (1 - \alpha_N)\{(F_R - F_N) - (M_R - M_N)\}. \tag{10}$$

Consequently,

$$|b| \le |\alpha_N - \alpha_R| + (1 - \alpha_N)\big(|F_R - F_N| + |M_R - M_N|\big), \tag{11}$$

and $|\alpha_N - \alpha_R| \le \mathrm{TV}(p_N, p_R)$.

Proof. Partitioning the policy-control expectation according to $A = a^+$ gives Equation (8). Subtracting C_z from F_z gives Equation (9). Adding and subtracting $(1 - \alpha_N)(F_R - M_R)$ in $\tau_R - \tau_N$ gives Equation (10). The triangle inequality gives Equation (11); the total-variation claim follows by choosing the singleton event $\{a^+\}$ in its variational definition. □

Proposition A.7 (Sufficient-state transport)**.** *Let W contain the environment state, all policy memory, and any decoder or execution state that affects sampling at the intervention time. Suppose natural arrival and replay condition on the same value $W = w$; the action sampler is the same kernel $\pi(\cdot \mid w)$ in both contexts; and, for every action a, the conditional law of the future trajectory and outcome under $\mathrm{do}(A = a)$ depends on the pre-intervention context only through w. Then $p_N = p_R$, $q_N = q_R$, and $b = 0$.*

Proof. Equality of the action kernel at w gives $p_N = p_R$. Conditional sufficiency gives $q_N(a) = q_R(a)$ for every action. Therefore $F_N = F_R$ and $C_N = C_R$, so Equation (4) gives $b = 0$. □

Proposition A.8 (Approximate sufficient-state transport)**.** *Under Assumption A.1, for a finite action set and binary outcome, suppose $\mathrm{TV}(p_N, p_R) \le \epsilon_p$ and $\max_a |q_N(a) - q_R(a)| \le \epsilon_q$. Then*

$$|b| \le \epsilon_p \ \mathrm{osc}(q_N) + 2\epsilon_q \le \epsilon_p + 2\epsilon_q, \tag{12}$$

where $\mathrm{osc}(q_N) = \max_a q_N(a) - \min_a q_N(a)$.

Proof. The forced-response difference is at most ϵ_q. For the controls,

$$\begin{aligned} |C_R - C_N| &\le |\mathbb{E}_{p_R}(q_R - q_N)| + |\mathbb{E}_{p_R} q_N - \mathbb{E}_{p_N} q_N| \\ &\le \epsilon_q + \epsilon_p \,\mathrm{osc}(q_N), \end{aligned}$$

using the variational bound for total variation. Apply the triangle inequality to $b = (F_R - F_N) - (C_R - C_N)$ and use $\mathrm{osc}(q_N) \le 1$. □

These results require alignment of responses beyond the target action. Action-law equality is also not necessary: with actions $\{a^+, u, v\}$, $p_N = (0.2, 0.8, 0)$, $p_R = (0.2, 0, 0.8)$, $F_N = F_R = 0.8$, and response 0.3 for both alternatives in both contexts, action TV is 0.8 but $b = 0$. The counterexample concerns effect transport, not physical-state equivalence.

A.3 Stochastic repairs and finite-sample uncertainty

Proposition A.9 (Sharp frontier for a stochastic repair)**.** *Under Assumption A.1, let g be any declared repair distribution on a finite action set, let $F_N^g = \sum_a g(a) q_N(a)$, and let $D_R^g = F_R^g - C_R$ be the replay-local effect. Define*

$$L(p, g, F) = \min_{u \in [0,1]^{|\mathcal{A}|} : g^\top u = F} p^\top u, \tag{13}$$

$$U(p, g, F) = \max_{u \in [0,1]^{|\mathcal{A}|} : g^\top u = F} p^\top u. \tag{14}$$

Given (p_N, g, F_N^g, D_R^g) and no restriction on q_N, the sharp identified set for $b^g = \tau_R^g - \tau_N^g$ is

$$\left[D_R^g - F_N^g + L(p_N, g, F_N^g),\ D_R^g - F_N^g + U(p_N, g, F_N^g)\right]. \tag{15}$$

Both endpoints are computed by a fractional-knapsack rule: among actions with $g(a) > 0$, fill $u(a)$ in increasing order of $p_N(a)/g(a)$ for L and decreasing order for U; actions outside the support of g take zero for L and one for U. At most one positive-g coordinate is fractional; ratio ties may be broken arbitrarily.

Proof. The feasible response vectors are exactly those in the box $[0,1]^{|\mathcal{A}|}$ consistent with the observed repair mean. Their possible natural-control means are the linear image $p_N^\top u$, hence the interval $[L, U]$; every interior value is attained by convexity. Since $b^g = D_R^g - (F_N^g - C_N)$, translating this interval gives Equation (15), proving sharpness. The stated optimizer is the fractional-knapsack solution to a one-constraint box linear program. □

The singleton repair is $g = \delta_{a^+}$, giving $L = \alpha_N F_N$ and $U = \alpha_N F_N + 1 - \alpha_N$. The following result propagates simultaneous uncertainty in the observed moments instead of treating a plug-in identified set as a confidence interval.

Proposition A.10 (Finite-sample outer region)**.** *Let $D_R = F_R - C_R$. Suppose intervals $[\ell_\alpha, u_\alpha]$, $[\ell_N, u_N]$, and $[\ell_D, u_D]$ simultaneously cover α_N, F_N, and D_R, respectively, with probability at least $1-\eta$. Then, with probability at least $1-\eta$, the entire population identified set in Equation (2) is contained in*

$$[\ell_D - (1-\ell_\alpha)u_N,\ u_D + (1-\ell_\alpha)(1-\ell_N)] \cap [-2, 2]. \tag{16}$$

Thus simultaneous finite-sample binomial intervals, or any other valid joint moment region, can be propagated without treating a plug-in identified set as a confidence interval.

Proof. On the simultaneous-coverage event, the lower endpoint of Equation (2) is $D_R - (1-\alpha_N)F_N$ and is bounded below by $\ell_D - (1-\ell_\alpha)u_N$. Its upper endpoint is $D_R + (1-\alpha_N)(1-F_N)$ and is bounded above by $u_D + (1-\ell_\alpha)(1-\ell_N)$. Finally, $\tau_N, \tau_R \in [-1, 1]$ implies $b = \tau_R - \tau_N \in [-2, 2]$, so intersecting with this logical range preserves coverage. □

A.4 Estimation and decision certificates

For the fixed-population estimator, a primitive outcome in cell x, arm a and replicate j has coefficient $c_{xaj} = w_x s_a / n_{xa}$, where s_a is the contrast sign. The following bound conditions on the observed usable counts and assumes independent draws from the declared arm laws.

