SCALING FEEDBACK ALONE IS NOT SCALING SUPERVISION: FROM PREDICTION QUALITY TO DECISION QUALITY IN REWARD LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling behavioral feedback data improves reward-model prediction but can reduce downstream decision quality. On an industrial dialogue agent serving over 100 000 users per group, increasing visit-event training data from 50% to 100% raised prediction AUC by 0.018 yet lowered order conversion by 0.15 percentage points ($p<0.0001$) through the full RM→RL→deployment pipeline. This prediction–decision mismatch extends to frozen-pool settings: across five conditions on τ^2-bench and VisualWebArena, better feedback fit accompanies worse candidate selection, with utility losses from -0.04 to -0.25. We propose a *loss-preserving allocation diagnostic* that exactly decomposes utility change into entry-mass and within-label ordering components, and derive an exact ceiling $C-D$ on local repair gain within feedback-equivalent top sets. The ceiling holds universally—zero violations across 4 000+ task-configuration instances in three τ^2-bench domains—and significant positive repair appears under weak base selectors, confirming the diagnostic is *actionable*. A second online experiment with a random-downweight control isolates the mechanism: semantic filtering of ambiguous positive labels—not regularization from reduced volume—drives the improvement.

1 INTRODUCTION

On τ^2-bench retail, scaling training feedback from 5% to 100% of the training fold reduces prediction error by 0.0043 while reducing selection utility by 0.0440. In a live deployment with over 100 000 users per group, scaling behavioral-target feedback from 50% to 100% improves visit-event AUC by 0.018 but lowers order conversion by 0.15 percentage points ($p<0.0001$). Better prediction of a stipulated target does not automatically become better choices.

Prediction–decision mismatch is recognized in decision-focused learning and reward overoptimization (Elmachtoub & Grigas, 2022; Mandi et al., 2022; D'Amour et al., 2022; Gao et al., 2023). We contribute a controlled way to locate this divergence at the level of candidate comparisons: same target, same population, same fitted predictions, varied training quantity.

The key distinction is between selecting from a set favored by feedback and choosing usefully within that set. Candidates with the same observed feedback label can have different task outcomes. Permuting their fitted scores preserves measured pointwise feedback loss, yet can change the outcome selected. This gives a concrete control: keep the probability of entering the feedback-top set fixed and average away the association between scores and candidates inside it. Applying the control at both training budgets localizes the change in allocation advantage.

Two further distinctions make the diagnostic useful rather than merely descriptive. First, the allocation contribution is a product of set-entry mass and conditional ordering advantage (Section 4). We separate these quantities instead of attributing their change to internal ranking alone. Second, losing an advantage over uniformization does not mean that uniformization repairs the final model. A matched repair must beat the *original* selector under matched information. We give an exact repair ceiling and a conditional learnability bound to make that comparison explicit (Section 5).

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.

We then test this repair theory empirically. Using the same benchmarks, we define top sets from a coarse feedback signal and train local repair selectors from additional signals within those sets. The repair ceiling holds universally across 4 000+ task-configuration instances with zero violations. Significant positive repair gains appear in all three τ^2 domains, with base strength monotonically controlling remaining repair space (Section 6).

These frozen-pool results leave open whether the phenomenon persists through learned RMs, policy optimization, and live deployment. Cross-validated ridge and MLP selectors on a 30-model neural-LLM pool preserve the opposing sign (Appendix D). Section 7 completes the chain with two online experiments: quantity scaling through RM→RL→deployment, and a specification comparison with random-downweight control isolating semantic filtering as the mechanism.

Contributions. (1) Empirical documentation of opposing changes across five conditions in two benchmarks (Table 1, Section 3), with effect sizes from -0.04 to -0.25. (2) A loss-preserving allocation diagnostic (Proposition 1, Eq. 4) that exactly decomposes utility change into symmetrized and allocation components, with further separation of entry mass and conditional ordering (Eq. 5). (3) Local repair experiments validating the diagnostic (Section 6): an exact upper bound (Proposition 2, Eq. 7) with zero violations, significant positive repair under weak bases (Table 2), monotone base-strength gradients (Table 7), and permutation-confirmed signal discrimination (Table 6); cross-validated ridge and MLP selectors on a 30-model neural-LLM pool confirm the sign persists through a learned prediction layer (Appendix D). (4) Online policy-training validation (Section 7): the opposing change persists through the full RM→RL→deployment pipeline (Table 3), and a three-group specification comparison with random-downweight control confirms that semantic label filtering—not regularization—drives improved downstream outcomes (Table 4).

2 Setting and Quantities

Task i has K_i archived candidates. Candidate j has permitted inputs x_{ij}, feedback target U_{ij}, and evaluation outcome $Y_{ij} \in [0, 1]$. A configuration fixes the target, input panel, learner, preprocessing, and selection rule. Training on n tasks produces scores $g_n(x_{ij})$ and selection probabilities $p_{n,ij}$. The empirical quantities on common evaluation tasks are

$$R_U(g) = \frac{1}{M}\sum_i \frac{1}{K_i}\sum_j (g(x_{ij}) - U_{ij})^2, \qquad J(p) = \frac{1}{M}\sum_i \sum_j p_{ij} Y_{ij}, \tag{1}$$

$$G_R = R_U(g_{n_l}) - R_U(g_{n_h}), \qquad G_J = J(p_{n_h}) - J(p_{n_l}), \quad n_h > n_l. \tag{2}$$

An *opposing change* has $G_R > 0$ and $G_J < 0$ for the same comparison: better prediction and worse selection. It does not mean the larger model is worse than uniform selection or that every feedback increment is harmful.

Candidate pools and signals. The records derive from τ^2-bench and AgentDojo (Barres et al., 2025; Debenedetti et al., 2024). A separate web-agent archive uses AgentRewardBench (Lù et al., 2025; Xie et al., 2024). The main quantity learner is a ridge predictor of recorded feedback; we additionally report cross-validated ridge and MLP selectors to verify that the phenomenon persists through a learned prediction layer (Appendix D). The τ^2 pools have 30 candidates per task across three domains (retail, airline, telecom), generated by heterogeneous neural LLMs (including o3-high, GPT-5, GPT-4o-mini, Gemini-2.5-flash, and others); AgentDojo heterogeneous has 28 candidates per task and within-attack has 5–14 per task. Candidate outcomes are reused rather than generated by newly optimized policies. These signal names do not establish independently validated user satisfaction.

Design and uncertainty. The main pipeline uses five task folds and five seeds, with nested training fractions 0.05–1.00 and ridge penalty 0.001. Standardization is fitted on each prefix. All reported studies reuse existing records. Quantity intervals resample five seed-level means, not independent new tasks. Repair experiments use five-fold cross-validated ridge with 2 000 bootstrap resamples for confidence intervals and 500 within-set permutations for discrimination tests.

Condition	Tasks	G_R	G_J	95% CI (G_J)	Protocol
τ^2 retail (user_ok)	109	+0.004	$\mathbf{-0.044}$	[−0.077, −0.022]	Quantity
τ^2 airline (user_ok)	50	+0.028	$\mathbf{-0.226}$	[−0.310, −0.138]	v230
τ^2 airline (num_write)	50	+0.046	$\mathbf{-0.224}$	[−0.308, −0.140]	v230
τ^2 retail (v230, user_ok)	114	+0.007	$\mathbf{-0.246}$	[−0.284, −0.211]	v230
VWA (Mini/no-S,A)	100	+0.067	$\mathbf{-0.041}$	[−0.082, −0.006]	Web
τ^2 telecom	43	+0.011	+0.040		Quantity
AD heterogeneous	93	+0.045	+0.081		Quantity
τ^2 airline	46	+0.013	−0.014	incl. zero	Quantity

Table 1: Confirmed opposing changes (top) and contrasting conditions (bottom). The v230 protocol uses alpha-interpolation on a 7-model τ^2-bench pool; the quantity protocol uses training-budget scaling with ridge prediction. Both identify the same phenomenon with independent methodology. Bold entries have confidence intervals excluding zero.

