SCREENING IS NOT RANKING: RESOLUTION-AWARE MODEL SELECTION UNDER COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fidelity metrics are one-sided certificates that a model was not badly damaged; they don't necessarily rank candidates that all pass the screen. We reframe compressed-model selection as a decision problem whose quality is *selection regret* (the capability gap to the best available) and identify four structural limits: whether the fidelity bound orders candidates (proxy limit), whether the pool confounds rate with design quality (pool limit), whether the budget resolves the capability gap (target limit), and whether a scalar target is appropriate (target-definition limit). We prove one formal result per limit and test all four on pre-registered compressed-model pools. On both pre-registered pools, fidelity screens broken candidates reliably (AUROC 0.97–1.00), but among viable Qwen2.5-3B candidates, nine metrics induce the same ranking because they primarily track compression rate rather than design quality. The empirically estimated resolution bound is not violated in any of $11{,}734$ pair-by-budget checks; resolvable pairs invert 0.005% of the time against 22.8% for unresolvable ones. Screening before capability evaluation cuts regret by up to an order of magnitude. Within resolvable pairs, teacher KL orders 92.5–100% of comparisons correctly, so the ranking failure we document is local: at fixed bytes, restoring row norms improves capability ($+0.0089$) while every proxy ranks it the wrong way. The reversal replicates on an independent architecture (SmolLM2-1.7B) but not on a fresh-Qwen redraw, whose training-seed variance is comparable to the effect size; broader screening and resolution findings are also supported in vision. Our prospective evidence is output-head and lexical-interface compression. We propose a protocol: screen with fidelity, control for rate and family, evaluate survivors, and report "unresolved at this budget" when that is the honest answer.

1 INTRODUCTION

A compression run produces a pool of candidates – varying bit widths, ranks, codebooks, and training recipes – and someone has to pick one to deploy (Zhu et al., 2024). Because full downstream evaluation is expensive (Polo et al., 2024), the pick is usually made with a cheap fidelity measurement: reconstruction error (Hsu et al., 2022; Frantar et al., 2022), perplexity (Jin et al., 2024), teacher KL (Hinton et al., 2015), or output agreement. The standard justification is that these metrics correlate with downstream accuracy (Jin et al., 2024; Jaiswal et al., 2024), and this correlation is often used as evidence that better fidelity identifies better candidates.

We argue that correlation is the wrong object: the practitioner wants the best candidate, not an accuracy estimate. What matters is *selection regret*: the capability lost by deploying the metric's choice instead of the true best (Audibert & Bubeck, 2010). Correlation and regret come apart in both directions: a metric can correlate strongly and still pick badly, or correlate weakly simply because the benchmark cannot tell the candidates apart (a fact about the budget, not the metric) (Card et al., 2020).

Fidelity is a one-sided certificate: "small error implies small damage" is exactly what is needed to reject broken candidates, but it has no converse and says nothing about which intact candidate is better. This leads to three regimes a pooled correlation conflates (Figure 1): **screening** broken candidates, where fidelity is reliable; **coarse ranking** by compression rate, where fidelity looks predictive but merely recovers bytes spent; and **local decisions** within a fixed budget, where certificates stop ordering and evaluation becomes hard.

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.

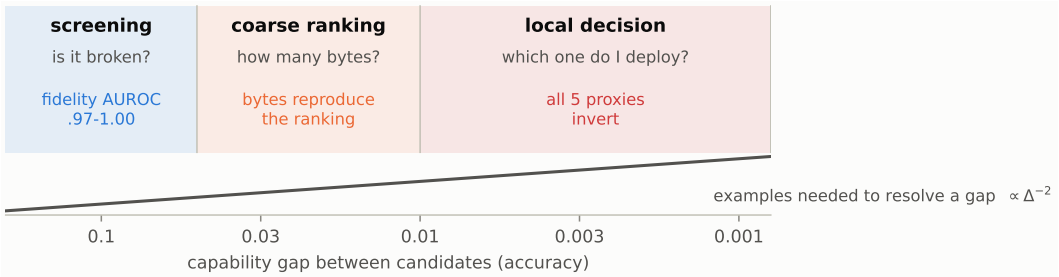

Figure 1: **Screening is not ranking.** The three regimes a practitioner faces when selecting among compressed candidates. As capability gaps shrink, the number of evaluation examples needed to resolve them grows like Δ^{-2}, while fidelity's one-sided geometry means it can reject broken candidates but cannot order intact ones (Audibert & Bubeck, 2010). Each band is annotated with a measured result from our data.