Proposition A.11 (Fixed-weight bounded-outcome interval)**.** *Condition on the fixed cell weights and observed usable arm counts. Suppose the primitive outcomes $Y_i \in [0,1]$ entering a panel contrast are independent and the target contrast can be written $\widehat{\theta} = \sum_i c_i Y_i$ with nonrandom coefficients. Then*

$$\Pr\{|\widehat{\theta} - \mathbb{E}\widehat{\theta}| \geq r\} \leq 2\exp\left(-2r^2 / \sum_i c_i^2\right). \tag{17}$$

Thus $\widehat{\theta} \pm \sqrt{\frac{1}{2}\sum_i c_i^2 \log(2/\gamma)}$ is a two-sided $1-\gamma$ interval and $n_{\text{eff}} = 1/\sum_i c_i^2$ is its Hoeffding effective sample size.

Proof. Each summand $c_i Y_i$ lies in an interval of length $|c_i|$. Hoeffding's inequality therefore gives the displayed upper tail and lower tail bounds with denominator $\sum_i |c_i|^2$; their union yields the factor two. Solving the right-hand side for γ gives $r_\gamma = \sqrt{\frac{1}{2}\sum_i c_i^2 \log(2/\gamma)}$. □

For simultaneous intervals $I_R = [L_R, U_R]$ and $I_b = [L_b, U_b]$, let

$$I_N = I_R - I_b = [L_R - U_b, U_R - L_b]. \tag{18}$$

Proposition A.12 (Replay decision certificate)**.** *Suppose $I_R = [L_R, U_R]$ and $I_b = [L_b, U_b]$ simultaneously cover τ_R and $b = \tau_R - \tau_N$ with probability at least $1-\eta$. Then $I_N = [L_R - U_b, U_R - L_b]$ covers τ_N with probability at least $1-\eta$. For a fixed threshold δ, adopting only when $\inf I_N > \delta$ or rejecting only when $\sup I_N < \delta$ has error probability at most η on the corresponding false-decision event. Furthermore, for two repairs g, h sharing the same policy control, $\tau_z^g - \tau_z^h = F_z^g - F_z^h$; their ranking therefore needs forced-outcome contrasts but not the common control mean.*

Proof. On the joint coverage event, $\tau_N = \tau_R - b \in [L_R - U_b, U_R - L_b]$. A false certified adoption or rejection can occur only outside that event. The ranking identity follows by cancelling C_z from $(F_z^g - C_z) - (F_z^h - C_z)$. □

The recovery-set extension uses the same simultaneous-coverage argument. For restoration set S and query q, define

$$J_S^q = I(\theta_S^q) - I(b_S^q). \tag{19}$$

The antichain $\mathcal{M}^q$ contains inclusion-minimal restoration sets whose transported and direct-live intervals give the same non-abstaining decision at the declared threshold.

Proposition A.13 (Simultaneous minimal recovery contract)**.** *Fix a finite, preregistered family of restoration sets $S \subseteq \mathcal{K}$ and decision queries q. Suppose all effect intervals $I(\theta_S^q)$ and $I(\theta_N^q)$ cover simultaneously except on an event of probability at most η_E, and all bias intervals $I(b_S^q)$ cover simultaneously except on an event of probability at most η_B. Then, with probability at least $1-\eta_E-\eta_B$, every interval J_S^q in Equation (19) contains θ_N^q. Consequently, any non-abstaining decision selected from the data-dependent antichain $\mathcal{M}^q$ is correct at the declared threshold, provided it agrees with the non-abstaining direct-live interval decision.*

Proof. On the intersection of the two simultaneous coverage events, $\theta_S^q \in I(\theta_S^q)$ and $b_S^q \in I(b_S^q)$ for every declared (S, q). Therefore $\theta_N^q = \theta_S^q - b_S^q \in J_S^q$ simultaneously for every pair. The union bound gives probability at least $1 - \eta_E - \eta_B$ for this event. Because coverage holds for the entire finite family before the antichain is selected, choosing its inclusion-minimal members from the observed intervals does not invalidate the statement. A false certified adoption, rejection, or matching direct-live decision can occur only outside the joint event. □

The certificate need not improve precision when all estimates share data: $\widehat{\tau}_R - \widehat{b} = \widehat{\tau}_N$ exactly. Rectangular interval subtraction is valid on joint coverage but discards that dependence. Our empirical comparisons do not establish a query-efficiency advantage.

B Experimental Protocols and Reproducibility

This section specifies the populations, interventions and inference behind the principal claims. Supporting estimates are collected in Appendix C; peripheral studies use separate data and decision rules.

B.1 Primary ALFWorld population and live/replay execution

The policies are step-8 Qwen3-8B GRPO/GiGPO checkpoints trained with `verl-agent` revision `20bd331`. The original recipe uses learning rate 10^{-6}, batch size 16, eight rollouts per prompt, eight epochs, KL coefficient 0.01, BF16 and FSDP; prompt/response limits are 2,048/512 tokens. Measurement retains each policy's native prompt, tokenizer, two-step history, sampling kernel and invalid-action semantics.

Discovery uses 72 `valid-seen` games, 12 per family, disjoint from the earlier `valid-unseen` pools. Each of six checkpoints receives 24 trajectories per game, for 10,368 trajectories. A candidate needs at least 16 exact policy-history arrivals, a failed arrival and an executable verified expert continuation. Each family contributes at most two templates; a checkpoint needs eight templates from four families. Five checkpoints pass, yielding 56 checkpoint–template cells over 26 unique templates. GiGPO-201 has a maximum support of 11 and no qualifying cell; it remains an unsupported policy. Selection is blind to forced/control intervention outcomes, but the failed-arrival condition means it is not outcome-free.

At each candidate, the observable signature is

$$x_t = \phi(h_t) = (\text{game}, o_t, \mathcal{A}_t, \text{score}_t, \text{done}_t), \tag{20}$$

with executable actions $\mathcal{A}_t$; the serialized native history must also match. The frozen repair a^+ is the next action of the verified expert suffix, which is not supplied to the policy. First exact natural arrival defines T_x for the complete matching rule. Arrival frequency and $\alpha_N = P(A_{T_x} = a^+ \mid T_x < \infty)$

are estimated from natural rollouts, with one action sample per arrival. For binary success, the probability chain $P(T_x < \infty)\alpha_N P(Y = 1 \mid T_x < \infty, A_{T_x} = a^+)$ is observational. A forced continuation estimates its last factor only under the relevant randomized-action/no-confounding condition.

Live forced/control attempts are deterministically allocated before rollout and start from reset. At first exact arrival both query the policy; forced discards the draw and executes a^+, while control executes the draw. Only the executed action enters memory. No reset, prefix replay, snapshot restoration or environment replacement occurs after arrival. Replay arms instead execute the frozen discovery prefix in fresh simulators before the same two interventions. Outcomes use disjoint seeded attempts; shared labels or unique seeds do not prove independence or common context laws. The arm label is absent from policy input.

Ordinary cells receive 128 attempts per live arm and 64 per replay arm. Three previously outcome-selected local outliers receive 512 attempts on each of the four primary arms. Their new outcomes alone enter the reported cell and panel estimates, and the fixed aggregation weights are unchanged. The resulting 26,496 primary attempts include 586 live non-arrivals, leaving 25,910 usable outcomes (Table 4). Non-arrivals are counted as attempts but excluded from arrival-conditional means. Ten separately chosen history-stress cells receive 128 forced and control outcomes each, adding 2,560 records. All records complete without transport error. Arrival floors are 48 per ordinary live arm and 384 per targeted arm.

