WHEN IS LOCAL VALIDATION CHEAPER FOR A TEXT-UPDATE COMPOSITION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Can evaluations of individual text updates validate their composition more cheaply than direct execution? We study both whether local results determine the full gain and how combining their estimates affects sampling cost, for paired binary outcomes and a fixed composition of k updates. Under a degree-r mean-interaction model, we derive explicit weights w_j that recover the full gain from average gains over local subsets of each size (shell means). With local prices c_j and full-query price c_F, the minimax expected validation cost is $\Theta(\min\{(\sum_j |w_j|\sqrt{c_j})^2, c_F\}\epsilon^{-2}\log(1/\delta))$. The lower bound covers adaptive mixtures, task reuse and random stopping; a cost-aware local sampler or direct execution attains the rate. We distinguish mean additivity (gains add in expectation) from taskwise additivity (gains add on each task): their singleton-query factors are k^2 and k, even with identical mean responses. Without a verified structural identity, we correct cached predictions using sampled full outcomes to obtain an exact finite-pool gain interval. On 13 complete public GEPA quartets, additive proxies overstate every full gain, while a fixed retrospective audit reduces new full-composition queries from 3,400 to 2,780 (18.24%) with either branch or clipped predictions. Across 1,000 paired replays, mean query savings remain 15.52–16.06%. These results identify how execution prices, paired structure and cached predictions govern validation reuse.

1 INTRODUCTION

Two modified versions of a prompt-based system have already been evaluated on the same tasks, alongside the system from which they were developed. Their evaluation results are available. The changes are then combined into a single candidate, raising a new question: does this combined version improve on the original system? Evaluating the combined version directly would answer this question, but it requires additional executions. Can the existing results reduce how much new evaluation is needed?

This is the situation we study in released GEPA prompt-merge records (Agrawal et al., 2026). The original candidate is the *parent*, the two modified candidates are its *branches*, and their combination is the *merge*. The branch results describe the candidates evaluated separately; they do not directly measure the merge's performance. Reusing those results therefore requires understanding what they can tell us about the combined candidate.

We study whether component outcomes determine the combined gain (*identification*) and how estimation errors affect the cost of using them. For a fixed composition of k text updates, *local* evaluations use proper subsets of the updates; *full* evaluations use all of them. Figure 1 summarizes the validation routes.

When local execution savings offset sampling cost. Under a model allowing mean interactions among at most r updates, explicit weights recover the full gain from averages over subsets of sizes $1, \ldots, r$. We derive matching upper and adaptive lower cost bounds that quantify when cheaper local executions compensate for the extra sampling required.

Why gains on individual tasks matter. Mean additivity requires gains to add in expectation; taskwise additivity requires this on each task. For queries executing one update at a time, their

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.

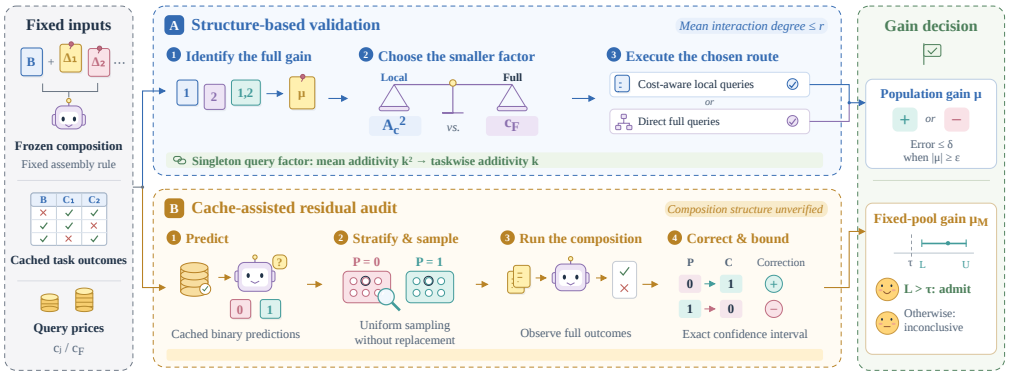

Figure 1: Validation routes for a fixed text-update composition. (A) Structure-based validation: identify the population gain μ from local shell means, compare A_c^2 with c_F, and execute the route with the smaller linear-design factor. Taskwise additivity reduces the singleton factor from k^2 to k. (B) Cache-assisted residual audit: stratify tasks by cached predictions, sample full outcomes without replacement, and correct prediction errors to obtain an exact confidence interval for the fixed-pool gain μ_M. Admit when its lower endpoint exceeds τ.

worst-case cost factors are k^2 and k. Binary response families with identical mean gains but different paired outcomes establish this separation.

How cached predictions reduce full executions. Without a verified structural relation, we predict full outcomes from the cache, execute the composition on sampled tasks, and correct prediction errors to obtain an exact finite-pool gain interval. On 13 complete GEPA quartets, additive proxies overstate every full gain. Yet cached branch and clipped predictions reduce aggregate full queries in all 1,000 paired replays, with mean savings of 15.52–16.06% and different per-case outcomes.

Related work. SkillEvolBench studies acquired skills under frozen deployment, including composition and context shifts (Lei et al., 2026). Our cost analysis builds on linear-functional allocation (Elfving, 1952) and polynomial extrapolation without an intercept (Dette et al., 2021). Multi-fidelity identification uses known fidelity-bias constraints (Poiani et al., 2024); combinatorial exploration (Cao & Krishnamurthy, 2019) and low-rank factorial design (Wen et al., 2026) study selection under other structural models. Active Testing (Kossen et al., 2021), prediction-powered inference (Angelopoulos et al., 2023), and its stratified extension (Fisch et al., 2024) combine surrogate predictions with measured outcomes. Our audit applies this correction principle with exact hypergeometric intervals for a fixed binary pool.

2 A FIXED COMPOSITION AND ITS OBSERVABLE EVIDENCE

Freeze a parent artifact, $k \geq 2$ distinct text updates and a deterministic assembly rule. For $S \subseteq [k]$, the rule constructs the artifact using exactly those updates. It can specify an order and resolve conflicts; commutativity is not assumed. A task handle Z contains the task, environment and coupled random seed. Its binary verifier outcome is $C_S(Z) \in \{0, 1\}$, with parent $B(Z) = C_\varnothing(Z)$. Write

$$D_S = C_S - B, \qquad g(S) = \mathbb{E}D_S, \qquad \mu = g([k]). \tag{1}$$

The expectation is over a fixed task distribution. Section 6 distinguishes this population gain from the gain μ_M conditional on a finite recorded pool.

Restricted versus direct queries. A restricted query reveals $C_S(Z)$ for $1 \leq |S| \leq r < k$ and costs one candidate execution. The algorithm may choose S adaptively, obtain fresh independent task handles, or revisit a handle. Repeating a revealed query does not reveal a new outcome. The

Table 1: The evidence contracts compared in this paper. Population sign-test costs suppress $\epsilon^{-2}\log(1/\delta)$. Costs count queries except in the priced row, which uses expected charged cost. The last row targets a conditional finite-pool quantity.

Outcome assumption	Candidate access	Worst-case cost
None	All proper subsets	Target not identified
Mean interaction degree r	Subsets of size at most r	$\Theta(A_r(k)^2)$
Taskwise target additivity	Single updates	$\Theta(k)$
None	Full composition	$\Theta(1)$
Mean interaction degree r	Priced local and full queries	$\Theta(\min\{A_c^2, c_F\})$
Fixed finite binary pool	Cached prediction plus sampled full outcomes	Error-dependent; census is exact

lower-bound interface gives parent outcomes for free and exposes uninformative task identities. Upper bounds use fresh independent tasks and at most one parent execution per candidate query, so charging those parents changes the order only by a constant.

Direct queries of $C_{[k]}$ need only $O(\epsilon^{-2}\log(1/\delta))$ fresh executions because $D_{[k]} \in [-1, 1]$. Section 4 therefore allows both local and full candidates and charges their declared prices in the same model. The comparison therefore balances statistical amplification against execution price.

Table 1 summarizes the access models for validating the frozen composition.

An identification boundary. Without structure, even complete knowledge of every proper-subset outcome can leave the sign of μ unknown. Let all proper subsets equal their parent, with B uniform. In one world the full composition changes a fraction 2ϵ of the $B = 0$ tasks to success; in another it changes the same fraction of $B = 1$ tasks to failure. The gains are $+\epsilon$ and $-\epsilon$, yet all restricted observations have the same law. This holds for $0 < \epsilon \leq 1/2$. No increase in restricted evaluations can resolve the difference. The next sections quantify what explicit interaction assumptions add to this interface.

3 Mean interactions identify the target but amplify noise

Assume the mean gain has multilinear interaction degree at most r:

$$g(S) = \sum_{\substack{\varnothing \neq A \subseteq S \\ |A| \leq r}} \theta_A. \tag{2}$$

This is a restriction on expectations, not on each task's response. For $j = 0, \ldots, k$, let m_j average $g(S)$ over all j-subsets. If $b_a = \sum_{|A|=a} \theta_A$, counting subsets containing A gives

$$m_j = \sum_{a=1}^{r} b_a \frac{\binom{j}{a}}{\binom{k}{a}}, \qquad m_0 = 0. \tag{3}$$

The shell averages are polynomial even when g is not permutation invariant. Estimating m_j requires uniform sampling over size-j subsets.

Proposition 1 (Explicit extrapolation)**.** *For $1 \leq r < k$, assumption (2) implies*

$$\mu = \sum_{j=1}^{r} w_j m_j, \qquad w_j = (-1)^{r-j}\binom{k}{j}\binom{k-j-1}{r-j}. \tag{4}$$

These are the unique linear weights using only $m_1, \ldots, m_r$. Writing $A_r(k) = \sum_j |w_j|$, for fixed r,

$$A_r(k) = \frac{2^r - 1}{r!}k^r + O_r(k^{r-1}). \tag{5}$$

The interpolation proof is given in Appendix B.1.

