DOES BETTER ESTIMATION IMPLY BETTER SELECTION? A DECISION-THEORETIC AUDIT OF LLM AGENT UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating LLM-agent updates with low-cost proxies aims to reduce benchmarking expense, yet lower surrogate estimation error does not ensure better adoption decisions. We present a decision-theoretic audit framework that separates finite-pool mean squared error (MSE), same-pool decision regret, observed held-out gain, and acquisition cost. Our empirical benchmark evaluates three language models and twelve fixed update pairs on LiveCodeBench, using exact combinatorial integration across all target-sampling configurations to eliminate Monte Carlo acquisition noise over 12,384 recorded runs. In the post hoc subgroup of eight statically changed updates (Changed8), four-pair cross-fitting (CF4; two folds of two target task pairs) exhibits severe small-sample degeneracy at a 512-token proxy ceiling: both fitted slopes are zero in 95.71% of joint sampling configurations, yielding exactly the target-only mean on the same sample. Across all configurations in this regime, CF4 alters the corresponding target-only decision in only 0.076% of cases while requiring, on average, 2.40× as many cold-start acquisition completion tokens. At a 2,048-token proxy ceiling in the same subgroup, CF4 reduces mean MSE relative to unit-coefficient uniform correction but increases same-pool regret; externally calibrated correction reduces regret relative to CF4 but achieves lower observed held-out gain. Across all twelve updates, simultaneous 95% confidence bounds contain zero for both prespecified pooled utility contrasts. These findings reveal a low-budget regime where proxy adjustment adds substantial token overhead with little effect on decisions, demonstrating that surrogate evaluators must be benchmarked against matched target-only baselines through decision regret, held-out utility, and total acquisition cost rather than surrogate MSE alone.

1 INTRODUCTION

Automated agent search frameworks (Hu et al., 2025; Zhang et al., 2025) iteratively propose candidate updates, such as inserting a verification node, toggling a critique step, or appending edge-case reminders. Before adopting a child program, the developer must determine whether empirical evidence justifies deployment. However, evaluating candidates on full target benchmarks incurs substantial compute and token expenditures. Because local mutations dominate the search space and often produce narrow performance margins, candidate evaluation forms the primary computational bottleneck in automated workflow discovery. We isolate the post-generation selection decision, analyzing how surrogate proxy signals and finite target budgets jointly govern update adoption.

To reduce evaluation cost, practitioners frequently run candidate updates on low-budget proxy tasks, reserving expensive target evaluations for a promising subset. In prediction-powered and active inference, auxiliary predictions significantly improve estimation precision (Angelopoulos et al., 2023; 2024; Zrnic & Candès, 2024). Yet lower estimation error on proxy tasks does not necessarily produce better adoption decisions on unseen target benchmarks, particularly when accounting for the upfront cost of executing the full proxy table. While decision-focused learning formalizes the divergence between surrogate prediction error and downstream decision loss (Elmachtoub & Grigas, 2022), how this trade-off manifests in budgeted LLM-agent workflows remains unexamined under finite computational resources.

For any candidate update and recorded outcome table, the difference in held-out utility between two selection policies equals their difference in adoption probability multiplied by the candidate's

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.

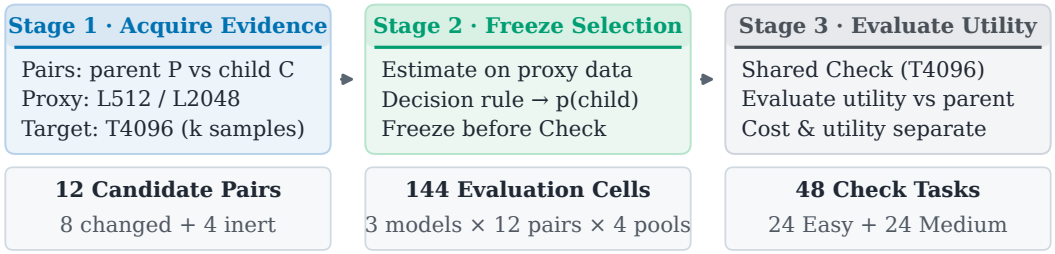


Figure 1: Evaluation after candidate generation is fixed. Direct-target rules bypass proxies. A sampled rule thresholds its estimate; the evaluator exactly averages these decisions before freezing probabilities. All four pools share Check. Exactness is conditional on recorded executions, not fresh service runs; KEEP's zero acquisition does not remove deployment cost.

held-out margin. This identity clarifies why estimation gains may fail to improve decisions: variance reductions that leave the sign of the estimated margin unchanged provide zero decision gain, whereas shifting the decision threshold on a negative-margin update degrades performance. This failure mechanism is intensified by discrete, sparse outcomes in code evaluation, where paired pass/fail differences frequently lack variation across small subsets. In this setting, fitting an adaptive regression slope on sparse target pairs frequently collapses to zero, causing cross-fitted corrections to incur proxy generation costs while algebraically reducing to the sample target mean.

