POST-SELECTION CERTIFICATION OF GRADIENT CORESETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient sketches make coreset optimization tractable; valid post-selection certificates require accounting for how the residual was chosen. We give an exact characterization for Gaussian sketches. For least-squares reweighting over a fixed k-dimensional affine family, the selected residual's sketch-to-exact squared-norm ratio is a scaled product of independent chi-square and beta variables. This law supplies an exact model-specific correction and identifies two distinct thresholds: relative distortion vanishes when $k = o(m)$, while uncorrected fixed-vector coverage can vanish when $\sqrt{m} \ll k \ll m$. For arbitrary selectors, a fresh Gaussian projection after commitment gives exact conditional intervals with rank-independent audit dimension. We also account for support-search multiplicity, numerical error, and fidelity to the deployed update. Experiments on Qwen2.5-1.5B/0.5B and GSM8K/MBPP tasks complement the theory with constrained real-model diagnostics: at 512 sketch rows, median residual distortion is 3.4% and ranking correlation is 0.95. Reweighting lowers the primary joint-subset exact residual by 23.6%, while increasing sketch optimism in 43 of 45 matched cells. Uniform bounds cover all 27 primary residuals, with six meeting the full-update tolerance. Together, these results distinguish accurate optimization from useful certification and provide calibrated routes to certifying the selected coreset update.

1 INTRODUCTION

Gradient-based data selection replaces a large training pool by a small weighted subset whose update approximates a declared target (Killamsetty et al., 2021a). Random projections make this matching tractable (Xia et al., 2024). But the same sketch often plays two roles: it defines the optimization objective and measures the residual of the resulting coreset. These roles require different guarantees. A norm estimate calibrated for a vector fixed before sketching is not automatically calibrated for a residual whose subset or weights were chosen using that sketch.

The distinction is more consequential than a worst-case kernel example suggests. Consider even a fixed support with unconstrained least-squares reweighting. We derive an exact distribution for the *actual selected residual*, building on Gaussian sketch-regression identities (Ahfock et al., 2021; Browne & Andrews, 2024). If m is the number of sketch rows and k the number of fitted directions, estimation accuracy and inferential validity have different thresholds: $k = o(m)$ suffices for vanishing relative distortion, whereas retaining nominal fixed-vector coverage requires $k = o(\sqrt{m})$ in this model. Between these scales, the sketch becomes arbitrarily accurate in relative terms while its uncorrected upper certificate covers with probability tending to zero. Accurate optimization diagnostics alone therefore cannot validate post-selection inference.

The exact law also enables valid same-sketch inference for the fixed affine family. A frozen support menu permits simultaneous correction, and uniform embeddings control larger reachable classes. These solutions pay for structural knowledge or simultaneous coverage. Alternatively, sealing fresh Gaussian rows until the deployed coreset is committed makes the selected residual fixed relative to those rows. This changes the inferential task from controlling an optimized family to estimating one Gaussian scale.

Our real-model study distinguishes these issues empirically. At the primary setting, a visible sketch estimates residual norms accurately and ranks selectors well. Its *uniform* upper bounds cover the exact residual in all 27 primary cells, and six also lie below the full-update scale. This separates numerical

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.

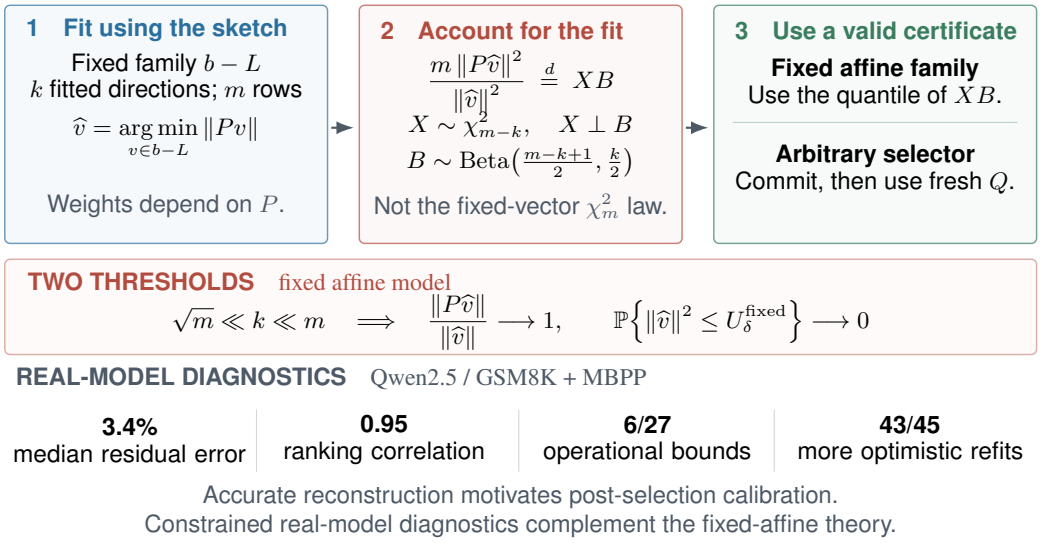

Figure 1: **Optimization accuracy, certificate validity, and certificate usefulness are different.** The affine-reweighting theorem gives an exact separation: for $\sqrt{m} \ll k \ll m$, relative distortion vanishes but uncorrected fixed-vector coverage vanishes too. A model-specific correction is valid for a fixed affine family; a fresh audit covers a residual chosen by an arbitrary selector. The bottom row summarizes constrained real-model diagnostics; the affine-law and independent-audit guarantees are theoretical.

coverage from usefulness at a declared tolerance. Reweighting the same selected subset lowers its exact residual by 23.6% while increasing optimistic sketch underestimation in most matched cells. The paired comparison shows that improved reconstruction and post-selection calibration are complementary requirements.

The paper contributes an exact selected-residual law and its two-threshold consequence; a calibrated comparison of model-specific reuse, simultaneous guarantees, and independent auditing; and an evidence-grounded diagnosis on real language models. The chi-square audit, regression distributional ingredients, and simultaneous-correction principles are classical. Our focus is their implication for the *post-optimization coreset residual*, including which guarantees survive changes of support, coefficients, information access, and numerical representation.

2 The Target and the Information Available

At a frozen anchor θ_0, let $G = [g_1, \ldots, g_n] \in \mathbb{R}^{d \times n}$ contain declared additive update features. A fixed preconditioner may be absorbed into g_i. Nonnegative full-pool weights a define $g_{\mathcal{V}} = Ga$. A feasible subset S and deployed weights $w \in \mathcal{W}_S$ obey a declared token or cardinality budget. Define

$$v(S, w) = g_{\mathcal{V}} - G_S w, \qquad R(S, w) = \|v(S, w)\|^2 . \tag{1}$$

Here $\|\cdot\|$ is the Euclidean norm. The visible sketch has independent and identically distributed (iid) entries $P_{ij} \overset{\text{iid}}{\sim} N(0, 1/m)$; it reveals PG and any permitted non-audit information. A selector returns $(\widehat{S}, \widehat{w})$, giving $\widehat{v} = v(\widehat{S}, \widehat{w})$ and visible estimate $\widehat{R}_P = \|P\widehat{v}\|^2$. Unless stated otherwise, G, the feasible class, and the full-pool target are fixed independently of P. Write $q_{\nu,\alpha}$ for the lower α-quantile of χ^2_ν.

Validity and usefulness. A $1-\delta$ upper certificate U satisfies $\mathbb{P}\{R \le U\} \ge 1-\delta$ under its declared information contract. A tolerance τ makes it useful when $\sqrt{U} \le \tau$. The experiments use $\tau = \|g_{\mathcal{V}}\|$: such a certificate rules out a reconstruction error exceeding the full update's own magnitude. A valid bound above this scale is operationally vacuous, not a coverage failure. Ranking correlation, relative distortion, coverage, and usefulness are four different quantities.

Two meanings of conditioning. A same-sketch marginal guarantee averages over P for a fixed problem, even though $\widehat{v}$ depends on P. A transcript-conditional, distribution-free guarantee instead fixes the exposed transcript and must cover every residual still compatible with it. These are not interchangeable. Fresh auditing conditions on the entire pre-audit state, but averages over independent *new* randomness. Exact feature access is compatible with this independence; it merely makes a compressed audit unnecessary when the exact residual is cheap to compute.

3 Exact reweighting laws and the limits of sketch reuse

3.1 Continuous adaptation on a fixed affine family

Fix $b \in \mathbb{R}^d$ and a k-dimensional subspace L independently of P. The residual family is $b - L$. For a fixed coreset support and affine weight constraints $w = w_0 + Nt$, take $b = g_{\mathcal{V}} - G_S w_0$ and $L = \mathrm{col}(G_S N)$. Thus k counts *fitted directions*, not necessarily selected examples or full-pool rank. Let

$$\widehat{v} = \operatorname*{arg\,min}_{v \in b - L} \|Pv\|^2, \qquad \sigma^2 = \mathrm{dist}(b, L)^2 > 0. \tag{2}$$

For $m > k$, the selected residual is almost surely unique.

Theorem 3.1 (Exact selected-residual law)**.** *Under Equation (2), for $1 \le k < m$,*

$$T_{m,k} := \frac{m\,\|P\widehat{v}\|^2}{\|\widehat{v}\|^2} \overset{d}{=} XB, \quad X \sim \chi^2_{m-k}, \quad B \sim \mathrm{Beta}\left(\frac{m-k+1}{2}, \frac{k}{2}\right), \quad X \perp B. \tag{3}$$

Moreover, $m\,\|P\widehat{v}\|^2 / \sigma^2 = X$, $\sigma^2 / \|\widehat{v}\|^2 = B$, and

$$\mathbb{E}\left[\frac{\|P\widehat{v}\|^2}{\|\widehat{v}\|^2}\right] = \frac{(m-k)(m-k+1)}{m(m+1)}. \tag{4}$$

For $k = 0$, set $B = 1$ and $X \sim \chi^2_m$. For $m \le k$ the sketched optimum is zero almost surely, while every feasible exact residual has norm at least $\sigma > 0$.

Why the product appears. Write $b = Ua_0 + e$ with $U^\top U = I_k$, $e \perp L$, and $\|e\| = \sigma$. The standardized sketch design $A = \sqrt{m}\,PU$ and response noise $z = \sqrt{m}\,Pe/\sigma$ are independent standard Gaussian arrays. Least squares leaves $m - k$ residual degrees of freedom, producing X. At the same time, fitted coefficients absorb part of the projected e into L, increasing the *exact* residual from σ^2 to $\sigma^2(1 + \|\eta\|^2)$, where $\eta = (A^\top A)^{-1} A^\top z$. The latter inflation produces B. Independence and the beta law follow from Gaussian regression; Section A.1 gives a self-contained derivation. The factorization targets the selected residual, not only the fixed-family optimum σ^2.

3.2 Accurate estimation can coexist with vanishing coverage

The ordinary fixed-vector squared-norm upper bound is

$$U^{\mathrm{fixed}}_\delta(Pv) = \frac{m\,\|Pv\|^2}{q_{m,\delta}}. \tag{5}$$

Its exact coverage after affine fitting is $\mathbb{P}\{T_{m,k} \ge q_{m,\delta}\}$, strictly less than $1 - \delta$ for $k > 0$. Merely changing the chi-square degrees of freedom from m to $m - k$ does not suffice: the additional beta factor remains.

The mechanism is a scale mismatch: the calibration width is of order $m^{-1/2}$, whereas fitting induces relative bias of order k/m. The bound $\mathrm{Var}(T_{m,k}) \le 2m$ and its exact mean prove the vanishing-coverage regime; a chi-square central limit theorem gives Equation (6). Full proofs are in Section A.2. These are Gaussian affine-family results, not a claim that the constrained selectors in the real experiments follow this exact law.

Theorem 3.2 (Different thresholds for accuracy and uncorrected inference)**.** *Consider a sequence of fixed affine problems satisfying Theorem 3.1, with $m \to \infty$, dimension $k = k_m$, and fixed $0 < \delta < 1$. If $k = o(m)$, then $\|P\widehat{v}\| / \|\widehat{v}\| \to 1$ in probability. If also $k/\sqrt{m} \to c \in [0, \infty)$, the coverage of Equation (5) converges to*

$$1 - \Phi\left(\Phi^{-1}(\delta) + \sqrt{2}\, c\right), \tag{6}$$

where Φ is the standard normal cumulative distribution function (CDF). If $\sqrt{m} \ll k \ll m$, that coverage converges to zero despite vanishing relative distortion. If $k/m \to \kappa \in (0, 1)$, the norm ratio converges to $1 - \kappa$ and uncorrected coverage again tends to zero.