Across these regimes, we identify four structural limits on what a proxy ranking can tell us: pool (rate confounds design quality), proxy (the bound is one-sided), target (finite budget may not resolve the gap), and target-definition (no scalar fits all deployments). We prove one formal result per limit. For the pool, pooled Spearman splits exactly into a between-rate term and a within-rate design term. For the proxy, we construct a case where the smaller error produces the larger distribution damage. For the target, we derive a regret bound and a matching rank-agreement bound that makes reliability a property of the pool, not the benchmark. For target-definition, we show that Pareto-incomparable candidates admit no universal scalar proxy.

We test all four on compressed-model pools. Fidelity screening before capability evaluation cuts regret by up to an order of magnitude; the resolution index separates resolvable from unresolvable comparisons; and a byte-free training change improves capability while all five measured proxies rank it the wrong way – a reversal confirmed in a pre-registered SmolLM2 replication, with the broader screening and resolution findings also supported in vision. We close with a five-step protocol for compressed-model selection.

2 Compressed-model selection as a decision problem

We formalize the selection task as follows. Given a candidate pool $\mathcal{C} = \{c_1, \ldots, c_K\}$, let $M(c)$ denote a cheap proxy loss and $A(c)$ downstream capability. Because $A(c)$ cannot be observed exactly, we estimate it from n evaluation examples as $\hat{A}_n(c)$. A selector is any map from measurements to a candidate; we write $\hat{c}$ for the candidate it selects. Its quality is its regret $\mathrm{Reg}(\hat{c}) = A(c^\star) - A(\hat{c})$, where $c^\star = \arg\max_{c \in \mathcal{C}} A(c)$ is the true best. Two selectors matter throughout: the *proxy selector* $\arg\min_c M(c)$, which costs no benchmark examples, and the *capability selector* $\arg\max_c \widehat{A}_n(c)$, which costs Kn scored examples. A claim about a metric is always conditional on the pool, the evaluation suite, and the budget (changing any of these can change the verdict without touching the metric, and Section 4 shows all three happening).

We keep five diagnostics separate: screening AUROC for capability drops exceeding 0.05; pooled and within-rate/family Spearman; split-half target reliability (Spearman–Brown corrected); honest regret with disjoint selection and scoring splits; and target-definition via Pareto incomparability and deployment-weighted winners. Decision budgets are scored documents; Section 4.3 reports CPU selector cost separately.

We draw evidence from four candidate pools, labeled by epistemic role: P1 and P2 are pre-registered, R is retrospective, and L is a controlled local experiment. **P1** (48 candidates on Qwen3.5-0.8B) and **P2** (22 on Qwen2.5-3B) are pre-registered: all fidelity is measured before any capability number exists. P1 spans round-to-nearest quantization (Frantar et al., 2022), truncated SVD (Eckart & Young, 1936), product quantization (Jegou et al., 2010), sparse PCA (Zou et al., 2006), and magnitude pruning (Han et al., 2015); P2 concentrates on high-rate candidates and collapses to a single quantization ladder after viability filtering, which lets us test whether apparent proxy skill is explained by compression rate. **R** (47 trained interfaces on Qwen2.5-0.5B, plus bit allocators, GPTQ (Frantar et al., 2022),

AWQ (Lin et al., 2025), and compositional codes (Egiazarian et al., 2024)) is retrospective but retains per-document scores on 7,767 development and 5,321 frozen documents, making budget curves possible. $\mathbf{L}$ is a byte-matched sparse-interface design matrix at 29.36–29.73 MB in which representation, recipe, and served bytes are all fixed, and the baseline and norm-restoration arms share eight training-seed IDs.