If every shell mean has deterministic error at most η, the resulting linear estimate can have error $A_r(k)\eta$. The linear error bound is attained when errors align with the signs of w_j. The following statistical result extends the squared amplification factor to adaptive restricted-query estimators.

With equal local prices, restricted-query complexity is $\Theta(A_r(k)^2\epsilon^{-2}\log(1/\delta))$, including adaptive choices and task reuse (Theorem 6, Appendix A). Section 4 derives the general cost model allowing both local and full queries.

Scale. $A_1(k) = k$, $A_2(k) = k(3k-5)/2$, and $A_{k-1}(k) = 2^k - 2$. For degree two the identity is

$$\mu = \binom{k}{2} m_2 - k(k-2)m_1. \tag{6}$$

At $k = 8$ the coefficients are 28 and -48, giving $A_2(8) = 76$. The negative singleton term removes effects counted repeatedly in the pair mean. Increasing r enlarges both the response class and the available local queries; the rates compare these joint changes. We next compare this statistical amplification with execution prices.

4 When is local validation cheaper than full execution?

Now allow both local subsets of sizes $1, \dots, r$ and the full composition. A size-j query has a known fixed price $c_j > 0$; a full query costs $c_F > 0$. Task handles and parent bits carry no additional charge. Prices are fixed inputs to the cost model. The resource is worst-case *expected total charged cost*, allowing adaptive choices, task reuse and almost-surely finite random stopping. Define

$$A_c = \sum_{j=1}^{r} |w_j|\sqrt{c_j}, \qquad B_c = \sum_{j=1}^{r} \frac{|w_j|}{\sqrt{c_j}}, \qquad G = \min\{A_c^2, c_F\}. \tag{7}$$

Theorem 2 (Local and full queries in one cost model)**.** *Under (2), the minimax expected charged cost of deciding the sign of μ with error at most δ whenever $|\mu| \geq \epsilon$, for $0 < \epsilon \leq 1/8$ and $0 < \delta \leq 1/4$, is*

$$\Theta\big(G\epsilon^{-2}\log(1/\delta)\big)\,. \tag{8}$$

The constants are universal. A pure local or pure full strategy attains this order; the lower bound permits arbitrary adaptive mixtures.

Attaining the cost factor. For the local route, draw a shell with $q_j = |w_j|/(\sqrt{c_j}B_c)$, then a uniform subset on a fresh task. The signed observation $W = (w_j/q_j)D_S$ satisfies

$$\mathbb{E}W = \mu, \quad \mathbb{E}W^2 \leq A_c B_c, \quad \mathbb{E}[\text{price}] = A_c/B_c. \tag{9}$$

Median-of-means amplification therefore costs $O(A_c^2\epsilon^{-2}\log(1/\delta))$ in expectation. Direct paired differences give the same bound with c_F. Choose the smaller factor before observing outcomes.

A hard family with full execution available. The lower bound constructs two jointly admissible binary response laws with opposite gains. Every newly revealed bit carries KL information bounded by $16\epsilon^2$ times its price divided by G, including on reused tasks. A stopped-transcript testing argument then gives the matching expected-cost bound. Appendix C provides the interpolating construction and complete proof.

Proposition 3 (Exact comparison within the linear sampling design)**.** *Among fixed unbiased linear estimators that sample one candidate on each fresh task and sample uniformly within a chosen local shell, the minimum worst-case variance times expected price is exactly G. Thus local sampling has a strictly smaller optimal design factor precisely when $c_F > A_c^2$.*

Indeed, if the full-response coefficient is t, unbiasedness forces the local coefficients to be $(1-t)w_j$. Optimal allocation gives $(|1-t|A_c + |t|\sqrt{c_F})^2$, minimized at a pure route. This is classical linear-functional allocation (Elfving, 1952); the explicit boundary follows from the composition weights. The proposition gives the exact linear-design boundary, while Theorem 2 characterizes minimax sign-test cost up to universal constants.

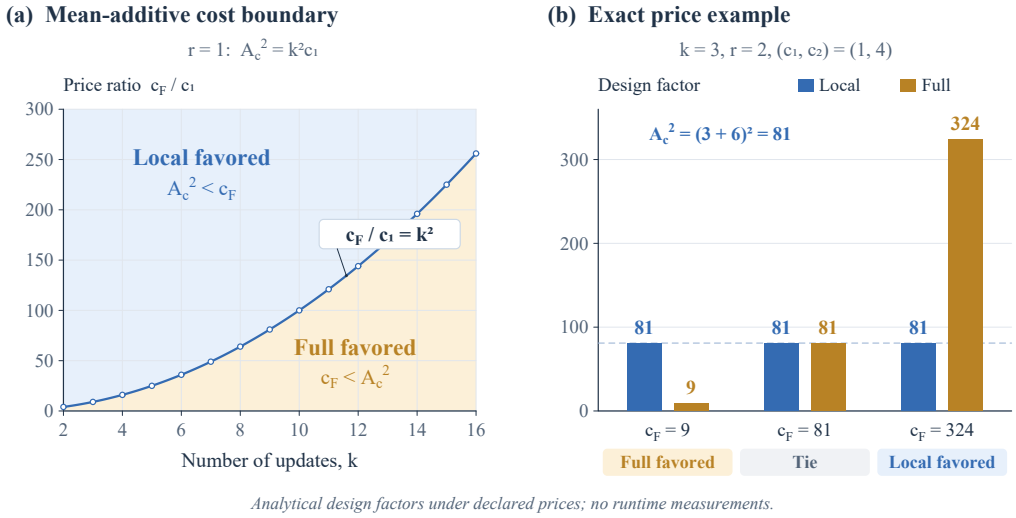

Figure 2: Local/full boundaries for the optimal linear-design factor. (a) Under mean additivity, local sampling is favored above $c_F/c_1 = k^2$ and full sampling below it. (b) With $k = 3$, $r = 2$ and $(c_1, c_2) = (1, 4)$, the local factor is $A_c^2 = 81$: full prices 9, 81 and 324 give full, tied and local preferences. Curves show analytical comparisons at the declared prices.

What execution savings must overcome. For mean-additive singletons, $A_c^2 = k^2 c_1$. More generally $|w_1| \geq k$, so $A_c^2 \geq k^2 c_1$. If $c_F \leq k c_1$, including nonnegative affine per-update prices, the local design factor is at least k times the full factor. For $k = 3, r = 2$ and $(c_1, c_2) = (1, 4)$, the weights are $(-3, 3)$, the local allocation is $(2/3, 1/3)$, and $A_c^2 = 81$. Full prices $9, 81, 324$ therefore select full, tie, and local respectively. Figure 2 illustrates these analytical comparisons. Section 6 treats a finite cache separately, using its recorded predictions to reduce the need for fresh full queries.

5 Taskwise structure changes the query cost

Consider the stronger target identity

$$D_{[k]}(Z) = \sum_{i=1}^{k} D_{\{i\}}(Z) \quad \text{almost surely.} \tag{10}$$

It need not impose additivity on every intermediate subset. Since binary differences relative to one parent all have the same sign, their sum lies in $\{-1, 0, 1\}$ only if at most one singleton differs from the parent. Thus

$$\sum_i D_{\{i\}}^2 = D_{[k]}^2 \leq 1. \tag{11}$$

If additivity holds for every subset, each task's Boolean response is constant or depends on one coordinate, an elementary degree-one case of classical low-degree Boolean structure (Nisan & Szegedy, 1994). The assumption concerns paired outcomes, including their dependence on the parent.

Theorem 4 (Taskwise-additive singleton complexity)**.** *Under (10), the worst-case singleton-query cost for the same sign test and parameter range as Theorem 6 is*

$$\Theta\big(k\epsilon^{-2}\log(1/\delta)\big)\,. \tag{12}$$

The lower bound still permits repeated-task queries and adaptive choices.

For the upper bound, choose a fresh task and uniform $I \in [k]$ and return $W = kD_{\{I\}}$. Now $\mathbb{E}W = \mu$, $\mathbb{E}W^2 \leq k$, and $|W - \mathbb{E}W| \leq 2k$. Bernstein's moment argument yields

$$\mathbb{P}(|\overline{W} - \mu| \geq \epsilon) \leq 2\exp\left[-\frac{N\epsilon^2}{2(k + 2k\epsilon/3)}\right]. \tag{13}$$

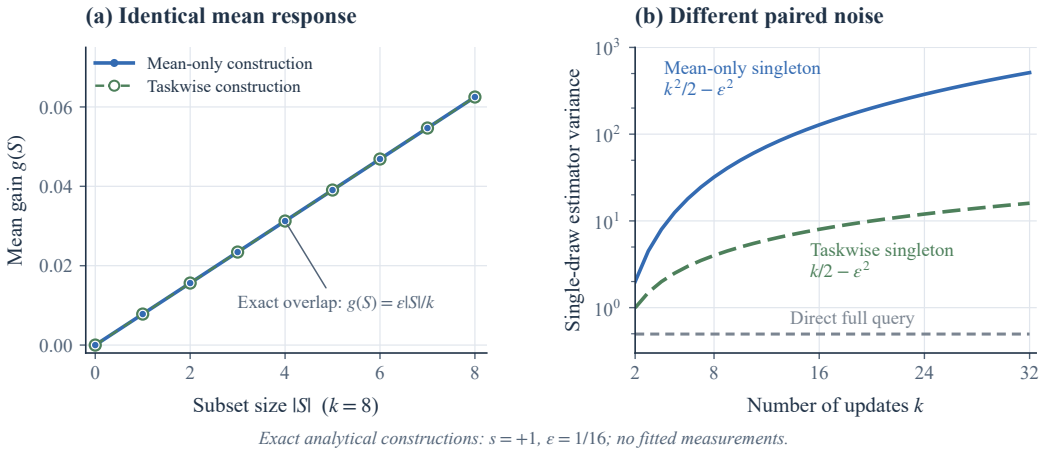

Figure 3: Identical mean responses, different paired noise ($s = +1$, $\epsilon = 1/16$). (a) Both constructions satisfy $g(S) = \epsilon|S|/k$; their curves coincide at $k = 8$. (b) The singleton estimator has variance $k^2/2 - \epsilon^2$ and $k/2 - \epsilon^2$ in the mean-only and taskwise constructions, respectively. Direct full queries have variance $1/2 - \epsilon^2$ in both. Lines connect exact analytical values at integer k.