3.3 When reuse is valid, and what selection adds

Let $h_{m,k,\alpha}$ be the lower α-quantile of the product law in Equation (3). For $m > k$, the model-specific endpoint

$$U_\delta^{\text{affine}} = \frac{m\,\|P\widehat{v}\|^2}{h_{m,k,\delta}} \tag{7}$$

has exact $1 - \delta$ *marginal* coverage. This is constructive same-sketch inference. It requires a fixed affine family, exact least-squares fitting, and no inequality-induced active-set selection. Plugging a support or active-set dimension selected using P into this formula is not justified.

For a frozen finite family of affine problems indexed by j, assign risks $\delta\pi_j$ with $\pi_j > 0$ and $\sum_j \pi_j = 1$ before drawing P. Each problem may fit its coefficients using P. A union bound makes their corrected endpoints simultaneously valid, so a subsequent choice among them remains covered (Corollary A.1). Uniform embeddings offer another route: they cover every feasible weight vector, including constrained fits, at the cost of controlling the entire declared class. A frozen union of low-dimensional support spaces can be cheaper than embedding the full gradient span (Proposition A.5). The full-span rank is therefore a sufficient complexity measure, not a universal necessity.

Proposition 3.3 (Support search without correction)**.** *For $M \le d$ orthonormal unit residuals fixed before a Gaussian sketch, let $J = \arg\min_j \|Pv_j\|$. Then*

$$\mathbb{P}\{\|v_J\|^2 \le U_\delta^{\text{fixed}}(Pv_J)\} = (1 - \delta)^M. \tag{8}$$

The proof is the minimum of M independent chi-square variables. This standard multiplicity mechanism is separate from fitting k continuous directions; neither result is an empirical model of every selector. At $\delta = .05$, coverage is .630 for $M = 9$, .250 for $M = 27$, and below 10^{-9} for $M = 405$.

Finally, Theorem A.2 identifies the scope of a stronger guarantee. Once a transcript h and observation $Pv = y$ are fixed, a distribution-free deterministic upper norm certificate must dominate $\sup\{\|v\| : v \in \mathcal{R}_h,\ Pv = y\}$. Here $\mathcal{R}_h$ must contain only residuals compatible with *all* exposed information. This identification statement does not rule out the marginal correction in Equation (7), and becomes trivial if the transcript already determines the exact residual.

4 Exact audits after arbitrary coreset selection

4.1 Commit the deployed residual, then draw or reveal fresh rows

SplitCIC (Selection-Split Compressed Influence Certification) separates selection from audit randomness. The selection stage may use any subset search, constrained reweighting, validation-only tuning, or exact feature computation. Before opening the audit, commit the deployed indices and full-precision weights, the feature definition and anchor, tolerance τ, risk δ, and numerical error contract. Independently generate $Q \in \mathbb{R}^{m_a \times d}$ with iid $N(0, 1/m_a)$ entries. Either project the committed residual afresh, or precompute Qg_i in a service that reveals no Q-dependent information before commitment. Return

$$z = Qg_{\mathcal{V}} - \sum_{i \in \widehat{S}} \widehat{w}_i Q g_i = Q\widehat{v}. \tag{9}$$

Precomputation avoids another gradient pass but stores m_a coordinates per candidate. Post-commit projection avoids the sealed per-item table but may require recomputation. Rank independence concerns audit rows, not total gradient-extraction cost or demonstrated wall-clock speedup.

Theorem 4.1 (Exact conditional audit). *Let $\mathcal{F}_0$ contain the exact problem, deployed residual, and every pre-audit decision, with $\widehat{v}$ measurable in $\mathcal{F}_0$ and $Q \perp\!\!\!\perp \mathcal{F}_0$. Conditional on $\mathcal{F}_0$, for $\widehat{v} \neq 0$,*

$$\frac{m_a \|z\|^2}{\|\widehat{v}\|^2} \sim \chi^2_{m_a}, \qquad \frac{m_a \|z\|^2}{q_{m_a,1-\delta/2}} \leq \|\widehat{v}\|^2 \leq \frac{m_a \|z\|^2}{q_{m_a,\delta/2}} \tag{10}$$

is an exact $1-\delta$ interval. The one-sided upper endpoint is $U_\delta = m_a \|z\|^2 / q_{m_a,\delta}$. For $\widehat{v} = 0$, the interval is $[0,0]$ with coverage one.

Each row is an independent normal with variance $\|\widehat{v}\|^2 / m_a$. Conditioning fixes the outcome of selection, so no selector-complexity penalty appears. The guarantee does not become conditional on the *revealed* Q.

Proposition 4.2 (Tolerance decision in the Gaussian scale channel). *For fixed $\tau > 0$, the rule*

$$\textsc{Certify} \iff U_\delta \leq \tau^2 \tag{11}$$

has probability at most δ of certifying any $\|\widehat{v}\| \geq \tau$. Among level-$\delta$ tests observing only $z \sim N(0, \theta I_{m_a}/m_a)$, it is uniformly most powerful against $\theta < \tau^2$, with power

$$\pi_{m_a}(\theta) = F_{\chi^2_{m_a}}\left(q_{m_a,\delta}\tau^2/\theta\right), \qquad \theta > 0. \tag{12}$$

This is optimality in the stated scale channel, not against procedures with exact-gradient access. Squared-norm relative accuracy needs $O(\varepsilon^{-2}\log(1/\delta))$ audit rows, with a matching Gaussian-channel lower bound (Corollary 4.3 and Proposition A.7). Fixed-residual estimation and one-sided testing are classical; their relevance here is the absence of an adaptation penalty *after* a valid commitment.

4.2 Repeated decisions, precision, and update meaning

One simultaneous interval may be interpreted at several tolerances without invalidating its coverage. What requires new control is changing the residual using audit information, searching uncorrected alternatives, or making repeated claims. A frozen list of residuals can share an audit with predeclared risks $\delta\pi_j$; adaptive revisions instead use unopened independent blocks with predictable risks whose sum is at most δ (Corollary A.6 and Theorem A.10). These guarantees include selection of a reported result only when it is covered by the corresponding simultaneous event.

A numerical audit returns $\widetilde{z} = Q\widetilde{v} + e$, where $\|e\| \leq \rho$. If the committed proxy residual is independent of Q, the same proxy feature $\widetilde{g}_i$ is used in both updates, and $\|\widetilde{g}_i - g_i\| \leq b_i$, write $\bar{w}_i = \widehat{w}_i$ on $\widehat{S}$ and zero otherwise. Then the valid upper *norm* certificate is

$$\|\widehat{v}\| \;\leq\; \sqrt{\frac{m_a}{q_{m_a,\delta}}}\left(\|\widetilde{z}\| + \rho\right) + \sum_i |a_i - \bar{w}_i|\, b_i. \tag{13}$$

The last term is the sharp worst-case deterministic proxy radius, accounting for cancellation of shared features; separate bounds on full and selected errors would double-count that cancellation (Proposition A.13). Conditional audit leakage of total variation at most κ reduces coverage by at most κ; mutual-information and sequential versions appear in Section A. No empirical error radius or leakage bound is assumed without verification.

For an additive step of size $\eta \geq 0$,

$$\theta^+ = \theta_0 - \eta g_{\mathcal{V}}, \quad \widehat{\theta}^+ = \theta_0 - \eta G_{\widehat{S}}\widehat{w}, \quad \left\|\widehat{\theta}^+ - \theta^+\right\| = \eta\,\|\widehat{v}\|. \tag{14}$$

Thus certification concerns fidelity to a *declared* update, not its desirability. Local functional consequences require known smoothness constants; multi-step guarantees require recertification and regularity (Section A.11). Adam training is not the additive update in Equation (14).

4.3 Choosing an information contract

If M residuals and all their weights are fixed before reveal, apply Theorem 4.1 at level δ/M and union bound. A post-audit choice among those M candidates remains covered. If the audit creates

Table 1: **Certification requirements and Gaussian row scales.** Sufficient rates for relative squared-norm accuracy ε and failure probability δ. Here r is the residual-span rank. For a frozen list, $\pi_j > 0$, $\sum_j \pi_j = 1$; the shared audit size must satisfy the displayed rate for every candidate j.

Choice class	Valid information contract	Row scale
One committed residual	Fresh independent audit	$\varepsilon^{-2}\log(1/\delta)$
Frozen list, prior π_j	Shared audit with weighted simultaneous correction	$\varepsilon^{-2}[\log(1/\pi_j) + \log(1/\delta)]$
Unrestricted rank-r family	Uniform reused-sketch Gaussian embedding	$\varepsilon^{-2}[r + \log(1/\delta)]$

a new residual or changes a weight, the finite-list argument no longer applies. Together with the affine-family and union-of-spaces alternatives above, Table 1 summarizes useful sufficient contracts, not an exhaustive hierarchy.

The one-residual rate is minimax-optimal within the Gaussian channel. The frozen-list line is a sufficient weighted-union-bound rate, with exact no-correction failure on the orthogonal witness. The full-span line is a standard Gaussian oblivious subspace embedding sufficient rate; only $m \geq r$ is asserted as a general injectivity lower bound. The supporting guarantees appear in Corollaries A.6 and 4.3 and Propositions A.4 and A.7.

Corollary 4.3 (Fixed-residual relative accuracy)**.** *For $0 < \varepsilon < 1$, if $m_a \geq 8\varepsilon^{-2}\log(2/\delta)$, then after arbitrary pre-audit adaptation,*

$$(1-\varepsilon)\,\|\widehat{v}\|^2 \leq \|Q\widehat{v}\|^2 \leq (1+\varepsilon)\,\|\widehat{v}\|^2$$

with probability at least $1-\delta$.

Proof. Use the chi-square concentration consequence of Laurent & Massart (2000), $\mathbb{P}(|X/m_a - 1| \geq \varepsilon) \leq 2\exp(-m_a\varepsilon^2/8)$ for $X \sim \chi^2_{m_a}$ and Theorem 4.1. □

5 Experiments on real models and tasks

5.1 Protocol and inferential scope

The completed study uses Qwen2.5-1.5B-Instruct and 0.5B-Instruct (Qwen et al., 2025), with exact gradients in a rank-8 frozen-A/trainable-B adapter on the last attention output projection. The corpus has 64 Grade School Math 8K (GSM8K) mathematics tasks (Cobbe et al., 2021) and 64 Mostly Basic Programming Problems (MBPP) code tasks (Austin et al., 2021), split into 72 train, 16 validation, and 40 untouched final-test tasks. The visible-sketch grid is $m \in \{64, 128, 256, 512, 1024\}$, with token budgets 10, 20, and 35 percent and three projection seeds. The primary setting is $m = 512$, 20 percent. Matched-compute downstream training uses ten conditions, three training seeds, and 64 AdamW steps; the base model is evaluated separately. Final evaluation uses three generations per task. Intervals use 10,000 crossed task-by-seed bootstrap replicates; ten joint-versus-reference comparisons are Holm corrected. Additional results appear in Section B; exact configurations, implementation details, and result rows are provided in the supplementary package.

Selectors and matched information. All selectors share the target-token budget. Comparisons include random selection, loss ranking, k-center, behavior and gradient facility location, and influence-based orthogonal matching pursuit with nonnegative least-squares refitting. The joint selector combines behavior and gradient facility-location gains equally. Its inherited-weight, reweighted, and residual-refined variants isolate the effects of coefficient fitting and support refinement. The full-pool baseline shares the optimizer-step and padded-position budget.

Practical fitting uses nonnegative regularized coefficients followed by probability normalization. These constrained fits differ from the unconstrained fixed-affine model in Theorem 3.1. Rare-mode recall is defined by explicit low-frequency failure tags. Exact solver parameters, tie-breaking rules, and deployed-weight checks are specified in the supplementary implementation note.

