CONTRAST-FIRST CALIBRATION: REUSING LABELS TO COMPARE SYNTHETIC SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

A cheap synthetic source is useful when its errors can be corrected with affordable real labels. Choosing that correction also uses labels, which should remain available for training. We study how to evaluate these choices when the labels are reused. Our central protocol conditions on a finite, paid source bank and allows its data to determine the correction subspaces. For hard corrections with a common real-data count, a pairwise risk error separates into two random quantities and a deterministic retention factor. One confidence event therefore covers every later real-data budget with the same radius as the corresponding fixed-budget construction. The assessment extends to unknown Gaussian noise variance through regression residuals. For a broader pooled-data selector, a risk-localized oracle bound replaces a discrepancy-dependent square-root penalty with an additional cost on the learning-risk scale. Label-level experiments use 29 recipes, a learned correction basis, estimated variance, and cost-matched holdout certificates. They identify both successful source correction and a shared-error failure mode. A matched diagnostic separates the probability-allocation cost of a general confidence event from its coefficient-envelope cost. The resulting framework provides computable risk comparisons for retaining calibration labels and choosing subsequent acquisition plans.

1 INTRODUCTION

An approximate simulator can be wrong in a simple, repeatable direction, while a more accurate source spreads its smaller errors over many directions. Real labels can repair the first source efficiently, but the learner must find a useful correction and pay for that discovery. Residual learning already exploits simple discrepancies (Ramakrishnan et al., 2015; Lin & Reimherr, 2024; Zhou, 2026); our question is how to *evaluate a source-and-correction choice without discarding the labels that informed it*.

A common workflow has a reusable source bank and an expanding target dataset. A practitioner trains a few inexpensive predictors, examines their disagreement to propose correction subspaces, and collects a first real-data batch. That batch both ranks the corrections and improves the predictor. A later decision may fix the next batch size after inspecting the assessment, subject to a remaining acquisition allowance. Reporting the uncertainty of each source separately can hide a clear comparison: noise shared by two loss estimates may cancel when they are subtracted. Treating retained labels as independent training data causes a different problem, because the selected predictor retains the noise that made it look attractive.

A worked two-direction example. Let the frozen source predictions be $G_1 = (1, 0.3)$ and $G_2 = (0, 0.5)$, with target $f = (0, 0)$ for illustration. Their raw squared errors are 1.09 and 0.25. Correcting the first coordinate of source 1 using real data leaves squared source error 0.09, whereas source 2 still has 0.25. With a real pilot $Y = f + \varepsilon$ of covariance vI, both corrected candidates retain the same first-coordinate pilot noise. Their estimated risk difference has error $0.4\varepsilon_2$; the term involving ε_1^2 cancels. If more real labels are later pooled, the common first-coordinate contribution changes in both risks by the same amount. Thus the comparison can be checked once and retained across later budgets. With different correction strengths, an additional centered quadratic contrast survives; the main theorem bounds it explicitly.

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.

For a numerical example, suppose $v = 0.01$ is known and the observed second pilot coordinate is $Y_2 = 0.02$. The estimated corrected-risk difference is approximately -0.15. Its standard deviation is 0.04, so the standard two-sided 95% Gaussian interval is approximately $[-0.23, -7.36 \times 10^{-2}]$. This observed bank favors the first correction at that confidence level despite its worse raw error, and the same interval applies after any common real-data extension. The example illustrates what a certificate means: evidence for a comparison among specified predictors, not a statement that the chosen predictor is error-free. Its bank values and noise scale are supplied for illustration; the experiments charge for fitting the bank and estimate the real noise.

What the paper establishes. We develop this contrast-first approach in two acquisition protocols. The principal one freezes the paid synthetic predictors before real calibration and learns hard correction subspaces from source-only data. Its conditional comparison needs only two random error quantities, yielding budget-uniform coverage without an additional uniformity penalty. The second permits later synthetic acquisition as well. For this broader protocol, we give restricted-domain certificates and a risk-localized expected-risk comparison with a full-budget recipe oracle. The two protocols have different comparators; Table 1 makes the distinction explicit.

A fixed budget should use a fixed-budget assessment unless additional flexibility is needed. Our uniform result is useful when the next real batch is chosen *after* the pilot, or when one existing source bank is reused under several later acquisition budgets. We measure this situation directly, alongside a decomposition showing how much an unnecessarily broad event costs. Transport optimization is optional: any feasible coupling gives a valid bound on the same confidence event.

Connection to prior guarantees. Risk estimation after model selection is classical (Mallows, 1973; Stein, 1981; Kneip, 1994; Birgé & Massart, 2001). Corrected oracle results for estimators tuned by Stein's unbiased risk estimate (SURE) already localize excess risk (Cauchois et al., 2021); ordered-smoother aggregation can achieve a cost independent of menu cardinality in its structured setting (Bellec & Yang, 2020). Unknown-variance estimator selection is also established (Baraud et al., 2014). We do not replace those results. At full retention, our isotropic hard-correction score is exactly SURE for an affine estimator. We instead study an *observable conditional comparison of future pooled predictors*, with a source-conditioned learned menu, explicit acquisition accounting, and a budget chosen after calibration. Appendix E compares these objects at theorem level.

The experiments include actual Gaussian regression labels, not only coefficient draws. Source noise is finite, the correction basis is learned, and variance is estimated from residuals. A strong holdout certificate is sometimes tighter, especially at small dimension. Source disagreement can also miss a bias shared by all sources. These controls identify when the assessment supports a useful correction and when it merely certifies that the available synthetic recipes should be declined.

2 A PAID SOURCE BANK AND A REUSABLE TARGET PILOT

Let $f \in \mathbb{R}^d$ represent the target in a prediction metric where squared coefficient error is squared prediction error. A source bank contains fitted vectors G_h. A recipe a selects a source $h(a)$ and an orthogonal projector P_a: its retained source component is $P_a G_{h(a)}$, and its corrected component is learned from real labels in the complementary subspace. Projectors need not be diagonal or mutually commuting. They may be chosen by any procedure using the source data, including source disagreement. Write $G_a = G_{h(a)}$.

What is conditioned on. We condition on the full source bank and its recipe-construction procedure before analyzing an independent real pilot $Y = f + \varepsilon$, $\varepsilon \sim N(0, S)$. This treats the bank as observed, not noiseless: its estimation error remains inside $G_a - f$, and every source label is charged. Its distribution need not be Gaussian. Learning the menu from real calibration labels would require a separate construction split or a larger simultaneous family; the theorem below does not cover this choice.

For clarity, first take $S = vI$ with $v = \sigma^2/p$. For balanced fixed-design regression, p is an actual label count and the pilot is its least-squares sufficient statistic. If the total real count later becomes $N \geq p$, all recipes share the same N and retention $\alpha = p/N$. The final predictor is

$$\widehat{f}_a(N) = P_a G_a + (I - P_a)\overline{Y}_N. \tag{1}$$

The source bank is frozen, while $\overline{Y}_N$ pools the pilot with independent new real observations. Its conditional noise covariance is $\alpha(1-\alpha)S$. Define

$$Q_a(\alpha) = \mathbb{E}[\left\|\widehat{f}_a(N) - f\right\|^2 \mid \text{bank}, Y]$$
$$= \|P_a(G_a - f)\|^2 + \alpha^2 \|(I - P_a)\varepsilon\|^2 + v\alpha(1-\alpha)\,\mathrm{tr}(I - P_a), \tag{2}$$
$$\widehat{Q}_a(\alpha) = \|P_a(G_a - Y)\|^2 - v\,\mathrm{tr}\,P_a + v\alpha\,\mathrm{tr}(I - P_a). \tag{3}$$

The score is unbiased for $\mathbb{E}Q_a$ before selection, but Q_a is the realized-pilot target. At $N = p$, there is no future fitting noise: $Q_a(1)$ is the actual prediction loss of the already fitted candidate.

What is reported. A simultaneous contrast radius $\rho_{ab}(\alpha)$ gives the observable certificate

$$C_a(\alpha) = \max_b\{\widehat{Q}_a(\alpha) - \widehat{Q}_b(\alpha) + \rho_{ab}(\alpha)\}, \qquad \rho_{aa} = 0. \tag{4}$$

On its confidence event, $Q_a - \min_b Q_b \le C_a$ for any data-selected recipe and permitted retention. The maximum includes the self-comparison, so $C_a \ge 0$. This is regret relative to the feasible conditional menu. To compare with skipping source acquisition, the experiments additionally report the full-budget real-only risk, with all source expenditure converted into real labels.