Why task reuse does not remove the factor. The lower-bound family assigns each task a hidden uniform update index that can change its parent outcome. Even after previous misses, the next singleton query carries at most $256\epsilon^2/k$ conditional KL information about the gain's sign. This yields the matching adaptive lower bound (Appendix B.3).

The same mean surface can support different paired evidence. Specialize the construction of Theorem 6 to $r = 1$ and compare it with the hidden-active-update family. Both have $g(S) = s\epsilon|S|/k$, candidate success marginal $\mathbb{E}C_S = 1/2 + s\epsilon|S|/k$, and the same full-difference distribution. Their singleton second moments differ:

$$\begin{aligned} \mathbb{E}D_{\{i\}}^2 &= 1/2 \qquad \text{in the mean-only construction,} \\ \mathbb{E}D_{\{i\}}^2 &= 1/(2k) \quad \text{in the taskwise construction.} \end{aligned} \tag{14}$$

The variance of the uniform-singleton estimator is consequently $k^2/2 - \epsilon^2$ or $k/2 - \epsilon^2$, respectively. The direct full difference has variance $1/2 - \epsilon^2$ in both. Figure 3 evaluates these formulas, isolating the effect of paired dependence.

Activity is observable; identification is additional. For arbitrary paired binary outcomes, define

$$h = \mathbb{E}\sum_i D_{\{i\}}^2, \quad \mu_{\text{proxy}} = \mathbb{E}\sum_i D_{\{i\}}, \quad \text{Var}(kD_{\{I\}}) = kh - \mu_{\text{proxy}}^2. \tag{15}$$

The variance identity needs no additivity. Identifying the target further requires $\mu_{\text{proxy}} = \mu$, or a bound on their difference. In particular $h \le 1$ does not establish taskwise or mean additivity. The indistinguishable worlds in Section 2 even have $h = 0$.

For N independent uniform-singleton observations, write $b = \mu_{\text{proxy}} - \mu$. The exact mean-squared error for the target is

$$\mathbb{E}(\overline{W} - \mu)^2 = b^2 + \{kh - \mu_{\text{proxy}}^2\}/N. \tag{16}$$

More local samples remove the second term but leave the first. Thus the identification premise controls the bias term, while paired activity controls sampling variance. Section 7 measures both on recorded compositions.

Appendix B.6 gives the corresponding conditional sign test for μ_{proxy}. Transferring its conclusion to μ requires the identifying relation.

Approximate taskwise structure. If (10) fails on a set of probability at most ζ, then

$$|\mu_{\text{proxy}} - \mu| \le k\zeta, \qquad \mathbb{E}W^2 \le k[1 + (k-1)\zeta]. \tag{17}$$

The bias factor k is attainable for $k \ge 2$, so a failure frequency ζ translates into a target-error tolerance of $k\zeta$. The first bound follows by bounding the taskwise residual by k on the failure set; the second counts changed singletons. Appendix B.4 gives the attaining example. Directly auditing complete-composition residuals provides an alternative to specifying ζ.

6 A RESIDUAL AUDIT WHEN COMPOSITION STRUCTURE IS UNVERIFIED

Cached outcomes can reduce audit uncertainty by predicting the full composition. We combine these predictions with sampled full outcomes using stratified prediction correction (Angelopoulos et al., 2023; Fisch et al., 2024). For a fixed binary pool, hypergeometric tail inversion gives an exact gain interval whose efficiency depends on the residual errors.

Freeze M task records, parent outcomes B_t, the composition, and binary predictions P_t computed from cached records. Here the target is $\mu_M = M^{-1}\sum_t (C_{[k]}(t) - B_t)$, conditional on this entire pool. Let $G_b = \{t : P_t = b\}$, $M_b = |G_b|$, and define

$$H = \frac{1}{M}\sum_t (P_t - B_t), \quad E_0 = \sum_{t \in G_0} C_{[k]}(t), \quad E_1 = \sum_{t \in G_1} (1 - C_{[k]}(t)). \tag{18}$$

The first error count is missed success; the second is predicted success that fails. The identity

$$\mu_M = H + (E_0 - E_1)/M \tag{19}$$

holds for every binary prediction.

Before observing new composition outcomes, choose sample sizes q_b and sample uniformly without replacement in each G_b. The observed error count X_b has law $\text{Hypergeom}(M_b, E_b, q_b)$. Let $[L_b, U_b]$ be an exact integer confidence interval for E_b with noncoverage at most $\alpha/2$, obtained by inverting its two tails.

Proposition 5 (Finite-pool gain gate)**.** *Conditional on the frozen pool, predictions and stated sampling design,*

$$\left[H + \frac{L_0 - U_1}{M},\ H + \frac{U_0 - L_1}{M}\right] \tag{20}$$

contains μ_M with probability at least $1 - \alpha$. Admitting only when its lower endpoint exceeds a fixed threshold τ controls false declarations of $\mu_M > \tau$ at level α.

A union bound and (19) prove coverage (Appendix B.5). An unobserved stratum uses $[0, M_b]$; a census determines the target exactly. Across candidates, methods and looks, error levels sum to the desired total, without requiring independent decisions.

How prediction quality enters. With no sampled errors, its lower endpoint is $H - U_1/M$: missed successes can only help, while falsely predicted successes can invalidate the cached gain. The exact finite-population expression for U_1 is given in Appendix B.7. The upper bound accounts for unobserved errors before a census.

What determines the corrected estimator's variance? For $q_b > 0$ in every nonempty stratum, the unbiased corrected point estimate is $\widehat{\mu} = H + (M_0 X_0/q_0 - M_1 X_1/q_1)/M$. Under independent sampling in the two strata,

$$\text{Var}(\widehat{\mu} \mid \text{pool}, P) = \sum_{b : M_b > 1} \frac{E_b(M_b - E_b)(M_b - q_b)}{q_b(M_b - 1)M^2}. \tag{21}$$

This is the hypergeometric variance formula; empty or fully observed singleton strata contribute zero. It shows why predictability within strata can help even when the cached gain is biased. The gate uses exact tail inversion. The variance formula supplies a separate diagnostic computed from complete recorded pools; the audit itself accesses full outcomes only through its queries.

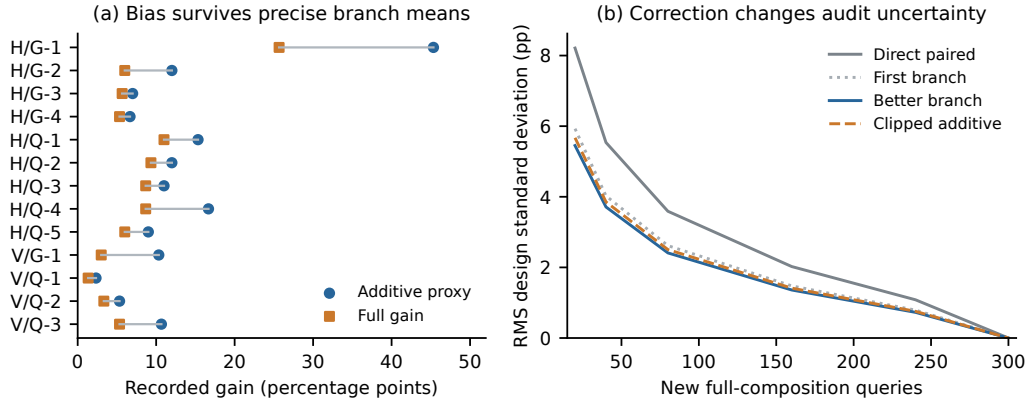

Figure 4: Proxy bias and audit uncertainty on 13 recorded quartets. Left: additive proxies and full gains on 300-task pools. Right: square root of the mean conditional design variance in (21) across quartets; all methods reach zero at the census. Codes denote HotpotQA (H), HoVer (V), GPT-4.1 mini (G), and Qwen3-8B (Q). Uncertainty is conditional on each fixed pool.

What to predict from the cache. For two branches, natural choices are $P = B$, $P = C_{\{1\}}$, the branch with better cached accuracy, and $P_{\text{clip}} = \text{clip}(C_{\{1\}} + C_{\{2\}} - B, 0, 1)$. The first choice is direct paired validation with known parents. Clipping removes predictions outside the binary range, and the audit estimates the remaining discrepancy. In fact, writing C_1, C_2, C_{12} for the four-artifact case,

$$C_{12} - C_1 - C_2 + B = (C_{12} - P_{\text{clip}}) - (C_1 + C_2 - B - P_{\text{clip}}). \tag{22}$$

The second term is known range saturation; the first is the remaining composition residual measured by the audit.

The guarantee targets the fixed-pool gain μ_M. Predictions, thresholds and allocations are fixed before the audit observations under the stated design. Section 7 specifies these choices for the replay; Section 8 discusses generalization beyond the pool.

7 Validation reuse in released prompt-merge records

We measure additive-proxy bias and audit query savings by replaying the GEPA artifact at commit `cbefbc1` (Agrawal et al., 2026).

Data and protocol. The 13 complete binary quartets span nine merge-enabled runs, HotpotQA and HoVer, and GPT-4.1 mini and Qwen3-8B. Each has 300 aligned tasks, giving 3,900 quartet–task rows (provenance in Appendix D). The replay conditions on these search-used pools; its design was frozen before execution.

We compare the four predictions in Section 6 at budgets $20, 40, 80, 160, 240, 300$ and threshold $\tau = 0$. A fixed random permutation per quartet gives nested uniform samples within each prediction stratum, with $q_0 = \lfloor qM_0/M \rfloor$ and $q_1 = q - q_0$. The global error budget 0.05 is divided across 13 quartets, four methods and six looks. We count newly revealed full outcomes, with parent and branch outcomes cached; 300 queries gives a census.