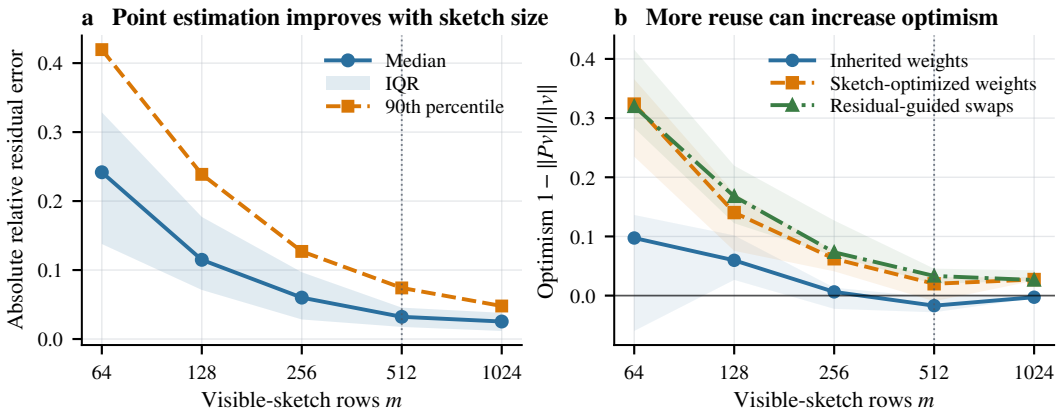

Figure 2: **Point estimates improve, while further optimization can increase optimism.** (a) Visible-sketch residual error over all primary-model methods, three budgets, and three projection seeds. (b) Optimism $1 - \|Pv\| / \|v\|$ for the same joint subset before and after reweighting, and after residual-guided swaps. Curves are medians; bands are interquartile ranges (IQR). The dotted line marks $m = 512$. The component comparison is post hoc and descriptive, not an independent-replicate test or a fit of Equation (3).

The experiments study constrained sketch-based selectors and uniform certificates on real gradients. Their capped, regularized, normalized weights and sketch-selected supports differ from the fixed-affine model, so the observations complement rather than calibrate its exact law. Independent-audit power is evaluated analytically from the measured residuals; a separate independent-audit experiment is not part of this study.

5.2 Accurate residual estimates and tolerance-dependent certificates

At the primary setting, median absolute relative residual distortion is .0340, its 90th percentile is .0594, and median exact-versus-sketched selector-ranking Spearman correlation is .95. All 27 conservative Gaussian residual-span upper bounds cover their exact residuals; six also lie below $\|g_{\mathcal{V}}\|$, giving operational fraction .2222. This distinction identifies how a useful certificate depends on its tolerance in addition to estimation accuracy. These are descriptive cell fractions, not estimates of a common repeated-sampling coverage probability.

The 0.5B transfer has median distortion .0567, 90th percentile .1092, ranking correlation .90, and operational fraction .0741. The same distinction between estimation accuracy and tolerance-dependent usefulness therefore appears in both models.

5.3 Reweighting improves reconstruction and changes calibration

The inherited-weight joint subset, its sketch-reweighted version, and its residual-refined version form a matched component comparison. Reweighting increases optimism in 43 of 45 cells (median paired increase .0499); swaps increase it in 35 of 45 (median .0143). At $m = 64$, median optimism is .097, .324, and .320, respectively. The budgets and dimensions reuse the same gradient pools, so these signs are descriptive, not 45 independent statistical replicates.

At the primary setting, reweighting also lowers the joint subset's exact residual by 23.6% relative to inherited weights. These observations are compatible: optimizing a sketch can improve the true objective while changing the calibration of the diagnostic used to assess it. The affine theorem identifies the latter effect relative to the fixed-family *exact optimum*; it does not predict that every practical reweighting step worsens an inherited coreset. Joint residual is 24.8% below influence-only OMP, with full registered behavior coverage; swaps provide a further 1.65% reduction. Behavior-only has the lowest primary residual, while random has the lowest transfer-model residual, illustrating the objective- and model-dependence of selector comparisons.

Table 2: **Primary reconstruction and untouched-test reward.** Exact residual is $\|v\| / \|g_{\mathcal{V}}\|$; reconstruction metrics average three projection seeds at the 20% budget and $m = 512$. Reward averages three seeds on 40 tasks. OMP denotes orthogonal matching pursuit. Bold marks observed extrema only, not significance. All ten registered joint-versus-reference reward comparisons have Holm-adjusted $p = 1.0$. All selectors, intervals, and transfer results appear in Section B.

Method	Exact residual ↓	Behavior ↑	Rare recall ↑	Reward ↑
Behavior facility location	**.4743**	1.000	1.000	.1417
Random	.4901	.652	.667	**.1889**
Gradient facility location	.5624	.843	1.000	.1417
Joint + swaps	.5779	1.000	1.000	.1556
Joint	.5876	1.000	1.000	.1667
Joint, inherited weights	.7695	1.000	1.000	.1528
Influence / OMP	.7813	.845	1.000	.1500
Full-pool matched	–	–	–	.1472
Base model	–	–	–	.1361

5.4 Downstream evaluation and update fidelity

Random has the highest observed final reward (.1889), versus joint (.1667) and base (.1361). Joint minus random is $-.0222$, with 95 percent interval $[-.0722, .0250]$; every registered joint-versus-reference Holm-adjusted p-value equals 1.0. The intervals do not distinguish downstream superiority or equivalence, so final reward is reported separately from local update fidelity.

The additive one-step identity holds to floating-point precision (maximum relative error 8.99×10^{-8}). The functional diagnostics reach the measurement floor of the 32-bit floating-point (FP32) adapter/bfloat16 (BF16) forward path: primary-step deltas are zero, and the largest measured changes at the larger step are of order 10^{-7}. Accordingly, the empirical fidelity result concerns parameter updates; Section B reports the functional measurements and their numerical resolution. A single anchor perturbation gives joint residual inflation .9994, a local stability diagnostic for that perturbation.

A separate *synthetic* 2,700-cell rank study has median distortion .0769 in registered well-embedded cells and .9498 when under-ranked. The contrast isolates the role of sketch dimension relative to rank. The rank/vacuity rule flags every under-ranked cell by construction; this flag frequency is distinct from a coverage estimate. Additional transfer and robustness results are reported in Section B.

5.5 Audit design from measured residuals

Applying Equation (12) to the 27 protected primary residual ratios, with $\tau = \|g_{\mathcal{V}}\|$ and per-residual $\delta = .05$, gives the certification probabilities in Table 3. These analytic probabilities under an ideal independent Gaussian audit connect the measured decision margins to audit size.

Table 3: **Audit power from measured residuals.** Exact design-stage probabilities conditional on the 27 primary residual norms at $m = 512$, not realized audit outcomes. Each value uses per-residual risk $\delta = .05$ and tolerance $\tau = \|g_{\mathcal{V}}\|$. Values displayed as 1.000 are rounded.

Audit rows	Mean power	Median power	Worst cell	Best cell
32	.801	.991	.416	1.000
64	.925	.99999	.686	1.000
128	.990	$> .999999$	.929	1.000
256	.99977	1.000	.99785	1.000
512	$> .999999$	1.000	$> .999999$	1.000

The monotone gain shows that audit rows trade storage for decision power; they cannot manufacture a decision margin, but in this observed regime 128 correctly sealed rows would attain at least .929 *individual* power in every cell under the declared tolerance and exact Gaussian model.

Each displayed power is for one committed residual at nominal risk .05. The minimum across 27 such values is not the probability that all 27 residuals would be certified in one realization. Simultaneously

reporting their certificates would require a simultaneous procedure, such as the predeclared risk allocation in Corollary A.6. The row-count comparison is a storage/dimension calculation, not measured wall-clock speedup.

6 Related work

Coresets and language-model data selection. Gradient matching, bilevel selection, and validation alignment are established approaches (Mirzasoleiman et al., 2020; Killamsetty et al., 2021a;b; Borsos et al., 2020). Recent data-selection methods use optimizer-aware features, influence functions, checkpoint trajectories, and target-aligned paths (Xia et al., 2024; Deng et al., 2024; San Joaquin et al., 2024; Chen et al., 2025; Wang et al., 2026; Liu et al., 2026; Tang et al., 2026; Yang et al., 2026). Controlled comparisons emphasize the roles of representation and selector (Nayak et al., 2026). Our contribution is certification of the resulting update residual, complementary to these selection objectives. Behavior facility location and optimized reconstruction remain different objectives; only the former has an unconditional submodularity guarantee under nonnegative similarities (Nemhauser et al., 1978; Wei et al., 2015; Das & Kempe, 2011; Elenberg et al., 2018).

Inference from randomized numerical algorithms. Gaussian sketched-regression coefficients and residual sums have exact distributional laws (Ahfock et al., 2021; Browne & Andrews, 2024); bootstrap methods estimate sketching error under their own assumptions (Lopes et al., 2018; Narahara et al., 2026). Building on these distributional ingredients, the selected-residual ratio in Equation (3) combines them for a different target: the exact norm of the residual *returned by sketch-space optimization*. Its two-threshold consequence isolates why vanishing distortion need not recover fixed-vector coverage. Independent Gaussian residual probes are also classical (Smetana et al., 2019; Martinsson & Tropp, 2020); our audit formulation specifies their post-selection target, commitment boundary, and deployment-error accounting.

Adaptivity and uniform preservation. Fixed-vector and subspace embeddings provide different simultaneous guarantees (Johnson & Lindenstrauss, 1984; Woodruff, 2014). Adaptive-query failures and reusable robust sketches study richer interaction models (Hardt & Woodruff, 2013; Cherapanamjeri & Nelson, 2020; Cohen et al., 2022; Ahmadian et al., 2025). Holdouts and algorithmic stability offer other routes to post-selection validity (Dwork et al., 2015; Zrnic & Jordan, 2023). Our exact affine law assumes one fixed family, while the independent audit allows arbitrary pre-audit selection but does not provide unlimited reuse. The fiber characterization is an identification argument under a deliberately stronger conditional contract, not a substitute for those marginal or reusable-query theories.

7 Discussion and conclusion

The exact selected-residual law provides a calibrated foundation for certifying gradient coresets after optimization. In a fixed affine family, it identifies distinct dimensional thresholds for estimation accuracy and uncorrected coverage, and supplies the corresponding same-sketch correction. Uniform guarantees extend to a declared reachable family, while fresh auditing gives exact conditional inference for an arbitrarily selected committed residual. These alternatives make the required structural knowledge and information separation explicit.

The real-model study complements this characterization with accurate compressed residual estimates and a 23.6% reconstruction improvement from reweighting the primary joint subset. At the same time, the matched-cell analysis shows why the resulting diagnostic needs its own post-selection calibration. All primary uniform endpoints cover numerically, with tolerance-dependent usefulness. The empirical scope is Qwen2.5-1.5B/0.5B, 128 tasks, one adapter location, and a fixed short training schedule; the exact affine calibration and independent-audit guarantees are established theoretically.

A useful certificate specifies the residual, the information used to select it, the probability law supporting the bound, and the update actually deployed. This separation supports principled choices among model-specific correction, uniform preservation, and independent auditing, while keeping optimization quality and inferential validity aligned.

Reproducibility Statement

We provide anonymized source code for the results reported in this paper in the supplementary material.

AI Use Statement

The author developed the research idea and proofs. AI tools provided coding, debugging, writing, and editing support, including polishing the exposition and presentation of the proofs. The author reviewed and verified the resulting material and retains responsibility for the research questions, methods, proofs, experiments, interpretation, and final content.

References

Daniel C. Ahfock, William J. Astle, and Sylvia Richardson. Statistical properties of sketching algorithms. *Biometrika*, 108(2):283–297, 2021. doi: 10.1093/biomet/asaa062.

Sara Ahmadian, Edith Cohen, and Uri Stemmer. The cost of compression: Tight quadratic black-box attacks on sketches for ℓ_2 norm estimation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0863. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/249f73e01f0a2bb6c8d971b565f159a7-Abstract-Conference.html.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. arXiv:2108.07732, 2021. URL https://arxiv.org/abs/2108.07732.

Zalán Borsos, Mojmír Mutný, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aa2a77371374094fe9e0bc1de3f94ed9-Abstract.html.

Ryan P. Browne and Jeffrey L. Andrews. Statistical inference for sketching algorithms. *Information and Inference: A Journal of the IMA*, 13(3):iaae023, 2024. doi: 10.1093/imaiai/iaae023.