3 Contrast first: one event for every later real budget

The main simplification is algebraic. For a pair of recipes let $D_{ab} = P_a - P_b$ and $g_{ab} = -P_a(G_a - f) + P_b(G_b - f)$. Writing $e_a = \widehat{Q}_a - Q_a$, direct expansion gives

$$e_a(\alpha) - e_b(\alpha) = 2g_{ab}^\top\varepsilon + (1+\alpha^2)\{\varepsilon^\top D_{ab}\varepsilon - \mathrm{tr}(D_{ab}S)\}. \tag{5}$$

The first error is linear and the second is centered quadratic. Neither random quantity depends on α. Hard corrections remove a bias–retained-noise cross term; the common real count collects the two retained quadratic terms into one contrast. Bounding absolute risks first would miss these cancellations.

The remaining linear scale is unknown because g_{ab} contains f. It has the observed estimate $\widehat{g}_{ab} = -P_a(G_a - Y) + P_b(G_b - Y) = g_{ab} + D_{ab}\varepsilon$. For known full covariance define $T_{ab} = S^{1/2}D_{ab}S^{1/2}$, and let U_{ab} be the observable upper bound on $\left\|S^{1/2}g_{ab}\right\|^2$ from Lemma 4, applied to $S^{1/2}\widehat{g}_{ab}$. This uses a debiased squared norm rather than simply adding the entire noise norm.

Theorem 1 (A reusable contrast event)**.** *Condition on a paid source bank and $K \ge 2$ source-measurable orthogonal-projector recipes. Suppose the independent real pilot has known covariance S. With $t = \log(5\binom{K}{2}/\delta)$, define*

$$\rho_{ab}(\alpha) = 2\sqrt{2tU_{ab}} + (1+\alpha^2)\{2\sqrt{t}\,\|T_{ab}\|_F + 2t\,\|T_{ab}\|_{\mathrm{op}}\}. \tag{6}$$

With probability at least $1-\delta$, every pair satisfies $|e_a(\alpha) - e_b(\alpha)| \le \rho_{ab}(\alpha)$ for all $\alpha \in [0,1]$. Consequently (4) *bounds conditional regret for any recipe and later real count selected after the pilot. The same event and formula are obtained by the corresponding fixed-count proof using the linear and quadratic bounds separately.*

Proof idea. A Gaussian linear tail, a noncentral mean-norm bound, and a Gaussian quadratic tail fail with total probability at most $5e^{-t}$ per pair. Their intersection bounds the two random quantities in (5), so changing their deterministic multiplier requires no additional event. Appendix A gives all constants and the proof.

Why the conditioning matters. The theorem can be applied after training a noisy synthetic predictor because it targets the usefulness of that *observed* predictor, not the unobserved infinite-data source mean. Source fitting error therefore requires no extra confidence radius once the bank is conditioned on, but it still increases the predictor's loss and costs labels. This distinction is useful when a source bank is expensive to replace: the immediate question is which existing correction can be supported by the target measurements already available. Conditioning also permits a source-learned rotation. It cannot reveal errors shared by all sources, which motivates the shared-bias control below.

Table 1: The guarantees answer different questions. All final counts are fixed before observing their new fitting labels. "Bank" includes its paid observations and learned projectors.

Result	Allowed decision and assumptions	Comparator and output
Theorem 1	Frozen bank; hard projections; common real count chosen after pilot	Conditional menu oracle; observable certificate for every later count
Corollary 2	Same protocol; residual interval for scalar Gaussian variance	Same oracle, with uncertainty in the noise variance
Reference gate	Same protocol; compare only to all-real	Same-paid-data all-real plan; observable no-harm gate
Theorem 3	Both streams may grow; diagonal contractions; counts fixed before pilot	Full-budget recipe oracle plus paid overhead; expected-risk bound
Appendix C	Separate retentions in predeclared rectangles	Conditional menu oracle; more general, often wider certificate

When $P_a = P_b$, $D_{ab} = 0$ and the quadratic term disappears completely. With different corrections, the quadratic cost depends on their difference, not on the full dimension separately. The equality of fixed- and variable-budget events is specific to this protocol and this construction; an optimized fixed-budget confidence interval can still exploit more distributional information.

Estimated variance rather than assumed variance. For actual labels with $X^\top X = pI$, least squares has covariance vI. Gaussian residuals supply a chi-square interval $[l, u]$ for v, independent of the fitted mean. Use $(l + u)/2$ in the score. The upper endpoint controls tail scales, while the lower endpoint must be used when subtracting noise from a squared norm.

Corollary 2 (Residual-calibrated assessment)**.** *If* $\mathbb{P}(v \notin [l, u]) \le \delta_v$*, the radius in* (15)*, computed with* $t = \log(5\binom{K}{2}/\delta_m)$*, gives the conclusions of Theorem 1 with confidence* $1 - \delta_v - \delta_m$*. The radius includes the score-centering allowance* $\frac{1}{2}(u - l)(1 + \alpha)|\operatorname{tr}(P_a - P_b)|$.

The practical effect of the extra trace allowance is largest when two corrections retain very different numbers of directions. It vanishes for projectors of equal rank, even though their subspaces may differ. Simply substituting an estimated variance into a known-variance certificate does not provide this protection. Our implementation records the residual degrees of freedom and both interval endpoints, so the user can see the additional uncertainty.

The interval protects against uncertainty in a scalar Gaussian noise level. Arbitrary covariance underestimation remains outside this guarantee; known full covariance uses Theorem 1 directly.

An acquisition assessment in practice. After the pilot, enumerate allowable next batch counts, compute scores and certificates, and choose the smallest count meeting a supplied loss tolerance. A continuous budget range is also covered. If none is certified, report that fact rather than interpreting a wide bound as a high loss. Alternatively, compare each recipe only against the all-real candidate and deploy a synthetic correction only if its upper contrast is nonpositive. This reference gate needs $K - 1$, rather than $\binom{K}{2}$, comparisons. Its protected baseline uses the same already paid target data.

The output distinguishes a prediction from a certification status. Failure to certify is not an instruction to purchase arbitrarily many labels; a user can cap the allowed counts or choose a holdout protocol instead. An externally imposed tolerance can be absolute prediction loss, a fraction of the observed noise level, or a task-specified accuracy budget. None requires the evaluator's true discrepancy profile. A zero menu-regret certificate means the chosen plan is best in that menu; it does not imply zero prediction error or certify that the original source-bank purchase was economical.

4 Localizing the cost of selecting a pooled estimator

The preceding assessment starts with a paid source bank. To analyze a fuller acquisition protocol, let both real and synthetic streams receive fresh observations after calibration. Fix their future counts for each recipe before the pilot, allow diagonal weights in $[0, 1]$, and select the smallest pooled-count score. Bounding selection error by the largest possible source discrepancy incurs a $P^{-1/2}$ penalty,

Algorithm 1: evaluate a frozen bank, then choose a real batch

1. **Build independently.** Fit source anchors, construct the projectors using source data, and include $P_0 = 0$. Record every source cost.
2. **Fit and calibrate.** Collect p real labels. Compute Y, a residual variance interval, and the scores in (3). Reserve error probability for both residual and pairwise events.
3. **Evaluate a proposed count.** For any $N \geq p$, set $\alpha = p/N$, select the smallest score, and compute its $K - 1$ comparisons in (4). Use the all-pairs confidence parameter because the selected index depends on Y.
4. **Choose before sampling.** Select the first allowed N meeting the supplied tolerance. If none qualifies, return the current fit with an uncertified status. A reference-only gate may instead compare with all-real.
5. **Execute once.** Acquire the chosen new real batch and pool it in (1). Inspecting these new labels to revise N requires a new analysis.

which can overwhelm a B^{-1} fitting risk and push the planned pilot fraction to one. The following comparison instead localizes to the actual reference risk.

Theorem 3 (Risk-localized paid selection)**.** *Let ν be the largest independent pilot coefficient variance and let r be the expected risk of any fixed, feasible paid recipe. Define $L = \log(2K)$,*

$$u = 4\nu(L+1), \quad V = \nu\{d + 2\sqrt{d\log K} + 2\log K\}, \quad H = 6\nu\sqrt{dL} + 4\nu L.$$

For the pooled-score selector, with $a = 2\sqrt{u}$ and $b = 2\sqrt{uV} + H$,

$$\mathbb{E}\left\|\widehat{f} - f\right\|^2 \leq \Psi(r, \nu, d, K) := \left\{\frac{a + \sqrt{a^2 + 4(r+b)}}{2}\right\}^2. \tag{7}$$

In particular, $\Psi - r$ is at most a numerical constant times $\sqrt{r\nu\log(2K)} + \nu\sqrt{d\log(2K)} + \nu\log(2K)$.

The proof bounds the selected bias by the selected conditional error plus retained noise, then solves a self-bounding inequality. It integrates over all Gaussian outcomes. For the full-budget recipe oracle, use its risk plus its paid allocation overhead as r (Appendix B.4). If the pilot variance and oracle risk are both $O(B^{-1})$, the selection surcharge is now on that scale for fixed d, K.