Additive proxies overstate the recorded gain. Every additive proxy exceeds its positive full gain by 1–19.67 percentage points (Figure 4, left). Taskwise violations range from 5.33% to 23%, despite activity h of 0.203–0.480, illustrating the distinction between activity and bias in (16).

Residual correction links proxy bias to audit precision. For H/G-1, the first quartet in saved order, clipping changes the proxy from $136/300$ to $75/300$. Stratum sizes $(104, 196)$ and error counts $(5, 3)$ then recover the full gain: $75/300 + (5-3)/300 = 77/300$. At 80 queries, the clipped

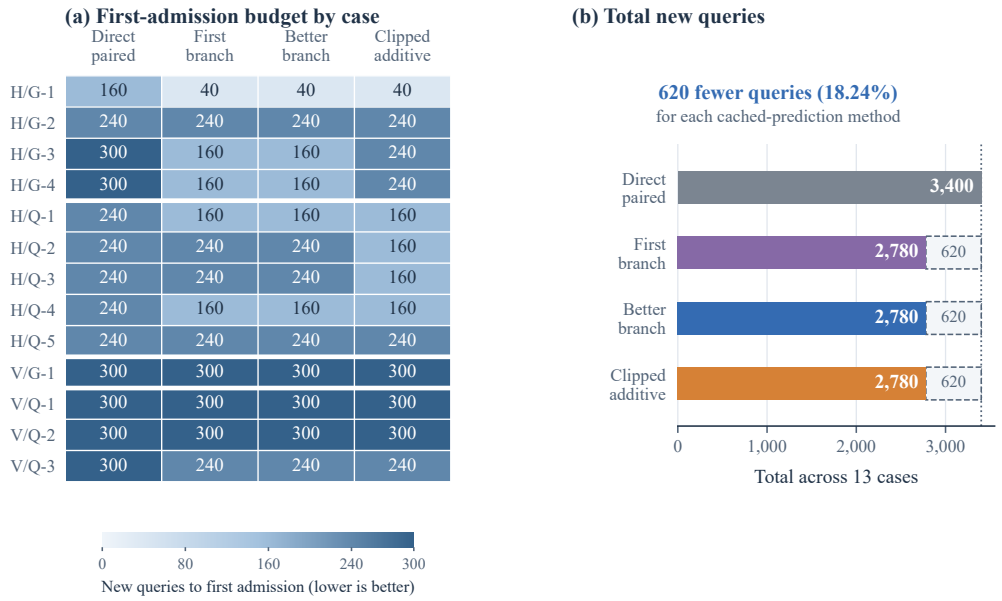

Figure 5: First-admission budgets and aggregate query savings. (a) Earliest prescribed budget with lower gain bound above zero; 300 is a census. Case codes follow Figure 4. (b) Each cached-prediction method uses 2,780 new full queries versus 3,400 for direct validation, saving 620 (18.24%); dashed extensions show the difference. Results are from one fixed retrospective replay with parent and branch outcomes already cached.

interval $[41/300, 19/50]$ clears zero, whereas the direct interval $[-1/60, 109/300]$ does not. The first branch also admits. This is a shared-budget illustration, not a first-admission comparison.

Both the better branch and clipping reduce design variance at 80 queries in all 13 pools. Figure 4 averages these conditional variances; admission additionally depends on discrete tail inversion.

Cached predictions save queries across sampling orders. The original replay in Figure 5 uses 3,400 queries for direct validation versus 2,780 for each cached method (18.24% fewer). We repeat the frozen comparison under 1,000 prespecified independent permutations per pool, shared across methods. Mean paired aggregate savings are 15.52% for the first branch, 16.06% for the better branch and 15.58% for clipping; all three reduce total queries in every replay. Residual auditing therefore retains its query advantage as task order changes, while predictor rankings vary (Appendix D.1).

8 Discussion and Conclusion

Validation reuse depends on identification, paired uncertainty and price. Under a degree-r mean-interaction model, comparing A_c^2 with c_F determines the optimal linear design and the minimax cost order. Taskwise structure further reduces the singleton factor from k^2 to k. An existing cache offers a complementary route: auditing prediction residuals can certify a finite-pool gain before a census, with mean query savings of 15.52–16.06% across the repeated GEPA replays. A validation system can therefore assess structural identification and predictive reuse separately.

Limitations and future work. The theory assumes binary outcomes, an outcome-query interface and fixed positive prices; fractional scores, semantic information and history-dependent costs need further analysis. The replay measures conditional query savings on search-used pools with recurring tasks and uncoupled generation noise. Independent pools and runtime measurements would assess transfer and total costs, including cache acquisition.

AI USE STATEMENT

Generative AI tools were used to assist with checking mathematical proofs, writing and debugging code, and improving the language and readability of the manuscript. The authors assume full responsibility for the accuracy and integrity of the final manuscript, including all mathematical derivations, experimental results, and code.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

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://doi.org/10.1126/science.adi6000.

Tongyi Cao and Akshay Krishnamurthy. Disagreement-based combinatorial pure exploration: Sample complexity bounds and an efficient algorithm. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 558–588, 2019. URL https://proceedings.mlr.press/v99/cao19a.html.

Holger Dette, Viatcheslav B. Melas, and Petr Shpilev. Some explicit solutions of c-optimal design problems for polynomial regression with no intercept. *Annals of the Institute of Statistical Mathematics*, 73:61–82, 2021. doi: 10.1007/s10463-019-00736-0. URL https://www.ism.ac.jp/editsec/aism/pdf/s10463-019-00736-0.pdf.

Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952. doi: 10.1214/aoms/1177729442.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c9fcd02e6445c7dfbad6986abee53d0d-Abstract-Conference.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Testing: Sample-Efficient Model Evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Yingtie Lei, Zhongwei Wan, Jiankun Zhang, Samiul Alam, Zixuan Zhong, Peizhou Huang, Xin Wang, Jingxuan Zhang, Donghao Zhou, Yunta Hsieh, Zhihao Dou, Hui Shen, Yan Xu, Dimitrios Dimitriadis, Tuo Zhang, and Mi Zhang. SkillEvolBench: Benchmarking the Evolution from Episodic Experience to Procedural Skills. arXiv:2605.24117v1, 2026. URL https://arxiv.org/abs/2605.24117v1.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Noam Nisan and Mario Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994. URL https://web.vu.lt/mif/s.jukna/boolean/Nisan-Szegedy.pdf.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.03033.

Xin Wen, Xi Chen, Will Wei Sun, and Yichen Zhang. Policy-Aware Design of Large-Scale Factorial Experiments. arXiv:2604.08804v1, 2026. URL https://arxiv.org/abs/2604.08804v1.

A THE RESTRICTED-QUERY SPECIALIZATION

Theorem 6 (Restricted-query complexity)**.** *For $0 < \epsilon \le 1/8$ and $0 < \delta \le 1/4$, the worst-case number of restricted candidate queries needed to distinguish $\mu = +\epsilon$ from $\mu = -\epsilon$ under (2), with error at most δ under either sign, is*

$$\Theta\big(A_r(k)^2 \epsilon^{-2} \log(1/\delta)\big)\,. \tag{23}$$

The lower bound permits adaptive subset choices and repeated-task queries.

An attaining sampling rule. Draw J with probability $|w_J|/A_r(k)$, choose a uniform J-subset S, and evaluate it on a fresh task. Then

$$W = A_r(k)\,\mathrm{sign}(w_J) D_S, \qquad \mathbb{E}W = \mu, \qquad |W| \le A_r(k). \tag{24}$$

For N independent repetitions, the bounded-variable inequality (Hoeffding, 1963) gives

$$\mathbb{P}(|\overline{W} - \mu| \ge \epsilon) \le 2\exp\left(-\frac{N\epsilon^2}{2A_r(k)^2}\right). \tag{25}$$

This allocation is an ordinary signed functional-estimation design. Its role is to attain the explicit cost for the permitted queries.

A joint response lower bound. Interpolate $P_r(0) = 0$ and $P_r(j) = \mathrm{sign}(w_j)$ for $1 \le j \le r$. Then $P_r(k) = A_r(k)$ and $|P_r(t)| \le A_r(k)$ at every integer $0 \le t \le k$. The surfaces $g_\pm(S) = \pm\epsilon P_r(|S|)/A_r(k)$ belong to the degree-r class. Draw a fair parent and independent candidate bits with success probabilities $1/2 + g_\pm(S)$. Every newly revealed allowed bit carries at most $16\epsilon^2/A_r(k)^2$ KL information, even on an old task; repeated entries carry none. The transcript testing bound is therefore

$$N \ge \frac{A_r(k)^2\,\mathrm{kl}(1-\delta, \delta)}{16\epsilon^2}. \tag{26}$$

Appendix B.2 verifies the entire binary response table and conditional information bound. A variance calculation for one estimator would not establish this adaptive lower bound.

B COMPLETE MATHEMATICAL ARGUMENTS

All statements concern the frozen response interface of Section 2. We use N for the number of candidate queries and M for the finite-pool size. In a lower bound, a task handle exposes its identifier and, if requested, its parent outcome. It does not reveal the latent table of all candidate outcomes. The identifier distribution is identical under the two hypotheses. Additional white-box information about an executing program is outside the stated oracle.

B.1 SHELL INTERPOLATION

We prove Proposition 1.

Proof. Interpolate (3) at $0, 1, \ldots, r$. The coefficient of m_j at k is $\prod_{0\le a\le r, a\ne j}(k-a)/(j-a)$, which simplifies to (4). The coefficient of m_0 disappears because its value is known. The nonzero-node basis is nonsingular. Each $|w_j|$ has leading coefficient $1/[j!(r-j)!]$; summing proves (5). □

B.2 Mean-degree lower bound, including adaptive task reuse

Let $L_j(t) = \prod_{0\le a\le r, a\neq j}(t-a)/(j-a)$ for $0 \le j \le r$, and define

$$P_r(t) = \sum_{j=1}^{r} \operatorname{sign}(w_j) L_j(t). \tag{27}$$

At 0 it is zero, and at each $j \in \{1,\dots,r\}$ it is $\operatorname{sign}(w_j)$. For integer $t > r$, the numerator of every $L_j(t)$ is positive and its denominator has $r-j$ negative factors. Hence $\operatorname{sign}(w_j)L_j(t) = |L_j(t)|$. Every such absolute product is nondecreasing for $t \ge r$, as its factors $t-a$ are nonnegative there. Therefore

$$P_r(k) = \sum_{j=1}^{r} |w_j| = A_r(k), \qquad |P_r(t)| \le A_r(k) \quad (t = 0,\dots,k). \tag{28}$$

At nodes $0,\dots,r$ the bound follows from $A_r(k) \ge 1$; between those integers no bound is needed.