Ziliang Chen, Yongsen Zheng, Zhao-Rong Lai, Zhanfu Yang, Cuixi Li, Yang Liu, and Liang Lin. Quadratic coreset selection: Certifying and reconciling sequence and token mining for efficient instruction tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0073. URL https://papers.neurips.cc/paper_files/paper/2025/hash/02e4145752691bb16e970bd71b2c3253-Abstract-Conference.html.

Yeshwanth Cherapanamjeri and Jelani Nelson. On adaptive distance estimation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/803ef56843860e4a48fc4cdb3065e8ce-Abstract.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Edith Cohen, Xin Lyu, Jelani Nelson, Tamas Sarlos, Moshe Shechner, and Uri Stemmer. On the robustness of CountSketch to adaptive inputs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 4112–4140, 2022. URL https://proceedings.mlr.press/v162/cohen22a.html.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning*, 2011. URL https://icml.cc/2011/papers/542_icmlpaper.pdf.

Zhiwei Deng, Tao Li, and Yang Li. Influential language data selection via gradient trajectory pursuit. arXiv:2410.16710, 2024. URL https://arxiv.org/abs/2410.16710.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/bad5f33780c42f2588878a9d07405083-Abstract.html.

Ethan R. Elenberg, Rajiv Khanna, Alexandros G. Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568, 2018. doi: 10.1214/17-AOS1679.

Moritz Hardt and David P. Woodruff. How robust are linear sketches to adaptive inputs? In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pp. 121–130, 2013. doi: 10.1145/2488608.2488624.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pp. 189–206. American Mathematical Society, 1984. doi: 10.1090/conm/026/737400. URL https://www.stanford.edu/class/cs114/readings/JL-Johnson.pdf.

Samuel Karlin and Herman Rubin. The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, 27(2):272–299, 1956. doi: 10.1214/aoms/1177728259.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5464–5474, 2021a. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021b. doi: 10.1609/aaai.v35i9.16988.

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.

Hetian Liu, Jin Cui, Mengcheng Shi, Yanbin Hu, Xinyue Long, Boran Zhao, and Pengju Ren. GLOBE: Trajectory-aligned gradient matching with structured sparseoptimization for coreset selection. arXiv:2608.02690, 2026. URL https://arxiv.org/abs/2608.02690.

Miles E. Lopes, Shusen Wang, and Michael W. Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3217–3226, 2018. URL https://proceedings.mlr.press/v80/lopes18a.html.

Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020. doi: 10.1017/S0962492920000021.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6950–6960, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Akito Narahara, Takayuki Kawashima, and Takafumi Kanamori. Bootstrap error estimation and sketch-size selection for sketched ridge regression. arXiv:2608.15478, 2026. URL https://arxiv.org/abs/2608.15478.

Nihal V. Nayak, Paula Rodriguez-Diaz, Neha Hulkund, Sara Beery, and David Alvarez-Melis. A critical look at targeted instruction selection: Disentangling what matters (and what doesn't). In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.14696.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions–I. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. arXiv:2412.15115v2, 2025. URL https://arxiv.org/abs/2412.15115v2.

Ayrton San Joaquin, Bin Wang, Zhengyuan Liu, Nicholas Asher, Brian Lim, Philippe Muller, and Nancy F. Chen. In2Core: Leveraging influence functions for coreset selection in instruction finetuning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10324–10335, 2024. doi: 10.18653/v1/2024.findings-emnlp.604. URL https://aclanthology.org/2024.findings-emnlp.604/.

Kathrin Smetana, Olivier Zahm, and Anthony T. Patera. Randomized residual-based error estimators for parametrized equations. *SIAM Journal on Scientific Computing*, 41(2):A900–A926, 2019. doi: 10.1137/18M120364X.

Tianhao Tang, Haoyang Li, and Lei Chen. GRACE: A dynamic coreset selection framework for large language model optimization. arXiv:2604.11810, 2026. URL https://arxiv.org/abs/2604.11810.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

Fangxin Wang, Peyman Baghershahi, Langzhou He, Henry Peng Zou, Sourav Medya, and Philip S. Yu. Filter-then-weight: Online data selection and reweighting for LLM fine-tuning. arXiv:2604.00001v2, 2026. URL https://arxiv.org/abs/2604.00001v2.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1954–1963, 2015. URL https://proceedings.mlr.press/v37/wei15.html.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54104–54132, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Huitao Yang, Hengzhi He, and Guang Cheng. Let the target select for itself: Data selection via target-aligned paths. arXiv:2605.09404, 2026. URL https://arxiv.org/abs/2605.09404.

Tijana Zrnic and Michael I. Jordan. Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691, 2023. doi: 10.1214/23-AOS2303.

A Complete Proofs and Additional Theory

A.1 Exact law of the residual selected by affine reweighting

This section derives Theorem 3.1 for fixed b and L. We build on the Gaussian sketch-regression coefficient law of Ahfock et al. (2021) and the coefficient/residual-sum independence in Browne & Andrews (2024, Section 2.1). The product law below follows by combining these ingredients with the orthogonal decomposition of the full residual returned by the sketch fit; the two-threshold consequence concerns this selected target, not a new generic regression distribution.

Orthogonal reduction. Choose an orthonormal basis $U \in \mathbb{R}^{d\times k}$ for L and write $b = Ua_0 + e$, $U^\top e = 0$, $\|e\| = \sigma > 0$. Every residual is $e - Ut$, after translating coefficients by a_0. Set $A = \sqrt{m}\,PU$ and $z = \sqrt{m}\,Pe/\sigma$. Orthogonality and Gaussianity make all entries of A and z independent $N(0,1)$. For $m > k$, A has full column rank almost surely. With

$$\eta = (A^\top A)^{-1}A^\top z, \qquad H = A(A^\top A)^{-1}A^\top,$$

the optimizer is $\widehat{v} = e - \sigma U\eta$, so

$$\|\widehat{v}\|^2 = \sigma^2(1 + \|\eta\|^2), \qquad m\,\|P\widehat{v}\|^2 = \sigma^2\,\|(I-H)z\|^2\,. \tag{15}$$

Independence of fitting error and sketch residual. Conditional on A, Hz and $(I-H)z$ are independent Gaussian projections. Thus η is independent of $X = \|(I-H)z\|^2$ conditionally on A, and $X \mid A \sim \chi^2_{m-k}$. Since the latter law does not depend on A, X is independent of the pair (A,η), hence of $\|\eta\|^2$ unconditionally. This last step is necessary: conditional independence alone would not suffice without the constant conditional law of X.

Marginal coefficient radius. Conditional on A, $\eta \sim N(0,(A^\top A)^{-1})$. The joint density in (A,η) is proportional to

$$\det(A^\top A)^{1/2}\exp\{-\tfrac{1}{2}\operatorname{tr}[A^\top A(I+\eta\eta^\top)]\}.$$

For fixed η, change variables $C = A(I+\eta\eta^\top)^{1/2}$. The matrix Jacobian contributes $\det(I+\eta\eta^\top)^{-m/2}$, and the determinant prefactor contributes its additional power $-1/2$. After integrating out C and normalizing,

$$f(\eta) = \frac{\Gamma((m+1)/2)}{\pi^{k/2}\Gamma((m-k+1)/2)}(1+\|\eta\|^2)^{-(m+1)/2}. \tag{16}$$

Consequently $W = \|\eta\|^2$ has beta-prime density

$$f_W(w) = \frac{w^{k/2-1}(1+w)^{-(m+1)/2}}{\mathrm{B}(k/2,(m-k+1)/2)}, \quad w > 0.$$

It follows that $B = (1+W)^{-1} \sim \mathrm{Beta}((m-k+1)/2, k/2)$ and $B \perp X$. Substitution in Equation (15) proves the product law. Taking expectations gives Equation (4). The mean norm ratio has a particularly simple form:

$$\mathbb{E}\frac{\|P\widehat{v}\|}{\|\widehat{v}\|} = \left(1 - \frac{k}{m}\right)c_m, \qquad c_m = \sqrt{\frac{2}{m}}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)}.$$

Indeed, independence gives $\mathbb{E}\sqrt{XB} = \mathbb{E}\sqrt{X}\,\mathbb{E}\sqrt{B}$; the gamma factors cancel, leaving the fixed-vector mean c_m times $1 - k/m$. In addition,

$$\mathbb{E}\,\|P\widehat{v}\|^2 = \sigma^2\frac{m-k}{m}, \qquad \mathbb{E}\,\|\widehat{v}\|^2 = \sigma^2\frac{m-1}{m-k-1} \quad (m > k+1). \tag{17}$$

For $k \geq 1$, the latter expectation is infinite at $m = k+1$; the product law itself still holds. For $k = 0$, there is no fitted direction and the usual fixed-vector law follows directly. For $m \leq k$, the $m \times k$ Gaussian matrix A has full row rank almost surely, so there is a t with $At = \sigma z$. The sketched minimum is then zero although every exact residual in $b - L$ has norm at least σ. If instead $\sigma = 0$ and $m > k$, the unique minimizing residual is zero, with degenerate coverage one rather than a nonzero ratio distribution. □

A.2 THE TWO THRESHOLDS: PROOF AND FINITE-DIMENSIONAL BOUNDS

The product representation yields

$$\mu_{m,k} := \mathbb{E}T_{m,k} = \frac{(m-k)(m-k+1)}{m+1} = m - 2k + \frac{k(k+1)}{m+1}, \tag{18}$$

$$\mathrm{Var}(B) = \frac{2k(m-k+1)}{(m+1)^2(m+3)},$$

$$\mathrm{Var}(T_{m,k}) = \mathbb{E}[B^2]\,\mathrm{Var}(X) + (\mathbb{E}X)^2\,\mathrm{Var}(B) \le 2(m-k) + 2k = 2m. \tag{19}$$

For $k = o(m)$, $\mu_{m,k}/m \to 1$ and Equation (19) makes $T_{m,k}/m \to 1$ in probability. Taking square roots proves vanishing norm distortion. For $k/m \to \kappa \in (0,1)$, the same argument gives $T_{m,k}/m \to (1-\kappa)^2$.

For $k = o(m)$, the chi-square central limit theorem gives $(X - (m-k))/\sqrt{2m} \Rightarrow N(0,1)$. Also $B \to 1$ and

$$\mathrm{Var}\left(\frac{(m-k)(B - \mathbb{E}B)}{\sqrt{2m}}\right) \le \frac{k}{m} \longrightarrow 0.$$

Therefore, by independence (or directly by the displayed variance bound and Slutsky's theorem),

$$\frac{T_{m,k} - \mu_{m,k}}{\sqrt{2m}} = B\frac{X - (m-k)}{\sqrt{2m}} + \frac{(m-k)(B - \mathbb{E}B)}{\sqrt{2m}} \;\Rightarrow\; N(0,1).$$

For fixed δ, $(q_{m,\delta} - m)/\sqrt{2m} \to \Phi^{-1}(\delta)$. If $k/\sqrt{m} \to c < \infty$, Equation (18) gives $(\mu_{m,k} - m)/\sqrt{2m} \to -\sqrt{2}c$, proving Equation (6). When $\sqrt{m} \ll k \ll m$, the difference $q_{m,\delta} - \mu_{m,k} \sim 2k$ is positive. Chebyshev's inequality gives

$$\mathbb{P}\{T_{m,k} \ge q_{m,\delta}\} \le \frac{2m}{(q_{m,\delta} - \mu_{m,k})^2} \longrightarrow 0. \tag{20}$$

When $k/m \to \kappa \in (0,1)$, $q_{m,\delta}/m \to 1$ while $T_{m,k}/m \to (1-\kappa)^2 < 1$, proving coverage collapse. □

Strict finite-dimensional undercoverage. For $k > 0$, $0 < B < 1$ almost surely and $X \sim \chi^2_{m-k}$. Couple an independent $Z \sim \chi^2_k$ to obtain $X + Z \sim \chi^2_m$. For every positive threshold t, $\mathbb{P}(XB \ge t) < \mathbb{P}(X + Z \ge t)$: the inequality follows by ordering, and strictness follows from the continuous positive densities. At $t = q_{m,\delta}$ the right side is $1 - \delta$. Replacing the denominator by $q_{m-k,\delta}$ also undercovers, because $XB < X$ and the corresponding tail is strictly less than $1 - \delta$.