What this changes for pilot planning. With stream costs c_h and noise scales σ_h, equalizing pilot variances gives $\nu = \Lambda/P$, where $\Lambda = \sum_h c_h\sigma_h^2$. A fully supplied objective replaces r by the fresh all-real bound $dc_0\sigma_0^2/(B - P)$. This objective needs no supplied bias norm, and its minimizing fraction P/B is exactly independent of B. In our 29-recipe sweep it chooses 68.93% at every budget; a discrepancy-dependent global-bound baseline rises from 67.17% to 91.44%. This corrects the rate problem, but the finite-sample bound is still conservative. Because its supplied reference already subtracts pilot expenditure, it is not an initial certificate that buying a pilot beats skipping it. We evaluate the selected fractions empirically rather than identifying an upper-bound minimizer with an optimal predictor.

5 Choosing the scope of an assessment

Fixed, restricted, or fully uniform. When the final budget is known, use a direct fixed-plan contrast. When only a plausible range is known, declare a retention rectangle before the pilot and reparameterize it onto $[0, 1]^2$. The same nine Bernstein coefficients cover that rectangle, with different matrices. Distinct coefficient errors should be contrasted first; any inexpensive feasible coupling of their weights yields a valid bound. Transport optimization can tighten that bound but is optional. Appendix C provides the exact restricted matrices and proof.

Measured price of generality. We use $d = 256$, real precision $p = 256$, source precisions $q_1 = q_2 = 1024$, and budget $B = 1382.4$. Real cost and noise standard deviation are one; both source costs and noise standard deviations are 0.01. The source biases are $0.35e_5$ and $e_1 + 0.30e_5$, and both recipes

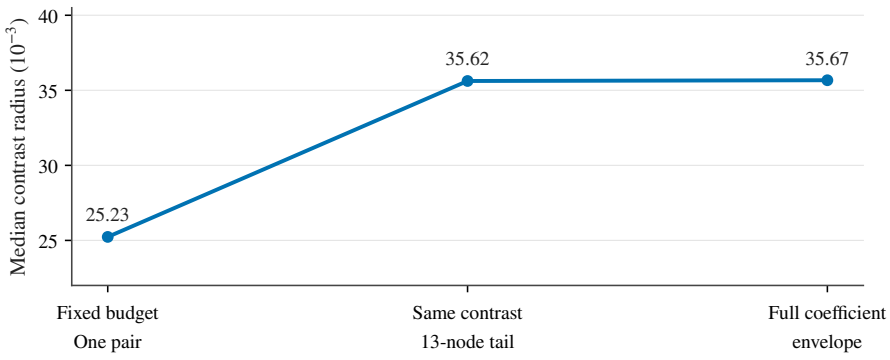


Figure 1: The price of a general two-stream confidence event, decomposed on the same 1,000 pilots. Moving the fixed contrast from a one-pair tail parameter to the 13-node simultaneous parameter increases its radius by about 41.17%. The Bernstein/coefficient envelope adds about 0.14% beyond that. These are contrast radii, not the smaller regret certificates obtained after subtracting observed score gaps. The frozen-bank theorem uses a different acquisition protocol.

correct the first four coordinates. Counts follow the pooled allocation in Appendix B.4. The same 1,000 Gaussian pilots (seed 963101) are reused for every construction, at $\delta = 0.05$. In this comparison, the direct fixed-budget radius, the direct radius at the general event's probability parameter, and the coefficient-envelope radius have medians 2.52×10^{-2}, 3.56×10^{-2}, and 3.57×10^{-2}. Thus almost all inflation comes from the union-bound size, not corner over-approximation (Figure 1). Restricting the acquisition budget to a predeclared $\pm 5\%$ interval does not rescue this example: lost coefficient deduplication slightly worsens the bound.

Computational path. The frozen-bank assessment needs one pairwise computation per eligible pair. For a completed score-selected plan, evaluate only its $K - 1$ certificate comparisons, while the probability parameter still accounts for all $\binom{K}{2}$ potentially selected pairs. In a common learned basis with hard masks this costs $O(Kd)$ per decision and $O(Kd)$ storage for anchors and masks. Budget-dependent scores cost $O(K)$ after residual norms are cached. A reference-only gate uses $O(Kd)$ work and a smaller predeclared event. No transport solver is involved. Arbitrary dense projectors and covariance require additional matrix preprocessing; the diagonal-in-a-learned-basis experiment is the inexpensive special case.

6 Experiments: Labels, Certificates, and Acquisition Choices

Experiments use deterministic seeds and per-trial evaluation. We report prediction error, certificate distributions normalized by a common risk scale, oracle-normalized widths, and paired Monte Carlo intervals. Thresholded useful fractions remain secondary. Each method pays for its source labels, basis construction, target fitting, and any validation split. Only the evaluator uses the true target.

6.1 A learned menu with estimated noise and a holdout certificate

The principal test generates Gaussian labels on balanced cosine designs. Both source and real noise standard deviations are one, and the real noise level is *estimated*, not supplied to the assessment. Dimensions are $128, 512, 1024, 2048$, with 400 independent confirmation seeds per geometry. Each source receives $8d$ labels to construct a disagreement direction and another $32d$ labels; all $40d$ are retained in its anchor. A Householder rotation aligns the first coordinate with the measured source disagreement. The 29-recipe menu contains two sources, 14 hard correction strengths each, and all-real. Its basis is therefore learned from finite source data rather than provided by the target oracle.

Every method buys $4d$ real labels. The reuse method pools all of them. The holdout method fits with $3d$ and validates with d, using a Gaussian pairwise confidence interval for the *certificate itself*. Both estimate variance from the same independent regression residuals and charge its 1% error probability; comparisons receive 4%. Source labels cost 0.01 and real labels cost one, so total budget is $B = 4.8d$

and the continuous-count real-only benchmark is $d/B \simeq 0.21$. Each method's certificate compares against its own feasible fitted bank. The 83.3% initial real-data share is used entirely for the current fit by reuse; it is not a tiny preliminary pilot.

Separating a useful predictor from a useful certificate. A low predicted risk and a small certificate are measured separately. The normalized width distributions show how much precision the assessment supplies before any particular tolerance is chosen. The holdout method deliberately keeps validation labels out of its fitted bank. Refitting on those labels while retaining the same independent Gaussian interval would invalidate that interval. Conversely, forcing the reuse predictor to discard them would erase the efficiency mechanism being tested. Matching purchased observations and prediction families, while allowing their legitimate fitting protocols to differ, is therefore the appropriate comparison. Cross-fitting or aggregation could provide additional competitors; these experiments do not exhaust all ways to exploit a split.

A completed-fit oracle is computable only by the evaluator and is used solely as a reporting denominator; the algorithm receives the scores, variance interval, and confidence radii. We also retain cases where the selected predictor has exactly zero menu regret, reporting their frequency rather than forming an unstable certificate-to-zero ratio.

In the disagreement geometry the two sources err mainly in different signed amounts along one dense direction. Reuse has lower mean prediction error than holdout at every tested dimension, but holdout certificates are substantially tighter at the two smaller dimensions (Table 2). At $d = 2048$, their paired risk difference is -1.86×10^{-3}, with pointwise 95% interval $[-2.60 \times 10^{-3}, -1.12 \times 10^{-3}]$. Both beat full-budget real-only learning. Reuse's median oracle-normalized regret certificate is 0.56, with the complete distribution in Figure 2. Against the same frozen bank and real labels, leaving the basis unlearned gives mean error 0.25 instead of 5.75×10^{-2}. The paired learned-minus-fixed interval is $[-19.29, -19.14] \times 10^{-2}$.

A failure that the learned basis cannot remove. In the shared-bias control, both sources have the same additional error orthogonal to their disagreement. Learning from disagreement cannot reveal it. At $d = 2048$, the selector returns the feasible all-real recipe, whose mean error is 0.25; the full-budget real-only relaxation is 0.21. Its zero menu-regret certificate correctly evaluates the available recipes but does not recover the sunk source cost. This distinguishes an informative assessment from a claim that source acquisition was worthwhile. The reference gate likewise rejects synthetic corrections in this control. It accepts them on all 400 disagreement trials at $d = 2048$, but only 43.25% at $d = 1024$. Holdout's reference gate is substantially less conservative in that intermediate case. At $d = 1024$, the reference-gated reuse predictor has mean risk 0.17, compared with 5.75×10^{-2} for unconstrained score selection. Both are below the full-budget real-only benchmark in this geometry, but the gate pays a substantial predictive price for requiring evidence of improvement on each pilot. This is why gate acceptance, prediction loss, and oracle-regret certification are reported as separate outputs. A user seeking only minimum empirical error and a user requiring a supported improvement need not prefer the same rule.