The three targets are exactly the earlier reconstruction-study cells with $|\widehat{\tau}_R - \widehat{\tau}_S| > 0.10$, where S is source-prefix reconstruction rather than live natural arrival (Table 3 and Appendix C.1). The new allocation was locked on 16 September 2026 at 11:16:45 UTC, with zero outcome calls recorded. The earlier outcomes select the targets; the confirmation estimator reads only the new checkpoint event ledgers, whose task IDs match the new assignments. It excludes the earlier selection outcomes. Each target receives 512 new attempts per primary arm, with live means conditional on arrival. Equal cell-within-checkpoint and checkpoint-within-panel weights remain fixed. Thus each GRPO-104 target has weight $1/33$ and the GiGPO-202 target has weight $1/22$. An outcome's contrast coefficient is $\pm w_x/n_{xa}$, so unequal arm counts enter the Hoeffding variance proxy as $\sum_{x,a} w_x^2/n_{xa}$.

Table 3: Targets verified from the locked assignments. Candidate indices are zero-based within checkpoints; prior bias compares two reconstructed contexts and is used only for selection. Every target has 512 attempts per new primary arm.

Checkpoint / index	Stable template identifier	Prior bias
GiGPO-202 / 1	`look_at_obj_in_light-CD-None-DeskLamp-314`	-0.15625
GRPO-104 / 1	`look_at_obj_in_light-AlarmClock-None-DeskLamp-323`	-0.11159
GRPO-104 / 4	`pick_clean_then_place_in_recep-DishSponge-None-CounterTop-403`	$+0.14063$

Table 4: Primary ALFWorld allocation and exact arrivals. Live entries are arrivals/attempts, not successes. The replay column gives the usable count in each of its two arms. Totals are 56 cells and 25,910 usable primary outcomes.

Checkpoint	Cells	Live forced	Live control	Replay per arm
GRPO-101	12	1,447/1,536	1,466/1,536	768
GRPO-103	10	1,232/1,280	1,239/1,280	640
GRPO-104	11	2,150/2,176	2,143/2,176	1,600
GiGPO-202	11	1,674/1,792	1,697/1,792	1,152
GiGPO-203	12	1,502/1,536	1,504/1,536	768

The primary estimand weights cells equally within each fixed checkpoint, then checkpoints equally within a panel. Bootstrap analyses resample outcomes within the relevant fixed arm strata, not tasks or checkpoints as population draws. Bounded-outcome intervals assume independent usable draws under the declared laws, conditional on observed arrival counts. The preregistered component

family uses four 98.75% intervals; the bias family uses two 97.5% intervals. These families are not one joint 95% statement. The main-text joint audit is explicitly post-hoc and assigns $0.05/6$ to each of the six panel estimands. The absolute equivalence margin is ± 0.10; ± 0.05 and relative scales are descriptive sensitivities.

B.2 WebShop reconstruction omissions

The WebShop study pins the official product, goal and search-index artifacts and qualifies a Qwen3-4B ATOD policy independently of intervention outcomes. Of two screened policies, the 4B model passes a 256-episode capability panel; the Qwen3-30B policy fails. Natural discovery runs 10,240 trajectories over 320 goals, observes 3,753 checkout-ready visits and 1,172 signatures, and selects 12 cells spanning 12 product categories. A separate small-catalog feasibility failure does not contribute outcomes.

Each cell receives five contexts, three arms and 64 replicates, for 11,520 outcomes. The arms are `Buy Now`, `Back to Search`, and policy control. Complete replay restores prefix, environment state, policy history, visible page, prompt and tokens. The other replay conditions omit policy history, one frozen goal-relevant backend option, or both. Live intervention occurs at exact natural arrival. The primary outcome is terminal bounded task score; exact success is secondary. These manipulations are declared failure probes, not a sample of deployment errors.

Surface and action diagnostic. For each fixed candidate, the rule compares natural arrival N with backend-option omission R_T. Pass requires exact canonical equality of observation, ordered action menu, prompt and token IDs across all recorded replay interventions and the fixed live input, together with

$$\widehat{\mathrm{TV}} = \tfrac{1}{2} \sum_u |\widehat{p}_N(u) - \widehat{p}_{R_T}(u)| \le 0.15.$$

The two empirical distributions use 64 policy-control draws each. The parser lowercases output and strips whitespace around the extracted action-tag content. Case-insensitive menu matches use the menu's canonical string; unmatched projections retain their string, including parameterized searches. Projection validity is checked separately: missing action or reasoning tags, or Han characters, produce a shared `__INVALID__` bucket, regardless of whether the extracted command is executable. No invalid draw is discarded or resampled. Thus the diagnostic uses projected-action buckets, not exclusively admissible actions. The verified configuration fixes the inclusive threshold at 0.15. Reapplying the original function to the saved events reproduces 12/12 exact input matches and 11/12 passes; candidate `atod_qwen3_4b-607314ca7c84934a` fails at 0.296875. The separate cross-method submitted-string diagnostic also uses 0.15, but retains action-string distinctions when projection validity is false (Appendix C.3); it is a different representation.

The final serving allowance is 5,120 total tokens with a 512-token response budget. An earlier incomplete stream is excluded; this allowance and fresh seeds were fixed before effect inspection, without changing the scientific rules. Analysis resamples events within each fixed cell–context–arm stratum, using 20,000 draws, seed 20260917 and simultaneous maximum-absolute-studentized bootstrap intervals within the declared families. Neither cells nor policies are resampled. Main-text bootstrap intervals use family error 0.05; the primary bounded recovery certificate separately assigns 0.025 each to its effect and bias families. Its threshold is 0.05 and bias tolerance ± 0.10. Both queries abstain under that certificate, including with complete replay.

B.3 Secondary designs and their comparators

The cross-method study uses 18 fresh Qwen3-8B training runs: three seeds (5101–5103), three methods and two environments, with checkpoints at iterations 4, 8 and 16. The fixed reference is one checkpoint without additional task RL. Learning rate is 10^{-7}, training batch size 8, rollouts per prompt 16 and KL coefficient 0.05 for all methods. Agent-DAPO adds constant-group filtering and asymmetric clipping 0.20/0.28, with at most eight generation batches per iteration; it retains KL and omits overlong-response shaping. Equal iterations are not equal interaction budgets. Native histories retain two steps, temperature is 0.4, and response length is at most 512 tokens.