Exact correction and numerical evaluation. The product law is continuous with positive density on $(0, \infty)$, so its lower quantile $h_{m,k,\delta}$ is positive and unique. Its distribution function can be evaluated from

$$H_{m,k}(t) = \int_0^1 F_{\chi^2_{m-k}}(t/b)\frac{b^{(m-k+1)/2-1}(1-b)^{k/2-1}}{\mathrm{B}((m-k+1)/2, k/2)}\,db. \tag{21}$$

By construction, $\mathbb{P}\{T_{m,k} \ge h_{m,k,\delta}\} = 1 - \delta$, proving Equation (7). A central interval replaces the lower and upper chi-square quantiles in Equation (10) by $h_{m,k,1-\delta/2}$ and $h_{m,k,\delta/2}$ and uses $m\,\|P\widehat{v}\|^2$ in the numerator. This is coverage of a random selected target over repeated sketches, not conditional coverage given the fitted coefficients or full sketch. The supplement supplies deterministic quadrature and inversion code; its numerical tolerance is not part of the exact mathematical statement.

Corollary A.1 (A frozen menu of adaptively fitted affine families)**.** *Fix M affine families $b_j - L_j$ and weights $\pi_j > 0$, $\sum_j \pi_j = 1$, independently of P. Let $k_j = \dim L_j < m$ and let $\widehat{v}_j$ minimize the sketched norm on family j. Assume exact fitting and either $\mathrm{dist}(b_j, L_j) > 0$ or the zero-residual case described above. Then*

$$\mathbb{P}\left\{\|\widehat{v}_j\|^2 \le \frac{m\,\|P\widehat{v}_j\|^2}{h_{m,k_j,\delta\pi_j}} \text{ for all } j\right\} \ge 1 - \delta.$$

Any index selected from the menu using the same sketch inherits the bound.

Proof. Each fixed family has failure probability at most $\delta\pi_j$ by Theorem 3.1 and the quantile construction. Summing these probabilities proves simultaneous coverage without requiring independence between families. The selected endpoint is valid on that simultaneous event. □

This differs from a frozen *residual* menu: each residual here is fitted using P. Correcting only the families inspected after seeing the sketch is not the stated guarantee. A complete fixed support menu may be combinatorially large. Regularization and inequality constraints also change the single-family law; the observed active set cannot be substituted without a valid selection argument.

A.3 TRANSCRIPT-CONDITIONAL SKETCH-FIBER BARRIER

We restate the information contract explicitly. Fix a realized visible transcript h, including the sketch P and every item of side information exposed to the certificate. Let $\mathcal{R}_h$ be the residual class over which the certificate promises distribution-free validity. Auxiliary randomness ξ is permitted, but conditional on (h, Pv) its law is the same for every residual in the fiber: it contains no fresh residual-dependent measurement.

Theorem A.2 (Sharp conditional identification radius)**.** *For a fixed transcript h and residual class $\mathcal{R}_h$ compatible with it, define the nonempty fiber and its radius by*

$$\mathcal{R}_{P,h}(y) = \{v \in \mathcal{R}_h : Pv = y\}, \qquad \Gamma_{P,h}(y) = \sup_{v \in \mathcal{R}_{P,h}(y)} \|v\| .$$

Suppose a measurable upper norm bound $U(h, y, \xi) \in [0, \infty]$ has conditional coverage at least $1 - \delta$ for every residual in each fiber. Its auxiliary randomness has the same conditional law for every such residual. Then

$$\mathbb{P}_\xi\{U(h, y, \xi) \geq \Gamma_{P,h}(y)\} \geq 1 - \delta.$$

The deterministic bound $\Gamma_{P,h}$ is pointwise minimal. A confidence set covering each compatible residual with probability at least $1 - \delta$ has diameter at least $\|v_0 - v_1\|$ with probability at least $1 - 2\delta$, for any two points in the same fiber. An unbounded kernel ray makes a finite bound impossible under this contract; positive lower preservation of an entire r-dimensional subspace requires at least r sketch rows.

The compatibility class includes only underlying problems consistent with every revealed quantity, not arbitrary residuals after ignoring information already known. If exact G, a, S, and w are revealed, the residual is known and the class is a singleton. The theorem concerns conditional distribution-free identification, not all marginal post-selection procedures.

Proof of Theorem A.2. Fix $y \in P\mathcal{R}_h$. Choose a sequence $v_k \in \mathcal{R}_{P,h}(y)$ whose norms increase to $\Gamma_{P,h}(y)$; for an infinite radius, choose $\|v_k\| \uparrow \infty$. Validity for each v_k gives $\mathbb{P}_\xi\{U(h, y, \xi) \geq \|v_k\|\} \geq 1 - \delta$. The events decrease as k increases. Continuity of probability from above yields

$$\mathbb{P}_\xi\{U(h, y, \xi) \geq \Gamma_{P,h}(y)\} = \lim_k \mathbb{P}_\xi\{U(h, y, \xi) \geq \|v_k\|\} \geq 1 - \delta.$$

For a confidence set, let $A_j = \{v_j \in \mathsf{C}(h, y, \xi)\}$ for two points in the same fiber. Since $\mathbb{P}(A_j) \geq 1-\delta$, the union bound gives $\mathbb{P}(A_0 \cap A_1) \geq 1 - 2\delta$. On that event the set diameter is at least $\|v_0 - v_1\|$.

For deterministic certificates, the function $\Gamma_{P,h}$ is valid by definition and every valid certificate dominates it, so it is pointwise minimal in the extended nonnegative reals. If $\mathcal{R}_h$ contains $\{tv : t \geq 0\}$ for nonzero $v \in \ker(P)$, then $\Gamma_{P,h}(0) = \infty$. If it contains a radius-ρ ball in an r-dimensional subspace L and $m < r$, rank–nullity gives a unit $x \in L \cap \ker(P)$, hence $\rho x \in \mathcal{R}_{P,h}(0)$ and $\Gamma_{P,h}(0) \geq \rho$. Finally, a positive lower embedding $\|Px\| \geq c\,\|x\|$ on an r-dimensional residual span makes P injective on that span, which is impossible for $m < r$. □

The theorem does not say every practical under-ranked fiber is unbounded, nor does it rule out problem-specific inference under additional structural assumptions. Cardinality, sign, simplex, and coefficient caps may bound a fiber. It says that a transcript-conditionally valid, distribution-free same-sketch certificate with no fresh residual-dependent observation must solve the worst-case feasible inverse problem in that fiber; average distortion of observed cells is insufficient.

A.4 UNIFORM RESIDUAL-SPAN EMBEDDING

Let $\mathcal{R} = \{v(S,w) : (S,w) \text{ feasible}\}$ and $\mathcal{L}_{\mathcal{R}} = \operatorname{span}(\mathcal{R})$, with exact algebraic rank $r_{\mathcal{R}}$. The problem and these spaces are fixed before drawing P.

Assumption A.3 (Residual-span embedding)**.** *For $0 < \varepsilon < 1$ and $0 < \delta < 1$, with probability at least $1 - \delta$, the visible sketch satisfies*

$$(1-\varepsilon)\|x\|^2 \le \|Px\|^2 \le (1+\varepsilon)\|x\|^2 \quad \text{for every } x \in \mathcal{L}_{\mathcal{R}}.$$

Proposition A.4 (Uniform reused-sketch certificate)**.** *On the event in Assumption A.3, simultaneously for every feasible (S,w),*

$$(1-\varepsilon)R(S,w) \le \|P(g_{\mathcal{V}} - G_S w)\|^2 \le (1+\varepsilon)R(S,w).$$

The statement remains valid when (S,w) is selected after observing the entire sketch.

Proof. Every feasible residual lies in the fixed space $\mathcal{L}_{\mathcal{R}}$. On the one embedding event, the inequalities therefore hold simultaneously. Selection after observing P does not require a union bound. □

For a fixed subset, if $\widehat{w}_S$ minimizes the sketched residual and $w_S^\star$ minimizes the exact residual over the same feasible weight set, then

$$\|g_{\mathcal{V}} - G_S\widehat{w}_S\|^2 \le \frac{1+\varepsilon}{1-\varepsilon}\|g_{\mathcal{V}} - G_S w_S^\star\|^2.$$

This follows by applying the lower embedding inequality to $\widehat{w}_S$, sketched optimality, and the upper inequality to $w_S^\star$. The result applies to least squares, nonnegative least squares (NNLS), or capped-simplex weights, but only when the uniform embedding event actually holds.

Proposition A.5 (A union of support spaces need not pay full-pool rank)**.** *Suppose the fixed reachable class is contained in a union of M fixed linear spaces L_j of dimensions ℓ_j. Choose fixed $\pi_j > 0$ summing to one. For $0 < \varepsilon < 1$ and $0 < \delta < 1$, a Gaussian sketch gives simultaneous squared-norm preservation within $1 \pm \varepsilon$ throughout the union, with probability at least $1 - \delta$, whenever*

$$m \ge 9\varepsilon^{-2}\max_j\left(\sqrt{\ell_j} + \sqrt{2\log(2/(\delta\pi_j))}\right)^2.$$

In particular, uniform π_j gives the sufficient scale $O(\varepsilon^{-2}[\max_j \ell_j + \log(M/\delta)])$.

Proof. For an orthonormal basis U_j, standard Gaussian extreme-singular-value bounds imply that every singular value of PU_j is in $[1-a_j, 1+a_j]$, except with probability at most $\delta\pi_j$, where $a_j = \sqrt{\ell_j/m} + \sqrt{2\log(2/(\delta\pi_j))/m}$ (Woodruff, 2014; Martinsson & Tropp, 2020). The stated row count makes $a_j \le \varepsilon/3$. Since $(1-a_j)^2 \ge 1-\varepsilon$ and $(1+a_j)^2 \le 1+\varepsilon$, a union bound proves the claim. □For affine families $b_j - L_j$, use the linear span of $\{b_j\} \cup L_j$, whose dimension is at most $k_j + 1$. The guarantee remains valid for any constrained fit within those families. This is a sufficient bound, not an asserted minimax characterization of every reachable set. Replacing exact rank by numerical or stable rank requires controlling the neglected directions; rank diagnostics alone are not such a proof.

A.5 EXACT FINITE-MENU POST-SELECTION LAW

Proof of Proposition 3.3. For one row $p^\top$ of P, the vector $(p^\top v_1, \dots, p^\top v_M)$ is Gaussian with covariance I_M/m because the v_j are orthonormal. Hence all mM coordinates are independent and

$$X_j = m\|Pv_j\|^2 \overset{\text{iid}}{\sim} \chi^2_m.$$

Because ties occur with probability zero, $X_J = \min_j X_j$. Therefore

$$\mathbb{P}(X_J \le t) = 1 - \mathbb{P}(X_1 > t, \dots, X_M > t) = 1 - [1 - F_{\chi^2_m}(t)]^M.$$

Since every selected residual has squared norm one, the naive upper bound covers iff $X_J \ge q_{m,\delta}$. Thus

$$\mathbb{P}\{1 \le U_\delta^{\text{fixed}}(Pv_J)\} = \mathbb{P}\{X_1 \ge q_{m,\delta}, \dots, X_M \ge q_{m,\delta}\} = (1-\delta)^M.$$

The median $t_{1/2}$ of $\|Pv_J\|^2 = X_J/m$ solves $1 - [1 - F(mt_{1/2})]^M = 1/2$, giving $t_{1/2} = q_{m,1-2^{-1/M}}/m$. For exact joint coverage $1-\delta$, choose a per-candidate lower-tail level α_M satisfying $(1-\alpha_M)^M = 1-\delta$, namely $\alpha_M = 1 - (1-\delta)^{1/M}$. □

Interpretation. This witness separates two effects. Even though each fixed residual has an exact pivot, searching for a small sketch norm changes the distribution to an order statistic. For this orthogonal family, the exact no-correction collapse forces a multiplicity adjustment if one insists on the same fixed-vector construction. The full rank dependence appears when the post-sketch choice class is an unrestricted linear family rather than a frozen list. We do not claim a universal finite-list minimax lower bound for every structured procedure.

Corollary A.6 (Prior-weighted frozen-menu audit)**.** *Let $v_1, \dots, v_M$ and their numerical representations be fixed before an independent Gaussian audit is revealed. Choose predeclared weights $\pi_j > 0$ with $\sum_{j=1}^{M} \pi_j = 1$, and give candidate j the one-sided upper endpoint*

$$U_j = \frac{m_a \|Qv_j\|^2}{q_{m_a,\delta\pi_j}}.$$

Then, with probability at least $1-\delta$, $\|v_j\|^2 \le U_j$ simultaneously for all j. Consequently, any index $J(Q)$ chosen after seeing all audited endpoints remains covered. Applying the fixed-vector concentration bound at levels $\delta\pi_j$ gives simultaneous ε-relative preservation once a shared dimension satisfies $m_a \ge \max_j 8\varepsilon^{-2}\log(2/(\delta\pi_j))$. The uniform choice $\pi_j = 1/M$ recovers the ordinary sufficient finite-list rate.