We measure capability as teacher-forced multiple-choice accuracy, macro-averaged over nine tasks (ARC, HellaSwag, OpenBookQA, WinoGrande, BoolQ, MMLU, RACE, CommonsenseQA, and COPA (Clark et al., 2018; Zellers et al., 2019; Mihaylov et al., 2018; Sakaguchi et al., 2021; Clark et al., 2019; Hendrycks et al., 2020; Lai et al., 2017; Talmor et al., 2019; Roemmele et al., 2011)) on a development suite and a suite frozen before any capability score is seen. The fidelity battery spans matrix error, activation-weighted logit error, perplexity, teacher KL, and top-1 agreement. The primary language-model pools use Qwen models (Yang et al., 2025); we additionally report cross-architecture language-model and vision replications, discussed in Section 5.1. Full protocols, task suites, byte accounting, and statistics are in Appendix A; pre-registrations are in Appendix C.

3 Four limits on compressed-model selection

We identify four structural limits. The *proxy limit* (Section 3.1): fidelity bounds damage from one side only, so it does not in general guarantee a correct ranking. The *pool limit* (Section 3.2): a pooled correlation conflates within-design signal with between-rate confounding. The *target limit* (Section 3.3): a finite budget resolves only pairs whose capability gap exceeds a threshold set by the pool geometry. The *target-definition limit* (Section 3.4): no scalar proxy fits all deployment weightings when candidates are Pareto-incomparable. Proofs are in Appendix A.2.

3.1 Proxy limit: a certificate is not an ordering

We ask whether a fidelity metric on the output embedding table can rank two compressed models by their downstream behavior. The answer is no in general, because the map from representation error to distribution damage is one-sided.

For an output table E, hidden state h, output distribution $p_E = \mathrm{softmax}(Eh)$, and perturbation $E \mapsto E + \Delta$, write $\mathbf{C}_h = \mathbb{E}[hh^\top]$ and $Q(\Delta) = \frac{1}{4}\,\mathrm{Tr}(\Delta \mathbf{C}_h \Delta^\top)$ (exact for output-only perturbations at fixed hidden states; direct output path of a tied table only).

Theorem 1 (One-sided fidelity hierarchy)**.** *For categorical softmax distributions,*

$$\mathbb{E}_h D_{\mathrm{KL}}(p_E \| p_{E+\Delta}) \;\le\; Q(\Delta) \;\le\; \tfrac{\lambda_{\max}(\mathbf{C}_h)}{4}\|\Delta\|_F^2. \tag{1}$$

Neither inequality has a converse of the form needed to rank candidates: the first admits no reverse bound even at fixed full-rank $\mathbf{C}_h$, and the second admits no covariance-independent reverse bound.

The failure is constructive: we exhibit two explicit perturbations where lower fidelity error produces strictly greater distribution damage, one exploiting a gauge degree of freedom and one exploiting the anisotropy of $\mathbf{C}_h$.

Corollary 1 (Two explicit reversals)**.** (i) Gauge. *A common-logit shift $\Delta_g = \mathbf{1}a^\top$ has $D_{\mathrm{KL}}(\Delta_g) = 0$ and $Q(\Delta_g) > 0$, so for any sufficiently small non-gauge Δ_n we have $Q(\Delta_n) < Q(\Delta_g)$ while $D_{\mathrm{KL}}(\Delta_n) > D_{\mathrm{KL}}(\Delta_g) = 0$: lower fidelity error, strictly greater distribution damage, even for activation-weighted error.* (ii) Anisotropy. *If $\mathbf{C}_h q_i = \lambda_i q_i$ then $\Delta_i = \sqrt{c}\,u q_i^\top$ satisfies $\|\Delta_i\|_F^2 = c$ and $4Q(\Delta_i) = c\lambda_i$ for any unit vocabulary direction u, so equal-MSE costs span $\kappa(\mathbf{C}_h) = \lambda_{\max}(\mathbf{C}_h)/\lambda_{\min}(\mathbf{C}_h)$ (the condition number of $\mathbf{C}_h$) and Frobenius and activation-weighted error can order two candidates oppositely.*