This motivates a decision-theoretic audit that connects incremental proxy signals, discrete threshold decisions, and end-to-end acquisition costs (Figure 1). We evaluate twelve fixed update pairs across three language models, comparing uniform, cap-guided, and trace-guided retesting; fixed and externally calibrated coefficients; cross-fitted and plug-in slopes; and direct target-only baselines. Exact combinatorial integration across all target-sampling configurations evaluates selection probabilities and held-out Check utility without Monte Carlo noise, while pre-specified non-parametric concentration bounds govern simultaneous statistical inference.

Our empirical results yield three core findings: (1) **Surrogate estimation gains decouple from downstream decision utility:** On All12/L2048, cross-fitting reduces finite-pool MSE across all models, yet produces divergent held-out utility ($+0.69$, $+0.17$, -0.27 pp on Q4, Q9, Mistral); on Changed8, Mistral suffers simultaneous estimation and utility degradation, while a within-family MSE oracle underperforms passive retention. (2) **Adaptive corrections suffer structural collapse under discrete sparsity:** At L512, paired proxy differences are sparse: in 95.71% of Changed8 configurations, both cross-fitted slopes clip to zero, collapsing the estimator to the 4-target mean and altering the adoption decision in only 796/1,048,320 cases (0.076%). (3) **Adaptive rules incur substantial acquisition overheads:** On Changed8/L512, cross-fitting consumes 2.40× the cold-start acquisition completion tokens of direct 4-target sampling (24,513 vs 10,202 completion tokens) while altering the decision in fewer than 0.1% of configurations. Meanwhile, unadjusted proxy selection (LOWALL) delivers a descriptive advantage (58.88% vs. 58.68% for KEEP) yet operates without within-cell covariance guarantees, showing that aggregate heuristics can mask local estimation pathologies.

2 RELATED WORK

Candidate search and update acceptance. Automated search frameworks iteratively generate and evaluate agent workflows using execution feedback (Hu et al., 2025; Zhang et al., 2025). Fixing candidate updates isolates the core decision problem: selecting or rejecting a mutation based on finite empirical evidence. PACE (Shawn, 2026) addresses update acceptance through paired sequential tests with anytime-valid guarantees. In contrast, we examine fixed-budget acquisition policies evaluated through exact combinatorial integration, focusing on how discrete, sparse outcomes affect proxy-assisted selection.

Auxiliary information and decision objectives. Classical survey sampling uses inverse-probability weighting and difference estimators to correct auxiliary proxies (Horvitz & Thompson, 1952; Cassel et al., 1976). Prediction-powered inference (PPI, PPI++) and active statistical inference (ASI) extend these principles to machine-learning predictions (Angelopoulos et al., 2023; 2024; Zrnic & Candès,

2024). We apply these principles to finite-pool workflow evaluation, testing whether the theoretical efficiency of difference estimation holds in discrete, small-sample agent benchmarks. In parallel, decision-focused learning formalizes the gap between surrogate prediction error and task-specific decision loss (Elmachtoub & Grigas, 2022). Rather than training predictors under decision loss, we quantify this gap by testing whether surrogate estimation gains translate into higher held-out selection utility when proxy acquisition is accounted for.

Resource allocation and execution budgets. Multi-fidelity optimization frameworks such as Hyperband and BOHB (Li et al., 2018; Falkner et al., 2018) allocate compute dynamically across competing candidates. In our setting, the allocation unit is a paired evaluation for a specific parent-child candidate. While test-time reasoning budgets optimize token consumption within single executions (Han et al., 2025), our accounting captures the end-to-end cost of candidate acquisition, including proxy generation. We synthesize these perspectives into a decision-theoretic audit of proxy signals, adaptive corrections, and total acquisition budgets under discrete paired outcomes.

3 METHOD

3.1 EVALUATION OBJECTS AND INFORMATION BOUNDARIES

As a concrete update, E05 modifies the workflow pipeline from `SOLVE-CRITIQUE-REFINE-VERIFY-FINALIZE` to `SOLVE-CRITIQUE-VERIFY-FINALIZE`, altering the verifier input and remaining call budget. Proxy conditions L512 and L2048 impose completion ceilings of 512 and 2,048 tokens, respectively, while the target condition T permits up to 4,096 tokens. These conditions define differing resource constraints rather than program correctness.

Let m index a fixed model, e a fixed update with parent P_e and child C_e, and j a selection pool $\mathcal{P}_j$ of size $n_P = 16$. A cell is $c = (m, e, j)$ ($\mathcal{C}$ contains $3 \times 12 \times 4 = 144$ cells). Let t denote a task, a a candidate, $\ell \in \{\mathrm{L2048}, \mathrm{L512}\}$ a proxy condition, and T the finite target condition. Write $Z^b_{m,a}(t) \in \{0, 1\}$ for official all-tests PASS under condition $b \in \{\ell, T\}$ when technically complete. Define

$$x^\ell_{c,t} = Z^\ell_{m,C_e}(t) - Z^\ell_{m,P_e}(t), \qquad y_{c,t} = Z^T_{m,C_e}(t) - Z^T_{m,P_e}(t), \qquad \theta_c = \frac{1}{n_P} \sum_{t \in \mathcal{P}_j} y_{c,t}. \tag{1}$$

The estimand θ_c defines the exact finite-pool contrast over the $n_P = 16$ tasks in $\mathcal{P}_j$, providing a benchmark free of asymptotic approximation error, where incomplete runs are recorded as missing rather than imputed as zero.