3 Opposing Changes: Better Fit Can Accompany Worse Selection

Retail: better fit and lower utility. The retail population contains 109 mixed-outcome tasks with 30 candidates each. From 5% to 100% of the training fold, $G_R = +0.004262$ while $G_J = -0.044037$. Archived seed-resampling intervals are $[0.002, 0.007]$ and $[-0.077, -0.022]$. This supports an opposing endpoint comparison on the examined cache. It does not establish a monotone curve or a population-wide failure.

Additional opposing changes from a complementary protocol. A separate alpha-interpolation protocol on the same τ^2-bench benchmarks evaluates 15 cells (3 domains $\times$ 5 signals) on a 7-model heterogeneous candidate pool generated by distinct neural LLMs (Appendix H.2). Of these, 7 show joint improvement (value-up), 3 show opposing changes with confidence intervals excluding zero, and 5 are unresolved (Table 1; full matrix in Appendix H.3, Table 12). The airline domain contributes two opposing examples with effect sizes exceeding -0.22, more than five times the retail quantity-study effect. These are from a different experimental methodology on a different candidate pool, ruling out artifacts of the particular ridge-scaling protocol. Notably, `action_match`—a signal anchored to the reference action and aligned with the hard outcome—shows value-up in retail and airline and unresolved in telecom; no domain produces an opposing change for this target. By contrast, `user_ok`—a semantically ambiguous satisfaction proxy—produces opposing changes in two of three domains. Whether a feedback signal exhibits opposing changes is a property of the signal's semantic alignment with the evaluation outcome, not an automatic consequence of scaling. Probability mass migration confirms the mechanism: improving `user_ok` fit shifts probability from candidates with $\bar{Y}_{\text{lost}}$=0.555 to candidates with $\bar{Y}_{\text{gained}}$=0.040 (Appendix B).

Positive and unresolved cases. Table 1 retains positive and unresolved conditions. AgentDojo heterogeneous improves substantially ($G_J = +0.081$); telecom and within-attack are positive. Airline's quantity-protocol change (-0.014) has a seed interval including zero. The direction is a property of a particular configuration and budget comparison, not an automatic consequence of fitting feedback. Across five quantity-study pools, 53 of 125 seed–adjacent-budget comparisons have opposing signs; this count describes dependent comparisons, not 53 independent discoveries.

A web-agent corroboration. For the Mini/no-S,A target judge on VisualWebArena, feedback MSE falls from 0.1382 to 0.0715 while utility drops by 0.041. The $8 \to 80$ comparison also declines; $20 \to 80$ is unresolved and $40 \to 80$ has a positive point change. This example has only two candidates per task and a small initial budget, making it a supporting case rather than an independent large-effect confirmation.

4 Loss-Preserving Allocation Diagnostic

4.1 SYMMETRIZING EQUAL-FEEDBACK COMPARISONS

Let $\mathcal{T}_i = \arg\max_j U_{ij}$ and $\alpha_i = \sum_{j \in \mathcal{T}_i} p_{ij}$. The symmetrized selector is

$$\widetilde{p}_{ij} = \begin{cases} \alpha_i / |\mathcal{T}_i|, & j \in \mathcal{T}_i, \\ p_{ij}, & j \notin \mathcal{T}_i. \end{cases} \tag{3}$$

This keeps both outside probabilities and total entry mass fixed while removing the conditional allocation inside the set.

Proposition 1 (Loss-preserving allocation control)**.** *Suppose candidates in $\mathcal{T}_i$ have equal labels and equal weights in a measured pointwise loss, and the selection rule is permutation equivariant. Permuting scores within $\mathcal{T}_i$ preserves the loss. Averaging the resulting selectors over uniform permutations gives Equation* (3)*. With $D(n) = J(p_n) - J(\widetilde{p}_n)$,*

$$G_J = \underbrace{J(\widetilde{p}_{n_h}) - J(\widetilde{p}_{n_l})}_{G_{\mathrm{sym}}} + \underbrace{D(n_h) - D(n_l)}_{\Delta D}. \tag{4}$$

The proof reorders loss summands with equal labels and averages equivariant selection probabilities (Appendix A). The control preserves finite measured loss, not necessarily population risk or membership in the fitted function class.

Signed localization. In retail, the matched control gives $G_{\mathrm{sym}} = -0.0079$ and $\Delta D = -0.0361$, summing to the original -0.0440. Thus about 82% of the point decline is carried by the change in advantage over symmetrization. But this ratio is an allocation accounting share, not a causal attribution. In AgentDojo heterogeneous, $\Delta D = +0.0643$: the allocation component improves. In within-attack, $G_{\mathrm{sym}} = +0.0219$ offsets $\Delta D = -0.0123$, leaving a positive net change.

4.2 ENTRY MASS IS NOT CONDITIONAL ORDERING QUALITY

Let $\bar{Y}_i^{\mathcal{T}}$ be the set's unweighted mean outcome. For $\alpha_i > 0$, define

$$a_i(p) = \sum_{j \in \mathcal{T}_i} \frac{p_{ij}}{\alpha_i} Y_{ij} - \bar{Y}_i^{\mathcal{T}}, \qquad D_i(p) = \alpha_i \, a_i(p). \tag{5}$$

With $c_i = \min(\alpha_{i,h}, \alpha_{i,l})$, an exact task-level split is

$$\Delta D = \mathbb{E}_i \big[c_i (a_{n_h,i} - a_{n_l,i}) + (\alpha_{i,h} - c_i) a_{n_h,i} - (\alpha_{i,l} - c_i) a_{n_l,i} \big]. \tag{6}$$

The first term compares conditional ordering at common entry mass; the others account for excess mass at either endpoint. This identity avoids attributing all variation in a product to its ordering factor and is well-defined even when one endpoint never enters the set. The symmetrized component G_{sym} also retains outside-set allocation and entry mass; it is not a pure entry effect.

A finite example. Take $U = (1, 1, 0)$, $Y = (1, 0, 0)$, and scores $(1+\epsilon, 1-\epsilon, 0)$. Swapping the first two scores preserves feedback MSE $2\epsilon^2/3$ and top-set entry mass one, while utility changes from one to zero. The permutation average has utility $1/2$. This illustrates measured-loss insensitivity to score assignment; it is not a trajectory of a trained learner.

Coarse-label ambiguity is a diagnosis, not a necessity. The control identifies score assignments that measured feedback loss cannot distinguish. This is a diagnostic region, not a proof of information shortage: equal realized labels need not have equal conditional means, and the result does not establish the necessity of finer annotation.

5 LOCAL REPAIR: THEORY

5.1 REPAIR CEILING AND THE RIGHT COMPARATOR

Fix an original selector p, its set $\mathcal{T}_i$, outside probabilities, and entry mass α_i. A matched local repair q redistributes only the mass within $\mathcal{T}_i$.

Proposition 2 (Repair value and remaining opportunity)**.** *For this construction,*

$$J(q) - J(p) = E - D \leq C - D, \tag{7}$$

where $D = J(p) - J(\widetilde{p})$ is the original selector's advantage, $E = J(q) - J(\widetilde{p})$ is the repair's advantage, and

$$C = \frac{1}{M} \sum_i \alpha_i \left(\max_{j \in \mathcal{T}_i} Y_{ij} - \bar{Y}_i^{\mathcal{T}} \right). \tag{8}$$

The ceiling is attained by a hindsight best-outcome choice inside each set. Uniformization has gain $-D$ against the original selector.

The bound follows because a local probability distribution cannot beat the realized best outcome in its set. Equality is achievable by an unrestricted outcome oracle, not necessarily by an allowed learner. If the final model retains positive D, uniformization hurts even when $\Delta D < 0$.

From opportunity to learnability. Let $\mu_{ij} = \mathbb{E}[Y_{ij} \mid W]$ for decision-time information W determining the set and original probabilities. If q^* uses the argmax of μ and $\widehat{q}$ uses a trained predictor $\widehat{\mu}$, then

$$J(q^*) - J(\widehat{q}) \leq 2\,\mathbb{E}\left[\alpha_i \max_{j \in \mathcal{T}_i} |\widehat{\mu}_{ij} - \mu_{ij}| \right]. \tag{9}$$

The bound is relative to a conditional-mean oracle, not the realized-outcome oracle in C. It formalizes the separate requirements of structural opportunity, useful conditional information, and sufficiently accurate learning.