6.2 A BUDGET SELECTED AFTER THE PILOT

After the same real pilot, a procurement rule considers five total counts $N = p + jd$, $j = 0, \ldots, 4$. Its supplied tolerance is $0.1\widehat{\sigma}^2 d/p$, based on observed residual noise rather than the hidden source gap. It chooses the first count whose selected recipe is certified, before seeing new labels; otherwise it leaves the original fit unchanged and marks it uncertified. Theorem 1 covers every considered count and the continuum between them on one event.

At $d = 2048$ in the disagreement case, 74 of 400 pilots meet the tolerance at some count; 63 require an additional batch. The corresponding Wilson intervals are $[15.00\%, 22.60\%]$ and $[12.51\%, 19.64\%]$. Thus the post-pilot budget is not constant across trials. No checked conditional-regret violation occurs. The observed certificate need not decrease monotonically with the budget: the identity of the best conditional plan can change. The shared-bias control is certified at the initial count because all-real is best within its menu, not because synthetic data help. The example demonstrates subsequent batch choice; it does not turn an already incurred pilot into a free initial test.

Table 2: Learned-menu disagreement experiment, $K = 29$, unit source and real noise. All rows use $p = 4d$, $q_h = 40d$ labels per source, and total cost $B = 4.8d$. The certificate is normalized by full-budget real-only risk 0.21; brackets give its empirical 10th–90th percentiles. Prediction risks are shown in units of 10^{-2}. All comparisons have 400 paired seeds.

d	Method	Risk (10^{-2})	Median C/R_0	10–90% interval
128	Reuse	7.02	3.31	[2.96, 3.68]
128	Holdout	8.60	1.42	[1.07, 1.75]
512	Reuse	5.97	0.77	[0.64, 0.89]
512	Holdout	6.64	0.57	[0.35, 0.71]
1024	Reuse	5.75	0.35	[0.29, 0.44]
1024	Holdout	6.28	0.36	[0.18, 0.48]
2048	Reuse	5.75	0.15	[0.13, 0.21]
2048	Holdout	5.94	0.19	[0.08, 0.31]

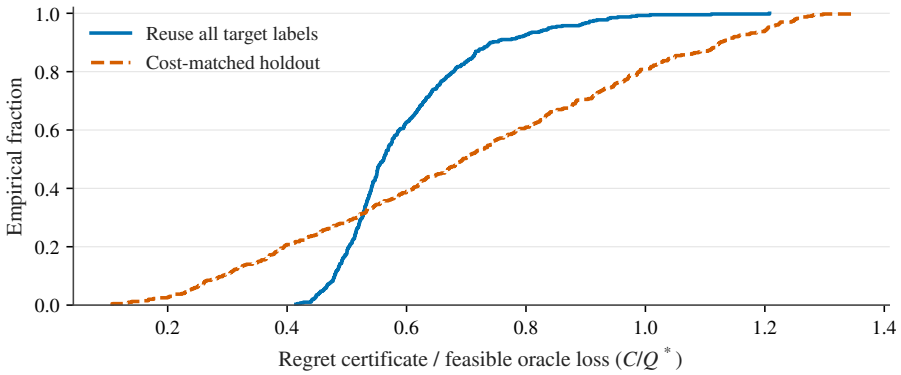


Figure 2: No gap-tuned tolerance is needed to read this comparison. Empirical distributions of the oracle-normalized regret certificate at $d = 2048$, with a learned 29-recipe menu and estimated noise. Each method's oracle is the minimum actual loss in its fitted bank. Reuse has a lower median; holdout has a longer lower tail as well as a longer upper tail. Zero-regret cases are reported separately rather than dividing by zero.

6.3 Local planning and the all-real fallback diagnostic

The sequence-model planning comparison uses $d = 64$, real cost 100, source costs one, and unit noise scales. Each source has hard prefix corrections of dimensions $0, 1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 40, 48, 56$; all-real gives 29 recipes. The target is zero. Source 1 has bias $0.4e_1$ in the correctable geometry and $0.4e_{64}$ in the misaligned geometry; source 2 has bias $\sqrt{0.01/32}$ in each of the first 32 coordinates. Budgets are $10^4, 10^5, 10^6, 10^8$, with 5,000 pilots per setting (seed $963400 + 10k + j$ for zero-based budget and geometry indices). The global baseline receives a discrepancy bound of one; the localized planner needs no such input. The localized planner selects a constant 68.93% pilot fraction; fixed 20% and 50% fractions and the discrepancy-dependent global planner are evaluated on common noise draws. At $B = 10^6$ in the correctable geometry, risks are 1.63×10^{-3}, 3.69×10^{-3}, 2.25×10^{-3}, and 1.47×10^{-3}, respectively. At $B = 10^8$, the localized planner outperforms the calibration-heavier global planner. Its numerical upper bound remains loose, and no method dominates every setting.

To test the effect of adding an all-real option, a matched diagnostic uses the two-source geometry and allocations in Section 5, with 12,000 independent pilots from seed 963301. The two source recipes, their counts, and each noise draw are identical across the menus with and without all-real. The new option is selected on 16.43% of trials. On those trials, the chosen source's score overstates its conditional risk by 0.15 on average; the extra real-only choice costs 9.95×10^{-2} in conditional risk. The mean paired increase is 1.63×10^{-2} with 95% interval $[1.57 \times 10^{-2}, 1.70 \times 10^{-2}]$. The conditional oracle does not change. This is selection noise in a quadratic score, not an adverse effect of enlarging the oracle or changing the existing allocations. A reference gate addresses that decision by requiring evidence for an improvement, at the cost of more frequent fallback.

Coverage. The 3,200 label-level confirmation datasets have no observed selected-certificate or reference-gate violations; the variance interval itself misses its nominal target on some trials, as expected for a 99% interval. Zero failures do not establish tightness.

7 DISCUSSION

Contrast-first calibration assesses what an observed source bank and real pilot can support. Freezing the bank and using hard corrections with a common target count make future-budget comparisons particularly simple; source-only menu learning and residual-based variance estimation bring the guarantee to actual regression labels. Risk localization removes a structural planning-rate defect in a broader protocol. The positive result is a concrete conditional workflow: a practitioner can learn a correction basis from finite source data, use all target labels in the current predictor, and obtain a simultaneous statement about its available alternatives and future batches. Holdout supplies a simpler competing assessment and may be preferable when its sample cost is affordable. The choice between them should be guided by their measured widths and prediction errors. The evidence supports these conditional assessments, not a universal initial source-acquisition policy. Remaining limits are Gaussian target noise, a fixed prediction representation, source-measurable correction construction, and conservative simultaneous bounds. Within that scope, the procedure tells a practitioner which correction is supported, whether an all-real fallback is safer, and which subsequent batch sizes meet a declared risk tolerance.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. doi: 10.1214/09-SS054. URL https://arxiv.org/abs/0907.4728.

Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the Gaussian setting. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(3):1092–1119, 2014. doi: 10.1214/13-AIHP539. URL https://arxiv.org/abs/1007.2096.

Pierre Bellec and Dana Yang. The cost-free nature of optimally tuning Tikhonov regularizers and other ordered smoothers. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 746–755, 2020. URL https://proceedings.mlr.press/v119/bellec20a.html.

Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001. doi: 10.1007/s100970100031.

Maxime Cauchois, Alnur Ali, and John Duchi. A comment and erratum on "excess optimism: How biased is the apparent error of an estimator tuned by SURE?". *arXiv preprint arXiv:2112.14353*, 2021. URL https://arxiv.org/abs/2112.14353.

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. doi: 10.1080/01621459.1995.10476626.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Michael B. Giles. Multilevel monte carlo path simulation. *Operations Research*, 56(3):607–617, 2008. doi: 10.1287/opre.1070.0496.

Alex A. Gorodetsky, Gianluca Geraci, Michael S. Eldred, and John D. Jakeman. A generalized approximate control variate framework for multifidelity uncertainty quantification. *Journal of Computational Physics*, 408:109257, 2020. doi: 10.1016/j.jcp.2020.109257.

Alex A. Gorodetsky, John D. Jakeman, and Michael S. Eldred. Grouped approximate control variate estimators, 2024. URL https://arxiv.org/abs/2402.14736.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176.

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

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012. doi: 10.1214/ECP.v17-2079.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://www.jmlr.org/papers/v17/kaufman16a.html.

Marc C. Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. doi: 10.1093/biomet/87.1.1.

Alois Kneip. Ordered linear smoothers. *The Annals of Statistics*, 22(2):835–866, 1994. doi: 10.1214/aos/1176325498.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142(1):201–211, 2008. doi: 10.1016/j.jeconom.2007.05.017.

Haotian Lin and Matthew Reimherr. Smoothness adaptive hypothesis transfer learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30286–30316, 2024. URL https://proceedings.mlr.press/v235/lin24q.html.