Discovery/evaluation frames contain 24/36 ALFWorld templates and 200/250 WebShop goals. Capability is measured independently; it does not remove primary policies. Support thresholds and

pre-outcome counts of 8–64 per eligible cell/context/arm are fixed. The post-recovery own-support stream contains 344,580 outcomes, with 11 capped-arrival shortfalls retained. A separate missing-state supplement adds 2,352 ALFWorld and 130,788 WebShop outcomes; it supplies declared states rather than substituting other states from the same task. The aggregate family has total error budget 0.05 and an individual allocation of $0.05/6195$. That denominator is inherited from the original post-recovery analysis and is not recomputed after the coverage supplement. It counts 987 estimated aggregate contrasts, 134 panels with no qualified support, and 5,074 missing declared cell–context–arm slots. The estimated contrasts comprise 251 context effects, 151 live-referenced biases, 51 omission increments, and 534 panel differences, taken over the locked manifests, method panels, and required contexts. The manifests cover environment-specific own support, strict or pairwise common-task support, capability and cost matches, and checkpoint stages. A panel contrast requires qualified cells and all required exact-state arm units; differences compare common bias or omission contrasts between eligible panels within a manifest. Missing slots are counted by manifest, cell, context and arm, not as unique physical states. Unsupported panels and missing slots remain in the denominator so that absent support cannot shrink it. The coverage supplement keeps this denominator even though it fills the missing slots and estimates 1,199 contrasts. Cellwise history-minus-live and omission-minus-full analyses are separate families of 1,140 and 504 contrasts, each with its own budget of 0.05 and no joint guarantee with the aggregate family. The 1,140-contrast family contains 636 full-replay-minus-live contrasts, 498 history-minus-live contrasts, and three each for backend and joint omissions; it supports the 12/236 and 20/262 history detections and the local case in Table 15. The 504-contrast family instead contains 498 history-minus-full contrasts and three each for backend and joint omissions; it supports the 11/236 and 21/262 detections in Section 3.6. Aggregate intervals support Table 13. Common tasks need not imply identical policy-specific states or repairs.

Downstream studies have independent populations and objectives (Appendix C.4). The same-token study uses eight training seeds and the same 128 teacher records across auxiliary arms. Modern-model studies use native `ms-swift`, BF16 LoRA rank 32/scale 64, two-observation memory and unconstrained decoding. Broad action-only SFT uses 32 updates of 64 examples at learning rate 10^{-5} and coefficient 0.1, with no RL/SDAR updates. Its calibrated-pool control panel instead uses eight DAPO updates, 16 rollout episodes per update, four generations and KL 0.05, followed by the optional 64-example correction. Uniform matches live-RTA weight mass; raw/conservative replay differ in mass and coverage.

Retail uses a Qwen3-30B actor and Qwen3.5-35B visible-history advisor before the first database write. Calibration uses 45 training tasks and 328 outcomes; evaluation retains all 40 test tasks, 32 paired seed blocks and four arms (5,120 episodes). Eleven tasks' language assertions are judged by GPT-5.6-Sol with method labels withheld. Two context-window failures score zero; 49 transient failures are retried under their original keys and seeds. Inference resamples complete paired seed blocks, not tasks.

B.4 Evidence files and implementation changes

Saved manifests identify model/tokenizer revisions, task frames, seeds, exact context laws, arm assignments and outcome ledgers. The source package's `EVIDENCE_INDEX.json` maps the retained results to their original analysis paths and hashes. The paper builds directly from saved tables and figures with Tectonic 0.17.0; model calls are not part of compilation. The original project retains analysis scripts and raw ledgers for reanalysis.

Only changes affecting interpretation are material here. An early same-seed-coupling study was excluded because distributed sampling did not reproduce suffixes exactly. The primary live confirmation uses fresh outcomes and independent-stream assumptions instead. Cross-method analyses are post-recovery evidence: an assistant-prefill/parser defect affected the earlier capability pilot, and the WebShop serving-context allowance was enlarged. Recorded format-invalid flags in the repaired cross-method streams still coexist with submitted executable action strings; validity buckets therefore cannot be treated as action-law agreement. Modern runs retain explicit zero-query port-retry and unstarted-job relocation records, without changing their data or scientific settings. Routine controller and hash-check logs are reproducibility aids. They do not change the interpretations above.

Table 5: Engineering measures and their inferential role. *Reported* refers to the ALFWorld live audit or 12-cell WebShop study described above; *potential* identifies a design option, not a verified feature of those experiments. No row alone establishes Assumptions A.1 or A.2, or the sufficient state in Proposition A.7.

Engineering measure	What it supports	What it does not establish
Match observations, menus, prompts and tokens (reported)	Equality of the checked visible inputs at the candidate.	Equality of hidden state, complete-context populations or responses to unmeasured actions.
Restore environment and external-session state	Reported prefix execution and WebShop backend restoration reproduce specified components; a fuller state audit is a potential extension.	That the restored fields exhaust relevant state, or that every action has the same continuation law.
Restore policy memory and decoder state	Reported serialized history preserves supplied memory. Restoring relevant KV-cache and decoder state is a potential numerical-fidelity measure.	Cache restoration alone does not preserve environment responses, all sampling state or the distribution of naturally reached contexts.
Deterministic or batch-invariant inference (potential)	Controls numerical variability under a fixed configuration; may improve repeatability of an action kernel.	Preservation of the original stochastic action law, continuation laws or sampled population. Determinism does not itself imply zero bias.
Preallocate or randomize arms	Reported deterministic preallocation fixes attempts before outcomes. Potential randomization supports comparability if independent of context and unaffected by arm-dependent arrival selection.	Deterministic allocation leaves common-law validity assumed. Randomization alone does not establish consistency, noninterference, stable protocols or equality across natural and replay contexts.
Disjoint seeds and fresh outcomes (reported)	Avoids deliberate seed reuse and separates confirmation outcomes from earlier selection outcomes.	Statistical independence, context-law equality, or a representative candidate population; unique seeds do not prove any of these.

C Supporting Results, Sensitivities, and Negative Findings

The supporting evidence is grouped by what it establishes: precision and heterogeneity of the primary measurement, scope across policies, and limits of downstream use. The final subsection summarizes investigations that constrain the scope of the conclusions.

C.1 Primary inference, heterogeneity, and sensitivity

The preregistered ALFWorld confidence families and the joint post-hoc family answer different coverage questions. Table 7 retains the original intervals; Table 1 reports the six-estimand joint family visualized in Figure 3c. Checkpoint values in Table 8 show heterogeneity hidden by panel averages. Twenty-seven of 56 cells have four zero empirical arm means; these are not proven zero population effects.

Absolute tolerance and relative scale. Table 6 separates descriptive ratios from interval conclusions. The declared absolute margin is ± 0.10. The post-hoc six-estimand joint weighted-Hoeffding family in Table 1 certifies this panel-level margin for both methods, but not the tighter descriptive ± 0.05 margin. The original preregistered two-bias family, with 97.5% intervals per panel, gives the same two conclusions (Table 7). These families are separate; they are not pooled into a larger joint claim.

Table 6: ALFWorld equivalence sensitivity on the fixed panels. Ratios use unrounded point estimates only. Interval conclusions in the last two rows use the post-hoc joint six-estimand family (at least 95% coverage).