Each selection rule operates strictly under its specified information boundary: proxy-based selectors access the full paired proxy table and their acquired target evaluations. Crucially, all selection rules, model configurations, and contrast definitions were fixed prior to querying the held-out Check set. We evaluate decision utility through exact combinatorial integration across all target-sampling subsets, computing expected policy performance without Monte Carlo sampling variance.

3.2 RETEST DESIGNS AND COEFFICIENT CONTROLS

For a rule r, let k be its target-pair budget and let $S \subseteq \mathcal{P}_j$ have size k. Suppress r, c, ℓ temporarily. The proxy indicators o_t, d_t, and f_t record, respectively, exhaustion of either candidate's output budget, disagreement in paired PASS, and disagreement in parse success. UNIFORM, CAP, and TRACE assign weights $w_t = 1, 1 + 2o_t$, and $1 + 2o_t + d_t + f_t$. On the size-k subsets $\mathcal{S}_k$, the law $q(S)$ equally mixes uniform sampling and sampling proportional to $\prod_{t \in S} w_t$. Each task has inclusion probability $\pi_t \geq k/(2n_P) > 0$; Equation (8) gives the exact law. Equal weights give uniform sampling without replacement. For proxy mean $\bar{x} = n_P^{-1} \sum_t x_t$ and a fixed coefficient λ, the generalized-difference estimate is

$$\widehat{\theta}_\lambda(S) = \lambda \bar{x} + \frac{1}{n_P} \sum_{t \in S} \frac{y_t - \lambda x_t}{\pi_t}. \tag{2}$$

We evaluate budgets $k \in \{4, 8\}$ and fixed shrinkage coefficients $\lambda \in \{0, 1/4, 1/2, 3/4, 1\}$ across all three designs. Standard difference estimation corresponds to $\lambda = 1$, while $\lambda = 0$ isolates target

sampling under non-uniform selection weights. Finite-pool MSE is defined as $\mathbb{E}_{S\sim q}[(\widehat{\theta}_\lambda(S) - \theta_c)^2]$ over the finite pool.

External calibration. DEVTUNED minimizes empirical mean squared error for Equation (2) on an external 16-task development probe, weighting models and updates equally. The resulting coefficients in $[0, 1]$ are fixed per proxy condition, design, and budget, then applied out-of-sample to all evaluation models without retuning on Check. Upfront token costs required to generate this probe are accounted for as an external calibration asset (Appendix B).

Purchased-label adaptation. CF-ADAPT4 draws four tasks uniformly and partitions them uniformly into ordered sets A, B of size two. Let β_A be the least-squares slope fitted on A, clipped to $[0, 1]$, with zero used when its denominator is zero. Define

$$\widehat{\theta}_A = \frac{1}{n_P}\left[\sum_{t\in A} y_t + \beta_A \sum_{t\notin A} x_t + \frac{n_P - 2}{2}\sum_{t\in B}(y_t - \beta_A x_t)\right], \quad \widehat{\theta}_{\mathrm{CF}} = \frac{\widehat{\theta}_A + \widehat{\theta}_B}{2}, \tag{3}$$

where $\widehat{\theta}_B$ swaps A and B. Exact combinatorial integration covers all $\binom{16}{4}\binom{4}{2} = 10{,}920$ configurations. PLUGINSLOPE4/8 divides sample covariance by the full-pool proxy variance, clips the resulting slope to $[0, 1]$, and adjusts the same sample. Six full-target oracles (ORACLEMSEONLY) provide non-deployable upper bounds on estimation accuracy. While four-pair cross-fitting allocates two target evaluations per slope, we also evaluate eight-pair cross-fitting ($k = 8$, Appendix J) to examine sensitivity to fitting budget.

3.3 DECISIONS, UTILITY, AND SHARED OUTCOMES

Let $\omega \in \Omega_{r,c}$ denote a complete subset/partition configuration with probability $\rho_{r,c}(\omega)$. A rule chooses the child only when its estimate is strictly positive:

$$p_{r,c} = \sum_{\omega\in\Omega_{r,c}} \rho_{r,c}(\omega)\mathbf{1}\{\widehat{\theta}_{r,c}(\omega) > 0\}. \tag{4}$$

Non-positive estimates retain the parent program. KEEP never adopts ($p = 0$); LOWALL thresholds the proxy sample mean; TARGETRANDOMk thresholds the sample mean of k target tasks; and FULLTARGET16 evaluates all 16 target tasks. In total, our audit spans 64 policy settings across 144 cells.