Zifan Lyu, Chahine Nejma, Tobias Wegel, Fanny Yang, and Florian E. Dorner. Cutting LLM evaluation costs with SySRs: A bandit algorithm that provably exploits model similarity, 2026. URL https://arxiv.org/abs/2606.07726. Published at the 43rd International Conference on Machine Learning.

Colin L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973. doi: 10.1080/00401706.1973.10489103.

Xuhui Meng and George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *Journal of Computational Physics*, 401:109020, 2020. doi: 10.1016/j.jcp.2019.109020.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376, 2023. doi: 10.52202/075280-2191.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Optimal model management for multifidelity monte carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194, 2016. doi: 10.1137/15M1046472.

Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D. Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A*, 473(2198):20160751, 2017. doi: 10.1098/rspa.2016.0751.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073. URL https://arxiv.org/abs/1803.00567.

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, pp. 121882–121927, 2024. doi: 10.52202/079017-3874.

Elizabeth Qian, Dayoung Kang, Vignesh Sella, and Anirban Chaudhuri. Multifidelity linear regression for scientific machine learning from scarce data. *Foundations of Data Science*, 7(1):271–297, 2025. doi: 10.3934/fods.2024049.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

Mark Rudelson and Roman Vershynin. Hanson–Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013. doi: 10.1214/ECP.v18-2865.

Daniel Schaden and Elisabeth Ullmann. On multilevel best linear unbiased estimators. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):601–635, 2020. doi: 10.1137/19M1263534.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19523–19536, 2022. doi: 10.52202/068431-1419.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

Yiming Xu, Vahid Keshavarzzadeh, Robert M. Kirby, and Akil Narayan. A bandit-learning approach to multifidelity approximation. *SIAM Journal on Scientific Computing*, 44(1):A150–A175, 2022. doi: 10.1137/21M1408312.

Yan Zhou. Black-box assisted regression: Phase transitions and minimax optimality, 2026. URL `https://arxiv.org/abs/2606.25743`. Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research, volume 306.

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121.

A PROBABILITY TOOLS AND THE FROZEN-BANK GUARANTEE

This appendix proves the main source-conditioned result without a transport construction. Throughout, source-bank data are conditioned on, and probabilities refer to independent real-data noise. A projector may be learned from the conditioned data. All norms on vectors are Euclidean, all matrix operator norms are spectral, and all covariance matrices are positive semidefinite.

A.1 AN OBSERVABLE UPPER BOUND FOR A GAUSSIAN MEAN NORM

Lemma 4 (Debiased upper scale)**.** *Let $y = m + \eta$ with $\eta \sim N(0, A)$. Write $F^2 = \operatorname{tr}(A^2)$, $\lambda = \|A\|_{\mathrm{op}}$, and $a = (\|y\|^2 - \operatorname{tr} A)_+$. For $t > 0$, with probability at least $1 - e^{-t}$,*

$$\|m\|^2 \le U(a, F^2, \lambda, t) := a + 4\lambda t + 2\sqrt{tF^2 + 2a\lambda t + 4\lambda^2 t^2}. \tag{8}$$

Proof. The Gaussian quadratic lower-tail inequality, including a deterministic mean, gives

$$\|y\|^2 \ge \|m\|^2 + \operatorname{tr} A - 2\sqrt{t\{\operatorname{tr}(A^2) + 2m^\top A m\}}.$$

For completeness, diagonalize A. For each eigenvalue $a_i \ge 0$, the logarithm of the moment generating function of the negative centered squared coordinate equals $\lambda' a_i - \frac{1}{2}\log(1+2\lambda' a_i) + 2\lambda'^2 a_i m_i^2/(1+2\lambda' a_i)$. For $\lambda' \ge 0$, this is at most $\lambda'^2(a_i^2 + 2a_i m_i^2)$. Summing, applying Chernoff's inequality, and optimizing λ' proves the displayed tail. This is a noncentral version of the Gaussian quadratic calculations used in Laurent & Massart (2000); Hsu et al. (2012).

Put $x = \|m\|^2$. Since $m^\top A m \le \lambda x$ and clipping the debiased observation upward is conservative, $x \le a + 2\sqrt{tF^2 + 2\lambda t x}$. If $x \le a$, the result is immediate. Otherwise square $x - a \le 2\sqrt{tF^2 + 2\lambda t x}$ and solve the resulting quadratic. The larger root is precisely (8). □

A.2 THE HARD-PROJECTION CONTRAST IDENTITY

Condition on all source observations, giving anchors $G_a \in \mathbb{R}^d$ and projectors $P_a = P_a^\top = P_a^2$, $a = 1, \dots, K$. Different recipes may reuse the same source anchor. The target pilot is $Y = f + \epsilon$, $\epsilon \sim N(0, S)$. If the final target precision is increased from p to N, the common retention fraction is $\alpha = p/N$. Assume the new independent sufficient statistic has covariance compatible with pooling, so that the conditional covariance of the pooled coefficient error is $\alpha(1-\alpha)S$. Repeated balanced designs satisfy this requirement.

Define $r_a = P_a(G_a - f)$ and $V_a = \operatorname{tr}\{(I - P_a)S\}$. The conditional error mean and the score are

$$m_a(\alpha) = r_a + \alpha(I - P_a)\epsilon, \tag{9}$$

$$Q_a(\alpha) = \|r_a\|^2 + \alpha^2 \|(I - P_a)\epsilon\|^2 + \alpha(1-\alpha)V_a, \tag{10}$$

$$\widehat{Q}_a(\alpha) = \|P_a(G_a - Y)\|^2 - \operatorname{tr}(P_a S) + \alpha V_a. \tag{11}$$

The orthogonality $P_a(I - P_a) = 0$ removes the cross term in (10); neither diagonal S nor commuting projectors are needed. Expanding the score and subtracting gives

$$e_a(\alpha) := \widehat{Q}_a - Q_a = -2r_a^\top \epsilon + \epsilon^\top P_a \epsilon - \operatorname{tr}(P_a S) - \alpha^2\{\epsilon^\top (I - P_a)\epsilon - V_a\}.$$

For $D_{ab} = P_a - P_b$, $g_{ab} = -r_a + r_b$, and $Z_{ab} = \epsilon^\top D_{ab}\epsilon - \operatorname{tr}(D_{ab}S)$,

$$e_a(\alpha) - e_b(\alpha) = 2g_{ab}^\top \epsilon + (1 + \alpha^2) Z_{ab}. \tag{12}$$

This separates the entire budget dependence from two random errors. Common target counts and hard projections are the reason the reduction holds. With different real retentions, the two separate retained-noise terms need not cancel into D_{ab}; with soft corrections, additional bias–noise terms depend on retention.

A.3 KNOWN FULL COVARIANCE

Set $T_{ab} = S^{1/2} D_{ab} S^{1/2}$ and

$$\widehat{g}_{ab} = -P_a(G_a - Y) + P_b(G_b - Y) = g_{ab} + D_{ab}\epsilon.$$

Then $S^{1/2}\widehat{g}_{ab}$ has mean $S^{1/2} g_{ab}$ and covariance T_{ab}^2. Lemma 4, with

$$a = \left(\left\|S^{1/2}\widehat{g}_{ab}\right\|^2 - \operatorname{tr} T_{ab}^2\right)_+, \quad F^2 = \operatorname{tr} T_{ab}^4, \quad \lambda = \|T_{ab}\|_{\mathrm{op}}^2,$$

provides an observable U_{ab} bounding $\left\|S^{1/2} g_{ab}\right\|^2$ except on probability e^{-t}. Separately,

$$|2 g_{ab}^\top \epsilon| \le 2\sqrt{2t}\left\|S^{1/2} g_{ab}\right\|$$

except on probability $2e^{-t}$. A two-sided Gaussian quadratic bound, valid for indefinite T_{ab}, gives

$$|Z_{ab}| \le 2\sqrt{t}\,\|T_{ab}\|_F + 2t\,\|T_{ab}\|_{\mathrm{op}}$$

except on probability $2e^{-t}$. These events need not be independent. Their intersection has probability at least $1 - 5e^{-t}$ and bounds (12) for *all* $\alpha \in [0,1]$ by

$$\rho_{ab}(\alpha) = 2\sqrt{2tU_{ab}} + (1+\alpha^2)\{2\sqrt{t}\,\|T_{ab}\|_F + 2t\,\|T_{ab}\|_{\mathrm{op}}\}. \tag{13}$$

Taking $t = \log(5\binom{K}{2}/\delta)$ proves the simultaneous event in Theorem 1. Zero contrasts can be omitted, and a predeclared subset of pairs can receive its own error allocation. Selecting the pair after seeing the pilot without charging all eligible pairs is not allowed.