Quantity or conclusion	GRPO	GiGPO
Natural-effect estimate $\widehat{\tau}_N$	0.1617	0.1426
Bias estimate $\widehat{b}$	$+0.0056$	-0.0034
Declared $0.10/\lvert\widehat{\tau}_N\rvert$, descriptive	$\approx 62\%$	$\approx 70\%$
$\lvert\widehat{b}\rvert/\lvert\widehat{\tau}_N\rvert$, descriptive	$\approx 3.5\%$	$\approx 2.4\%$
Joint bounded bias interval	$[-0.0558, 0.0669]$	$[-0.0777, 0.0709]$
Entire bias interval inside ± 0.10	Yes	Yes
Entire bias interval inside ± 0.05	No	No

Small descriptive point ratios are not certified relative-error guarantees. Such a guarantee would require joint inference for the numerator and denominator; dividing a bias confidence limit by a natural-effect point estimate does not provide it. The reported inference retains the fixed-panel weights and arrival-conditional scope.

The three-arm identification width. The natural repair probability is estimated below 0.1 in 53/56 cells; the repair is never observed in 48. Mean plug-in widths are 0.941 for GRPO and 0.958 for GiGPO. These are identified-set estimates, not confidence intervals. A post-hoc one-sided Hoeffding analysis allocates error $0.05/56$ to each natural repair probability. With simultaneous coverage of at least 95%, the lower bound on $1 - \alpha_N$ exceeds 0.80 in 53/56 cells and 0.90 in 3/56. This characterizes the missing-control uncertainty under the factorization assumption.

Table 7: Original preregistered RTA-L1 fixed-panel analysis. Component intervals are simultaneous 98.75% weighted-Hoeffding intervals within their four-effect family; bias intervals are 97.5% within their two-panel family.

Panel	Cells	τ_N (Hoeffding CI)	τ_R (Hoeffding CI)	b (bootstrap CI)	b (Hoeffding CI)
GRPO	33	0.1617 [0.1270,0.1963]	0.1672 [0.1194,0.2151]	0.0056 [−0.0123,0.0234]	0.0056 [−0.0493,0.0605]
GiGPO	23	0.1426 [0.1001,0.1851]	0.1392 [0.0817,0.1967]	−0.0034 [−0.0148,0.0079]	−0.0034 [−0.0699,0.0631]

Table 8: Primary ALFWorld checkpoint effects. Values are point estimates. The Zero column counts cells with zero observed success in all four arms.

Checkpoint	Cells	$\widehat{\tau}_N$	$\widehat{\tau}_R$	$\widehat{b}$	Zero
GRPO-101	12	0.059	0.053	-0.005	7
GRPO-103	10	0.214	0.234	0.020	5
GRPO-104	11	0.212	0.214	0.002	2
GiGPO-202	11	0.199	0.192	-0.007	3
GiGPO-203	12	0.086	0.086	-0.0002	10

Same-data decisions and cost. Table 9 uses the joint-six intervals at post-hoc thresholds. Corrected replay and direct live have the same point estimate; rectangular subtraction is wider because it ignores dependence between contrasts computed from the same arms.

Table 9: Post-hoc same-data decision comparison. Replay-only rows have no transport guarantee. Thresholds were selected after outcomes.

Panel	Procedure	Point	Interval	$\delta = 0$	$\delta = .05$	$\delta = .10$
GRPO	Replay only (unprotected)	0.167	[0.118, 0.217]	ADOPT	ADOPT	ADOPT
GRPO	Rectangular RTA	0.162	[0.051, 0.273]	ADOPT	ADOPT	ABSTAIN
GRPO	Direct live / joint contrast	0.162	[0.126, 0.198]	ADOPT	ADOPT	ADOPT
GiGPO	Replay only (unprotected)	0.139	[0.079, 0.199]	ADOPT	ADOPT	ABSTAIN
GiGPO	Rectangular RTA	0.143	[0.008, 0.277]	ADOPT	ABSTAIN	ABSTAIN
GiGPO	Direct live / joint contrast	0.143	[0.098, 0.187]	ADOPT	ADOPT	ABSTAIN

A retrospective task-ID-prefix truncation assigns each procedure 96 usable outcomes per cell (5,376 total), counting live non-arrivals as costs (Table 10). Rectangular RTA abstains at all three thresholds in both panels. This is not a randomized equal-query comparison or evidence of query-efficiency gains; startup and offline verification are excluded.

Table 10: Matched usable-outcome sensitivity with realized task-ledger costs.

Method	Attempts	Model requests	Prompt tokens	Completion tokens
Replay only	5,376	119,675	56,656,258	33,954,796
Direct live	5,666	160,723	76,533,785	39,553,247
Rectangular RTA	5,516	139,995	66,519,616	36,637,599

History and reconstruction sensitivities. The ten-cell history-0 stress detects 0/10 effects under both its historical rule and a separate post-hoc simultaneous Hoeffding analysis (error $0.05/10$; median radius 0.375). It is inconclusive about history insensitivity. The distinct source-prefix positive control reconstructs both contexts: source law S executes each logged prefix, while R executes the canonical prefix. Of 6,858 arrivals, 6,701 (97.7%) have the same prefix; all prefixes coincide in 51/56 cells. Its 20,884 outcomes therefore test an intentionally favorable reconstruction setting, not live arrival. The 97.5% bounded bias intervals are $[-0.0664, 0.0641]$ for GRPO and $[-0.0815, 0.0762]$ for GiGPO. A post-hoc shared-template bootstrap gives $[-0.0331, 0.0311]$ and $[-0.0340, 0.0216]$; deleting a family, deleting a checkpoint or changing aggregation shifts the means by at most 0.0103, 0.0027 and 0.00035. These sensitivities retain the fixed-population scope and do not replace the live confirmation.

C.2 WebShop effects and decision sensitivity

Table 11 gives the effects underlying Figure 3a,b. Full-replay bias intervals come from the eight-estimand component family; omission bias intervals come from the separate eight-estimand recovery-bias family. Both use family error 0.05 and the fixed-cell bootstrap. The primary bounded recovery certificate instead allocates 0.025 each to effects and biases. Its live-effect intervals cross the threshold 0.05, and its complete-replay bias intervals extend outside ± 0.10, so no recovery set is certified. For the aggregate certificate, the ten context effects (five contexts by two queries) each receive error $0.025/10$; the eight live-referenced biases (four replay contexts by two queries) each receive $0.025/8$. The weighted-Hoeffding radius uses the fixed outcome coefficients (Proposition A.11), and the transported natural interval subtracts the simultaneous replay-effect and bias intervals. For complete replay, direct-live intervals are $[-0.0770, 0.1096]$ for τ_g and $[-0.0571, 0.1295]$ for ρ; the corresponding bias intervals are $[-0.1276, 0.1318]$ and $[-0.1369, 0.1226]$. Thus neither the decision-threshold nor bias-tolerance requirement is met. The cellwise certificates also abstain. These are properties of the declared interval rule on these data, not a sample-size-only explanation.