For $s \in \{-1,1\}$ set $g_s(S) = s\epsilon P_r(|S|)/A_r(k)$. Any degree-$r$ polynomial with value zero at zero has an expansion $P_r(t) = \sum_{a=1}^{r} \gamma_a \binom{t}{a}$. Since $\binom{|S|}{a} = \sum_{A\subseteq S, |A|=a} 1$, this is a mean surface of the form (2), with coefficients $\theta_A = s\epsilon\gamma_{|A|}/A_r(k)$. Its full gain is $s\epsilon$.

On each new task draw $B \sim \mathrm{Ber}(1/2)$. Independently for every nonempty $S \subseteq [k]$, draw

$$C_S \sim \mathrm{Ber}(1/2 + g_s(S)), \qquad C_\varnothing = B. \tag{29}$$

This defines one jointly valid binary response table on a finite set of subsets. Its candidate-parent mean difference is exactly $g_s(S)$. All probabilities lie in $[3/8, 5/8]$ for $\epsilon \le 1/8$ by (28). For an allowed subset, its candidate probability is $1/2 \pm a$, where $a = \epsilon/A_r(k)$, with the sign determined by the shell and the hypothesis. Parent outcomes have identical law and are independent of every candidate outcome in both worlds.

For $|a| \le 1/8$ the Bernoulli relative entropy is

$$\mathrm{kl}(\mathrm{Ber}(1/2+a), \mathrm{Ber}(1/2-a)) = 2a \log \frac{1+2a}{1-2a} \le 16a^2. \tag{30}$$

To verify the inequality, use symmetry in a and, for $x \in [0,1)$, $\log[(1+x)/(1-x)] \le 2x/(1-x)$. This gives $8a^2/(1-2a) \le (32/3)a^2 \le 16a^2$.

Condition on any observed transcript and the algorithm's private randomness. A new subset on an old task still exposes an independent Bernoulli entry from (29). A repeated task/subset pair exposes a previously determined value and contributes zero additional KL. A new task is independent of the previous tasks. The conditional KL per candidate query is therefore at most $16\epsilon^2/A_r(k)^2$, regardless of adaptive choices. Summing by the transcript chain rule gives this constant times N as an upper bound on the total KL.

Let T be the event that the algorithm returns the positive hypothesis. Correctness under both hypotheses gives $\mathbb{P}_+(T) \ge 1-\delta$ and $\mathbb{P}_-(T) \le \delta$. Applying the log-sum inequality after grouping transcripts into T and its complement gives

$$\mathrm{KL}(\mathcal{L}_+, \mathcal{L}_-) \ge \mathrm{kl}(1-\delta, \delta) = (1-2\delta)\log\frac{1-\delta}{\delta}. \tag{31}$$

This quantity is of order $\log(1/\delta)$ uniformly over $0 < \delta \le 1/4$, proving (26). The estimator in (24) and thresholding its mean at zero give the matching upper order. Its concentration bound follows by multiplying the moment inequalities $\mathbb{E}e^{t(X-\mathbb{E}X)} \le e^{t^2(b-a)^2/8}$ for $X \in [a,b]$, optimizing the Chernoff bound, and applying it to both tails.

B.3 Taskwise-additive upper and lower bounds

If $B = 0$, every $D_{\{i\}} \in \{0,1\}$ and $D_{[k]} \in \{0,1\}$. If $B = 1$, all these differences are in $\{-1,0\}$. Thus the target identity (10) implies (11). If the identity holds for all subsets, the same sign argument shows that all singleton coefficients are zero or exactly one is nonzero. The response on every subset is then constant or the corresponding coordinate (possibly complemented). This elementary observation is not a new Boolean structure theorem.

For a uniform singleton on a fresh task, $W = kD_{\{I\}}$ has mean μ and second moment $k\mathbb{E}\sum_i D_{\{i\}}^2 \le k$. Put $X = W - \mathbb{E}W$. Then $|X| \le 2k$ and $\mathbb{E}X^2 \le k$. For $q \ge 2$, $|\mathbb{E}X^q| \le (2k)^{q-2}\mathbb{E}X^2$. Expanding the exponential and using $q! \ge 2\,3^{q-2}$ gives, for $0 \le t < 3/(2k)$,

$$\mathbb{E}e^{tX} \le \exp\left(\frac{t^2 k}{2(1-2kt/3)}\right). \tag{32}$$

Apply the product bound to N independent copies and set $t = \epsilon/(k + 2k\epsilon/3)$ in the Chernoff expression. This proves (13); for $\epsilon \le 1/8$, $N \ge 3k\epsilon^{-2}\log(2/\delta)$ suffices.

For the lower bound, draw a uniform parent bit. Under hypothesis $s \in \{-1, 1\}$, draw an activity bit A with

$$\mathbb{P}_s(A = 1 \mid B = b) = a_s(b) = 1/2 + s(1-2b)\epsilon. \tag{33}$$

Independently choose I^* uniformly in $[k]$ and set

$$C_S = B \text{ XOR } \big(A\mathbf{1}\{I^* \in S\}\big). \tag{34}$$

This is taskwise additive on every subset. Averaging first over the parent gives $\mathbb{E}D_{[k]} = s\epsilon$, and $\mathbb{E}D_S = s\epsilon|S|/k$. If t distinct singleton queries on a task have all agreed with the parent, the conditional probability of a change on an unqueried singleton is

$$p_s = \frac{a_s(B)}{k - ta_s(B)}. \tag{35}$$

Indeed the failure history has probability $1 - ta_s(B)/k$, and the joint event that this history occurs and the next singleton is active has probability $a_s(B)/k$. Once a hit occurs, all remaining singleton outcomes on the task are known and provide no more parameter information.

For opposite values $a, b \in \{1/2 - \epsilon, 1/2 + \epsilon\}$, let $p = a/(k - ta)$ and $q = b/(k - tb)$, where $0 \le t \le k - 1$. The bound $\mathrm{KL} \le \chi^2$ for Bernoulli laws yields

$$\mathrm{kl}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \le \frac{(p-q)^2}{q(1-q)} \tag{36}$$

$$= \frac{4\epsilon^2 k^2}{(k-ta)^2 b\,[k-(t+1)b]} \le \frac{16384}{81}\frac{\epsilon^2}{k} < 256\frac{\epsilon^2}{k}. \tag{37}$$

Here $a, b \in [3/8, 5/8]$, and both bracketed denominators are at least $3k/8$. The bound holds conditional on each possible parent and history, so adaptively switching tasks does not evade it. Repeated revealed queries contribute zero. The same transcript testing argument as above therefore gives $N \ge k\,\mathrm{kl}(1-\delta, \delta)/(256\epsilon^2)$, proving the lower order.

B.4 Bias, activity and the conditional proxy test

For arbitrary binary outcomes and uniform independent I,

$$\mathbb{E}(kD_{\{I\}}) = \mathbb{E}\sum_i D_{\{i\}}, \qquad \mathbb{E}(kD_{\{I\}})^2 = k\mathbb{E}\sum_i D_{\{i\}}^2. \tag{38}$$

Subtracting the squared expectation proves (15). Suppose the target identity fails on an event V of probability at most ζ. Outside V, the residual is zero and at most one singleton changes. On V, the common-parent sign constraint implies $|\sum_i D_{\{i\}} - D_{[k]}| \le k$, and at most k singletons change. Averaging proves both bounds in (17).

For attainability of the bias when $k \ge 2$, use a fraction ζ of tasks with $B = 0$, every singleton successful, and the full composition unsuccessful. All remaining tasks have constant-zero outcomes. The target gain is zero and the proxy mean is $k\zeta$. Thus the factor k cannot be removed from a guarantee using only this failure probability.

For the conditional test, let $p_+ = \mathbb{P}(D_{\{I\}} = 1)$ and $p_- = \mathbb{P}(D_{\{I\}} = -1)$. Independent draws give, conditional on $m = n_+ + n_- > 0$, a binomial count $n_+ \sim \mathrm{Bin}(m, p_+/(p_+ + p_-))$. Under $\mu_{\mathrm{proxy}} \le 0$, this success probability is at most $1/2$. Monotonicity of the binomial upper tail makes (42) a valid upper-tail p-value conditional on m; averaging over m gives unconditional validity. The zero-change event uses $p = 1$. This argument is fixed-horizon and iid; it does not apply without correction to adaptively selected search scores or optional stopping on repeated p-values.

B.5 Exact finite-pool interval and simultaneous decisions

For a stratum of size m, unknown error count K, and uniform sample size q, the error count X has mass

$$\mathbb{P}_K(X = x) = \frac{\binom{K}{x}\binom{m-K}{q-x}}{\binom{m}{q}}, \quad \max(0, q - m + K) \le x \le \min(q, K). \tag{39}$$

For a two-sided interval of noncoverage at most a, define

$$L(x) = \min\{K : \mathbb{P}_K(X \ge x) > a/2\}, \tag{40}$$
$$U(x) = \max\{K : \mathbb{P}_K(X \le x) > a/2\}. \tag{41}$$

These tails are monotone in K, so the retained integer set is an interval. For fixed true K, the probability that either of its own tail probabilities at the observed X is at most $a/2$ is at most a, by the two discrete tail tests and a union bound. Hence $\mathbb{P}_K(L(X) \le K \le U(X)) \ge 1 - a$. When $q = 0$ the interval is $[0, m]$; when $q = m$ it is $\{X\}$. An empty stratum has $m = q = X = 0$.