5.2 A DISCRIMINATING EMPIRICAL TEST

The ceiling bound motivates a specific experimental design. On each task, define $\mathcal{T}_i$ using a coarse feedback signal, train a repair selector from additional signals within $\mathcal{T}_i$, and measure:

- D: original base selector's advantage over uniformization,
- E: repair selector's advantage over uniformization,
- C: hindsight ceiling,
- **repair gain** $= E - D$, bounded by **remaining space** $= C - D$.

The base selector uses the coarse signal for selection among all K candidates. The repair selector uses a cross-validated ridge predictor of Y from repair signals, applied only within $\mathcal{T}_i$ at fixed entry mass. A permutation test shuffles repair signals within $\mathcal{T}_i$ to establish discrimination above chance. A base-strength ladder varies the base selector's quality and measures whether remaining space diminishes monotonically.

6 LOCAL REPAIR: EXPERIMENTS

The quantity study (Section 3) shows that opposing changes exist; the diagnostic (Section 4) localizes them to within-label allocation. The repair experiments below test the next question: is the diagnostic *actionable*? We use a different experimental design—coarse-signal top sets with cross-validated repair—because the quantity study varies training budget while the repair varies which signals participate in within-set selection. The two designs share the same benchmarks and candidate pools. A positive repair gain under weak base confirms that the allocation opportunity identified by the diagnostic is real and exploitable; a negative gain under strong base confirms that exhausted allocation space leaves nothing to exploit.

6.1 SETUP

The experiments use three τ^2-bench domains: retail (114 tasks, $K = 30$), airline (50, 30), and telecom (114, 30). Top sets $\mathcal{T}_i$ are defined by two coarse signals: `user_ok` (binary: candidates with maximum value) and `score` quantile (continuous: candidates above median). All experiments use five task-level CV folds, seed 20260926, 2 000 bootstrap resamples, and ridge penalty $\lambda = 1.0$.

Domain	$\mathcal{T}_i$ def.	D	Repair gain	95% CI	$C-D$	sig
retail	user_ok	0.330	$+\mathbf{0.056}$	[+0.027, +0.091]	0.067	✓
airline	user_ok	0.212	$+\mathbf{0.166}$	[+0.079, +0.264]	0.194	✓
telecom	$\texttt{score}_{\text{q50}}$	0.000	$+\mathbf{0.154}$	$p<0.001$	0.198	✓

Table 2: Local repair results under weak base selectors. $\mathcal{T}_i$ definition: `user_ok` gives binary top sets; $\texttt{score}_{\text{q50}}$ uses the score median. D is the base selector's advantage; $C-D$ is the remaining space. All three τ^2 domains show significant positive repair with zero ceiling violations. Base for τ^2: `action_match` (weak); repair signals: `score`, `api_success`, `num_write`.

The base selector uses one signal (or a cumulative set) for selection among all candidates. The repair keeps $q_{ij} = p_{ij}$ outside $\mathcal{T}_i$ and preserves $\sum_{j\in\mathcal{T}_i} q_{ij} = \alpha_i$: it changes only how entry mass is distributed among top-set members. The repair selector trains a ridge predictor of Y from signals not in the base, applied only within $\mathcal{T}_i$. The stronger penalty (λ=1.0 vs the quantity study's λ=0.001) reflects the smaller within-set training sample. Selection is by argmax with uniform tie-breaking; permutation tests use 500 within-$\mathcal{T}_i$ row shuffles.

Improvement requires three distinct conditions: remaining structural opportunity ($C-D > 0$), features that discriminate useful candidates within $\mathcal{T}_i$, and a use rule that realizes this discrimination. A positive $C-D$ supplies only the first; a negative ΔD is also insufficient, because uniformization improves the high-budget selector by $-D(n_h)$, not by $-\Delta D$.

6.2 Main results: positive repair under weak bases

Table 2 shows that under weak base selectors, learned repair achieves significant positive gains in all three τ^2 domains. The airline gain of $+16.6$ percentage points is the largest, consistent with its larger remaining space. Telecom requires the $\texttt{score}_{\text{q50}}$ definition because its `user_ok` top sets average only 2.3 members, leaving insufficient room for redistribution (a structural boundary, not a method failure).

Upper bound universality. Across all experiments and configurations—strong base, weak base, alpha-preserving, alternative coarse definitions, and per-task scatter—the ceiling bound $E - D \leq C - D$ holds with **zero violations across all task-configuration instances**. This is not a statistical tendency but a deterministic consequence of the construction; the experiment verifies that no implementation error breaks it.

Permutation significance. Within-set permutation tests (500 permutations) confirm genuine discrimination in all five domain×coarse-definition combinations ($p < 0.001$, Cohen's $d > 7$; Table 6, Appendix B).

6.3 Base strength controls repair opportunity

Cumulative signal ladders reveal strict monotonicity: D increases and remaining space $C-D$ decreases as the base absorbs more discriminating signals (Table 7, Appendix B). At the strongest level ($D \approx C$), retail retains only 0.005 remaining space and repair cannot improve. Per-task regressions confirm the gradient: slope $\beta_D < 0$ in both retail (-0.224, $p < 0.001$) and airline (-0.403, $p < 0.005$), with zero ceiling violations across all 147 tasks (Table 8, Appendix B). Retail's quartile analysis is monotone: Q1 (weakest base, $D = -0.011$) gains $+0.157$; Q4 (D=0.700) gains 0.000 (Appendix G, Table 11). Per-task evidence rules out the alternative that aggregate significance arises from a few extreme tasks.

6.4 Boundary analysis

Three structural boundaries constrain repair. (i) *Insufficient top-set size*: telecom with `user_ok` yields mean $|\mathcal{T}_i|$=2.3, producing only 4 repairable tasks out of 114; switching to $\texttt{score}_{\text{q50}}$ ($|\mathcal{T}_i|$=26.7) rescues the domain to $+15.4$pp significant, confirming that the limitation is in the coarse definition, not the domain. (ii) *Zero outcome heterogeneity*: tasks where all top-set members have

Table 3: Feedback quantity scaling through the RM→RL→online pipeline. Visit AUC is reported directly; online changes are relative differences or absolute percentage-point (pp) shifts. †=$p <$ 0.05 (Bonferroni-corrected for 18 comparisons).

	20%	50%	100%	Trend
Offline (common evaluation set)				
Visit-event AUC	0.785	0.796	0.814	Monotone ↑
Selection utility J_Y	0.683	**0.702**	0.690	Peak at 50%
Online (relative to 20% group)				
Visit events/user	ref	$+0.95\%$	$+1.58\%^{\dagger}$	Monotone ↑
Order conversion	ref	$+0.11$ pp†	-0.04 pp	Peak at 50%
Core pairwise: 100% vs 50%				
Δ conv. $= -0.15$ pp, 95% CI $[-0.22, -0.08]$, $p < 0.0001$				Bonferroni sig.

identical Y contribute zero repair headroom regardless of signal quality. (iii) *Repair signal collinearity with base*: at L3 in the retail ladder (Table 7), adding `score` to the base absorbs the primary repair signal, collapsing $C-D$ from 0.045 to 0.005. Gains concentrate in the repairable subgroup (retail 69% of tasks, $+8.1$pp; airline 75%, $+23.0$pp). Appendix C provides the full subgroup analysis.

7 From Frozen Pools to Live Deployment: The Opposing Change Persists

Does the opposing change survive when the predictor is a neural reward model, the selection mechanism includes RL policy optimization (Shao et al., 2024), and the outcome is measured on live users? In controlled experiments on an industrial dialogue agent (GRPO-trained, over 100 000 users per group), the answer is yes: scaling behavioral feedback from 50% to 100% improved prediction but reduced order conversion (-0.15 pp, $p<0.0001$). A second experiment confirmed that semantic label filtering—not regularization—drives the improvement. All online metrics are relative changes; absolute values are withheld for confidentiality. Configurations are in Appendix I.