Let $\mu^{\mathrm{obs}}_{a,c}$ be candidate a's observed Check PASS mean, weighting Easy and Medium equally, and let $\gamma^{\mathrm{obs}}_c = \mu^{\mathrm{obs}}_{C_e,c} - \mu^{\mathrm{obs}}_{P_e,c}$. To isolate why estimation accuracy diverges from downstream decisions, we evaluate three retrospective descriptive layers on each cell: (1) finite-pool estimation error $\mathrm{MSE}_{r,c} = \mathbb{E}_\omega[(\widehat{\theta}_{r,c}(\omega) - \theta_c)^2]$ against pool contrast θ_c; (2) finite-pool decision regret $R_{\mathrm{pool},r,c} = \max(\theta_c, 0) - p_{r,c}\theta_c$; and (3) held-out Check gain $G_{\mathrm{check},r,c} = p_{r,c}\gamma^{\mathrm{obs}}_c$ over KEEP. For two frozen rules r_1, r_0, the pool decision contrast $\Delta_{\mathrm{pool}} = R_{\mathrm{pool}}(r_0) - R_{\mathrm{pool}}(r_1)$ and Check contrast $\widehat{\Delta}_{\mathrm{check}}$ satisfy the exact finite-sample identity:

$$\widehat{\Delta}_{\mathrm{check}} - \Delta_{\mathrm{pool}} = T(r_1, r_0) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}(p_{r_1,c} - p_{r_0,c})(\gamma^{\mathrm{obs}}_c - \theta_c). \tag{5}$$

Equation (5) links pool decision regret to held-out Check utility through the transfer discrepancy $T(r_1, r_0)$, separating internal selection consistency from external generalization without unverified domain-shift assumptions. The distances $D_{\mathrm{before}} = 144^{-1}\sum_c |\delta_c|$ and $D_{\mathrm{after}} = \frac{1}{2}\sum_v |g_v|$ characterize within-cell selection differences and cross-cell cancellation, respectively.

3.4 CONDITIONAL FINITE-FRAME INFERENCE AND COSTS

Index Easy and Medium difficulty strata by h. The eligible frame contains strata $\mathcal{F}_h$ of sizes $N_h \in \{36, 37\}$, from which $n_h = 24$ tasks were sampled without replacement. The estimand $\Delta_F = \frac{1}{2}\sum_h N_h^{-1}\sum_{t\in\mathcal{F}_h} W(t)$ represents the finite-frame contrast. Simultaneous 95% concentration bounds intersect finite-population concentration inequalities with support bounds on unobserved tasks, leveraging paired responses across shared evaluations. Under the two-sided null hypothesis

$H_0 : \Delta_F = 0$, intervals containing zero are classified as `NOT_IDENTIFIED`, enforcing rigorous pre-specified inferential criteria (Appendix C).

For resource dimension u, let $K_{c,t}^{\ell,u}$ and $K_{c,t}^{T,u}$ denote measured paired execution costs, χ_r indicate a full proxy table purchase, and $\eta_{r,c,t}$ denote target acquisition probability. Expected acquisition cost is:

$$A_{r,c}^{u} = \chi_r \sum_t K_{c,t}^{\ell,u} + \sum_t \eta_{r,c,t} K_{c,t}^{T,u}. \tag{6}$$

External development assets add amortized calibration cost B_r^u/J across deployment horizons $J \in \{1, 4, 16, 64, 256\}$. Downstream execution costs are weighted by adoption probabilities $p_{r,c}$ and retention probabilities $1 - p_{r,c}$. We track prompt tokens, completion tokens, API calls, and execution runtime, providing a comprehensive multi-dimensional ledger of acquisition expenditure.

4 EXPERIMENTS

4.1 SETUP, FIXED CANDIDATES, AND EVIDENCE STAGES

We use LiveCodeBench `code_generation_lite`, `release_v6` (Jain et al., 2025), with the pinned official evaluator. The models are Qwen3.5-4B (Q4), Qwen3.5-9B (Q9), and Mistral-Nemo-Instruct-2407 (MN12). L2048 and L512 use global completion ceilings of 2,048 and 512, with static node caps and quartered static caps, respectively. Target T uses a ceiling of 4,096 and divides remaining tokens among remaining calls. Runs permit at most four calls, with temperature 0 and top-p 1. These conditions model fixed token limits under bounded search depth. Appendix A retains the evaluator commit, precision, context, safety margin, and full runtime contract.

An outcome-independent frozen ID-hash rule selected twelve pairs from 24 inherited edges: four per class, twelve roots, and six of each variant. Four prompt patches append one shared contract/edge-case reminder, once repeating an existing copy; four toggle modules. E01–E04 edit ignored edge-type labels without altering adjacency, prompts, or caps, serving as statically inert controls. The runtime audit identifies 18 execution classes among the 24 candidate graphs; inert pairs exhibit minor non-determinism (29 of 576 comparisons disagree) due to cloud-service variance, and all pairs are retained in the primary benchmark. Appendices A.1 and I detail the complete structural audit.

The primary study spans 12,384 recorded executions, including 3,456 runs on the held-out Check set (24 Easy and 24 Medium tasks from a 36/37-task eligible frame). Candidate selection operates over four disjoint 16-task pools across 144 cells. All selection rules were pre-specified before querying held-out Check tasks. Exact combinatorial integration calculates expected policy outcomes over all possible target-sampling configurations without Monte Carlo error. Across the 24 update pairs in the changed stratum, four held-out margins are positive, fifteen are negative, and five are zero, yielding a pooled empirical headroom over KEEP of 0.8681 pp (Appendix Table 30). Because negative margins predominate, conservative selection rules inherently benefit by rejecting mutations, emphasizing the importance of auditing decision regret alongside raw utility.