Use $a = \alpha/2$ in each of the two prediction strata. With probability at least $1-\alpha$, both true error totals lie inside their intervals. Since the coefficient of E_0 in (19) is positive and that of E_1 is negative, the resulting extrema give (20). False admission at threshold τ implies failure of that interval whenever $\mu_M \le \tau$. For decisions d with allocated levels α_d, the probability of any false admission is at most $\sum_d \alpha_d$, including dependent decisions and predetermined repeated looks. This argument does not authorize arbitrary new looks or label-dependent stratum allocations.

The implementation evaluates (39) with integer binomial coefficients and rational division. Binary search locates the two endpoints. A separate check evaluates the accepted endpoints and their adjacent excluded integers directly, verifying both the tail thresholds and the extremality of every saved interval.

B.6 The conditional paired proxy test

In independent uniform-singleton draws, condition on the number $m = n_+ + n_-$ of nonzero differences. Under $\mu_{\text{proxy}} \le 0$, the classical one-sided conditional paired sign test uses

$$p = 2^{-m} \sum_{j=n_+}^{m} \binom{m}{j}, \qquad p = 1 \text{ when } m = 0. \tag{42}$$

This is the familiar conditional test underlying paired proportion comparisons (McNemar, 1947). It tests the proxy; an unverified composition assumption cannot change that target. Repeated looks and candidate selection also require their own error control.

B.7 Exact residual uncertainty with zero observed errors

The gate has no universal savings factor because it depends on both the cached gain H and the measured errors. For example, if a sample in a stratum of size m observes no errors, its lower error bound is zero and its upper bound is the largest integer K satisfying

$$\frac{\binom{m-K}{q}}{\binom{m}{q}} > \alpha/4, \tag{43}$$

where a binomial coefficient is zero when its lower argument exceeds its upper argument. This is the exact probability of avoiding all K errors in a sample of size q. With zero observed errors in both strata, the gain lower bound is $H - U_1/M$: missed successes can only improve the target, whereas incorrect predicted successes can invalidate a positive cached gain. A favorable prediction can therefore permit admission before a census, but observing no errors does not itself make U_1 zero. When $P = B$, $H = 0$ and positive evidence must instead come from measured changes. This finite arithmetic explains the role of the cache without equating prediction accuracy with an assumption about composing updates.

C Complete cost comparison with full execution available

C.1 Access, stopping, and the resource being minimized

Fix the same k, r, assembly rule and iid task law as in Section 2. The allowed candidates have size $1, \dots, r$ or k. Intermediate sizes $r+1, \dots, k-1$ are not additional query options. A local size-j candidate is charged the known deterministic price $c_j > 0$, and a full candidate the price $c_F > 0$. A task handle and its parent outcome have no additional charge. Prices must therefore reflect the execution accounting being modeled; history-dependent savings from sharing parent executions are a different pricing model. In particular, this theorem is not a claim about measured token counts, wall-clock times, or amortization of a finite cache.

A strategy can use private randomness, adaptively choose a candidate and a new or existing task handle, and stop at an almost-surely finite time T. Its cost is the sum of the charged prices of its queries. Repeated observations reveal the same response and still carry their declared price; allowing such uninformative repeats cannot improve the infimum. Among strategies whose sign error is at most δ for every admissible law with $|\mu| \geq \epsilon$, minimize the supremum over those laws of expected total cost. A strategy with infinite worst-case expected cost cannot improve the finite upper bound below. The lower bound grants the parent and task identities for free but no latent outcome table or program-semantic information.

C.2 Upper bound and a finite sampling schedule

Put $V = A_c B_c$ and draw J with probability $q_j = |w_j|/(\sqrt{c_j} B_c)$. Draw a uniform J-subset on a fresh task and return $W = w_J D_S / q_J$. By the interpolation identity,

$$\mathbb{E}W = \sum_j w_j m_j = \mu, \tag{44}$$

$$\mathbb{E}W^2 \leq \sum_j \frac{w_j^2}{q_j} = A_c B_c = V, \tag{45}$$

$$\mathbb{E}[\text{price}] = \sum_j q_j c_j = A_c / B_c. \tag{46}$$

For a group of $m = \lceil 8V/\epsilon^2 \rceil$ independent draws, Chebyshev's inequality bounds the probability that its average differs from μ by at least ϵ by $1/8$. Take an odd number $L \geq 4\log(1/\delta)$ of independent groups. The probability that at least half the groups fail is at most $\exp[-2L(1/2 - 1/8)^2] = \exp(-9L/32) \leq \delta$. Outside that event, the median of the group averages differs from μ by less than ϵ, so its sign is correct under the promised gap. A median equal to zero can return either fixed sign; any resulting error is included in the event already bounded.

By Cauchy–Schwarz, $V \geq (\sum_j |w_j|)^2 \geq 1$. Since $\epsilon \leq 1/8$,

$$mL\, A_c/B_c \leq (8V/\epsilon^2 + 1) L\, A_c/B_c \leq 9L A_c^2/\epsilon^2. \tag{47}$$

Selecting the smallest odd L above the stated threshold establishes the universal upper order. For full queries use $W = D_{[k]}$, $V = 1$ and price c_F in the same argument. Selecting the smaller analytic factor before sampling proves the upper bound in Theorem 2. The local sample count is fixed, but the realized cost varies with the sampled shells; the theorem bounds expected cost, not a hard spending cap. The exact reference implementation alternatively selects the smallest odd L for which the binomial upper tail with success probability $1/8$ is at most δ, a valid, sometimes smaller finite schedule.

C.3 An admissible binary family for the joint access model

Let $L_j(t) = \prod_{0 \leq a \leq r, a \neq j} (t-a)/(j-a)$ and define

$$P_c(t) = \sum_{j=1}^{r} \text{sign}(w_j) \sqrt{c_j}\, L_j(t). \tag{48}$$

It vanishes at zero, takes value $\operatorname{sign}(w_j)\sqrt{c_j}$ at node j, and takes value A_c at k. For $t > r$, every numerator factor of $L_j(t)$ is positive and its denominator has sign $(-1)^{r-j} = \operatorname{sign}(w_j)$. Each summand of $P_c(t)$ is therefore nonnegative and nondecreasing for $t \geq r$. For integer $r < t \leq k$ this yields $0 \leq P_c(t) \leq A_c$. At nodes $j \leq r$, $\sqrt{c_j} \leq A_c$, since $|w_j|$ is a positive integer. Together these facts prove

$$|P_c(t)| \leq A_c \quad (t = 0, \ldots, k). \tag{49}$$

No assertion about between-node values is needed.

Under sign $s \in \{-1, +1\}$ let $g_s(S) = s\epsilon P_c(|S|)/A_c$. Expansion of the degree-r polynomial in the basis $\binom{t}{a}$, with zero constant coefficient, supplies precisely the multilinear mean representation (2). Its full gain is $s\epsilon$. On each fresh task draw $B \sim \mathrm{Ber}(1/2)$ and, independently for every nonempty subset, draw $C_S \sim \mathrm{Ber}(1/2+g_s(S))$. Set $C_\varnothing = B$. Equation (49) places every candidate parameter in $[3/8, 5/8]$, including the unqueried intermediate sizes. This is a jointly valid binary table with the required candidate-parent mean surface, rather than only a collection of compatible observed shell averages.

Parents and task identities have identical laws in both worlds. Conditionally on the history and a fixed hypothesis, any unrevealed entry on an old task is still independent of that history with its original Bernoulli parameter. A repeated entry is already known and contributes zero conditional KL. By (30), a newly revealed local entry of size j contributes at most $16\epsilon^2 c_j/A_c^2$; a full entry contributes at most $16\epsilon^2$. Consequently, after every feasible history, the conditional KL of the next observation is at most $16\epsilon^2/G$ times its charged price.

C.4 Adaptive stopping and the matching lower bound

Include the strategy's private random choices in the transcript and append an absorbing stop symbol after termination. Its decision rule, randomization and choice of the next query have identical conditional mechanisms under the two hypotheses; only the response law contributes conditional KL. For every finite transcript prefix, the chain rule gives

$$\mathrm{KL}(\mathcal{L}_+^{(n)}, \mathcal{L}_-^{(n)}) \leq \frac{16\epsilon^2}{G}\mathbb{E}_+ \sum_{t=1}^{T\wedge n} c(Q_t). \tag{50}$$

Relative entropy on increasing finite-prefix sigma-fields converges to the entropy of the full transcript, and the nonnegative cumulative cost converges monotonically. Hence

$$\mathrm{KL}(\mathcal{L}_+, \mathcal{L}_-) \leq \frac{16\epsilon^2}{G}\mathbb{E}_+[\text{total cost}]. \tag{51}$$

The terminal decision is measurable because termination is almost sure. Its probability of returning positive is at least $1 - \delta$ in the positive world and at most δ in the negative world. Binary data processing yields

$$\mathbb{E}_+[\text{total cost}] \geq \frac{G}{16\epsilon^2}\,\mathrm{kl}(1-\delta, \delta). \tag{52}$$

For $0 < \delta \leq 1/4$ the last KL is uniformly of order $\log(1/\delta)$. The two signs form a subfamily of the promised class, which proves the minimax lower bound, including adaptive mixtures and random stopping. No exact finite-error leading constant or instancewise inferiority of mixed strategies is asserted.

C.5 Exact design identity and its attaining response table

Consider the specific class in Proposition 3. A draw chooses $h \in \{1, \ldots, r, F\}$ with probability p_h and returns $(b_h/p_h)D_h$ on a fresh task, uniformly within a local shell. Types with zero coefficient may be omitted. For fixed coefficients, unbiasedness for every admissible mean surface is equivalent to

$$b_F = t, \qquad b_j = (1-t)w_j. \tag{53}$$

To see necessity, substitute $m_F = \sum_j w_j m_j$. The allowed shell-mean vectors contain a neighborhood of zero: any sufficiently small node vector interpolates to a polynomial whose values at all finitely many subset sizes are sufficiently small, and the independent binary construction realizes

it. Thus no nonzero linear coefficient can vanish on every allowed shell-mean vector. Sufficiency follows directly from the same interpolation identity.