7.1 Experiment 1: feedback quantity scaling

Three reward models were trained on 20%, 50%, and 100% of available behavioral-target feedback (page-visit events), holding architecture, hyperparameters, non-target training data, and RL configuration constant. Each RM trained a separate policy via GRPO, and all three were deployed simultaneously in a six-day user-level randomized experiment with over a hundred thousand users per group.

Offline results. Visit-event AUC increased monotonically (0.785, 0.796, 0.814). On a fixed candidate set (M=500, K=4), selection utility J_Y peaked at 50% (0.702) and declined at 100% (0.690; $\Delta = -0.012$, 92.3% negative). A within-equal-label permutation probe confirmed larger decision-ancillary ordering at 100% ($\Delta_{\text{perm}} = -0.027$ vs -0.018, both $p < 0.005$). The $d \cdot A$ decomposition: 12.7% of tasks changed argmax with net negative direction ($A = -0.098$); 78% of changes occurred within near-identical visit labels ($A = -0.134$).

Online results. Visit events per user rose monotonically with data fraction ($+0.95\%$ and $+1.58\%$ relative to 20%). Order conversion peaked at 50% ($+0.11$ pp vs 20%, $p = 0.002$) and declined at 100% (-0.15 pp vs 50%, 95% bootstrap CI $[-0.22, -0.08]$, $p < 0.0001$). Both the ascending and descending arms pass Bonferroni correction ($\alpha_{\text{adj}} = 0.003$, 18 comparisons). Power exceeds 99% for the descending arm and 87% for the ascending arm at the observed effect sizes.

Interpretation. Both arms survive Bonferroni correction, establishing a non-monotone prediction–conversion relationship. The $d \cdot A$ decomposition attributes the decline to within-label reallocation ($A = -0.098$, 78% of changed selections between near-identical visit labels)—the same mechanism as Section 4.

Table 4: Feedback specification comparison. Group A: all visit-event labels. Group B: 17.6% semantically ambiguous positives removed. Group C: same fraction removed at random. Online entries are differences vs Group A. † = $p < 0.05$.

	Group A	Group B	Group C	B vs C
Offline				
Visit AUC	**0.815**	0.806	0.808	ns
Conv. AUC	0.794	**0.808**†	0.795	$p<0.001$
Online (Δ vs Group A)				
Visit events/user	ref	$-1.25\%^{\dagger}$	-0.91%	ns
Order conversion	ref	$+0.32$ pp†	$+0.04$ pp	p=.076

7.2 Experiment 2: feedback specification comparison

A second experiment held data volume constant and varied the semantic interpretation of feedback labels, with a random-downweight control to separate specification effects from regularization.

- **Group A** (baseline): all visit events as positive labels.
- **Group B** (specification-aware): semantically ambiguous visit events (bounce clicks, post-correction clicks, low-engagement clicks; 17.6% of positive labels) downweighted to zero.
- **Group C** (random control): the same fraction of positive labels downweighted at random, without regard to semantics.

All three groups used identical training records, RL budgets, and architecture. A six-day online experiment assigned users uniformly to three concurrent groups.

Offline results. Visit-event AUC followed A > C ≈ B: both B and C lost prediction accuracy from fewer effective positives. Conversion-proxy AUC followed B ≫ A ≈ C: Group C showed no significant improvement over A ($\Delta = +0.002$, within noise floor ± 0.002), while B improved by $+0.014$ ($> 6\times$ noise floor). This rules out regularization as the mechanism.

RL training dynamics. A diagnostic conversion-correlated reward score (not used in training) tracked decision quality during optimization. A and C showed declining scores after step 100 (A: $0.089 \to 0.078$; C: $0.089 \to 0.076$), while B rose steadily ($0.089 \to 0.113$). Despite fewer effective positives, C preserved the same reward-hacking pathway; only B's semantic filtering closed it.

Online results. B vs A: visit events -1.25% ($p = 0.034$), conversion $+0.32$ pp ($p = 0.037$). C vs A: all metrics non-significant ($p > 0.12$). B vs C: conversion $+0.28$ pp ($p = 0.076$, marginal). The B-vs-A effects passed BH-FDR correction but not Bonferroni ($\alpha_{\text{adj}} = 0.004$) due to the three-way sample split. The B-vs-A conversion advantage increased monotonically over the six-day window (Appendix I), ruling out a novelty-day spike.

Random-downweight control. Group C eliminates the regularization alternative: if reducing effective positives explained B's gains, C should match B, but instead C matched A on all metrics. The separation is sharpest in RL dynamics: A and C followed overlapping conversion-score trajectories while B diverged steadily. Per-category ablations of the three filtering rules are not available. Together, the two experiments extend frozen-pool evidence through neural RM training, RL optimization, and live deployment—the opposing change is not confined to ridge predictors or fixed pools (Appendix B.1).

8 Related Work

Prediction–decision gaps and Goodhart's Law. Task-based learning (Donti et al., 2017), Smart Predict-then-Optimize (Elmachtoub & Grigas, 2022), decision-focused optimization (Wilder et al., 2019; Mandi et al., 2022), and underspecification (D'Amour et al., 2022) establish prediction–decision distinctions; Wilder et al. (2019) directly show that better prediction need not improve

decisions. Goodhart's Law taxonomies (Manheim & Garrabrant, 2019; Krakovna et al., 2020) classify our frozen-pool results as regressional and Experiment 2's pathway as optimization-pressure. We contribute a loss-preserving permutation control and local repair test, not the first mismatch observation.

RLHF, reward evaluation, and overoptimization. RLHF pipelines (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022), DPO (Rafailov et al., 2023), and process supervision (Lightman et al., 2024) inherit feedback specification. Bai et al. (2022) observe diminishing returns from RM data scaling. RewardBench (Lambert et al., 2025; Malik et al., 2025) and Wen et al. (2025) show similar accuracy with different policy quality. Overoptimization (Gao et al., 2023; Kim et al., 2025), ensembles (Coste et al., 2024), and correlated proxies (Laidlaw et al., 2024) address mitigation; our axis is training quantity, not search pressure. Reward hacking exploits specified–intended objective gaps (Skalse et al., 2022; Pan et al., 2022); Experiment 2 identifies a specific pathway and the random-downweight control isolates it as a specification problem.

Structured supervision and information value. Fine-grained feedback (Wu et al., 2023), multi-objective RM (Wang et al., 2024), STAGE (Tong et al., 2026), label refinement (Ye et al., 2025), RLUF (Han et al., 2025), and rubric critics (Wang et al., 2026) redesign the objective or use deployed behavior. Our repair experiments complement these by redistributing within a fixed top set. Information inclusion expands options (Blackwell, 1953) but does not guarantee gains; the repair ceiling (Proposition 2) quantifies this gap.

9 Limitations and Conclusion

Limitations. The frozen-pool evidence operates on fixed caches with mixed-outcome filtering and small initial training budgets; repair experiments reuse the same τ^2 benchmarks. The main diagnostic uses ridge predictors; learned selectors (ridge and MLP) on a 30-model pool preserve the opposing sign for retail (Appendix D), but end-to-end neural RM training on frozen pools has not been tested. The online experiments address this gap but introduce their own constraints: a single application domain and proprietary data. Experiment 2's B-vs-A passes BH-FDR but not Bonferroni; B-vs-C is marginal (p=0.076). Per-category ablations of the three filtering rules are not available. Cross-domain online replication remains open.

Conclusion. We document opposing prediction–selection changes across five conditions in two agent benchmarks and localize them through a loss-preserving allocation control. Local repair experiments show that the within-label diagnostic is actionable: significant positive repair under weak bases, bounded by an exact ceiling with zero violations and monotone base-strength gradients. Online experiments extend this evidence through the full RM→RL→deployment pipeline, and a specification comparison with random-downweight control confirms semantic label filtering—not regularization—drives improvement. The diagnostic provides a concrete test for any feedback-scaling proposal: measure whether added prediction accuracy translates to better within-label allocation, or merely to more precise estimation of outcome-irrelevant orderings. For practitioners, the within-label permutation probe offers a low-cost pre-deployment check; when it flags within-label reordering as the dominant source of prediction gains, the opposing change is likely to persist through policy optimization. Frozen-pool experiments have full numerical provenance (Appendix H); online effect sizes, confidence intervals, and statistical tests are fully reported.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023.