Table 11: Which reconstruction omissions change WebShop action effects? H denotes retained policy history and T retained hidden backend state. Effects use bounded task score. "Backend option omitted" removes one frozen goal-relevant field from the backend state and leaves policy history in place; it does not remove a tool. T in the restored column is that backend state. Brackets are 95% simultaneous fixed-cell bootstrap intervals within the separately declared component and bias families, not a joint guarantee for the entire table. The primary bounded-outcome decision analysis uses a separate error allocation and abstains.

Context	Restored	τ_g	b_g	ρ	b_ρ
Live N	live object	0.016	—	0.036	—
Full replay	H, T	0.018	0.002 [−0.021,0.025]	0.029	−0.007 [−0.037,0.023]
Backend option omitted	H	−0.029	−0.045 [−0.068,−0.023]	−0.129	−0.165 [−0.195,−0.135]
History omitted	T	0.556	0.540 [0.524,0.555]	0.556	0.520 [0.498,0.542]
Both omitted	$\varnothing$	0.394	0.378 [0.362,0.393]	0.394	0.358 [0.336,0.380]

The motivating example is candidate `atod_qwen3_4b-77a441639a8f349b`, goal 152, product `B083GGC2N9`. Its frozen backend has `style = d17 (dedicated right, black)`. The manipulation removes `server_session.options.style`; the saved proof records identical observation, action menu, prompt and tokens. Figure 1 abbreviates these labels and depicts the intervention, without assigning an individual-cell outcome to it. The editable figure package includes the candidate record and manipulation proof alongside the exact aggregate plotting data.

Table 12: WebShop decision sensitivity on 12 fixed cells at threshold 0.05. The first column compares repair-versus-policy threshold decisions; the remaining columns compare repair-versus-alternative threshold decisions and strict signs. These are descriptive disagreements; the primary bounded certificate abstains in every cell.

Replay context	Repair threshold	Ranking threshold	Ranking sign
Complete	0	1	0
Backend omitted	1	2	3
History omitted	8	6	2
Both omitted	7	5	2

C.3 Cross-method scope and diagnostic failures

Table 13: Complete replay on common-task support. Each RL method has three training seeds; Base is one fixed reference checkpoint. Intervals belong to the separate post-hoc aggregate family. All three RL bias intervals lie within ± 0.10 in both environments; the baseline intervals do not. Common tasks retain policy-specific states and repairs.

Environment	Method	Cells	$\hat{\tau}_N$	$\hat{\tau}_R$	$\hat{b}$ [interval]
ALFWorld	Base	21	0.106	0.111	0.005 [-0.144, 0.155]
ALFWorld	GRPO	63	0.165	0.157	-0.008 [-0.095, 0.078]
ALFWorld	GiGPO	63	0.111	0.107	-0.005 [-0.091, 0.082]
ALFWorld	Agent-DAPO	63	0.071	0.060	-0.010 [-0.097, 0.076]
WebShop	Base	40	0.066	0.072	0.006 [-0.124, 0.136]
WebShop	GRPO	120	0.077	0.071	-0.006 [-0.081, 0.069]
WebShop	GiGPO	120	0.089	0.084	-0.004 [-0.079, 0.071]
WebShop	Agent-DAPO	120	0.116	0.119	0.003 [-0.071, 0.078]

Support qualification and measurement caps are distinct. Across late policies and the fixed reference, ALFWorld qualifies 244 of 360 policy–task rows and measures 236; WebShop qualifies 2,863 of 4,040 policy–task–stage rows and measures 400. The WebShop denominator is the discovered 404-key stage union, not all possible stages of its evaluation goals. Hidden-state and joint omissions have only three eligible WebShop own-support cells and none in ALFWorld. Their wide intervals cannot establish hidden-state irrelevance.

On common tasks, Table 13 reports every method and the fixed reference. Capability-matched ALFWorld comparisons admit two three-method triples, but their intervals extend beyond the equivalence margin; WebShop has no qualifying capability match. Episode-cost matching admits GRPO/GiGPO but not Agent-DAPO under the 10% caliper. On the 11-template common ALFWorld stage population, early-minus-late bias is -0.009 for GRPO ($[-0.335, 0.317]$), 0.037 for GiGPO ($[-0.289, 0.364]$), and 0.039 for Agent-DAPO ($[-0.287, 0.365]$). WebShop's stage intersection is empty. These results leave method and stage orderings unresolved.

Action-law proxies do not achieve the declared 0.02 held-checkpoint MAE improvement (Table 14). The 636-cell analysis uses 76,810 control draws; features and observed-bias targets share outcomes. All format-validity flags are false under the recorded kernel, although 22,457/26,425 ALFWorld and 43,590/50,385 WebShop submitted actions are admissible. Consequently validity-bucket TV is identically zero, while submitted-string TV retains behavioral distinctions. Among matched surface/input cells with submitted-string TV at most 0.15, 18/188 ALFWorld and 23/365 WebShop cells have point $|\widehat{b}| > 0.10$; these counts are descriptive, not simultaneous tests or prevalence estimates.

Table 14: Descriptive leave-one-checkpoint-out prediction of absolute observed cell bias. Positive MAE improvement favors the feature over the training-fold mean baseline; 0.02 is the declared materiality threshold. Cells are equally weighted. Features and target estimates share control draws, so predictive accuracy is not validation against noise-free true transport bias.

Environment	Feature	Cells	Baseline MAE	Feature MAE	Improvement
ALFWorld	Submitted-action TV	236	0.0390	0.0393	-0.00025
ALFWorld	Validity-bucket TV	236	0.0390	0.0390	0.00000
ALFWorld	Target-probability gap	236	0.0390	0.0390	0.00003
WebShop	Submitted-action TV	400	0.0292	0.0291	0.00014
WebShop	Validity-bucket TV	400	0.0292	0.0292	0.00000
WebShop	Target-probability gap	400	0.0292	0.0293	-0.00004

A post-hoc tomato-task example shows why aggregate bias and coarse validity flags can miss local changes (Table 15). For GRPO seed 5101 at iteration 16, the target is `go to microwave 1` after taking a tomato from a fridge. Removing the two retained history pairs preserves observations and admissible actions but removes the forced-target advantage. The omission-minus-live effect is -0.9643, with interval $[-1.5833, -0.3453]$ in the separate 1,140-contrast family. Submitted-string TV is 0.4643 despite validity-bucket TV zero. The example demonstrates sensitivity under the executed kernel, not a pure memory-only mechanism or failure despite identical action distributions.

Table 15: Recorded tomato-task case: successes out of 56 outcomes per arm. These are cell-specific observations, not panel-level estimates.

Context	Forced target	Policy control	Estimated action effect
Live natural arrival	54/56	0/56	0.9643
Full replay	55/56	2/56	0.9464
History omitted	0/56	0/56	0.0000

In the distinct 504-contrast omission-minus-full family, history shifts are detected in 11/236 ALFWorld and 21/262 WebShop cells, while aggregate omission intervals include zero. In the 1,140-contrast omission-versus-live family, the material-shift counts are 12 and 20. These are different contrasts and error families, not interchangeable counts. Neither the hidden-state omission nor the joint omission produces a detected shift; each condition has only three eligible cells.