For each such design, $\mathrm{Var}(W) \le \sum_h b_h^2/p_h$. This is its exact worst-case variance on the whole degree-r class: draw a fair parent and set $C_S = 1 - B$ for every nonempty S. All mean gains are zero, hence admissible, while each sampled difference is a symmetric sign. Therefore $\mathbb{E}W = 0$ and $\mathrm{Var}(W) = \sum_h b_h^2/p_h$. This worst-case design statement includes zero-gain instances; it is distinct from the gap-restricted sign-testing minimax statement. It also uses a fresh task on every draw, as required by the stated design class. It is not the independent-entry lower-bound family above.

Cauchy–Schwarz now gives

$$\inf_p \left(\sum_h b_h^2/p_h \right) \left(\sum_h p_h c_h \right) = \left(\sum_h |b_h| \sqrt{c_h} \right)^2 \tag{54}$$

$$= \left(|1-t| A_c + |t| \sqrt{c_F} \right)^2 . \tag{55}$$

Equality holds for $p_h \propto |b_h|/\sqrt{c_h}$ on the nonzero coefficients. Minimizing over real t gives exactly G: if $A_c < \sqrt{c_F}$ the unique minimizer is $t = 0$; if the inequality is reversed it is $t = 1$; at equality every $t \in [0, 1]$ is optimal. Values outside that interval cannot improve the result. This proves Proposition 3 using classical cost-aware allocation.

C.6 Consequences and exact verification

The explicit coefficient $|w_1| = k\binom{k-2}{r-1} \ge k$ implies $A_c^2 \ge k^2 c_1$. If $c_F \le k c_1$, then $A_c^2/c_F \ge k$. In particular, for $c_j = a + bj$ and $c_F = a + bk$ with $a, b \ge 0$ and $a + b > 0$, we have $c_F \le k(a+b) = k c_1$. At fixed r and fixed positive local prices, the leading expression is

$$A_c^2 = k^{2r} \left(\sum_{j=1}^{r} \frac{\sqrt{c_j}}{j!(r-j)!} \right)^2 + O_{r,c}(k^{2r-1}). \tag{56}$$

These statements quantify the savings that a cheaper local execution must overcome in the worst-case mean class. They neither rule out favorable instances nor convert an already acquired finite cache into a fresh-task population oracle.

The exact reference accepts rational cost roots, so its represented prices are their squares; the proof covers all positive real prices. An independently implemented checker verifies 456 plans with $2 \le k \le 12$ and $r \le 4$, 3,960 integer polynomial bounds and 3,960 binomial-basis identities, 6,600 discrete allocation comparisons and 2,736 attaining mixture designs. It also enumerates full small binary tables, verifies 2,098 conditional unrevealed-entry cases, checks 152 affine-price instances and all 456 finite median schedules, and exercises 18 exact sampler draws and 10 invalid inputs. These finite checks probe the formulas, implementation and boundary conventions; the complete arguments above establish the general claims.

D Data provenance and the fixed retrospective protocol

Source: https://github.com/gepa-ai/gepa-artifact.

Commit: `cbefbc1aa0f43dd39874ec4bf42211365dbda42e`.

Public experimental archive: 1,931,042,598 bytes.

SHA-256: `0e1db10331bbbdd48701443b0690d5b7b8abef841a73b76882d26853b3f38dd7`

The complete download matches the release's LFS pointer. Inventorying the 26,090 archive members retains names and sizes. We separately save 78 state/configuration files and 27 merge-run logs, with member hashes. The analysis does not execute program cloudpickles. A restricted reader accepts primitive state data and replaces the sole allowed global, `random.Random`, with an inert local state holder.

Inclusion and alignment. Every merge-enabled run is inventoried, including runs with no merge attempts. An accepted trace supplies branch IDs, common ancestor and new candidate ID. The recorded lineage must agree, both branches must descend from the ancestor, and neither branch can be an ancestor of the other. All four candidate program files must appear as regular archive members. Their per-task validation vectors must have equal nonzero lengths. Complete binary quartets additionally require all four values to be zero or one and their means to agree with recorded aggregate percentages to within 0.011 percentage points, allowing the release's aggregate rounding. No fractional score is silently converted to a binary success.

All nine initial vectors alias the saved Pareto vector and disagree with their initial aggregate means. The analysis retains the raw states, original logs and source versions for provenance. It clones arrays locally and recovers an initial vector only from a complete first validation block, sorted by validation counter, containing all task indices exactly once with counters $1, \ldots, M$ and a matching initial mean. Duplicate or missing indices prevent recovery; a rounded aggregate is never used to invent task scores. Three initial vectors pass, but only one such initial-parent quartet contributes to the complete binary analysis. Other complete quartets have nonzero ancestors whose original vectors pass their checks.

Table 2: Complete inventory of the nine merge-enabled runs. Missing-parent and nonbinary cases are kept as exclusions, not negative outcomes.

Dataset	Model	Attempts	Accepted	Binary kept	Excluded
HotpotQA	GPT-4.1 mini	7	4	4	0
HotpotQA	Qwen3-8B	8	8	5	3
IFBench	GPT-4.1 mini	3	3	0	3
IFBench	Qwen3-8B	5	4	0	4
LiveBench	GPT-4.1 mini	0	0	0	0
Papillon	GPT-4.1 mini	0	0	0	0
Papillon	Qwen3-8B	9	5	0	5
HoVer	GPT-4.1 mini	11	7	1	6
HoVer	Qwen3-8B	6	6	3	3

Pair-specific composition. The ancestor is $S = \varnothing$, each branch is its corresponding singleton, and the released merge is $S = \{1, 2\}$. The pair-specific assembly rule follows the recorded merge choices, yielding the four artifacts needed for the diagnostic. Validation scores were available during GEPA search, and task-level differences include the generation noise of those executions. The diagnostic therefore describes the recorded quartet rather than an isolated causal interaction effect or a model for unobserved compositions.

Replay choices and costs. All complete binary quartets have $M = 300$. Their canonical order is the order in the saved audit: run name and then trace order. Quartet j uses seed $20260912 + j$, starting at $j = 0$, to permute all task indices. Each prediction stratum uses the restriction of that same permutation. The prediction methods are parent, first branch, better branch by cached accuracy (ties select branch one), and clipped additive prediction. Budget $q \in \{20, 40, 80, 160, 240, 300\}$ allocates $q_0 = \lfloor qM_0/M \rfloor$, $q_1 = q - q_0$; all allocations are determined without audit labels. The threshold is zero. Each of the $13 \cdot 4 \cdot 6 = 312$ decisions receives $\alpha = 1/6240$, with $\alpha/2$ for each stratum interval and $\alpha/4$ per tail.

The replay simulates query access by hiding recorded full outcomes and counts each newly revealed outcome as one full-composition query. The first-admission budget is cumulative within a method because the samples are nested as its budget increases. Per-method totals are separate comparisons; they are not summed as joint physical expenditure. Parent and branch outcomes are assumed cached, and their original acquisition belongs to a separate cost account. Sharing records across other compositions might alter operational cost, but is not credited here.

Frozen and development evidence. The inclusion protocol preceded examination of archive outcomes. The residual-gate replay was specified after descriptive inspection and before executing the fixed permutations, thresholds and allocations. The original comparison uses this single design,

Table 3: All complete binary quartets. IDs are ancestor, branch 1, branch 2, merge. Gains and violation rates are percentages; activity is on its original scale.

Case	Artifact IDs	Proxy	Full	Residual	Activity	Violation
H/G-1	0,2,3,4	45.33	25.67	-19.67	0.480	23.00
H/G-2	3,12,14,16	12.00	6.00	-6.00	0.267	13.33
H/G-3	1,16,18,19	7.00	5.67	-1.33	0.203	5.33
H/G-4	1,17,18,21	6.67	5.33	-1.33	0.213	5.33
H/Q-1	1,6,8,9	15.33	11.00	-4.33	0.220	8.00
H/Q-2	3,8,10,11	12.00	9.33	-2.67	0.300	13.00
H/Q-3	3,8,12,13	11.00	8.67	-2.33	0.290	13.33
H/Q-4	3,13,14,15	16.67	8.67	-8.00	0.387	19.33
H/Q-5	6,10,14,17	9.00	6.00	-3.00	0.317	17.00
V/G-1	2,14,17,20	10.33	3.00	-7.33	0.350	20.67
V/Q-1	2,4,6,7	2.33	1.33	-1.00	0.337	18.33
V/Q-2	4,7,8,9	5.33	3.33	-2.00	0.373	19.67
V/Q-3	2,8,10,11	10.67	5.33	-5.33	0.380	21.00

Table 4: First admitting query budget in every fixed replay. A value of 300 is a census of the recorded pool.

Case	Direct	Branch 1	Better branch	Clipped
H/G-1	160	40	40	40
H/G-2	240	240	240	240
H/G-3	300	160	160	240
H/G-4	300	160	160	240
H/Q-1	240	160	160	160
H/Q-2	240	240	240	160
H/Q-3	240	240	240	160
H/Q-4	240	160	160	160
H/Q-5	240	240	240	240
V/G-1	300	300	300	300
V/Q-1	300	300	300	300
V/Q-2	300	300	300	300
V/Q-3	300	240	240	240
Total	3,400	2,780	2,780	2,780

without subsequent selection of seeds, thresholds or methods. Its guarantee is conditional on each recorded pool and the specified sampling procedure.

D.1 Query savings across sampling orders

The repeated-order analysis tests whether the query savings depend on the particular task order in the original replay. We retain its 13 pools, four cached predictors, six budgets, zero threshold, allocation rule and error budget, and freeze 1,000 new paired replays before running them. Each replay independently permutes the 300 tasks in every pool; all four methods share that permutation, restricted to their prediction strata. The master seed is 2026092601, and the artifact supplies every derived seed and permutation. No new model outcomes are generated. The original 18.24% comparison remains unchanged.