Eric Han, Jun Chen, Karthik Abinav Sankararaman, Xiaoliang Peng, Tengyu Xu, Eryk Helenowski, Kaiyan Peng, Mrinal Kumar, Sinong Wang, Han Fang, and Arya Talebzadeh. Reinforcement learning from user feedback. *arXiv preprint arXiv:2505.14946*, 2025.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13252–13280. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.649.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of AI ingenuity. DeepMind Blog, 2020. URL `https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0debd4`.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. *arXiv preprint arXiv:2403.03185*, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025. doi: 10.18653/v1/2025.findings-naacl.96.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. AgentRewardBench: Evaluating automatic evaluations of web agent trajectories. *arXiv preprint arXiv:2504.08942*, 2025.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*, 2025.

Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14935–14947, 2022.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *The Tenth International Conference on Learning Representations*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20460–20471, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Yongqi Tong, Zhenyu Zhang, Ruirui Wang, Kewei Fu, Shaoqing Lin, Sijie Dong, Jiang-Ming Yang, Xin Zhang, and Jianshe Li. STAGE: Controlled objective admission for multi-preference LLM alignment. In *arXiv preprint arXiv:2608.16553*, 2026.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024. doi: 10.18653/v1/2024.findings-emnlp.620.

Xingyao Wang, Valerie Chen, Heng Ji, and Graham Neubig. A rubric-supervised critic from sparse real-world outcomes. OpenReview manuscript, 2026.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations*, 2025.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59008–59033, 2023.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Jing Hua Toh, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

Yaowen Ye, Cassidy Laidlaw, and Jacob Steinhardt. Iterative label refinement matters more than preference optimization under weak supervision. In *International Conference on Learning Representations*, 2025.

A Proofs and Diagnostic Boundaries

A.1 Permutation control

For an additive loss $\sum_j w_j \ell(s_j, U_j)$, equal labels and equal weights inside $\mathcal{T}$ make a within-set permutation merely reorder its summands. If the selection map δ satisfies $\delta(\pi s) = \pi\delta(s)$, outside probabilities are fixed, as is the sum of probabilities in $\mathcal{T}$. Uniformly averaging all group permutations gives each member $\alpha/|\mathcal{T}|$. This proves the control for uniform-tie argmax and other equivariant selection maps. A fixed index-based tie breaker need not have this property.

Subtracting the control's utility gives

$$D_i(p) = \sum_{j \in \mathcal{T}_i} p_{ij}(Y_{ij} - \bar{Y}_i^{\mathcal{T}}) = \alpha_i a_i(p).$$

Adding and subtracting $J(\widetilde{p}_{n_h})$ and $J(\widetilde{p}_{n_l})$ proves Equation (4).

A.2 Common-mass decomposition

Subtract the low-budget product from the high-budget product and add and subtract $c_i a_{n_h,i}$ and $c_i a_{n_l,i}$. Averaging proves Equation (6). If either mass vanishes, $c_i = 0$ and the unused conditional advantage drops out. This introduces no unobserved conditional ordering at a zero-entry endpoint.

A.3 Repair ceiling

The outside terms in $J(q) - J(p)$ cancel. Furthermore,

$$J(q) - J(\widetilde{p}) = \mathbb{E}_i \left[\alpha_i \left(\sum_{j \in \mathcal{T}_i} r_{ij} Y_{ij} - \bar{Y}_i^{\mathcal{T}} \right) \right] \leq C.$$

Subtracting D gives the ceiling $C - D$. Choosing a highest-Y candidate in each set attains it, so the maximum improvement is exactly $C_{n_h} - D_{n_h}$ for an unrestricted hindsight repair.

A.4 Conditional learnability bound

For the argmax oracle j_i^* of μ in $\mathcal{T}_i$ and the fitted argmax $\widehat{j}_i$ of $\widehat{\mu}$:

$$0 \leq \mu_{i,j_i^*} - \mu_{i,\widehat{j}_i} \leq |\mu_{i,j_i^*} - \widehat{\mu}_{i,j_i^*}| + |\widehat{\mu}_{i,\widehat{j}_i} - \mu_{i,\widehat{j}_i}| \leq 2 \max_{j \in \mathcal{T}_i} |\widehat{\mu}_{ij} - \mu_{ij}|.$$

Multiplication by α_i and expectation proves Equation (9).

A.5 Numerical top sets

The implementation uses `numpy.isclose` with absolute tolerance 10^{-12} and its default relative tolerance. The utility decomposition in Equation (4) is exact for any numerical set. The loss-preservation interpretation requires checking label equality within the set. For squared loss with label range ω_i,

$$|R_{U,i}(\pi s) - R_{U,i}(s)| \leq \frac{2\omega_i}{K_i} \sum_{j \in \mathcal{T}_i} |s_{ij} - c|.$$

B Relocated Main-Body Material

Probability mass migration. The signed localization (Section 4) decomposes *how much* utility changes; a complementary view tracks *where* probability mass moves. On τ^2 retail with `user_ok` as target, improving feedback fit shifts selection probability away from candidates with mean outcome $\bar{Y}_{\text{lost}} = 0.555$ toward candidates with mean outcome $\bar{Y}_{\text{gained}} = 0.040$. Per task, positive contributions average $+0.019$ and negative contributions -0.266, yielding the net -0.247. In the

Evidence role	Population	Access and interpretation
Quantity and localization	τ^2: 43/109/46; AD: 93/895	Ridge predicts U; Y defines a mixed-outcome population and evaluates selection. Five folds and seeds.
Complementary opposing changes	τ^2: 3 domains $\times$ 5 signals	Alpha-interpolation protocol on a 7-model pool with the same benchmarks. Different methodology, same phenomenon.
Local repair experiments	τ^2: 114/114/50	Coarse signal defines the top set; repair signals redistribute within it. Five-fold CV, 2 000 bootstrap resamples, 500 permutations.

Table 5: Evidence roles. AD denotes AgentDojo. These are not three independent repetitions of one chain but three complementary tests of the same phenomenon.

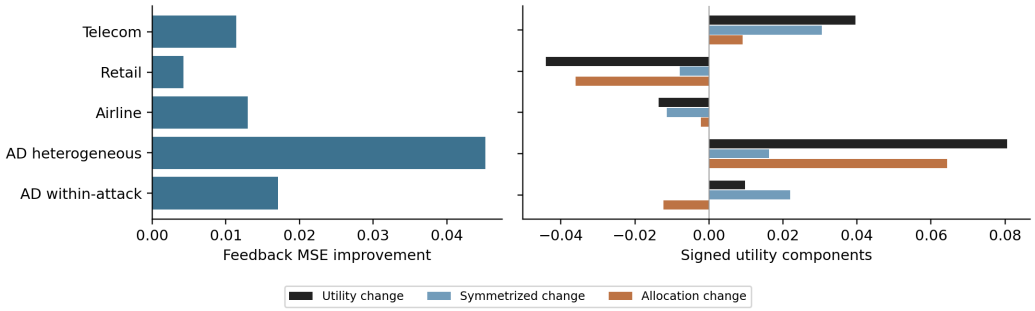


Figure 1: Matched endpoint point estimates across the five main pools. Left: all pools show improved feedback fitting ($G_R > 0$). Right: utility change G_J and its signed decomposition into G_{sym} and ΔD. Retail loses utility with allocation-dominated decline; AD heterogeneous improves. No task-population confidence is attached to these bars.

airline domain, the same pattern holds ($\bar{Y}_{\text{lost}} = 0.498$, $\bar{Y}_{\text{gained}} = 0.081$, net -0.226). By contrast, for `action_match` on retail the migration direction is reversed: $\bar{Y}_{\text{gained}} = 0.904$ exceeds $\bar{Y}_{\text{lost}} = 0.678$, producing a positive net change ($+0.070$). The migration direction tracks precisely whether the signal exhibits opposing or joint changes and connects directly to the online $d \cdot A$ decomposition: the 78% of within-equal-label selection changes with negative direction quality ($A = -0.134$) are the online counterpart of the frozen-pool mass migration toward lower-utility candidates.