C.4 Downstream comparisons

Same-token training and Retail advice. The same-token study compares eight seeds (9501–9508), eight GRPO steps and the same 128 teacher records over a fixed 144-game held-out frame. All 32 runs complete. Live-audited-minus-uniform AULC is 0.0330 (one-sided 97.5% lower bound

-0.0054; Holm $p = 0.214$); live-audited-minus-replay is 0.0060 (lower bound -0.0785; Holm $p = 0.429$). Final-success differences are 0.0347 and -0.0477, with lower bounds below the fixed -0.005 noninferiority margin. The joint rule fails.

Table 16: Same-token auxiliary training. AULC is the mean area under the five- point held-out success curve; final is step-8 success.

Arm	Mean AULC	Mean final success
RL-only	0.1321	0.2700
Uniform auxiliary	0.0530	0.0660
Replay-weighted auxiliary	0.0800	0.1484
Live-audited auxiliary	0.0859	0.1007

Retail comparison. Retail retains all four advice comparators (Table 17). RTA-minus-base is 0.0344 with family-adjusted 97.5% bootstrap interval $[0.0063, 0.0633]$ and Holm $p = 0.0172$. RTA-minus-always is -0.0289 ($[-0.0539, -0.0023]$), missing the two-point noninferiority margin. Replay-minus-base is 0.0445 ($[0.0141, 0.0766]$); RTA-minus-replay is -0.0102 ($[-0.0398, 0.0180]$). Calibration admits the same cancel/return operation set under both gates. Selective advice helps relative to the actor; an RTA-specific benefit is not established.

The task definitions identify eleven of forty tasks with language assertions. The scorer retains 844 judgments from the alias `gpt-5.6-sol`; an immutable backend version is not saved. Requests use temperature 0, medium reasoning effort, a 2,048-token output cap, JSON format and seed $202610030000 +$ (rollout seed $\bmod 100000$). The payload includes user, assistant and tool messages (content and tool-call fields), plus numbered assertion texts. System messages, method labels, gate decisions, model names, rewards and reference answers are omitted; the transcript itself can still contain evidence of behavior.

The judge is instructed to evaluate each criterion only from the transcript, return one Boolean judgment and concise evidence per criterion, and count communication requirements only when the assistant communicates to the user; tool output is context. The response schema is `results: [{id, satisfied, evidence}]`. The exact prompt and payload constructor accompany the source package. Database failure sets reward to zero without calling the judge. A passing database check with no assertions sets reward to one; otherwise all assertions must be satisfied. Requests permit six transport attempts in total. Malformed or missing judgments remain unresolved and block completion; the saved batches have none unresolved. One final judgment is retained per eligible episode. No human reliability check, repeat-judgment comparison or adjudication record was found. Resampling these scores does not establish judge accuracy.

A post-hoc sensitivity uses all 29 test tasks whose definitions have no language assertions. This includes task 79: its assertion field is null, although all 128 saved episodes fail the database check. The subset contains 3,712 episodes, or 928 per method, across the same 32 paired seed blocks. Success is 0.4935 for the base actor, 0.5399 for always-advisor, 0.5377 for replay gating and 0.5291 for RTA. RTA-minus-base is 0.0356 with a 97.5% percentile interval $[-0.0043, 0.0754]$; RTA-minus-replay is -0.0086 ($[-0.0474, 0.0313]$). These individual-comparison intervals use 20,000 resamples of complete seed blocks at analysis seed 202610020001, with tasks fixed. The intervals are reported separately and are not adjusted jointly across contrasts. All methods remain in the comparison. This subset neither validates the judge nor replaces the unchanged forty-task result.

Table 17: Retail full-test utility. Success is the complete environment-and- language score over 1,280 episodes per row.

Method	Success	Δ base	Advisor calls	Interventions
Base actor	0.501	—	0	0
Always advisor	0.564	$+0.063$	1,130	783
Replay gate	0.545	$+0.045$	375	240
RTA gate	0.535	$+0.034$	378	228

C.4.1 MODERN MODELS: WEIGHTING, INTERPOLATION, AND TRANSFER

The calibrated-pool RL study uses 40 Gemma and 39 Qwen corrections, three training seeds (6201–6203), five arms and eight updates. Each run evaluates 144 episodes at steps 0/2/4/6/8. AULC is the normalized trapezoidal average of these rates. Tables 18 and 19 retain every control and paired RTA contrast. The fixed rule requires positive endpoint and AULC differences in every seed against every control and positive endpoint seed-interval lower bounds; it fails. Raw replay differs in weight mass and coverage, so its higher point estimates do not isolate weighting shape.

Table 18: All 30 calibrated-pool RL/control runs. Final counts aggregate 432 episodes per arm and model; AULC is a percentage. These are not matched-budget controls for broad-data SFT.

	Gemma4-31B		Qwen3.6-27B	
Arm	Final	AULC	Final	AULC
RL only	170	39.00	167	37.59
Uniform	174	40.39	205	42.27
Conservative replay	177	40.51	176	38.31
Raw replay	186	41.58	248	43.89
Live RTA	185	41.15	229	42.71

Table 19: Live RTA minus each native control. All differences are percentage points; intervals are paired-training-seed Student-t 95% intervals with two degrees of freedom. They assume independent normal paired effects and are unadjusted for multiple comparisons.

Model	Control	Final gain	Seed interval	AULC gain	Seed interval
Gemma4-31B	RL only	+3.47	[+0.02, +6.92]	+2.14	[+0.95, +3.33]
Gemma4-31B	Uniform	+2.55	[-1.04, +6.14]	+0.75	[-0.88, +2.39]
Gemma4-31B	Conservative replay	+1.85	[-5.12, +8.82]	+0.64	[-2.60, +3.87]
Gemma4-31B	Raw replay	-0.23	[-3.82, +3.36]	-0.43	[-3.48, +2.61]
Qwen3.6-27B	RL only	+14.35	[+7.38, +21.32]	+5.12	[+2.27, +7.97]
Qwen3.6-27B	Uniform	+5.56	[-1.96, +13.08]	+0.43	[-2.76, +3.63]
Qwen3.6-27B	Conservative replay	+12.27	[+8.68, +15.86]	+4.40	[-0.69, +9.48]
Qwen3.6-27B	Raw replay	-4.40	[-9.67, +0.87]	-1.19	[-2.37, +0.00]

Broad-data SFT instead uses 390 verified correction states for Gemma and 155 for Qwen from a separate 24-template training frame (22/20 represented templates; six families). Only the semantic action-command span is supervised; tags, reasoning, EOS and padding receive zero loss. Uniform weights match the scheduled reference mass rather than being one. The reference is half the positive live-effect weight and half its pool mean, treating unidentified effects as zero. The loss coefficient multiplies the full-batch mean action-token NLL. On the development screen, the weighted/uniform endpoints are 18/21 successes for Gemma and 15/28 for Qwen over 48 episodes; the common weighting-benefit rule fails.