For any selected a and best conditional-risk recipe b, the event implies $Q_a - Q_b \le \widehat{Q}_a - \widehat{Q}_b + \rho_{ab}$. Taking the maximum over b, including the zero self-comparison, gives the observable certificate. It holds for every selected retention because the events on the linear and quadratic errors did not depend on retention. Thus the fixed-N proof using these same two primitive bounds and the continuum proof have *identical* t and $\rho_{ab}(\alpha)$. This equality is between the stated constructions, not a comparison with an optimal fixed-N confidence interval.

Conditional validity given the source bank implies unconditional validity by the law of total probability. The bank and projector-learning algorithm can have arbitrary distributions, provided target-pilot noise remains independent. The sources need not have known covariance, known biases, or Gaussian training noise. Their finite-sample errors are present in r_a and therefore in prediction risk.

A.4 UNKNOWN SCALAR NOISE: WHY BOTH INTERVAL ENDPOINTS ARE NEEDED

Suppose $S = vI$ and an independent residual statistic provides $[l, u]$ with $\mathbb{P}(v \notin [l,u]) \le \delta_v$. The interval may be random; conditional on it, target coefficient noise is still Gaussian. Use the midpoint $\bar{v} = (l+u)/2$ in the score. For $D = D_{ab}$ define

$$a_l = \left(\|\widehat{g}_{ab}\|^2 - l\,\|D\|_F^2\right)_+, \qquad q = \|D\|_{\mathrm{op}}^2, \qquad h = \operatorname{tr}(D^4).$$

On the variance event, Lemma 4 and monotonicity give

$$U_g = a_l + 4uqt + 2\sqrt{u^2 ht + 2a_l uqt + 4u^2 q^2 t^2} \;\ge\; \|g_{ab}\|^2 \tag{14}$$

outside an event of probability e^{-t}. In detail, the exact bound uses $(\|\widehat{g}\|^2 - v\,\|D\|_F^2)_+$, $v^2 h$, and vq. Replacing v by l in the subtracted center increases that center; replacing it by u in the positive tail terms increases the upper bound. Replacing v by u everywhere would *decrease* the debiased center and is unjustified.

The score difference changes by $-(\bar{v} - v)(1+\alpha)\operatorname{tr} D$. Consequently the valid radius is

$$\begin{aligned}\rho_{ab}^{\mathrm{int}}(\alpha) = {} & 2\sqrt{2tuU_g} + (1+\alpha^2)u\{2\sqrt{t}\,\|D\|_F + 2t\,\|D\|_{\mathrm{op}}\} \\ & + \tfrac{1}{2}(u-l)(1+\alpha)|\operatorname{tr} D|.\end{aligned} \tag{15}$$

Allocate $t = \log(5\binom{K}{2}/\delta_m)$. Combining the coefficient and variance events with a union bound yields confidence at least $1 - \delta_m - \delta_v$. The construction remains uniform over α. A tolerance based on the residual variance estimate is permitted because the event simultaneously bounds every comparison and retention.

Actual regression labels. For a fixed full-rank design X and homoskedastic Gaussian labels $y = Xf + \xi$, least squares gives coefficient covariance $\sigma^2(X^\top X)^{-1}$. Its residual sum of squares (RSS) divided by σ^2 is χ^2_{n-d} and is independent of the coefficients. If $X^\top X = nI$, then $v = \sigma^2/n$ and the residual degrees of freedom are $\nu = n - d$. An exact interval is

$$[l, u] = \frac{1}{n}\left[\frac{\mathrm{RSS}}{\chi^2_{\nu,1-\delta_v/2}}, \frac{\mathrm{RSS}}{\chi^2_{\nu,\delta_v/2}}\right].$$

Independent residual sums of squares can be added and their degrees of freedom added. The experiments estimate σ^2 using two target blocks, then pool their coefficient estimates; independence continues to hold. For known S, a general non-diagonal design uses (13). Unknown, unrestricted covariance is not inferred from a scalar residual interval.

A.5 An observable reference gate

Include the all-real recipe $P_0 = 0$. Allocate comparisons only to the $K - 1$ predeclared pairs $(a, 0)$, so $t = \log(5(K-1)/\delta_m)$. Among recipes satisfying $\widehat{Q}_a - \widehat{Q}_0 + \rho^{\mathrm{int}}_{a0} \le 0$, choose the smallest score. Return recipe 0 if the set is empty. On the common event, $Q_{\widehat{a}}(\alpha) \le Q_0(\alpha)$ for every chosen retention. This protects against the *same-paid-data* all-real plan; the bank's cost has already been incurred. It does not compare against an all-real predictor purchased using that bank expenditure as additional target labels. The experiments report gate acceptance and absolute risk separately so the trivially safe fallback is not counted as an oracle-regret certificate.

B A risk-localized bound for paid pooling

The bank theorem conditions on synthetic observations and does not prescribe their purchase. This appendix addresses a different protocol: all source and real streams can receive additional observations, and each recipe's final counts are fixed before the pilot. The following expected-risk comparison applies to the actual pooled-score selector. The bound localizes to the reference risk rather than a maximum source-discrepancy norm, while retaining conservative dimension factors.

B.1 Model and exact error identity

Stack the real and synthetic pilot means as $x = \mu + \epsilon$, with independent isotropic blocks and block covariance S. Every pilot variance is at most ν. A recipe has diagonal weights $W \in [0, I]$, chooses one source, and specifies retention fractions $\alpha_s, \alpha_r \in [0, 1]$. Define maps

$$Jx = W(x_h - x_0), \qquad Lx = \alpha_s W x_h + \alpha_r (I - W)x_0.$$

Then $b_a = J_a\mu$ and its conditional risk is $Q_a = \|b_a + L_a\epsilon\|^2 + V^{\mathrm{new}}_a$. The score is $\widehat{Q}_a = \|J_a x\|^2 - \mathrm{tr}(J_a S J_a^\top) + V^{\mathrm{pool}}_a$, where $V^{\mathrm{pool}}_a = V^{\mathrm{new}}_a + \mathrm{tr}(L_a S L_a^\top)$. For each fixed recipe it is unbiased for $\mathbb{E}Q_a$. Expanding once gives

$$\widehat{Q}_a - Q_a = 2\epsilon^\top (J_a - L_a)^\top b_a + \epsilon^\top C_a \epsilon - \mathrm{tr}(C_a S), \qquad C_a = J_a^\top J_a - L_a^\top L_a. \tag{16}$$

B.2 Uniform noise bounds

Let $K \ge 2$, $L_K = \log(2K)$, and $\ell_K = \log K$. For nonzero b_a, the centered normal random variable $Z_a = \epsilon^\top (J_a - L_a)^\top b_a / \|b_a\|$ has variance at most 2ν. This follows coordinatewise: the real coefficient of $J - L$ has absolute value $w + \alpha_r(1 - w) \le 1$, and its source coefficient $(1 - \alpha_s)w \le 1$. Define $Z_a = 0$ when $b_a = 0$. A union bound and integration of $\mathbb{P}(\max_a |Z_a|^2 > s) \le \min\{1, 2Ke^{-s/(4\nu)}\}$ give

$$\mathbb{E}G^2 \le u := 4\nu(L_K + 1), \qquad G = \max_a |Z_a|. \tag{17}$$

The covariance of $L_a\epsilon$ has operator norm at most ν and trace at most νd, since $\alpha_s^2 w^2 + \alpha_r^2(1-w)^2 \le 1$. The Gaussian quadratic moment-generating-function bound and the log-sum-exp inequality imply

$$\mathbb{E}\max_a \|L_a\epsilon\|^2 \le V := \nu\{d + 2\sqrt{d\ell_K} + 2\ell_K\}. \tag{18}$$

To see the constants without an integration remainder, for $0 < \lambda < 1/(2\nu)$ the centered log moment-generating function is at most $\lambda^2\nu^2 d/(1-2\nu\lambda)$. Apply Jensen to the maximum, add ℓ_K/λ, and optimize.

For $H_a = S^{1/2}C_aS^{1/2}$, $\|H_a\|_{\mathrm{op}} \le 2\nu$ and $\|H_a\|_F \le 3\nu\sqrt{d}$. Indeed, $S^{1/2}J_a^\top J_a S^{1/2}$ has rank at most d, operator norm at most 2ν, and $S^{1/2}L_a^\top L_a S^{1/2}$ has the corresponding bounds d, ν. The difference of these positive semidefinite matrices has operator norm at most the larger of the two operator norms; the Frobenius triangle inequality gives the other claim. Applying the same log moment-generating-function argument to both signs gives

$$\mathbb{E}\max_a |\epsilon^\top C_a\epsilon - \mathrm{tr}(C_aS)| \le H := 6\nu\sqrt{dL_K} + 4\nu L_K. \tag{19}$$

No independence across recipes is used.