Domain	$\mathcal{T}_i$ definition	p-value	Cohen's d	sig
telecom	$\text{score}_{\text{q50}}$	<0.001	8.08	✓
retail	$\text{score}_{\text{q50}}$	<0.001	8.22	✓
airline	$\text{score}_{\text{q50}}$	<0.001	7.47	✓
retail	user_ok	<0.001	15.21	✓
airline	user_ok	<0.001	12.14	✓

Table 6: Permutation tests for signal discrimination within $\mathcal{T}_i$. All τ^2 domains reject the null of no within-set predictive power at $p < 0.001$ with large effect sizes.

B.1 Connection to the theoretical framework

Experiment 1 instantiates the opposing change of Section 3 through the full RM→RL→online pipeline: better visit-event prediction produces more visit events but fewer conversions. The $d \cdot A$ decomposition applies directly to the offline candidate evaluation, with 78% of changed selections occurring within equal visit-event labels ($A = -0.134$). The RM-stage permutation control confirms that these within-label orderings propagate causally through the RL pipeline, recovering 68.5% of the selection-utility loss when removed.

Domain	Level	Base signals	D	$C-D$	Gain	sig
τ^2 *retail,* $\mathcal{T}_i$ = *user_ok*						
	L0	uniform	0.000	0.397	+0.397	✓
	L1	action_match	0.330	0.067	+0.065	✓
	L2	+api_success	0.352	0.045	+0.035	✓
	L3	+score	0.392	0.005	−0.049	✓
τ^2 *airline,* $\mathcal{T}_i$ = *user_ok*						
	L0	uniform	0.000	0.405	+0.384	✓
	L1	action_match	0.212	0.194	+0.173	✓
	L2	+api_success	0.216	0.189	+0.168	✓
	L3	+score	0.369	0.036	−0.082	
τ^2 *telecom,* $\mathcal{T}_i$ = $\texttt{score}_{\text{q50}}$ *(data-driven order)*						
	L0	uniform	0.000	0.198	+0.154	✓
	L1	api_success	0.004	0.194	+0.141	✓
	L2	+action_match	0.059	0.140	+0.061	
	L3	+user_ok	0.137	0.061	−0.023	
	L4	+num_write	0.154	0.044	−0.154	✓

Table 7: Base-strength ladders. D is strictly increasing and remaining space $C-D$ strictly decreasing in all three domains. As the base absorbs more discriminating signals, repair gains diminish and eventually become negative when $D \approx C$.

Domain	n	Slope	t	R^2	Violations
retail	99	$\mathbf{-0.224}$	−3.38	0.105	0 / 99
airline	48	$\mathbf{-0.403}$	−2.93	0.157	0 / 48

Table 8: Per-task regression of repair gain on base advantage D. Both domains show significant negative slopes: tasks where the base is already strong leave less room for repair. Zero ceiling violations across all 147 tasks.

Experiment 2 connects to the within-label allocation analysis. Groups A and B see identical training contexts; the difference is which visit events receive positive labels. Group B's filtering removes ambiguous positives correlated with visit occurrence but not with conversion—analogous to the candidates that the permutation probe identifies as outcome-irrelevant within the feedback top set. The random-downweight control confirms that improvement comes from targeting the right labels, not from label reduction itself.

Experiment 2 also instantiates the local repair framework of Section 5 at the RM training level. Define the feedback top set $\mathcal{T}$ as training instances receiving positive visit-event labels. Group B's semantic filtering sets the weight of ambiguous members to zero—a redistribution within $\mathcal{T}$ that preserves total training volume but changes which positives contribute to the reward signal. The frozen-pool repair redistributes selection probability within $\mathcal{T}_i$; the online repair redistributes label weight within the analogous training set. Both operate on feedback-equivalent subsets and both improve downstream outcomes only when targeting semantically identified members (the random-downweight control confirms this).

C Boundary Analysis Details

Telecom with user_ok. The `user_ok` coarse definition yields $|\mathcal{T}_i| = 2.3$ on average in telecom, producing only 4 repairable tasks out of 114. This is a structural boundary: the coarse signal is too fine-grained to create meaningful top sets. Switching to $\texttt{score}_{\text{q50}}$ ($|\mathcal{T}_i| = 26.7$) rescues telecom to $+15.4$pp significant, confirming that the limitation is in the coarse definition, not the domain.

Matched-access ablation on OSWorld. OSWorld provides a direct check against attributing extra information to a hierarchical structure. The comparison matches the sixth input (`score`) supplied to a gated selector, using 311 tasks with 43 candidates each. The earlier gain is nearly reproduced by supplying the missing input to ordinary ridge. The residual gated advantage is unresolved, and

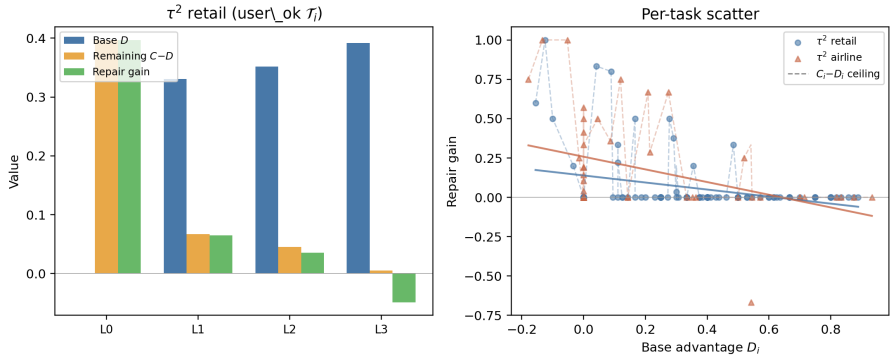


Figure 2: Left: base-strength ladder on τ^2 retail (`user_ok` $\mathcal{T}_i$). As base D grows, remaining space $C-D$ shrinks and repair gain diminishes. Right: per-task scatter of D_i vs repair gain for retail (circles) and airline (triangles). Both show significant negative slopes; all points lie below the C_i-D_i ceiling (dashed).

screening within the top set is worse than retaining all fine signals. This limits the architecture claim without proving that every gated rule is equivalent or ineffective: a local intervention can still be a useful mechanism test without outperforming every unrestricted learner.

Zero outcome heterogeneity. Tasks where all top-set members share the same Y value contribute zero repair headroom. In retail with `user_ok`, 31% of tasks have constant-Y top sets. These tasks are mechanically at the ceiling and should be distinguished from tasks where signals fail.

D Learned Reward Model Robustness

The main frozen-pool experiments use ridge regression as the prediction layer. This section tests whether the opposing-change sign persists when selection is based on a *learned* reward model rather than a raw signal value.

D.1 Design

On the τ^2-bench 30-model pool (retail 109 mixed-outcome tasks, K=30; telecom 43 mixed-outcome tasks, K=30), we train two learned selectors to predict the witness signal `user_ok` from the remaining process signals (`action_match`, `score`, `api_success`, `num_write`, `schema_valid`, `no_invalid_call`), excluding the witness itself and the evaluation outcome Y (`db_match`). Each selector uses five-fold cross-validation at the task level; only out-of-fold predictions are used for evaluation.

- **Ridge**: closed-form ℓ_2-regularized regression (λ=1.0) on standardized tabular features.
- **MLP**: one hidden layer with 16 tanh units, trained with L-BFGS for 500 iterations per fold (fixed seed).

Selection is by argmax of the learned score within each task. The CLSV statistic measures within-model residual selection advantage: CLSV $= \mathbb{E}_i[Y_{\text{selected}} - \bar{Y}_i]$ where Y is the residual after subtracting model-level means (controlling for model identity). Negative CLSV indicates that the learned selector systematically chooses lower-outcome candidates—the opposing-change sign.