Proof. For each fixed j, Theorem 4.1 gives failure probability at most $\delta\pi_j$. The union bound gives $\mathbb{P}\{\exists j : \|v_j\|^2 > U_j\} \le \sum_j \delta\pi_j = \delta$. The statement for $J(Q)$ follows on the simultaneous event. Substituting $\delta\pi_j$ into Corollary 4.3 and taking the maximum over j gives the description-length term. □

The prior must be frozen before reveal. It may encode a scientific preference, validation-only plausibility, or a prefix code over candidates, but it cannot be optimized using the audit outcomes it is meant to correct.

A.6 Exact independent audit interval

Proof of Theorem 4.1. Condition on $\mathcal{F}_0$. The residual $\widehat{v}$ is deterministic and Q retains its independent Gaussian law. For each audit row $q_j^\top$,

$$q_j^\top \widehat{v} \sim N\left(0, \frac{\|\widehat{v}\|^2}{m_a}\right),$$

independently over j. Hence

$$\frac{m_a \|Q\widehat{v}\|^2}{\|\widehat{v}\|^2} = \sum_{j=1}^{m_a} \left(\frac{q_j^\top \widehat{v}}{\|\widehat{v}\|/\sqrt{m_a}}\right)^2 \sim \chi^2_{m_a}.$$

Inverting the central and lower-tail chi-square events gives Equation (10). The zero-residual case is immediate. Since the conditional statement holds for every realized pre-audit state, averaging gives the same unconditional coverage. □

A.7 Uniformly most powerful tolerance decision

Proof of Proposition 4.2. Let $T = m_a \|Q\widehat{v}\|^2$ and $\theta = \|\widehat{v}\|^2$. Its density is

$$f_\theta(t) = \frac{t^{m_a/2-1}}{2^{m_a/2}\Gamma(m_a/2)\theta^{m_a/2}} \exp\left(-\frac{t}{2\theta}\right), \qquad t > 0.$$

For $0 < \theta_1 < \theta_0$, the likelihood ratio $f_{\theta_1}(t)/f_{\theta_0}(t)$ decreases in t. The family has a monotone likelihood ratio, so the Karlin–Rubin theorem (Karlin & Rubin, 1956) makes a lower-tail test uniformly most powerful for $H_1 : \theta < \tau^2$. The threshold $\tau^2 q_{m_a,\delta}$ has size δ at the least favorable boundary $\theta = \tau^2$ and smaller size for larger θ. Substituting $T = \theta X$, $X \sim \chi^2_{m_a}$, gives Equation (12). □

The density of the full vector depends on it only through T, and its direction is ancillary. Thus sufficiency (or the vector likelihood ratio directly) extends the conclusion from T-based tests to all vector-based tests. The test is uniformly most powerful among measurable level-δ tests based on the sealed Gaussian audit vector in this one-parameter scale family. It is not claimed optimal among arbitrary interactive or cryptographic protocols.

A.8 Audit complexity lower bound

Proposition A.7 (Gaussian-channel lower bound)**.** *Fix $0 < \varepsilon \le 1/8$ and $0 < \delta \le 1/8$. In the sealed audit channel that discloses only $Y = Qv$ (and not Q), any estimator based on m_a independent Gaussian audit coordinates that estimates $\|v\|^2$ to relative error ε uniformly over nonzero v with probability at least $1 - \delta$ requires $m_a = \Omega(\varepsilon^{-2} \log(1/\delta))$.*

Proof. Take squared norms $\theta_0 = 1$ and $\theta_1 = 1 + 4\varepsilon$. Their relative-error target intervals are disjoint. A uniformly accurate estimator induces a test with each error at most δ. Writing D_{KL} for Kullback–Leibler divergence, the audit vectors have laws $P_j = N(0, \theta_j I_{m_a}/m_a)$ and

$$D_{\mathrm{KL}}(P_0 \| P_1) = \frac{m_a}{2}\left[\frac{1}{1+4\varepsilon} - 1 + \log(1+4\varepsilon)\right] \le 4m_a\varepsilon^2.$$

The Bretagnolle–Huber inequality (Tsybakov, 2009) lower-bounds the sum of test errors by $\frac{1}{2}e^{-D_{\mathrm{KL}}}$. Requiring this to be at most 2δ gives $m_a \ge (4\varepsilon^2)^{-1}\log(1/(4\delta))$. □

A.9 Approximate independence and finite precision

Proposition A.8 (Leakage and deterministic numerical error)**.** *Suppose that, uniformly over all pre-audit states h in the support of $\mathcal{F}_0$, the actual conditional law of the audit given $\mathcal{F}_0 = h$ is within total variation κ of the ideal independent Gaussian law. If $\widetilde{z} = Q\widehat{v} + e$ and $\|e\| \le \rho$, then*

$$\widetilde{U}_{\delta,\rho} = \frac{m_a(\|\widetilde{z}\| + \rho)^2}{q_{m_a,\delta}}$$

satisfies $\mathbb{P}(\|\widehat{v}\|^2 \le \widetilde{U}_{\delta,\rho}) \ge 1 - \delta - \kappa$. A two-sided interval uses $[\max\{0, \|\widetilde{z}\| - \rho\}, \|\widetilde{z}\| + \rho]$ in the endpoints of Equation (10).

Proof. Under the ideal law, the exact audit event has probability at least $1 - \delta$. Total variation changes the probability of any event by at most κ. On the exact event, $\|Q\widehat{v}\| \le \|\widetilde{z}\| + \rho$, so the inflated endpoint is conservative. The lower endpoint follows from the reverse triangle inequality. □The decision $\widetilde{U}_{\delta,\rho} \le \tau^2$ has false-certification probability at most $\delta + \kappa$ for every $\|\widehat{v}\| \ge \tau > 0$: under the ideal law its rejection event is contained in the exact lower-tail rejection event, and total variation adds at most κ. For a specified total risk δ_{tot}, require $0 \le \kappa < \delta_{\mathrm{tot}} < 1$ and use Gaussian tail level $\delta = \delta_{\mathrm{tot}} - \kappa$. The central interval and one-sided decision each have their stated guarantee; these are not asserted to form a joint $1 - \delta - \kappa$ event.

Corollary A.9 (Mutual-information leakage bound)**.** *Let a random transcript H_0 generate the pre-audit sigma-field $\mathcal{F}_0$. Suppose Q has the ideal Gaussian marginal but may be statistically dependent on H_0, with mutual information $I(H_0; Q) \le \iota$ measured in nats. Then the unconditional coverage of the nominal interval in Equation (10) is at least*

$$1 - \delta - \sqrt{\iota/2}.$$

Proof. Let P be the actual joint law of (H_0, Q) and let $P_0 = P_{H_0} \otimes P_Q$ be the product of its marginals. Under P_0, conditioning on $\mathcal{F}_0$ and applying Theorem 4.1 gives coverage at least $1 - \delta$. Pinsker's inequality and the identity $D_{\mathrm{KL}}(P\|P_0) = I(H_0; Q)$ imply $\mathrm{TV}(P, P_0) \le \sqrt{\iota/2}$, so the coverage event can lose at most that amount under P. □

The total-variation statement in Proposition A.8 is transcript-uniform and stronger when a defensible worst-case κ is available; Corollary A.9 is an average, unconditional alternative. A usable ρ can come from interval arithmetic, deterministic accumulation bounds, or a protected high-precision duplicate with its own certified error bound. Higher precision alone does not establish a deterministic radius. A low-precision downstream forward pass does not automatically provide such a radius for logits.

A.10 SEQUENTIAL AUDIT BANKS AND AUDIT-ROW ALLOCATION

Theorem A.10 (Anytime validity from unopened audit blocks). *Let $Q^{(1)}, Q^{(2)}, \ldots$ be a countable bank of mutually independent audit blocks jointly independent of the initial selection-side state. Before block t is opened, residual $\widehat{v}_t$ and risk δ_t may depend arbitrarily on all selection information and outputs from blocks $1{:}t-1$, and $Q^{(t)}$ must retain its independent Gaussian law conditional on the complete history before reveal. If $0 < \delta_t < 1$ is predictable and $\sum_{t\geq 1} \delta_t \leq \delta$ almost surely, applying Theorem 4.1 to block t at risk δ_t gives simultaneous validity for every $t \geq 1$ with probability at least $1-\delta$. Consequently, the intervals remain jointly valid up to any stopping time adapted to the revealed history.*

Proof. Condition on the history before block t. The current residual and risk allocation are fixed, and the unopened block retains its independent Gaussian law, so Theorem 4.1 applies conditionally with failure probability at most δ_t. If F_t is the failure event, then $\mathbb{P}(F_t) \leq \mathbb{E}[\delta_t]$. The countable union bound and Tonelli's theorem give $\mathbb{P}(\bigcup_t F_t) \leq \sum_t \mathbb{E}[\delta_t] = \mathbb{E}[\sum_t \delta_t] \leq \delta$. Validity for all deterministic times immediately implies validity up to any adapted stopping time. □

Proposition A.11 (Continuous audit-row allocation). *For a finite integer $T \geq 1$, total budget $M > 0$, and $c_t > 0$, minimize $W = \sum_{t=1}^{T} c_t/\sqrt{m_t}$ over $m_t > 0$ with $\sum_t m_t = M$. The unique solution is*

$$m_t^* = M\frac{c_t^{2/3}}{\sum_s c_s^{2/3}}, \qquad W^* = \frac{(\sum_t c_t^{2/3})^{3/2}}{\sqrt{M}}.$$

Proof. The Lagrangian stationarity condition gives $m_t \propto c_t^{2/3}$; strict convexity of $m^{-1/2}$ gives uniqueness. □The coefficients must be frozen from selection-side information or estimated with a separate pilot; rows cannot be reallocated after opening the same audit.

A.11 UPDATE, FUNCTIONAL, PROXY, AND MULTI-STEP BOUNDS

Proposition A.12 (One-step and local functional consequences). *For the additive updates in Equation (14) with $\eta \geq 0$, $\left\|\widehat{\theta}^{+} - \theta^{+}\right\| = \eta \left\|\widehat{v}\right\|$ exactly. If f is L_f-Lipschitz on the segment joining the updates,*

$$\left|f(\widehat{\theta}^{+}) - f(\theta^{+})\right| \leq L_f \eta \left\|\widehat{v}\right\|.$$

If f is β_f-smooth,

$$f(\widehat{\theta}^{+}) - f(\theta^{+}) \leq \eta \left\|\nabla f(\theta^{+})\right\| \left\|\widehat{v}\right\| + \frac{\beta_f \eta^2}{2} \left\|\widehat{v}\right\|^2.$$

When the stated constants and derivative norms are known or independently upper bounded, a valid audit endpoint yields these functional bounds with the same probability. Without such constants, the bounds are implications, not computed guarantees.

Proof. Subtract the updates. The Lipschitz statement is immediate. For smooth f, apply the descent lemma at θ^{+} with displacement $\eta\widehat{v}$ and Cauchy–Schwarz. Apply the same argument to $-f$ for a two-sided absolute bound. □

Proposition A.13 (Sharp shared-feature proxy radius). *Suppose $\widetilde{g}_i = g_i + \xi_i$, $\|\xi_i\| \leq b_i$, and the same proxy is used for each occurrence of item i. Pad deployed weights by $\bar{w}_i = w_i$ on S and zero elsewhere, set $c_i = a_i - \bar{w}_i$, and let $\widetilde{v} = \sum_i c_i \widetilde{g}_i$. Then*

$$\|v - \widetilde{v}\| \leq B_c := \sum_i |c_i| b_i, \qquad \max\{0, \|\widetilde{v}\| - B_c\} \leq \|v\| \leq \|\widetilde{v}\| + B_c.$$

Given only the proxy vectors, coefficients, and individual norm bounds, the upper endpoint is worst-case sharp. It is no larger than the separate-error bound $\|\widetilde{v}\| + \sum_i a_i b_i + \sum_{i\in S} |w_i| b_i$.