All three cached predictors save aggregate queries in every replay. The better branch has mean saving 16.06%, with a 5th–95th percentile range of 11.75–20.59%; clipping has mean 15.58% and a wider range, 9.87–22.01% (Table 5 and Figure 6). Their order is not fixed: clipping uses fewer total queries than the better branch in 312 replays, more in 441 and the same number in 247. The consistent benefit is correction with cached predictions, rather than a uniformly best predictor.

Table 6 locates the benefit within the pools. For example, the better branch saves 109.54 and 121.08 queries on average in H/G-3 and H/G-4, whereas V/Q-1 reaches the same first-admission budget under every method and order. Individual paired comparisons can favor direct validation even though

Table 5: Sampling-order variation over 1,000 paired replays of the 13 fixed pools. Query totals are averaged over replays. Savings are computed against direct validation within each replay, then averaged. Brackets give the 5th–95th percentiles of the replay distribution.

Prediction	Mean total queries	Mean saving (%)	Saving 5th–95th (%)
Direct	3,254.66	—	—
Branch 1	2,747.90	15.52	[10.90, 20.48]
Better branch	2,730.08	16.06	[11.75, 20.59]
Clipped	2,746.00	15.58	[9.87, 22.01]

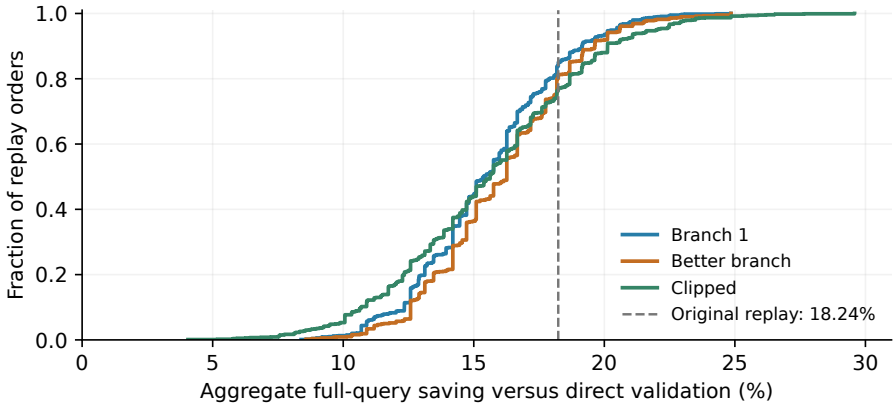

Figure 6: Distribution of paired aggregate query savings over 1,000 sampling orders of the same 13 pools. Each curve compares one cached predictor with direct validation within the same replay. The dashed line marks the original 18.24% result, retained as a single-order comparison.

aggregate savings remain positive. The artifact retains these outcomes, complete intervals and first budgets.

Table 6: Paired full-query savings in each pool over 1,000 sampling orders: mean [5th, 95th percentile]. Positive values mean fewer queries than direct validation in the same replay. The complete distribution, including occasional negative savings, is retained in the artifact.

Case	Branch 1	Better branch	Clipped
H/G-1	50.96 [40, 120]	50.96 [40, 120]	50.92 [40, 120]
H/G-2	24.94 [0, 80]	24.94 [0, 80]	40.70 [0, 140]
H/G-3	109.54 [80, 140]	109.54 [80, 140]	67.94 [0, 140]
H/G-4	121.08 [80, 140]	121.08 [80, 140]	68.28 [0, 140]
H/Q-1	34.16 [0, 80]	38.56 [0, 80]	52.88 [0, 160]
H/Q-2	9.08 [0, 80]	9.08 [0, 80]	58.28 [0, 80]
H/Q-3	4.74 [0, 80]	4.74 [0, 80]	39.70 [0, 80]
H/Q-4	76.50 [80, 80]	76.50 [80, 80]	50.02 [0, 80]
H/Q-5	16.46 [0, 60]	36.42 [0, 60]	31.90 [0, 60]
V/G-1	1.20 [0, 0]	0.36 [0, 0]	0.90 [0, 0]
V/Q-1	0.00 [0, 0]	0.00 [0, 0]	0.00 [0, 0]
V/Q-2	7.50 [0, 60]	1.80 [0, 0]	3.48 [0, 60]
V/Q-3	50.60 [0, 80]	50.60 [0, 80]	43.66 [0, 60]

The reported percentile ranges describe variation in randomized audit order, with each recorded pool held fixed. Each replay retains the original simultaneous error budget 0.05 over its $13 \cdot 4 \cdot 6$ decisions. Repetition characterizes the procedure's query cost; it does not combine the replays into an additional family of acceptance decisions.

E REPRODUCIBILITY, INTERPRETATION AND ETHICS

The implementation uses Python 3.12 and exact integer/rational arithmetic for the mathematical diagnostics and statistical gate. NumPy and Matplotlib are used for numerical figure generation. The manuscript uses the supplied ICLR 2027 style files. Public archive processing is CPU-only. The figure builder reads the recorded rational statistics; it does not rerun an LLM or replace missing outcomes. The complete original audit remains separate from any fresh reproduction outputs.

Verification layers. The retained mathematical and data-integrity checks cover 456 price plans, 91 rational statistics, 624 interval endpoints for 312 decisions, and 1,012 small-population coverage distributions. The separate mechanism checker verifies 312 design variances and 726 hypergeometric distributions. The composition checker verifies the shell weights by independent product and binomial representations, the subset-monomial averaging identity, extremal polynomial values, exhaustive small Boolean response tables, finite-distribution moments and both lower-bound information calculations. The KL checks use a rational logarithm series with a bounded remainder. Retained source versions document the correction of an earlier chi-square/KL mismatch in the checker.

For released records, the independent reduction checks all 49 merge events against separate text logs, reconstructs retained task rows from inert state data and eligible original log blocks, and recomputes 91 reported scalars through the 16 possible binary quartet cells. The residual gate's small-population coverage checks enumerate every underlying error count and feasible sample outcome for pool sizes zero through ten at two error levels. Separate endpoint checks directly evaluate each of the 624 saved stratum intervals and neighboring excluded totals. These are numerical and implementation checks; the general theorems rely on their proofs.

Reproduction path. The delivered sampling-order artifact starts from the 13 audited binary pool arrays. It includes the original replay, all prescribed seeds, input hashes and the exact interval routine. The replay first matches the original sample indices, interval endpoints and first-admission budgets, then regenerates the 1,000 paired replays, their intervals and cost summaries. An independent check reconstructs every sample from its saved permutation, checks all 312,000 intervals through the original gate interface, and generates the distribution figure and tables. The package uses relative paths and requires no model access or external data download. The accompanying mechanism checker, anonymous quartet inputs and expected results reproduce the 3,900 rows, 312 design variances and 726 small-population variance checks. The artifact documentation gives the execution instructions. Upstream archive extraction, initial-vector recovery and the other mathematical checks are documented above and in Appendix D.

Interpretation of verification. The checks establish consistency among the retained calculations, source records and implementation. Separate implementations provide computational cross-checks. The 3,900 quartet–task rows include recurring tasks across related quartets, and the inferential guarantee is conditional on each fixed pool and its sampling design.

Ethics and responsible interpretation. This work analyzes public research artifacts and performs no human-subject study or deployed intervention. A positive mean gain permits harmed tasks and does not establish safety or fairness of an agent. Data-integrity problems in a saved release are reported with a bounded, documented correction rather than attributed to an unverified current implementation. Appendix D records the inclusion criteria and exclusions; the artifact retains all method comparisons. The authors retain responsibility for the work and its disclosures.

F ERROR-MECHANISM DIAGNOSTICS AND INPUT RECOVERY

Four artifacts must be recoverable. A complete quartet contains a parent, two branches and their accepted merge, with aligned binary task outcomes. We verify branch ancestry, the merge trace, artifact membership, vector lengths and aggregate scores. Across nine merge-enabled runs, saved state and text logs agree on 49 attempts. Twelve lack a full candidate record. Of 37 accepted merges, 12 have nonbinary scores and 12 lack a reliably recoverable initial-parent vector, leaving 13 complete binary quartets and 3,900 task rows.

In all nine saved states, the initial score vector aliases a mutable validation vector and disagrees with its initial aggregate. We recover an initial vector only from a complete, consistent initial log block. Three runs pass; one recovered initial-parent quartet enters the analysis. The original-byte checks, recovery criteria and exclusions are documented in Appendix D. The current pinned source copies the list, so this saved-state finding is not attributed to its implementation.

F.1 Matched-cost interpretation of proxy error

On a frozen empirical task distribution, consider fresh independent draws with replacement. Let $V_L = kh - \mu_{\text{proxy}}^2$, $V_F = \text{Var}(D_{[k]})$ and $b = \mu_{\text{proxy}} - \mu$. For an exactly divisible charged budget T, $N_L = T/c_1$ local observations have MSE $b^2 + c_1 V_L/T$; $N_F = T/c_F$ full observations have MSE $c_F V_F/T$. The local mean has smaller MSE precisely when $Tb^2 < c_F V_F - c_1 V_L$. If the right side is nonpositive, it cannot help; if $b \neq 0$, increasing the budget eventually exposes the bias floor. This is an elementary estimator comparison, not an adaptive minimax theorem or the without-replacement audit protocol.

The exact saved-pool ratios V_L/V_F range from 2.68665 to 4.39285. Consequently, even a variance-only comparison requires full execution to be more expensive than that ratio before a local advantage is possible on these empirical laws; nonzero bias further limits it. These are conditional diagnostic ratios, not measured execution prices or evidence that the population interaction degree is one.

F.2 Portable verification

The supplied mechanism checker reads the anonymous quartet audit and recomputes all 3,900 binary rows, proxy and full moments, clipping residuals and 312 conditional design variances. It independently checks the hypergeometric variance expression by exact distribution enumeration for all populations $2 \leq M_b \leq 12$, error totals and positive sample sizes: 726 comparisons. The reported sampling variance comes from the stated stratified randomization; no new random permutation is generated. The earlier interval replay retains its own source and coverage checks. Original state/log recovery is a separate upstream step.