4.2 RQ1: DO SURROGATE ESTIMATION GAINS TRANSLATE TO DOWNSTREAM UTILITY?

Surrogate estimation accuracy decouples from downstream selection utility on these recorded tables. Under L2048, cross-fitting (CF) reduces estimation MSE across all three models in the pre-registered All12 sample relative to coefficient-one UNIFORM (Figure 2a). Yet this variance reduction produces divergent held-out utility shifts: $+0.6876$ and $+0.1684$ pp on Q4 and Q9, but -0.2682 pp on Mistral. CF lowers mean child adoption by 20.95, 18.82, and 14.28 pp on Q4, Q9, and Mistral (positive/negative/zero cells: 19/9/20, 15/10/23, 3/7/38). On Changed8, model heterogeneity emerges: while Q4 and Q9 retain MSE gains, Mistral's CF MSE increases by 0.000688 ($0.010181 \rightarrow 0.010869$, Appendix Table 20) alongside utility degradation (-0.3447 pp). Moreover, DEVTUNED achieves lower MSE than CF but lower Check utility, while the within-family MSE oracle underperforms KEEP (58.53% vs 58.68%; KEEP as zero-estimation has MSE 0.00415 vs oracle 0.01326). Appendix D–E retain all 18,432 records.

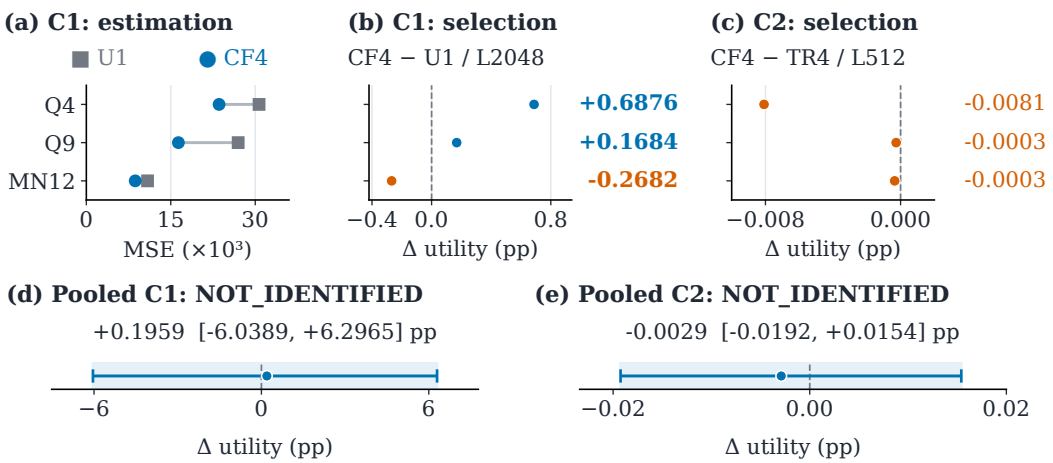
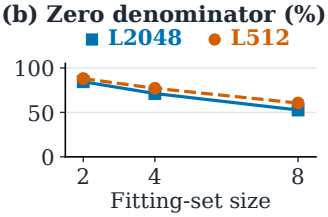


Figure 2: Estimation and selection on all twelve pairs. C1 compares CF4 to coefficient-one UNIFORM4 under L2048 (a: gray U1 squares, blue CF4 circles). C2 compares CF4/L512 to TARGET_RANDOM4. Model points are descriptive; only pooled points carry the original simultaneous finite-frame bounds, shown in full. Utility averages cellwise probability–margin products. Both simultaneous 95% bounds span zero (protocol: NOT_IDENTIFIED); separate axes preserve scale.

(a) Changed8: four disjoint joint events (%)

Proxy / rule	**Both slopes zero**	Nonzero slope; equal estimate	Changed estimate; same action	Changed action
L512 / CF4	**95.7074**	2.6013	1.6153	0.0759
L512 / CF8	**89.1819**	6.8098	3.6493	0.3591
L2048 / CF4	**88.7737**	4.1228	6.4545	0.6490
L2048 / CF8	**77.1828**	9.2199	11.7557	1.8417

(b) Zero denominator (%)



(c) CF4 → CF8 (post hoc)

Proxy	MSE (×10³)	U − KEEP (pp)	Completion (×10³)
L512	17.645 → 5.866	-0.2893 → -0.3243	24.51 → 34.72
L2048	17.920 → 5.967	-0.3034 → -0.3559	41.92 → 52.12

Figure 3: Changed8 fitting diagnostics. (a) Four disjoint joint events under the ordered disjoint-fold law. (b) The proxy-only denominator curve does not define a CF16 selector. (c) CF8 lowers MSE but increases token costs and worsens relative-KEEP utility; it is post hoc, not a primary replacement.