Proof. Because $v = \widetilde{v} - \sum_i c_i \xi_i$, the triangle and reverse-triangle inequalities give the bounds. For $\widetilde{v} \neq 0$, choose $\xi_i = -\operatorname{sign}(c_i) b_i \widetilde{v}/\|\widetilde{v}\|$; then v has norm $\|\widetilde{v}\| + B_c$. For $\widetilde{v} = 0$, use any common unit direction. Thus the upper endpoint is attainable under the declared information. Finally $|a_i - \bar{w}_i| \leq a_i + |\bar{w}_i|$. Combining this deterministic result with the numerical audit bound in Proposition A.8, applied to the committed $\widetilde{v}$, proves Equation (13). □

Theorem A.14 (Multi-step recertification). *For $t = 0, \dots, T-1$, let $\theta_{t+1} = \theta_t - \eta_t g_t(\theta_t)$ and $\widehat{\theta}_{t+1} = \widehat{\theta}_t - \eta_t \widehat{g}_t(\widehat{\theta}_t)$ start from $\theta_0 = \widehat{\theta}_0$, with $\eta_t \geq 0$. Let g_t be β_t-Lipschitz and define $e_t = \left\| g_t(\widehat{\theta}_t) - \widehat{g}_t(\widehat{\theta}_t) \right\|$. If unopened audit block t gives $e_t^2 \leq U_t$ with deterministic risk δ_t, then with probability at least $1 - \sum_{t=0}^{T-1} \delta_t$,*

$$\left\| \theta_T - \widehat{\theta}_T \right\| \leq \sum_{t=0}^{T-1} \eta_t \sqrt{U_t} \prod_{s=t+1}^{T-1} (1 + \eta_s \beta_s).$$

Proof. With $\Delta_t = \left\| \theta_t - \widehat{\theta}_t \right\|$, add and subtract $g_t(\widehat{\theta}_t)$ to obtain $\Delta_{t+1} \leq (1 + \eta_t \beta_t)\Delta_t + \eta_t e_t$. On the joint audit event, $e_t \leq \sqrt{U_t}$ for every t. Unroll from $\Delta_0 = 0$, interpreting an empty product as one. □

The exact identity assumes additivity in the declared space. A fixed preconditioner can be absorbed into g_i. Fully adaptive Adam moments are nonlinear in the aggregate; they require an optimizer-aware additive feature or a stated remainder.

A.12 Behavior coverage and reconstruction are different objectives

For nonnegative priorities q_u and similarities $k(u, i)$, defining the empty-set maximum as zero, the facility-location score $C(S) = \sum_{u \in \mathcal{U}} q_u \max_{i \in S} k(u, i)$ is normalized, monotone, and submodular. For $A \subseteq B$ and $j \notin B$, the current maximum for B is no smaller, so the marginal gain of j is no larger. Greedy therefore achieves the classical $1 - 1/e$ approximation under a cardinality budget.

In contrast, for a fixed target vector z and candidate matrix Z, when coefficient feasible sets are nested under support inclusion (as for unconstrained least squares or NNLS with zero padding), explained reconstruction energy

$$F(S) = \|z\|^2 - \min_w \|z - Z_S w\|^2$$

is monotone but not generally submodular when columns are correlated. Without nested feasible sets, even monotonicity need not hold. Restricted eigenvalue conditions imply a positive submodularity ratio for unconstrained least squares. A claim for constrained or regularized reconstruction must be proved for that feasible set and objective; it does not follow merely from the unconstrained statement. The paper therefore makes no unconditional approximation claim for reconstruction-guided selection.

A.13 Exact design factors

For a one-sided 95 percent squared-norm upper endpoint, the multiplicative factors relative to observed $\|Qv\|^2$ are approximately $2.010, 1.594, 1.374, 1.244, 1.164, 1.112, 1.077$ for $m_a = 16, 32, 64, 128, 256, 512, 1024$. The corresponding norm factors are $1.418, 1.263, 1.172, 1.115, 1.079, 1.054, 1.038$. These are exact chi-square design calculations, not experimental outcomes.

B Additional experimental results

This appendix reports complementary reconstruction, calibration, downstream, and robustness diagnostics. The supplementary implementation note provides the exact configuration, solver settings, and reproduction commands.

B.1 Experimental setup

We use rank-8 low-rank adaptation (LoRA) (Hu et al., 2022), with a frozen random input factor and a zero-initialized trainable output factor at the final attention output projection. The corpus contains 64 GSM8K and 64 MBPP tasks, split equally by domain into 72 training, 16 validation, and 40 final-test tasks. Selection uses training information only, validation is reserved for local functional diagnostics, and final-test generation follows checkpoint completion. MBPP reference tests are excluded from prompts. "Untouched" denotes this workflow split, not exclusion from model pretraining.

Each training task has three generated responses whose rewards are averaged into one pool record. Features are exact gradients in the declared adapter space at the frozen anchor, and target-token count defines the selection cost. The same normalized deployed weights define the coreset residual and stochastic training distribution. All methods use 64 optimizer steps with matched accumulation and padded model positions. Confidence intervals use the crossed task–training-seed bootstrap, with the familywise correction of Holm (1979).

B.2 Certificate implementation and the meaning of "coverage"

The completed run uses a Gaussian visible sketch scaled by $1/\sqrt{m}$. Numerical, stable, and effective ranks are computed from the exact protected matrix. Its conservative Gaussian singular-value factors are

$$\underline{\sigma} = 1 - \sqrt{r/m} - \sqrt{2\log(2/\delta)/m}, \qquad \overline{\sigma} = 1 + \sqrt{r/m} + \sqrt{2\log(2/\delta)/m}.$$

If $\underline{\sigma} \le 0$, the cell is vacuous; if $m < r$, it also receives RANK-TOO-HIGH. Otherwise the upper norm bound is the sketch residual divided by $\underline{\sigma}$. A finite upper bound larger than the full-update norm is scientifically vacuous.

A certificate is operational when its finite upper bound both covers the exact residual and does not exceed the full-update norm. All 27 primary endpoints cover numerically, and six satisfy this additional tolerance criterion. Across the full 405-cell primary grid, all 324 finite endpoints cover their residuals; the remaining cells have no finite endpoint. These descriptive comparisons do not estimate repeated-sampling nominal coverage because the cells share data and vary in method, budget, and dimension. The supplementary note maps this definition to the recorded fields. The uniform theorem requires the declared residual family to lie in the embedded space; a numerical rank estimate alone does not establish this assumption.

B.3 Full primary and transfer selector results

Metric definitions. The "Exact res." entries are $\|v\| \,/\, \|g_{\mathcal{V}}\|$, whereas "Sketch res." entries are $\|Pv\| \,/\, \|Pg_{\mathcal{V}}\|$. Their denominators differ. Distortion is instead $|\|Pv\| \,/\, \|v\| - 1|$, computed from the unnormalized row-level norms; it is not the ratio of the two normalized table columns. The "Med. dist." column is a median over projection seeds, while the other columns are means. "Beh. cov." is the archived normalized behavior-coverage score and rare recall uses the recorded low-frequency failure tags. The released construction combines structural task tags with derived failure-mode, reference-loss-tertile, target-length-tertile, and base-reward tags.

Table 4: Primary 1.5B model, $m = 512$, 20 percent budget. Means over three projection seeds, except median distortion. Residual normalizations are defined in this section.

Method	Exact res.	Sketch res.	Med. dist.	Beh. cov.	Rare recall	Operational
Behavior	.4743	.4693	.0322	1.000	1.000	1.000
Random	.4901	.4657	.0370	.652	.667	1.000
Gradient	.5624	.5405	.0186	.843	1.000	.000
Joint + swaps	.5779	.5472	.0455	1.000	1.000	.000
Joint	.5876	.5646	.0235	1.000	1.000	.000
Joint inherited	.7695	.7692	.0161	1.000	1.000	.000
k-center	.7742	.7533	.0358	.890	1.000	.000
Influence	.7813	.7496	.0508	.845	1.000	.000
Loss-top	.8214	.7769	.0329	.867	1.000	.000

Behavior and random have the lowest observed residuals in the primary and transfer models, respectively. On the primary model, visible-sketch reweighting reduces the joint-subset exact residual by 23.6 percent relative to inherited weights; the joint method improves on influence-only OMP by 24.8 percent while preserving full registered behavior coverage. Residual-guided swaps provide a further 1.65 percent reduction. These comparisons separate the contributions of the selection objective, coefficient fitting, and local support refinement.

Table 5: Transfer 0.5B model under the same primary sketch dimension and budget.

Method	Exact res.	Sketch res.	Med. dist.	Beh. cov.	Rare recall	Operational
Random	.5242	.5158	.0336	.695	.500	.333
Behavior	.5390	.5489	.0312	1.000	1.000	.333
Joint + swaps	.6306	.5851	.1141	1.000	1.000	.000
Joint	.6341	.5973	.1059	1.000	1.000	.000
Gradient	.6355	.6288	.0577	.809	1.000	.000
k-center	.7888	.7737	.0535	.839	.000	.000
Influence	.8225	.8111	.0321	.906	.833	.000
Joint inherited	.8479	.8182	.0723	1.000	1.000	.000
Loss-top	.8673	.8441	.0508	.911	1.000	.000

B.4 Point-estimation fidelity across dimensions

Across all 405 primary cells, median absolute relative distortion decreases monotonically with sketch dimension:

m	64	128	256	512	1024
Median distortion	.242	.115	.060	.032	.025
90th percentile	.420	.239	.127	.074	.048

At the primary budget and $m = 512$, aggregating 27 method–projection cells gives median distortion .0340, 90th percentile .0594, median Spearman .95, numerical rank 72, and operational certificate fraction .2222. The 0.5B transfer has median distortion .0567, 90th percentile .1092, median Spearman .90, and operational fraction .0741. These results quantify accurate compressed ranking and the separate tolerance requirement for an operational uniform certificate.

B.5 Reweighting and residual refinement

Define visible-sketch optimism

$$O(S, w) = \frac{\|v(S, w)\| - \|Pv(S, w)\|}{\|v(S, w)\|}.$$

Positive values mean the sketch understates the exact residual. We compare three matched variants: the joint-selected subset with inherited weights; the same subset with sketch-optimized weights; and the residual-refined subset.

Table 6: Median optimism over three budgets and three projection seeds. This analysis is post hoc and is used as mechanism evidence, not as a predeclared hypothesis test.

Variant / m	64	128	256	512	1024
Inherited weights	.097	.060	.006	-.017	-.003
Sketch-optimized weights	.324	.140	.062	.020	.027
Residual-guided swaps	.320	.168	.073	.033	.026

Across all 45 matched cells, optimized reweighting increases optimism in 43 cells, with median paired increase .0499. Swaps increase optimism in 35 cells, median .0143. Because budgets, sketch dimensions, and projection seeds reuse the same gradient pools, the 45 signs are reported descriptively rather than treated as independent statistical replicates. The matched comparisons consistently expose the calibration effect of sketch-based coefficient fitting within this pipeline.

B.6 Controlled effective-rank boundary

The controlled study generates 192 trajectories in dimension 2048 with eight behavior modes and crosses rank $r \in \{8, 32, 64\}$, redundancy $\{.2, .85\}$, rare prevalence $\{.02, .12\}$, sketch dimension $\{16, 32, 64, 128, 256\}$, budget $\{.10, .20, .35\}$, five methods, and three seeds, for 2,700 cells. The

registered summary reports median distortion .0769 and operational fraction .7929 in well-embedded cells. Under-ranked cells have median distortion .9498 and rank/vacuity failure fraction 1.000. The maximum numerical one-step identity error is 1.05×10^{-7}. This is a synthetic geometric stress test, separate from the real Qwen/GSM8K/MBPP evidence. The rank-too-high diagnostic is partly imposed by the rule $m < r$; its 100 percent frequency is not a measurement of nominal undercoverage. The large observed distortion is a separate result.

B.7 Parameter-update and functional diagnostics

The validation-only study applies exact full-pool and selected adapter updates at step sizes $10^{-4}, 10^{-3}, 3 \times 10^{-3}$ on 16 tasks. Across all registered cells, the maximum relative discrepancy in $\left\|\widehat{\theta}^{+} - \theta^{+}\right\| = \eta \left\|\widehat{v}\right\|$ is 8.99×10^{-8}. At $\eta = 10^{-3}$, parameter deviations range from 2.47×10^{-5} to 4.28×10^{-5}; every stored target-log-probability and negative log-likelihood (NLL) delta is exactly zero at that primary step.