B.3 Proof of Theorem 3

Let $R = \mathbb{E}Q_{\widehat{a}}$, where $\widehat{a}$ minimizes the pooled score. For any fixed reference recipe with expected paid risk r, score minimization and unbiasedness give $\mathbb{E}\widehat{Q}_{\widehat{a}} \le r$. Also $\|b_{\widehat{a}}\| \le \sqrt{Q_{\widehat{a}}} + \|L_{\widehat{a}}\epsilon\|$. Therefore (16)–(19) and Cauchy–Schwarz yield

$$R \le r + 2\sqrt{uR} + 2\sqrt{uV} + H.$$

All variables have finite moments: the menu is finite and each Q_a is a quadratic polynomial in Gaussian noise. Thus the expectations and Cauchy–Schwarz steps require no truncation. Set $a = 2\sqrt{u}$, $b = 2\sqrt{uV} + H$. Solving the quadratic in $\sqrt{R}$ gives

$$R \le \Psi(r,\nu,d,K) := \left\{\frac{a + \sqrt{a^2 + 4(r+b)}}{2}\right\}^2. \tag{20}$$

For $K = 1$, set $\Psi = r$. Expanding the quadratic bound, or using $\sqrt{a^2+4(r+b)} \le a + 2\sqrt{r+b}$ followed by Young's inequality, gives the rate

$$\Psi - r \le C\{\sqrt{r\nu\log(2K)} + \nu\sqrt{d\log(2K)} + \nu\log(2K)\}$$

for a numerical constant C. In particular, when $r = O(B^{-1})$ and $\nu = O(B^{-1})$ with fixed d, K, the excess is $O(B^{-1})$, rather than $O(B^{-1/2})$. This is a rate statement; the displayed finite-sample constants can still be large.

B.4 Full-budget comparator and scale-invariant pilot planning

Let $R^*_{\mathcal{A}}(B)$ be the best fresh-data recipe risk at budget B. Each recipe's counts are obtained by minimizing its pooled variance after retaining its relevant pilots and paying for all pilots. For the full-budget oracle recipe a^*, let $\Omega_{a^*}(B)$ denote the difference between its paid pooled variance and its full-budget fresh variance. This difference is nonnegative when the paid design wastes source labels from other recipes, and is always at most $k_{a^*}P/[B(B-P)]$ under the retained-pilot allocation described next. Then Theorem 3 applies with $r = R^*_{\mathcal{A}}(B) + \Omega_{a^*}(B)$.

For completeness, a recipe's variance before allocation is $\sigma_h^2\|w\|^2/n_h + \sigma_0^2\|1-w\|^2/n_0$. Minimizing it under $c_hn_h + c_0n_0 = T$ gives k_a/T, where $k_a = (\sigma_h\sqrt{c_h}\|w\| + \sigma_0\sqrt{c_0}\|1-w\|)^2$. The spent source fraction is the first term in this sum divided by their sum. After pilots q_h, p, write $A = \sigma_h^2\|w\|^2$ and $D = \sigma_0^2\|1-w\|^2$. The final real count lies in $[p, p + (B-P)/c_0]$, and the final source count is $q_h + (B - P - c_0(N_0 - p))/c_h$. The unconstrained variance minimizer is clipped to this interval, with pure-stream endpoints handled directly. Its variance is no greater than a fresh fit at budget $B - P$ because retaining the existing observations only increases counts. The extra cost relative to a full-budget fresh fit is at most $k_a/(B-P) - k_a/B$, as claimed.

Let $\Lambda = \sum_h c_h\sigma_h^2$ include the real stream. Equalizing pilot variances uses $n_h = P\sigma_h^2/\Lambda$ and gives $\nu = \Lambda/P$. A supplied all-real reference bound is $r_0(P) = dc_0\sigma_0^2/(B-P)$. The implementable planning objective is

$$G_B(P) = \Psi\left(\frac{dc_0\sigma_0^2}{B-P}, \frac{\Lambda}{P}, d, K\right), \qquad 0 < P < B. \tag{21}$$

For $P = \gamma B$, $G_B(P) = B^{-1}G_1(\gamma)$ exactly. It diverges at both endpoints, so a minimizer exists in the interior and its minimizing fraction is independent of B. No uniqueness assertion is needed for this conclusion. The implementation selects one minimizer numerically.

The reference $r_0(P)$ spends only $B - P$ on a fresh target fit, so this planning upper bound exceeds the full-budget all-real benchmark. It cannot justify an initial buy-versus-skip decision by itself. Its purpose is to remove a structural scaling defect from a conditional-on-purchasing planning calculation. The planning comparison reports actual risks of its selected fractions rather than interpreting the upper bound as a prediction.

C Restricted domains and the price of general uniformity

The frozen-bank protocol restricts the acquisition path to obtain a simple event. This appendix develops a general extension for independently retained real and synthetic streams, and measures why its extra scope can be expensive.

C.1 Exact rectangle coefficients

For a general recipe use the maps J, U, V from Appendix B, where U retains the source and V the real pilot. Declare a rectangle $\alpha_s \in [a_0, a_1]$, $\alpha_r \in [b_0, b_1]$ before observing the real pilot. Let $L_0 = a_0U + b_0V$, $U' = (a_1 - a_0)U$, and $V' = (b_1 - b_0)V$. The degree-two Bernstein basis on $[0, 1]$ is $\beta_0(s) = (1 - s)^2$, $\beta_1(s) = 2s(1 - s)$, $\beta_2(s) = s^2$. For indices $i, j \in \{0, 1, 2\}$, set

$$B_{ij} = (J - L_0 - \tfrac{i}{2}U' - \tfrac{j}{2}V')^\top J, \tag{22}$$

$$C_{ij} = J^\top J - L_0^\top L_0 - \tfrac{i}{2}(L_0^\top U' + U'^\top L_0) - \tfrac{j}{2}(L_0^\top V' + V'^\top L_0) - \mathbf{1}_{i=2}U'^\top U' - \mathbf{1}_{j=2}V'^\top V' - \tfrac{ij}{4}(U'^\top V' + V'^\top U'). \tag{23}$$

Substituting $\alpha_s = a_0 + (a_1 - a_0)s$ and $\alpha_r = b_0 + (b_1 - b_0)t$ into (16), then using $s = \sum_i (i/2)\beta_i(s)$ and $s^2 = \beta_2(s)$, proves the exact expansion

$$e_a(s, t) = \sum_{i,j} \beta_i(s)\beta_j(t)\{2\epsilon^\top B_{ij}\mu + \epsilon^\top C_{ij}\epsilon - \operatorname{tr}(C_{ij}S)\}.$$

There are still nine coefficients per recipe. Restriction changes their matrices and may change which nodes deduplicate. It does not automatically improve a newly allocated union bound.

C.2 Contrast first; a feasible coupling is sufficient

Index the distinct coefficient errors, including a zero anchor, by $z_1, \ldots, z_M$. Bound their *differences* before combining them. On the event $|z_i - z_j| \le d_{ij}(x)$ for all predeclared pairs, two recipe errors with coefficient weights p_i, q_j satisfy

$$|\textstyle\sum_i p_i z_i - \sum_j q_j z_j| \le \sum_{ij} \pi_{ij} d_{ij}(x) \quad \text{for every nonnegative } \pi \text{ with margins } p, q. \tag{24}$$

The proof is the identity $\sum_i p_i z_i - \sum_j q_j z_j = \sum_{ij} \pi_{ij}(z_i - z_j)$ and the triangle inequality. Product coupling $\pi = pq^\top$ is always feasible. Matching common node mass first, corresponding coefficient positions, or greedily matching low-cost edges are other inexpensive choices. Taking their minimum is valid on the *same* event: it does not cost additional confidence. Minimizing over all feasible couplings is an optional refinement.[1]

A general mixed form has error $2\epsilon^\top B\mu + \epsilon^\top C\epsilon - \operatorname{tr}(CS)$, with C symmetric and B possibly nonsymmetric. A bound obtained by adding a noise-norm allowance to the observed linear scale is

$$2\sqrt{2t}\{\left\|S^{1/2}Bx\right\| + \|T\|_F\} + 4t\,\|T\|_{\mathrm{op}} + 2\sqrt{t}\,\|H\|_F + 2t\,\|H\|_{\mathrm{op}},$$

where $T = S^{1/2}BS^{1/2}$ and $H = S^{1/2}CS^{1/2}$. Alternatively, Lemma 4 estimates the linear scale with covariance $TT^\top$, giving $2\sqrt{2tU} + 2\sqrt{t}\,\|H\|_F + 2t\,\|H\|_{\mathrm{op}}$. Both have failure at most $5e^{-t}$; taking

[1] When d is a pseudometric, transport duality identifies this optimum with the sharpest consequence of all recorded pairwise inequalities; see Peyré & Cuturi (2019). This finite-dimensional optimization property does not imply a statistically optimal confidence radius.

their minimum without accounting for both events is not done. Predeclared edge weights $\omega_{ij} > 0$ with sum at most one permit $t_{ij} = \log(5/(\delta\omega_{ij}))$. Formula (24) remains valid, but its nonuniform edge radii need not be a pseudometric, so the dual sharpness footnote is not automatically applicable.