4.3 RQ2: Can prespecified bounds confirm statistical gains?

Pre-specified simultaneous 95% concentration bounds establish the statistical resolution of the finite evaluation frame (Figure 2d–e). Contrast C1 yields a pooled point estimate of $+0.1959$ pp with an interval of $[-6.0389, +6.2965]$ pp, while C2 yields -0.0029 pp with an interval of $[-0.0192, +0.0154]$ pp. Because both simultaneous intervals contain zero, their directional signs are recorded as NOT_IDENTIFIED. These wide bounds demonstrate that under finite testing budgets, subtle performance differences between candidate selection policies remain statistically unresolvable without substantially larger evaluation frames (Appendix C).

4.4 RQ3: Why does small-sample cross-fitting collapse to target-only?

At L512, 82/144 cells have zero paired proxy differences ($x_t = 0$); only eleven have both candidates fail all proxy tasks. Before purchasing any target label, the exact probability that a fitting fold of size

Method	**Changed8** (96 cells) MSE ↓	R_{pool} ↓	G_{check} ↑	**All12** (144 cells) MSE ↓	R_{pool} ↓	G_{check} ↑	**Acq. Tokens** Changed8 / All12
KEEP	4.476	2.799	+0.000	4.150	2.648	+0.000	0 / 0
FULLTARGET16	0.000	0.000	-0.260	0.000	0.000	+0.072	40,808 / 40,390
TARGETRANDOM4	17.594	1.839	-0.285	15.923	1.736	-0.098	10,202 / 10,097
UNIFORM1	24.089	1.399	-0.684	22.841	1.399	-0.305	41,918 / 41,335
DEVTUNED†	16.453	1.281	-0.706	15.201	1.319	-0.318	41,918 / 41,335
CFADAPT4	17.920	1.825	-0.303	16.206	1.726	-0.109	41,918 / 41,335
LOWALL	7.161	1.693	-0.608	6.456	1.736	-0.362	31,716 / 31,238

Table 1: Three-layer diagnostic decomposition on fixed evaluation pools ($L = 2048$, pooled over models). MSE is same-pool error ($\times 10^3$); R_{pool} is same-pool decision regret (pp); G_{check} is observed held-out gain over KEEP (pp); Tokens are cold-start acquisition completion tokens. †: excludes external calibration assets (1,994,743 tokens). No subgroup interval is attached.

m_f has zero proxy variance is

$$\Pr(V_A = 0) = \frac{\binom{n_-}{m_f} + \binom{n_0}{m_f} + \binom{n_+}{m_f}}{\binom{N}{m_f}}, \qquad m_f \in \{2, 4\}, \tag{7}$$

where n_-, n_0, n_+ count differences $\{-1, 0, +1\}$ on pool size $N = 16$. We distinguish marginal fold causes from joint outcomes: single-fold zero variance ($V_A = 0$) or nonpositive covariance forces $\beta_A = 0$, but neither alone implies both slopes vanish. Under the disjoint-fold law, joint double-zero slopes $\mathcal{B} = \{\beta_A = \beta_B = 0\}$ imply target-mean collapse $\mathcal{E} = \{\widehat{\theta}_{\text{CF}} = \overline{y}_S\}$ and identical action $\mathcal{H} = \{\mathbb{I}(\widehat{\theta}_{\text{CF}} > 0) = \mathbb{I}(\overline{y}_S > 0)\}$. The inclusions $\mathcal{B} \subseteq \mathcal{E} \subseteq \mathcal{H}$ define four disjoint outcomes on Changed8: both slopes zero ($\mathcal{B}$, 1,003,320 configurations, 95.71%); nonzero slope with unchanged estimate ($\mathcal{E} \setminus \mathcal{B}$, 27,270, 2.60%); changed estimate with same action ($\mathcal{H} \setminus \mathcal{E}$, 16,934, 1.62%); and changed action ($\mathcal{H}^c$, 796, 0.0759%, 796/1,048,320 configurations). Thus both slopes vanish in 95.71% of configurations, while all equal-estimate configurations total 98.31% (1,030,590/1,048,320). Doubling the fitting budget to $m_f = 4$ ($k = 8$, Appendix J) still leaves 89.18% degenerate (Table 31).

4.5 RQ4: DOES EXECUTION EQUIVALENCE EXPLAIN TARGET-MEAN COLLAPSE?

We report all twelve pairs, the eight statically changed pairs (E05–E12), and the four statically inert pairs (E01–E04) in parallel, following the structural audit and retaining every frozen choice.

Table 1 isolates the three diagnostic layers. In Changed8/L2048, UNIFORM1 → CF4 lowers MSE ($24.09 \to 17.92 \times 10^{-3}$) but increases pool regret R_{pool} ($1.40 \to 1.82$ pp). DEVTUNED achieves lower pool regret (1.28 pp) than CF4, yet degrades on Check ($G_{\text{check}} = -0.71$ pp, $T = -0.95$ pp relative to CF4). Compared to direct TARGETRANDOM4, CF4 exhibits similar pool regret (1.82 vs 1.84 pp) and lower Check gain (-0.30 vs -0.28 pp), while purchasing the complete proxy table ($41,918$ vs $10,202$ tokens). FULLTARGET16 achieves zero pool regret but yields -0.26 pp on Check ($T = -3.06$ pp relative to KEEP). Changed and inert C1 means are $+0.3807$ and -0.1737 pp ($\frac{2}{3}(+0.3807) + \frac{1}{3}(-0.1737) = +0.1959$ pp matches All12). In Changed8, L512 has 1,003,320/1,048,320 double-zero configurations (95.71%) while requiring 2.40× direct target retesting completion tokens. The high collapse rate persists among statically changed updates. Consequently, the inclusion of statically inert pairs alone cannot explain the observed collapse. These results demonstrate that estimator collapse is driven by structural sparsity in proxy responses rather than the presence of inert mutations.