D.2 Results

Interpretation. In retail—the domain with the strongest frozen-pool opposing change—both learned selectors preserve the negative CLSV with confidence intervals excluding zero, and the effect is *amplified* relative to direct selection (-0.178 and -0.235 vs -0.100). This confirms that the opposing-change phenomenon survives a learned prediction layer: a cross-validated RM trained

Domain	Selector	CLSV	95% CI	Sign preserved
Direct (raw signal selection, reference)				
retail	`user_ok` direct	-0.100	$[-0.175, -0.028]$	✓
telecom	`user_ok` direct	-0.142	$[-0.240, -0.041]$	✓
Learned selector (5-fold cross-validated)				
retail	Ridge	-0.178	$[-0.227, -0.128]$	✓
retail	MLP	-0.235	$[-0.306, -0.164]$	✓
telecom	Ridge	$+0.093$	$[-0.033, +0.227]$	×
telecom	MLP	$+0.059$	$[-0.076, +0.207]$	×

Table 9: Learned-RM robustness test. CLSV is the within-model residual selection advantage; negative values indicate the opposing-change sign is preserved through the learning layer. Retail preserves the sign with both ridge and MLP (CIs exclude zero); telecom's sign is diluted to unresolved by the learning layer.

to approximate `user_ok` from other signals still systematically selects lower-outcome candidates. In telecom, the learned selectors dilute the signal to unresolved (positive point estimates with CIs spanning zero), consistent with telecom's weaker initial effect and its dual-control structure that limits agent-side signal discriminability.

Neural embedding scorer. A separate experiment uses frozen Qwen3-4B mean-pooled embeddings of candidate trajectories as input features, with a ridge head trained to predict `user_ok`. The embedding scorer achieves near-perfect in-sample fit correlation (r=0.997 in telecom), yet when used for intervention testing on the 7-model pool, it shows qualitatively different behavior from the tabular scorer: the telecom embedding scorer's baseline CLSV is -0.148 (CI excluding zero), compared to $+0.130$ for the tabular scorer. This suggests that the neural embedding captures trajectory-level information not present in the tabular process signals, but the small pool size (K=7) limits interpretation.

Relationship to online experiments. The learned-RM experiments occupy the middle of an evidence chain: raw-signal frozen-pool → learned-RM frozen-pool → neural RM + RL + live deployment. The online experiments (Section 7) complete this chain with production-scale neural RMs and GRPO policy optimization. The retail learned-RM result (CLSV $= -0.178$) is directionally consistent with the online finding that scaling visit-event feedback from 50% to 100% reduces conversion (-0.15 pp), both indicating that better proxy prediction selects worse downstream outcomes.

E Same-Input Web-Agent Controls

The web-agent archive tests alternative uses on the same tasks and inputs. Let h be the mean of available input judges and form $s_{n,\gamma} = (1-\gamma)g_n + \gamma h$. Half retention uses $\gamma = 0.5$. In VWA, it changes G_J from -0.041 to -0.004 and raises high-budget utility by $+0.040$ $[+0.006, +0.080]$. But input mean itself performs at least as well in the point comparison: paired difference -0.007 $[-0.024, +0.007]$. There is no established advantage over this simpler comparator.

The same operation attenuates beneficial learning on WA: G_J falls from $+0.021$ to $+0.004$. Across 21 conditions, half retention improves high-budget utility in eight, worsens it in nine, and leaves four unchanged. These are compatible with ordinary ensemble effects (Coste et al., 2024), not a specialized repair algorithm.

Calibration versus a different target. All 525 full-budget calibration slopes are positive: numerical recoding preserves selection. Direct outcome learning changes rankings and improves VWA utility by $+0.064$ $[+0.014, +0.120]$. Against input mean, the advantage is only $+0.017$ $[-0.012, +0.053]$. Extra labels are a resource, not evidence that coarse labels must be split.

Complete condition family. Across 21 conditions, feedback-MSE endpoint estimates decrease in all. The categories are three joint gains, one opposing change, eleven fitting gains with unresolved utility, and six unresolved fitting changes.

F Signal-Use Extensions

F.1 Training-selected subsets

Signal subsets selected on training-fold outcomes and evaluated on held-out outcomes show conditional value: $+0.150$ in τ^2 telecom and $+0.037$ in AD within-attack. Retail has zero gap (full aggregate already attains unit utility). These results support conditional signal-use value but are not matched repairs of the quantity failure: they change signal composition, not within-label allocation.

F.2 Context routing

A logistic router selecting context-specific subsets achieves $+0.020$ $[+0.011, +0.028]$ over the global subset on AgentDojo's 2 690-task pool. OSWorld's routing gain is unresolved: -0.001 $[-0.027, +0.024]$. These results share pool-wide preprocessing before folds and do not establish unseen-context generalization.

G Complete Repair Experimental Details

G.1 Alpha-preserving repair

An alpha-preserving variant constrains the repair selector to maintain $\alpha_i = P(p \text{ selects from } \mathcal{T}_i)$, changing only the conditional distribution within $\mathcal{T}_i$. Results: retail $+0.8$pp (sig), airline $+7.9$pp (sig). The smaller magnitudes confirm that entry mass changes contribute to the unconstrained repair gain.

G.2 Conditional subgroup analysis

Tasks are stratified by repairability: $|\mathcal{T}_i| \geq 3$ AND $\text{Var}(Y_{\mathcal{T}_i}) > 0$.

Domain	Repairable n (%)	Rep. gain	Boundary gain
retail	79 (69%)	+0.081 sig	+0.004 ns
airline	36 (75%)	+0.230 sig	-0.042 ns
telecom	4 (3.5%)	0.000 ns	-0.005 ns

Table 10: Subgroup analysis. Repair gains concentrate in repairable tasks. Telecom with `user_ok` has only 4 repairable tasks, motivating the $\texttt{score}_{\text{q50}}$ definition.

G.3 Multi-method repair comparison

Four repair methods were compared under strong base: ridge, mean of signals, best single signal, and uniform. Under strong base, all methods yield repair gain ≤ 0 in τ^2 domains, consistent with the near-zero remaining space. This confirms that method choice cannot overcome a structural absence of repair room.

G.4 Retail per-task quartile analysis

Quartile	n	Mean D	Mean gain	Mean remaining
Q1 (weakest)	25	-0.011	+0.157	0.157
Q2	24	0.213	+0.082	0.089
Q3	25	0.416	+0.021	0.021
Q4 (strongest)	25	0.700	0.000	0.000

Table 11: Retail per-task quartile analysis. Repair gain decreases monotonically from Q1 to Q4, with zero violations of the ceiling bound across all 99 tasks.

H Protocols and Numerical Provenance

H.1 Quantity and localization

The quantity pipeline uses five seeds, five task folds, six nested training fractions, and ridge penalty 0.001. Feature centering and scaling are estimated within each prefix. The mixed-outcome filter uses evaluation Y; non-mixed tasks have zero utility difference between selectors but can have nonzero risk difference.

H.2 v230 complementary protocol

The alpha-interpolation protocol evaluates 15 cells (3 domains $\times$ 5 signals) on a τ^2-bench 7-model heterogeneous candidate pool (K=7 per task; retail 114 tasks, airline 50, telecom 114). For each cell, the target signal U is predicted by ridge regression from the remaining signals. Let $g_n(x)$ denote the ridge predictor trained on n tasks and $h(x)$ the arithmetic mean of input signal scores (excluding the target). The interpolated score is

$$s_{n,\alpha} = (1-\alpha)\, g_n(x) + \alpha\, h(x), \qquad \alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}. \tag{10}$$

At α=0 the protocol reduces to the main ridge-scaling study; at α=1 it uses the input mean without any fitted component. The outcome anchor is `db_match` (binary task success). Five of seven process signals serve as targets (`action_match`, `score`, `user_ok`, `api_success`, `num_write`); two guardrail signals (`schema_valid`, `no_invalid_call`) are constant in this pool and serve only as input features. Confidence intervals are task-bootstrap resampled (2 000 resamples). Three cells are not entirely independent of the main study: retail/user_ok appears in both protocols, and the same benchmark pools are used. The airline cells are genuinely new conditions.

H.3 Complete v230 alpha-interpolation matrix

Table 12 reports the full 15-cell matrix from the alpha-interpolation protocol on the τ^2-bench 7-model pool. This is the same diagnostic protocol applied to an existing, previously explored candidate pool—not an independent confirmation on new data. All 15 cells show positive MSE gain (feedback fit improves); the utility response varies by signal and domain. Category distribution: 7 value-up, 3 value-down (CI excluding zero), 5 unresolved. The three value-down cells all involve semantically ambiguous feedback signals (`user_ok` or `num_write`); `action_match` (anchored to the reference action) is value-up or unresolved in every domain.