Neither case is exotic: on one real perturbation we find the gauge component carries 14.9% of the table's energy while moving log probabilities by at most 7.6×10^{-6}, and we measure $\kappa(\mathbf{C}_h) = 1.1 \times 10^8$ at 0.5B. We test case (ii) prospectively in Section 4.5. These are possibility results, not claims about every pool. A tied interface adds a second, unmeasured path,

$$\delta(Eh_E) = \underbrace{\Delta h_E}_{\text{direct output path}} + \underbrace{E\,Dh_E[\Delta]}_{\text{input/state path}} + O(\|\Delta\|^2), \tag{2}$$

which we audit in Section 4.5.

3.2 POOL LIMIT: A POOLED CORRELATION MIXES TWO SIGNALS

We ask whether a high pooled rank correlation between a proxy and capability is evidence that the proxy can identify the better design at a fixed byte budget. The answer is no in general, because the correlation conflates two unrelated signals: one that reflects which design is better within a rate stratum, and one that merely reflects how many bytes were spent.

Let U and V be the global rank transforms of a proxy (oriented so that higher ranks better) and of capability, and let S index a predefined rate/family stratum. Spearman correlation is Pearson correlation of U and V.

Proposition 1 (Within- and between-stratum association)**.**

$$\mathrm{Cov}(U,V) = \underbrace{\mathbb{E}_S[\mathrm{Cov}(U,V \mid S)]}_{\textit{within-stratum design signal}} + \underbrace{\mathrm{Cov}_S(\mathbb{E}[U \mid S], \mathbb{E}[V \mid S])}_{\textit{between-stratum rate/family signal}}. \tag{3}$$

Consequently pooled rank association can approach one while the proxy carries no design information inside any stratum.

This is the law of total covariance after the rank transform; we claim no novelty for the identity, only that it identifies which term compression papers typically report as proxy evidence and which they do not. The between-stratum term is the right thing to measure when the decision is *how many bytes to spend*, but it is not evidence about which design to prefer at fixed bytes.

3.3 TARGET LIMIT: REGRET, RANKINGS, AND WHAT A BUDGET CAN RESOLVE

We ask how many evaluation examples are needed to reliably identify the best candidate, and what happens to selection quality when that number is not available. The answer is that resolution is a property of the candidate pool's gap geometry, not of the benchmark alone.

Fix two trained candidates i and j. Let X_t be their paired per-example capability difference with $\mathbb{E}X_t = \Delta_{ij} > 0$, and assume $X_t - \Delta_{ij}$ is σ_{ij}^2-sub-Gaussian; bounded per-example scores satisfy this. We define the *resolution index*

$$\nu_{ij}(n) = \frac{n\Delta_{ij}^2}{2\sigma_{ij}^2}, \qquad \text{so that} \qquad \Pr\left[\widehat{A}_n(i) \le \widehat{A}_n(j)\right] \le e^{-\nu_{ij}(n)}. \tag{4}$$

All results in this subsection follow from (4). The concentration step is standard (Hoeffding, 1963; Albert, 2019); the point is that the decision-level quantities inherit its shape and are then directly measurable.

Theorem 2 (Selection regret of a capability run)**.** *Let $\hat{c}_n = \arg\max_c \widehat{A}_n(c)$ over K fixed candidates with unique best $c^\star$ and gaps $\Delta_i = A(c^\star) - A(c_i) > 0$. Write $\nu_i(n) = \nu_{\star i}(n)$ for the pairwise resolution index of* (4) *applied to the $c^\star$-vs-c_i pair. Then $\Pr(\hat{c}_n = c_i) \le e^{-\nu_i(n)}$ and*

$$\mathbb{E}[\mathrm{Reg}(\hat{c}_n)] \;\le\; \sum_{i \neq \star} \Delta_i \, e^{-\nu_i(n)}, \qquad \Pr[\mathrm{Reg}(\hat{c}_n) \ge \epsilon] \;\le \sum_{i:\, \Delta_i \ge \epsilon} e^{-\nu_i(n)}. \tag{5}$$

The bound reflects the structure of the problem: distant competitors contribute a large Δ_i with exponentially small error probability, and near-ties the reverse, so adding an obviously weak candidate to a pool barely changes expected regret.