4.6 RQ5: HOW DO ALTERNATIVES TRADE OFF DOWNSTREAM UTILITY AND COST?

CF/L2048 attains lower Check utility than KEEP, FULLTARGET16, and TARGETRANDOM4 despite a positive point estimate for C1 (Figure 4). Conversely, unadjusted proxy selection at L512 (LOWALL) achieves a positive descriptive Check gain over KEEP (+0.20 pp on All12, +0.28 pp on Changed8). However, this candidate-level result does not establish positive within-cell proxy–target association. In Changed8/L512 (96 cells), the complete recorded pools contain 37 zero-proxy-variance cells, 14 nonconstant-proxy cells with zero covariance, 18 with negative covariance, and 27 with positive covariance. Even in the 27 positive-covariance cells, 85.17% (251,114/294,840) of ordered disjoint-

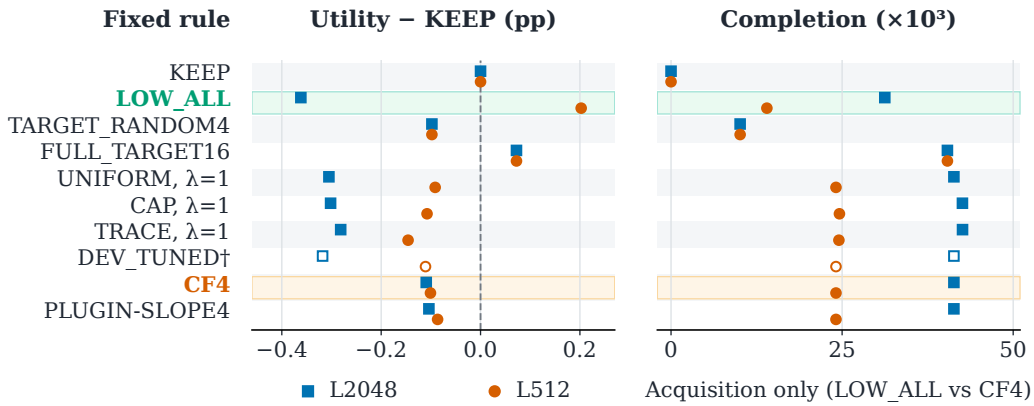


Figure 4: Utility and acquisition cost for ten fixed families across 144 cells. Aligned rows compare L2048 (squares) and L512 (circles); positions do not imply a ranking guarantee. Hollow DEV_TUNED† markers exclude external training assets (Appendix L); deployment is separate. Numerical tables retain all 64 settings.

fold configurations still produce double-zero slopes (Appendix Table 32). These categories confirm that within-cell proxy variation remains too sparse to sustain small-sample slope estimation.

Under cold-start accounting, CF4 consumes $\approx 24,513$ completion tokens/cell in Changed8/L512 versus $10,202$ for TARGETRANDOM4 ($2.40\times$; $24,103$ vs $10,097$ in All12, $2.39\times$), while deployment costs are essentially identical ($1,157.43$ vs $1,157.39$ tokens/task). DEVTUNED adds an upfront calibration burden of 1,994,743 completion tokens (1,475,102 at L512; Appendix G). Complete-proxy to full-target cost ratios are 0.7734 at L2048 and 0.3468 at L512. Uniform 4-pair correction saves acquisition tokens relative to FULLTARGET16 only when this ratio falls below $3/4$; 8-pair correction requires below $1/2$ (both exceeded at L2048; Appendix G).

4.7 RQ6: DOES LARGER FITTING BUDGET PREVENT DEGENERACY?

Increasing the fitting budget to four pairs per fold ($k = 8$) lowers estimation MSE but fails to eliminate degeneracy, leaving 89.18% of configurations with double-zero slopes in Changed8 (Figure 3). Moreover, selection gain relative to KEEP further degrades (utility drops from -0.2893 to -0.3243 pp), confirming that modest sample-size increases do not resolve the small-sample variance problem.

Four repeated execution rounds completed 4,608 runs across four candidate updates (E01, E02, E05, E09), selection pool P1, and sixteen Check tasks. Paired PASS outcomes diverged in 142 of 2,304 identical-candidate comparisons under zero-temperature decoding, demonstrating non-negligible platform-level non-determinism. Across rounds, the pooled C2 contrast shifts across zero, while Mistral's round-three policy achieves $+0.5060$ pp on round-four Check despite identical diagonal configurations (Figure 5). This round-to-round variation underscores that single-round empirical comparisons can be vulnerable to infrastructure-level drift (Appendix K).