C.3 A CONTROLLED DECOMPOSITION

At one fixed retention pair, let $r_{\mathrm{fixed}}(t)$ denote a direct mixed-form radius and $r_{\mathrm{box}}(t)$ the coefficient-coupled radius with the same primitive inequality. The diagnostic factors are

$$\underbrace{\frac{r_{\mathrm{fixed}}(t_M)}{r_{\mathrm{fixed}}(t_1)}}_{\text{union allocation}} \quad \text{and} \quad \underbrace{\frac{r_{\mathrm{box}}(t_M)}{r_{\mathrm{fixed}}(t_M)}}_{\text{coefficient envelope}}, \qquad t_1 = \log(5/\delta), \quad t_M = \log(5\tbinom{M}{2}/\delta).$$

Their product is exactly the total radius inflation on each pilot. The second factor includes bounding coefficient errors separately and their feasible coupling; it is the numerical cost of the Bernstein envelope construction after holding the probability parameter fixed. This decomposition is a diagnostic, not an assertion that the probability of the two events factors. Restricted rectangles are evaluated using their actual new node counts, so an increased deduplication burden is not hidden.

D COST-MATCHED HOLDOUT CERTIFICATES

Condition on a fitted prediction bank $v_1, \ldots, v_K$ constructed without the validation mean $Y_v = f + \epsilon_v$, where $\epsilon_v \sim N(0, S_v)$. Then

$$\|v_a - Y_v\|^2 - \|v_b - Y_v\|^2 - \{\|v_a - f\|^2 - \|v_b - f\|^2\} = -2(v_a - v_b)^\top \epsilon_v.$$

The exact Gaussian half-width of a two-sided pair interval is

$$2z_{1-\delta_m/(2\binom{K}{2})}\sqrt{(v_a - v_b)^\top S_v (v_a - v_b)}.$$

A union bound covers arbitrary subsequent selection. If $S_v = \sigma^2 I/n_v$ and an independent residual upper bound u_σ is available, replacing σ^2 by u_σ is conservative; there is no debiased center in this linear contrast. Charge its failure probability separately. A reference-only gate uses $K - 1$ comparisons instead.

The benchmark fits its bank with $3d$ real labels and validates with d labels. The reuse assessment fits and evaluates with all $4d$. Both purchase the same source observations and the same total real labels; neither receives free validation labels. Conditional comparators differ because the fitted banks differ. We therefore report both certificate size normalized by each bank's oracle loss and prediction error against a common full-budget real-only benchmark. Classical cross-validation does not require this particular Gaussian certificate, but sample splitting is a strong and transparent comparator (Arlot & Celisse, 2010).

E THEOREM-LEVEL COMPARISON AND LITERATURE CONTEXT

Exactly where SURE appears. At full real retention $\alpha = 1$, known covariance vI, and a conditioned source bank, the candidate is the affine estimator $\widehat{f}_a(Y) = P_a G_a + (I - P_a)Y$. Its Stein criterion is

$$\left\|\widehat{f}_a(Y) - Y\right\|^2 + 2v \operatorname{div} \widehat{f}_a - vd = \|P_a(G_a - Y)\|^2 + vd - 2v \operatorname{tr} P_a,$$

which is exactly (3). Risk estimation of this completed affine predictor is therefore not new. Stein (1981); Mallows (1973) provide the foundational perspective; Kneip (1994); Birgé & Massart (2001) give major model-selection developments.

Cauchois et al. (2021) correct an excess-optimism calculation and prove risk-localized oracle inequalities for SURE-tuned linear estimators. Their bounds contain the characteristic $\sqrt{R^* \sigma^2 \log K}$ term and a lower-order noise complexity term. We use that localization principle rather than retaining a global source-discrepancy norm. Our explicit constants and dimension dependence are conservative and are not an improvement over their common SURE specialization. Theorem 3 adds a charged pooling protocol and its allocation comparator, not a claim to originate localization.

For $\alpha < 1$, the conditional mean of the future predictor equals $P_a G_a + \alpha(I - P_a)Y + (1-\alpha)(I - P_a)f$. It contains the unknown f and is not an observable completed smoother of Y to which a usual plug-in SURE selection statement can simply be applied. Theorem 1 instead bounds pairwise *conditional future losses* through identity (5). Its one-event, all-count conclusion follows from that protocol-specific identity. These arguments still use classical Gaussian concentration; we do not suggest they lie outside the reach of existing proof techniques.

Aggregation and unknown variance. Bellec & Yang (2020) show that Q-aggregation of ordered smoothers can have excess risk at most $C\sigma^2$ independent of the number of smoothers. That procedure combines estimators and exploits ordering; our selector returns one source/correction recipe and the source-conditioned certificate permits unordered projectors. The experiments do not compare a broadly competitive aggregation predictor, so predictive superiority to such aggregation is not claimed. Baraud et al. (2014) analyze Gaussian estimator selection with unknown variance. Our residual interval is a simple exact-scale treatment for the conditional contrast assessment, not the first unknown-variance selection method. Cross-validation offers an accessible alternative with different sample-efficiency and selection properties (Arlot & Celisse, 2010); its Gaussian certificate is included explicitly rather than compared only as a point predictor.

Residual and transfer learning. The Δ-learning approach corrects approximate physical models (Ramakrishnan et al., 2015). Smoothness-adaptive hypothesis transfer quantifies different source and offset regularities (Lin & Reimherr, 2024), while black-box assisted regression addresses useful residual structure and phase transitions (Zhou, 2026). Wavelet adaptation illustrates why a suitable representation can make a function easier to estimate (Donoho & Johnstone, 1995). None of our acquisition results supplies a generally reliable way to discover the right target-error basis. The source-disagreement construction is one explicit, testable construction; the common-bias control exhibits its blind spot. Co-kriging and nonlinear multifidelity models use richer relationships between sources (Kennedy & O'Hagan, 2000; Perdikaris et al., 2017; Meng & Karniadakis, 2020), which our finite linear menu does not replace.

Multifidelity acquisition. Multifidelity linear regression already combines source outputs, control variates, and sample allocation (Qian et al., 2025). Optimal multifidelity Monte Carlo management, approximate control variates, multilevel estimators, and grouped estimators make cost-aware information reuse explicit (Peherstorfer et al., 2016; Gorodetsky et al., 2020; Giles, 2008; Schaden & Ullmann, 2020; Gorodetsky et al., 2024). The survey of Peherstorfer et al. (2018) places these methods in uncertainty propagation, inference, and optimization. Our simple square-root expenditure split is an instance of variance-optimal allocation, not a separate novelty claim. The source-conditioned protocol keeps bank-estimation noise inside prediction risk rather than pretending the bank is known without expenditure.

Comparison, label reuse, and adaptive acquisition. Bandit-learning multifidelity approximation and multifidelity best-arm identification study exploration costs (Xu et al., 2022; Poiani et al., 2024).

General best-arm complexity results emphasize decision gaps and information rather than raw estimation accuracy (Garivier & Kaufmann, 2016; Kaufmann et al., 2016). Shared model evaluation can use paired observations to reduce comparison cost (Lyu et al., 2026). Prediction-powered and cross-prediction-powered inference also use trusted labels to correct model outputs efficiently (Angelopoulos et al., 2023; Zrnic & Candès, 2024). Our confidence target is selected prediction risk, not a population-mean confidence interval. We reuse a fixed pilot across a continuum of *future plans*; this is not a confidence sequence for repeatedly inspecting newly acquired labels (Howard et al., 2021). Uniform oracle claims also require care near local alternatives, as familiar from oracle-property pathologies (Leeb & Pötscher, 2008).

Concentration, optional coupling, and scaling. Gaussian quadratic-form arguments and their sub-Gaussian generalizations are standard (Laurent & Massart, 2000; Hsu et al., 2012; Rudelson & Vershynin, 2013). The paper retains Gaussian assumptions because the noncentral norm inversion, residual independence, and exact label-level bridge are then explicit. Transport duality is a finite-dimensional envelope interpretation (Peyré & Cuturi, 2019); the contrast identity, rather than the transport solver, enables the frozen-bank guarantee. Finally, empirical neural scaling laws, compute-optimal allocations, data pruning, and repeated-data scaling motivate resource accounting (Kaplan et al., 2020; Hoffmann et al., 2022; Sorscher et al., 2022; Muennighoff et al., 2023). Our rates concern a stated Gaussian acquisition experiment; no universal neural scaling exponent is asserted.