Corollary 2 (Ranking resolution is a property of the pool)**.** *Let τ_n be Kendall's τ between the budget-n ranking and the true ranking, with no true ties and estimated ties counted as discordant. Then*

$$\mathbb{E}[\tau_n] \;\ge\; 1 - \frac{4}{K(K-1)} \sum_{i<j} e^{-\nu_{ij}(n)}. \tag{6}$$

Equation (6) makes precise what a single reliability number cannot express: a benchmark has no fixed reliability, and what it recovers depends on the *gap geometry of the candidate pool*. The same suite can rank catastrophic failures perfectly and be useless on a byte-matched pool without changing at all. We therefore report, beside every correlation, the *resolvable-pair fraction*

$$\mathrm{RPF}(n,\delta) = \binom{K}{2}^{-1} \left|\{(i,j) : \nu_{ij}(n) \ge \log(1/\delta)\}\right|, \tag{7}$$

the share of comparisons the budget can settle at confidence $1-\delta$, where $\delta \in (0,1)$ is the tolerated inversion probability.

Because our capability measure is a task macro-average, we replace σ^2/n with an effective variance $\sigma^2_{\text{eff}} = \sum_t \sigma_t^2/(T^2 n_t)$ that accounts for task-level heterogeneity; this reduces to (4) for equally sized tasks (Appendix A.2).

Fidelity remains valuable even when it cannot rank, because a screen does not have to identify the winner; it only needs to shrink K without discarding the best candidate.

Proposition 2 (Screen, then select)**.** *Let a screen that does not depend on the evaluation examples return a survivor set S with $|S| = m$, and define its screening regret $\epsilon_{\text{screen}} = A(c^\star) - \max_{c\in S} A(c)$. Evaluate each survivor on n examples with σ-sub-Gaussian estimates and let $\hat{c} = \arg\max_{c\in S} \widehat{A}_n(c)$. Then with probability at least $1-\delta$,*

$$A(c^\star) - A(\hat{c}) \;\le\; \epsilon_{\text{screen}} + 2\sigma\sqrt{\frac{2\log(2m/\delta)}{n}}. \tag{8}$$

At a fixed total budget $B = mn$ the second term grows like $\sqrt{m \log m / B}$ while the first is non-increasing in m: the bound favors screening when a one-sided certificate shrinks m enough to offset any loss of the optimum. Section 4.3 measures both terms and finds a regime where each wins.

3.4 Target-definition limit: no scalar proxy fits every deployment

We ask whether, even with unlimited evaluation, a scalar proxy can identify the best candidate. The answer is no when candidates differ across tasks, because "best" then depends on the deployment weighting; a scalar proxy cannot be right for all weightings at once.

Let $a(c) = (A_1(c), \ldots, A_T(c))$ be the vector of per-task capabilities and $w \ge 0$ a deployment weighting, so that $A_w(c) = w^\top a(c)$ is the deployed scalar target.

Proposition 3 (Pareto obstruction)**.** *If $a(c_i)$ and $a(c_j)$ are Pareto-incomparable, there exist weightings w, w' with $A_w(c_i) > A_w(c_j)$ and $A_{w'}(c_i) < A_{w'}(c_j)$. Hence a task-agnostic scalar proxy induces a universally correct ordering only if the candidates' capability vectors are totally ordered by Pareto dominance.*

The proof is immediate. Whether it matters is an empirical question: we find that 72–93% of candidate pairs in our viable pools are Pareto-incomparable (Section 4.5), which means "the best compressed model" has no answer until the deployment weighting is specified.