5 PRACTICAL IMPLICATIONS AND CONCLUSION

Engineering controls. Our findings suggest three concrete controls: (1) **pre-screen execution** via diff audits to prune inert mutations (E01–E04); (2) **audit proxy variance** $\Pr(V_A = 0)$ before purchasing target labels; and (3) **benchmark against matched controls**, separating matched-target sampling (TARGETRANDOM4, $k = 4$) from proxy-only selection (LOWALL, $k = 0$).

Limitations. This study evaluates candidate selection across fixed mutation pairs within LiveCodeBench. Future investigations should extend decision-theoretic audits to online iterative search, multi-turn agent interactions, and benchmarks exhibiting broader margin distributions.

Conclusion. Our audit demonstrates why surrogate estimation accuracy alone is insufficient to guide budgeted LLM-agent updates. Separating estimation error, decision regret, held-out gain, and

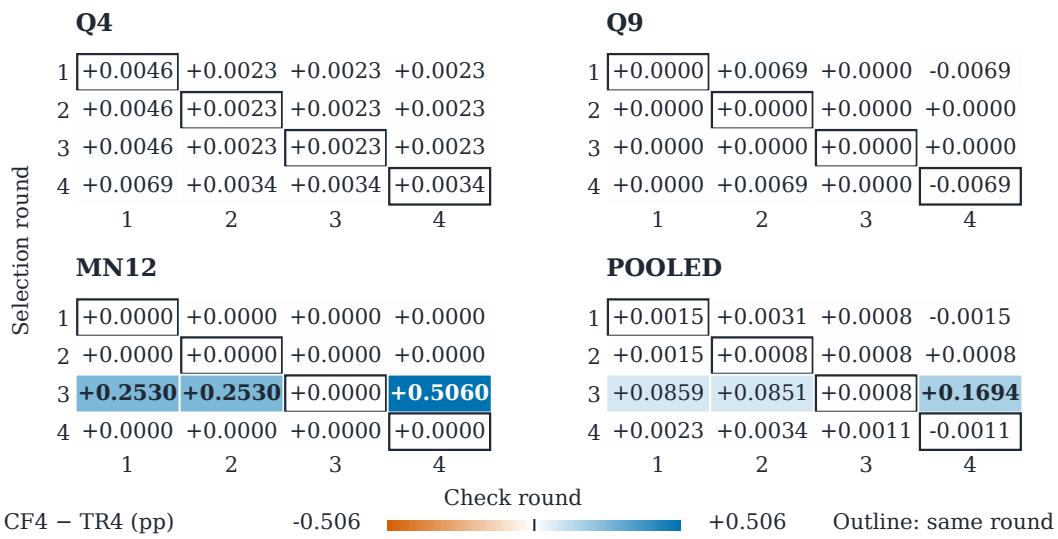


Figure 5: Four-round CF4/L512 minus TARGET_RANDOM4 crossgrid across four fixed pairs and sixteen exposed Check tasks (pp). Rows/columns fix selection/Check executions; outlines mark same-round contrasts. Shared observations are not independent replicates; the grid neither replaces C2 nor inherits its primary interval.

acquisition cost exposes critical trade-offs that MSE cannot capture. Under tight proxy ceilings (L512), cross-fitting collapses to the sample target mean in 95.71% of configurations despite paying for the full proxy table, altering decisions in only 0.076% of cases. These findings motivate pre-purchase proxy variance auditing and benchmarking against direct target sampling at matched budgets, ensuring that surrogate evaluation expenditures genuinely yield decision-relevant evidence for downstream deployment.

Reproducibility Statement

The accompanying review bundle contains candidate graph definitions and prompts, binary outcome and resource records, frozen query draws, and a standard-library CPU replay. Appendix H distinguishes recompiling the document, recomputing the numerical results, and executing new model runs. The default replay neither launches models nor loads benchmark task bodies or private tests. Public redistribution terms and final human release approval remain separate from numerical reproducibility.

AI Use Statement

The authors used AI-based assistance exclusively to aid and polish the writing of the manuscript, including grammar refinement, style consistency, and phrasing polish. All experimental designs, mathematical formulations, numerical computations, and scientific interpretations were conceived, executed, and verified by the authors.

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://arxiv.org/abs/2301.09633v4.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference, 2024. URL https://arxiv.org/abs/2311.01453v2. Version 2; original submission 2023.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. URL https://arxiv.org/abs/1309.4029v2. Corollary 2.5.

Claes M. Cassel, Carl E. Särndal, and Jan H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976. doi: 10.1093/biomet/63.3.615.

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

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1437–1446, 2018. URL https://proceedings.mlr.press/v80/falkner18a.html.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025. doi: 10.18653/v1/2025.findings-acl.1274. URL https://aclanthology.org/2025.findings-acl.1274/.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.08435v2.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. URL https://jmlr.org/papers/v18/16-558.html.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 2017. URL https://arxiv.org/abs/1610.04821. Section 2.1.

Zayx Shawn. PACE: Anytime-valid acceptance tests for self-evolving agents, 2026. URL https://arxiv.org/abs/2606.08106v1. Version 1, June 6; preprint.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/pdf?id=z5uVAKwmjf.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.