Interpolation scales saved LoRA B matrices by γ, holding A, base weights and decoding fixed (Hu et al., 2021). The exploratory selector chooses $\gamma = 0.25$ if both models gain five points, both selected seeds improve and each game-bootstrap lower bound is positive; otherwise it tries 0.5. Quarter strength fails Gemma's gain threshold and half strength qualifies. A reserved seen-frame check on the selected seeds gives gains 7.29/8.68 points; it is not fresh-seed replication. The complete fresh-seed comparison is below. Both task frames had already informed development, so new training seeds do not make them untouched.

Table 20: Fresh ALFWorld replication: all six trainings and 36 endpoints. Counts aggregate 432 episodes per arm. Half-gain intervals are paired game-cluster bootstrap 95% intervals (10,000 draws), conditional on the checkpoints and unadjusted. Seen/unseen frames have 72/36 games with 2/4 resets per seed. Full-strength controls are retained.

Frame	Model	Base	Full	Half	Half gain, pp [95%]
Unseen	Gemma	165	204	202	$+8.56$ [$+1.85, +16.20$]
Unseen	Qwen	160	134	228	$+15.74$ [$+7.87, +23.84$]
Seen	Gemma	216	263	239	$+5.32$ [$-0.46, +11.81$]
Seen	Qwen	232	179	269	$+8.56$ [$+3.93, +13.19$]

Half strength improves every fresh seed, but the common criterion fails on seen tasks because Gemma's game interval crosses zero. The seed-t 95% intervals are $[4.97, 12.16]$ and $[12.15, 19.33]$ points for unseen Gemma/Qwen, and $[1.34, 9.31]$ and $[5.93, 11.20]$ for seen tasks. They assume independent normal paired effects; three-pair sign-flip tests still have $p = .25$. Full strength changes unseen performance by $+9.03$ points for Gemma (game interval $[-3.70, 21.53]$) and -6.02 for Qwen ($[-20.83, 8.56]$); seen changes are $+10.88$ ($[2.55, 19.68]$) and -12.27 ($[-21.30, -3.01]$). These results establish useful uniform imitation with a conservative update on ALFWorld, not a causal-weighting advantage.

WebShop transfer. The new frame has 96 training, 24 development and 144 final goals across five categories, disjoint in goal and target-product identity from one another and the known earlier registry. The catalog is shared. A privileged title-first teacher with oracle product/option choice produces 272 verified click labels from 88 training goals; search actions remain in histories but are not labels. Training uses unit weights, the same 32-by-64 update budget, temperature 1, 128 response tokens and a 15-turn horizon. Every final checkpoint is evaluated at base/full/half strength without using development outcomes for selection.

Table 21: All six WebShop trainings and 18 endpoints: three seeds, 432 episodes per model/strength. Goal-bootstrap intervals use 10,000 draws and are unadjusted. Both full and half controls remain in the comparison.

Model	Base	Full	Half	Full gain, pp [95%]	Half gain, pp [95%]
Gemma	80	107	82	$+6.25$ [3.01,9.95]	$+0.46$ [0.00,1.16]
Qwen	104	0	0	-24.07 [$-30.09, -18.28$]	-24.07 [$-30.32, -18.06$]

The fixed common half-strength rule fails, as do the separately frozen secondary rules (full Gemma/half Qwen, or full both). Qwen has zero corrected successes in every seed. A post-hoc seed-6203 diagnostic finds tagged first turns on 142/144 base episodes, 0/144 full and 1/144 half; bare commands are rejected by the unchanged evaluator. The action-only mask does not supervise those tags. This identifies a format failure, not a demonstrated remedy or success under a changed parser. Gemma's full-strength gain does not rescue the common rule.

Recorded compute. Table 22 reports 1,235.02 allocated H20 GPU-hours, including 161.70 recovered from 36 ALFWorld endpoint jobs. Source maps count reused completed jobs once. Each duration is verified against its start and completion records; endpoint manifests assign four actor and four rollout-service GPUs.

Table 22: Recorded modern-study allocation, eight GPUs per completed job. Hours are not utilization-adjusted; the subtotal excludes the categories specified below. The subtotal uses unrounded durations.

Component	Completed jobs	GPU-hours
Fresh ALFWorld SFT training	6	25.46
Calibrated-pool RL and in-job evaluation	30	979.79
WebShop training	6	34.59
WebShop endpoint evaluation	18	33.48
ALFWorld seen endpoint evaluation	18	80.09
ALFWorld unseen endpoint evaluation	18	81.60
Recorded partial subtotal	96	1,235.02

These durations include model startup and the evaluations within the completed jobs; fresh ALFWorld training includes its two 24-episode development checks. They exclude data generation, teacher verification and calibration, earlier method search, primary measurement, cross-method studies, Retail, and allocation time outside these jobs, including unstarted port failures. The subtotal is neither active-device compute nor total project compute. No utilization-based adjustment is available.

C.5 Peripheral studies and excluded evidence

These investigations constrain the claims. The table and explanation below state their material consequences; additional ledgers serve reproducibility.

Table 23: Questions, outcomes and consequences of peripheral investigations. No failed gate is relaxed and no excluded outcome is pooled into the main results.

Question / study	Outcome	Consequence
Wrapper-to-native transport	Bridge fidelity retains 2/12 GRPO and 6/14 GiGPO templates; natural-control completion is inconclusive.	Historical motivation, not general transport.
Same-seed suffix equality	One outcome and 492 suffix traces differ among 559 target-action pairs.	Coupling study excluded; fresh streams used.
Proxy substitution for live data	MAE gain 0.0117 has an interval crossing zero and misses the 0.02 rule.	No material predictive advantage.
Sequential query savings	Both streams lack a justified stopping-time guarantee; the coverage and schedule issues are explained below.	No validated anytime-efficiency claim.
Earlier capability pilot	19/24 policies pass floors but no complete panel; parser/sampling defects affect interpretation.	Not an optimizer capability limit.
External mechanism checks	Sokoban fails serving parity; two Telecom attempts each score 0/8. Scripted Blackjack shows omission sensitivity in four constructed states.	No additional LLM transport confirmation or prevalence estimate.
Alternative correction objectives	Joint RL/SDAR, probability-weighted imitation, recovery suffixes and policy replay fail their declared rules.	Negative development evidence; no selected-branch novelty claim.

Sequential experiments target a fixed checkpoint–template–arrival mixture. Point-null tests on a finite grid, even after one-step hull expansion, do not prove coverage for intermediate means. The follow-up uses composite-null cells and fresh outcomes but copies the previously inspected unit schedule, leaving the required $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \mu$ unproved (Howard et al., 2021; Waudby-Smith & Ramdas, 2024). Neither stream supports valid stopping-time savings or contributes outcomes to the primary panel. Scripted Blackjack has zero action TV but omission biases from -0.0829 to 0.3417 in four fixed states; the deliberately incomplete private states are not a legal natural-deal population.