Domain	Target	Category	G_R	G_J	95% CI (G_J)
retail	action_match	value-up	+0.030	+0.070	[+0.039, +0.105]
retail	score	value-up	+0.039	+0.056	[+0.025, +0.090]
retail	user_ok	**value-down**	+0.007	−0.246	[−0.284, −0.211]
retail	api_success	value-up	+0.020	+0.112	[+0.086, +0.140]
retail	num_write	unresolved	+0.014	−0.060	[−0.125, +0.004]
airline	action_match	value-up	+0.169	+0.235	[+0.174, +0.297]
airline	score	value-up	+0.197	+0.237	[+0.142, +0.330]
airline	user_ok	**value-down**	+0.028	−0.226	[−0.310, −0.138]
airline	api_success	unresolved	+0.028	+0.039	[−0.076, +0.156]
airline	num_write	**value-down**	+0.046	−0.224	[−0.308, −0.140]
telecom	action_match	unresolved	+0.020	+0.011	[−0.016, +0.037]
telecom	score	value-up	+0.045	+0.097	[+0.058, +0.139]
telecom	user_ok	unresolved	+0.005	+0.037	[−0.000, +0.072]
telecom	api_success	value-up	+0.004	+0.065	[+0.025, +0.109]
telecom	num_write	unresolved	+0.021	+0.006	[−0.028, +0.039]

Table 12: Complete v230 alpha-interpolation matrix on the τ^2-bench 7-model pool. G_R: feedback MSE improvement (all positive). G_J: selection utility change. Category: value-up (G_J CI > 0), value-down (G_J CI < 0), unresolved (CI spans zero). Bold value-down entries are the opposing changes reported in Table 1.

H.4 Local repair experiments

The full 16-experiment series comprises four rounds: strong base, weak base, boundary fix, and root-cause repair. Each round produces a structured JSON result file containing all per-task quantities (D, E, C, repair gain, ceiling violation counts), bootstrap confidence intervals, and permutation test outcomes. All numerical results reported in the main text can be reproduced from these files; source code and result archives will be released upon publication.

I Online Experiment Details

I.1 Shared configuration

Both experiments use the same base language model (a large mixture-of-experts architecture), the same RL algorithm (GRPO with init_kl_coef $= 0.0$), and the same reward formula structure. RM training uses several million interaction records per experiment. Online evaluation uses user-level randomization with Horvitz–Thompson weighting for non-transaction sampling. All confidence intervals are user-level bootstrap percentile intervals (2 000 resamples).

I.2 Experiment 1: quantity scaling

Three RMs were trained on 20%, 50%, and 100% of the available visit-event feedback data. Non-visit training targets (conversion, continuation, retention) were held at full volume across all three groups. RL training used identical budgets (60 steps, rollout_batch $= 128$, group_size $= 8$). Online deployment ran for six days with $\sim$150K users per group.

Multiple comparison correction: 3 groups $\times$ 6 metrics = 18 tests. Bonferroni threshold $\alpha_{\text{adj}} = 0.00278$; BH-FDR target 0.05. Both inverted-U arms pass Bonferroni: the ascending arm (50% vs 20%: $p = 0.002$) and the descending arm (100% vs 50%: $p < 0.0001$). Power at the observed descending-arm effect size (-0.15 pp): $>99\%$; ascending arm: 87%.

Scaling curve fits. Visit-event AUC follows a log scaling law: $\text{AUC}(f) = 0.813 + 0.018 \ln f$ ($R^2 = 0.998$). Order conversion follows an inverted-U quadratic with peak at $f^* \approx 0.57$. The optimal data fraction for prediction ($f = 1.0$) differs from the optimum for conversion ($f \approx 0.57$).

Comparison	Metric	Δ	95% CI	p
100% vs 20%	Visit/user	$+1.58\%$	$[+0.7\%, +2.4\%]$	<0.0001
100% vs 20%	Conversion	-0.04 pp	$[-0.11, +0.03]$	0.24
100% vs 50%	Visit/user	$+0.62\%$	$[+0.1\%, +1.1\%]$	0.012
100% vs 50%	Conversion	-0.15 pp	$[-0.22, -0.08]$	<0.0001
50% vs 20%	Visit/user	$+0.95\%$	$[+0.3\%, +1.6\%]$	0.002
50% vs 20%	Conversion	$+0.11$ pp	$[+0.04, +0.18]$	0.002

Table 13: Complete pairwise confidence intervals for Experiment 1 online metrics. Both inverted-U arms pass Bonferroni correction ($\alpha_{\text{adj}} = 0.00278$).

Complete pairwise CI table.

I.3 Experiment 2: specification comparison

Three RMs were trained on identical data volumes. Group B downweighted 17.6% of visit-event positive labels identified by three semantic rules: bounce clicks ($<$ 5s return to dialogue, $\sim 9.7\%$), post-correction clicks (within 30s of user correction, $\sim 4.8\%$), and low-engagement clicks (no subsequent page interaction, $\sim 3.2\%$). Group C downweighted the same fraction uniformly at random (seed fixed for reproducibility). RL training: 200 steps, init_kl_coef $= 0.0$. Online deployment: six days, $\sim$35K users per group, 1:1:1 allocation.

Multiple comparison correction: 3 groups $\times$ 4 metrics $\times$ 3 pairwise = 12 tests. BH-FDR at 0.05; Bonferroni at $\alpha_{\text{adj}} = 0.004$.

Overlap between B and C filtered samples. Group C's random draw hit $\sim 17.7\%$ of Group B's semantically filtered samples—matching the random baseline expectation. Group C retained $\sim 82\%$ of the ambiguous labels that Group B removed, explaining why C's policy exhibited the same hacking pathway as A's.

Day-by-day trends. B vs A conversion effects increased monotonically over six days ($+0.18$ pp on Day 1 to $+0.44$ pp on Day 6). C vs A conversion fluctuated around zero with no consistent direction (3 positive, 3 negative days). This rules out a Day-1 novelty spike as the explanation for B's advantage.

J Proposed Confirmation Protocol

The repair results in Table 2 were obtained on existing data splits. The protocol below specifies three gates for independent confirmation on untouched tasks.

J.1 Gate A: cache identity

Preserve task IDs, candidate order, source hashes, feature names, targets, outcomes, split/seed IDs, fitted coefficients, scores, and selectors from the original experiment. If no prediction cache exists, reconstruct it under the original configuration and reconcile stored metrics before any intervention. Fix the primary budget contrast and retain all other comparisons. Specify exact tie semantics; audit any numerical tolerance rather than silently changing it.

J.2 Gate B: six-arm intervention

Load p_H from the verified cache. Train the local outcome predictor on original outer-training tasks with a fixed feature panel and training-only preprocessing. Any hyperparameter search belongs in an inner split. Evaluate six arms: (i) original p_H; (ii) within-set uniformization; (iii) learned local repair; (iv) shuffled repair assignment (permute fitted scores within each test set, keeping multiset and entry mass fixed); (v) an all-candidate-trained outcome model restricted to the same set and mass; (vi) a matched flat model with unrestricted selection. Primary contrasts are $J(q) - J(p_H)$ and improvement over uninformative assignment. Also report $J(q) - J(p_L)$ to distinguish partial repair from recovery of the low-budget utility.

J.3 Gate C: untouched-task evaluation

Before accessing confirmation outcomes, freeze the dataset/split manifest, features, targets, top-set rule, learners, budgets, annotation access, hyperparameter procedure, primary contrasts, multiplicity handling, minimal meaningful effect, sample-size rationale, and stopping rule. Average repeated seeds within each base task, then use paired uncertainty at the independent unit. Evaluate quantity, localization, and repair on the same held-out population to test the full chain. New seeds of already explored tasks are robustness checks, not untouched confirmation.

Interpretation gates. Effective repair above uniformization supports useful additional discrimination. Improvement only up to uniformization supports removing harmful assignment. A flat learner matching the repair limits architecture claims. Failure to improve leaves the specific repair unsupported. If quantity divergence is absent on new tasks, even a successful repair does not replicate the complete phenomenon.