At $\eta = 3 \times 10^{-3}$, eight of nine methods have at least one nonzero aggregate functional statistic, but the largest token-weighted target-log-probability root-mean-square (RMS) is only 9.67×10^{-7} and the largest absolute NLL difference is 5.54×10^{-7}. The corresponding parameter deviations reach 1.28×10^{-4}. The only defined across-method Spearman correlation is .567 at this largest step, below the registered .65 threshold; every log–log slope is undefined because each method has at most one nonzero step. The adapter difference was accumulated in FP32 but evaluated through a bfloat16 (BF16) base-model path, so the function-space statistic remains at the observed numerical floor. This is neither evidence of functional equivalence nor evidence for a Lipschitz prediction. The empirical certification result therefore concerns the parameter-update space.

B.8 Matched-compute final evaluation

The base model is not trained: its three seed labels index generation randomness only. For trained methods they index separately trained runs as well as evaluation generation. The released statistics implementation performs the crossed task–training-seed bootstrap with the frozen seed and 10,000 replicates, and applies the reported Holm adjustment. The intervals, point estimates, raw paired values, and corrected comparisons can be regenerated from the released rows.

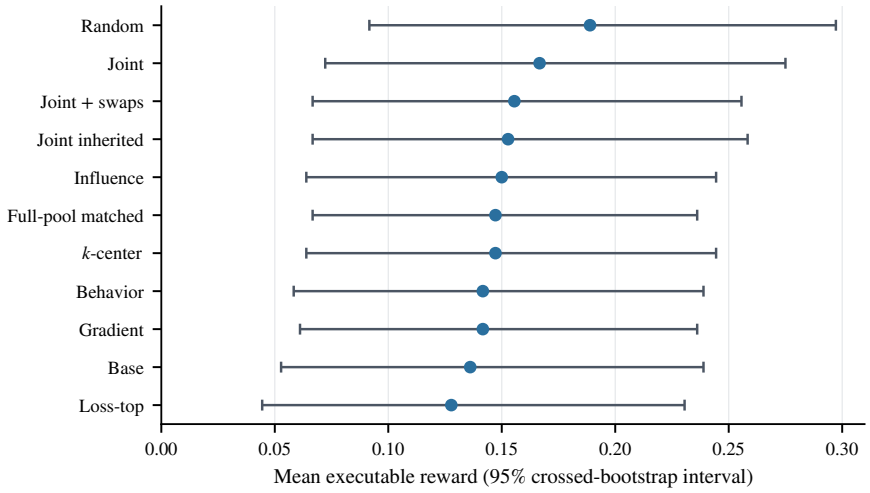

Figure 3: Mean executable reward and 95 percent crossed task–seed bootstrap intervals on the 40 untouched tasks. None of the ten registered joint-versus-reference comparisons is significant after Holm correction.

Joint versus random is -0.0222 with 95 percent interval $[-0.0722, 0.0250]$, raw $p = .329$, Holm $p = 1.0$. Joint versus base is 0.0306 with interval $[-0.0500, 0.1083]$, raw $p = .448$, Holm $p = 1.0$. All ten corrected comparisons have Holm $p = 1.0$. The test is small and code rewards are near floor,

Table 7: Untouched-test mean reward, 95 percent crossed bootstrap interval, and domain means.

Method	Mean	95% interval	Math	Code
Random	.1889	[.0917,.2972]	.3333	.0444
Joint	.1667	[.0722,.2750]	.2944	.0389
Joint + swaps	.1556	[.0667,.2556]	.2611	.0500
Joint inherited	.1528	[.0667,.2583]	.2556	.0500
Influence	.1500	[.0639,.2444]	.2500	.0500
Full-pool matched	.1472	[.0667,.2361]	.2500	.0444
k-center	.1472	[.0639,.2444]	.2278	.0667
Behavior	.1417	[.0583,.2389]	.2389	.0444
Gradient	.1417	[.0611,.2361]	.2444	.0389
Base	.1361	[.0528,.2389]	.2278	.0444
Loss-top	.1278	[.0444,.2306]	.2278	.0278

so these results neither establish superiority nor equivalence. The downstream comparison is reported separately from the reconstruction gains.

Table 8: All ten registered joint-minus-reference comparisons. Effects and intervals are reward differences; p-values are copied from the completed crossed task–seed bootstrap analysis. All Holm-adjusted p-values equal 1.0000.

Reference	Difference	95% interval	Raw p	Holm p
Base	0.0306	$[-0.0500, 0.1083]$	0.4478	1.0000
Random	-0.0222	$[-0.0722, 0.0250]$	0.3294	1.0000
Loss-top	0.0389	$[-0.0167, 0.1000]$	0.1742	1.0000
k-center	0.0194	$[-0.0333, 0.0750]$	0.4584	1.0000
Behavior	0.0250	$[-0.0361, 0.0917]$	0.4318	1.0000
Gradient	0.0250	$[-0.0361, 0.1000]$	0.4640	1.0000
Influence	0.0167	$[-0.0444, 0.0694]$	0.4984	1.0000
Joint inherited	0.0139	$[-0.0694, 0.0917]$	0.7163	1.0000
Joint + swaps	0.0111	$[-0.0250, 0.0472]$	0.5311	1.0000
Full-pool matched	0.0194	$[-0.0333, 0.0722]$	0.4030	1.0000

B.9 Anchor shift and contamination

Let G' denote the gradient matrix at the shifted anchor and define $v'(S, w) = G'a - G'_S w$. Write (S_0, w_0) for the subset and weights frozen at the original anchor and (S_1, w_1) for those recomputed at the shifted anchor. The stale/fresh residual ratio is

$$I_{\text{stale/fresh}} = \frac{\|v'(S_0, w_0)\|}{\|v'(S_1, w_1)\|}.$$

Both residuals use the shifted features and the same target weights a; the denominator is nonzero in the reported comparisons. Thus this ratio compares reuse against reselection at the shifted anchor, not residual norms at different anchors.

The registered anchor perturbation contains five methods. Joint selection Jaccard is .8519 and stale/fresh residual ratio .9994; behavior selects the same subset with ratio 1.0020. Influence changes more (Jaccard .8056) but has ratio .9395. The joint condition passes the registered inflation threshold 1.5. Since there is one shift signature, this supports stability only for that perturbation, not across training.

The contamination study contains 60 cells over five methods and three seeds. Slopes vary materially by seed and method; selected-contaminated fractions rise with injection in most cells, while behavior and rare-recall metrics often remain flat. We report it as descriptive stress evidence only. The resulting scope is the specified injection protocol.

Algorithm 1 Sealed audit with explicit numerical and leakage risk

Require: sealed table $\{q_i = Qg_i\}$, aggregate $q_{\mathcal{V}}$, unopened block

function COMMIT($H, S, w, \tau, \rho, \kappa, \delta_{\mathrm{tot}}$, analysis-spec)
  verify canonical hash, precision contract, and absence of prior reveal
  require $\tau > 0$, $\rho \geq 0$, $0 \leq \kappa < \delta_{\mathrm{tot}} < 1$; set $\delta = \delta_{\mathrm{tot}} - \kappa$
  store immutable commitment and justified error bounds; return audit id j
function REVEAL(j)
  compute $\widetilde{z} = q_{\mathcal{V}} - \sum_{i \in S} w_i q_i$ with $\|\widetilde{z} - Qv(S, w)\| \leq \rho$
  set $r_- = \max\{0, \|\widetilde{z}\| - \rho\}$ and $r_+ = \|\widetilde{z}\| + \rho$
  set $I = [m_a r_-^2 / q_{m_a, 1-\delta/2},\ m_a r_+^2 / q_{m_a, \delta/2}]$
  set $U = m_a r_+^2 / q_{m_a, \delta}$ and $D = \mathbf{1}\{U \leq \tau^2\}$
  mark block consumed; return I, U, D, risk δ_{tot}, diagnostics, signed manifest

C AUDIT PROTOCOL AND REPRODUCIBILITY

C.1 MINIMAL SEALED-AUDIT SERVICE

The guarantee depends on an information boundary rather than on a particular provider.

Feature service. Compute each declared update feature once and stream it to two projection workers. The selection worker exposes only Pg_i and permitted metadata. The audit worker stores Qg_i and $Qg_{\mathcal{V}}$ but returns nothing before commitment. Exact features may be visible to selection without violating the independent-audit theorem, provided they carry no information about Q. If the evaluator restricts feature access to assess a compressed-only selector, that is an experimental information-matching rule, not a requirement for audit independence. Computing $Q\widehat{v}$ only after commitment is an alternative to the sealed table.

Commitment record. Before reveal, serialize and hash: dataset and split hashes; immutable model/tokenizer revisions; anchor checkpoint; update representation and precision; selection-sketch seed and dimension; selector code/configuration hash; selected indices; raw and deployed weights; token and compute budgets; tolerance τ; confidence allocation; the complete finite candidate list, if any; primary metrics; and the reporting path; denote the resulting hash by H. The statistical commitment fixes the residual and the scope of claims before audit information is used. A new residual chosen using that information needs an unopened audit or valid simultaneous control. An unchanged valid interval can be compared with several tolerances; harmless metadata edits do not create a new statistical test.

Reveal. The service verifies the hash and computes the committed projection, returning an approximate vector $\widetilde{z}$ with a certified error radius ρ. It uses the inflated endpoint in Proposition A.8 for the tolerance decision, reports the total risk $\delta + \kappa$, and marks the block consumed. An unverified numerical or leakage bound must not be silently set to zero. It does not reveal Q, per-item audit vectors, uncommitted alternatives, or unopened blocks.

Sequential use. Generate a finite or countable bank of independent blocks $Q^{(1)}, Q^{(2)}, \ldots$ in advance. Block t is opened only after the current residual and risk δ_t are fixed, with a predeclared summable alpha-spending rule. The manifest records the stopping rule, alpha spending, consumed rows, invalid repeated queries, and unused blocks.

Threats to independence. Reusing selection randomness as audit randomness, previewing audit norms, or exposing per-item audit projections before commitment can destroy independence. Uncorrected post-audit selection of cells or residuals also invalidates the intended guarantee. Exact residual queries alone do not violate independence from a genuinely fresh Q; they may instead make a compressed certificate unnecessary. A quantitative total-variation correction is used only when a defensible κ is available; otherwise the audit is declared compromised.

The interval and decision each obey their individual total-risk guarantee in Proposition A.8; joint coverage of every returned claim requires the corresponding simultaneous control. The supplementary implementation includes the corresponding endpoint arithmetic.

C.2 Experimental controls

Model revisions, adapter features, precision, seeds, data splits, and selection budgets are fixed before final evaluation. Selected weights are checked for normalization and used consistently in residual computation and training. Comparisons share optimizer steps, accumulation, padded model positions, and evaluation budgets. Repeated generations are averaged within task–seed cells before crossed-bootstrap inference. The supplementary records retain all evaluated cells and their diagnostic codes.

C.3 Supplementary materials

The anonymous supplement contains the experiment implementation, bundled task subsets and fixed splits, complete result tables, and manuscript-facing analysis. Its implementation note gives exact configurations, selector details, metric definitions, and a map from reported quantities to the underlying records. Integrity manifests support checks of the distributed files.

C.4 Reproduction and numerical verification

The experiment pipeline covers response generation, exact adapter-gradient extraction, selection and reweighting, matched-compute training, final evaluation, and statistical aggregation. A separate analysis pipeline recomputes the reported descriptive contrasts and empirical figures from the result rows. Numerical tests check the affine-law calculations and audit endpoints; these calculations are distinct from the real-model experiments. Setup, launch, checkpoint/resume, and verification instructions are provided in the supplementary documentation.

C.5 Data, licensing, and safety

The completed study uses public GSM8K and MBPP subsets and open Qwen model weights under their respective licenses. The released evaluator parses candidate code, rejects unsafe syntax, and executes permitted code in a separate isolated Python process with time and resource limits; MBPP reference tests are withheld from model prompts. No private user data or human-subject data are used. Generated responses may contain unsafe, biased, or memorized content; any public release should preserve provenance and dataset-card documentation. A certificate of update fidelity is not a certificate of safety, fairness, or correctness.