4 Results

We report results in the order the four limits predict: the proxy limit (Section 4.1, Section 4.4), the pool limit (Section 4.1), the target limit (Section 4.2), and the target-definition limit (Section 4.5). Section 4.3 quantifies the joint payoff of the proxy and target limits.

4.1 Fidelity screens well, and a pooled ranking can be a rate ladder

On both pre-registered pools we find that every principal fidelity family separates broken from viable candidates: teacher KL and perplexity reach AUROC 0.986–1.000, reconstruction and activation-weighted error 0.970–1.000 (full intervals and threshold sweeps in Appendix B). Large behavioral damage leaves a strong fidelity signature, and we make no argument against using fidelity for this job. One exception marks the boundary: on the retrospective pool perplexity has AUROC 0.333, because six broken candidates were distilled on the same WikiText corpus on which perplexity is measured, improving PPL from 20.17 to 16.4–16.6. With nine broken candidates the interval includes chance, so we claim no systematic reversal; corpus fit is not a safety certificate when the candidates were fitted on that corpus. A 15-candidate Qwen2.5-1.5B scope check gives AUROC 1.000 for all seven proxies; screening to $m = 3$ at 1,590 candidate-documents reduces regret from 0.0049 to 0.0016, while the five viable 4-bit models remain unresolved (RPF = 0.30; Appendix B.4). Ranking is a different job. We find that high correlation among viable candidates is not by itself evidence of

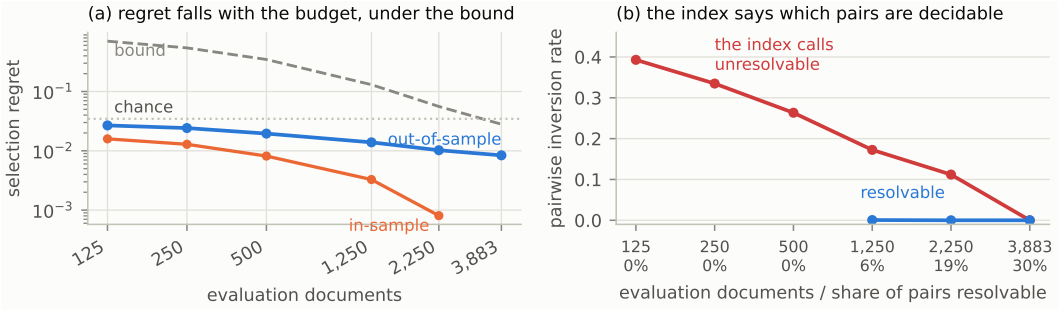

Figure 2: **Target resolution, measured.** (a) Selection regret against evaluation budget on the 47-candidate pool, with the bound of Theorem 2; the out-of-sample curve is scored on a disjoint half of the same suite. (b) Pairs the resolution index calls resolvable essentially never invert, while pairs it calls unresolvable invert about a fifth of the time. *Takeaway:* more evaluation documents reliably reduce regret, but only down to the floor set by the pool's gap geometry.

design-level selection: in P2 every viable candidate lies on one round-to-nearest rate ladder, so deployed bytes, reconstruction MSE, $\mathbf{C}_h$-weighted error, teacher KL, and five further metrics induce *the same* development ranking, $\rho = 0.828$. They are recovering compression rate. That is a useful thing to recover when the decision is how many bytes to spend, and it is the between-stratum term of Proposition 1, not the within-stratum one.

The within-stratum term can be near zero at the same time. On the largest homogeneous P1 stratum (round-to-nearest, $n = 18$), every fidelity metric sits near $\rho = -0.065$ while the pooled teacher-KL correlation is $+0.467$. The stratum's remaining capability spread is 0.034 and the suite's Spearman–Brown ranking reliability on those 18 candidates is only $+0.309$: pool structure and target precision are entangled in the raw number, and we need more evaluation documents to separate them. Across $4{,}000$ random viable subsets, teacher-KL ρ ranges from -0.619 to 1.000, all four high/low-ρ and high/low-regret regimes populated (Appendix B.5). Post-hoc SmolLM2-1.7B and Llama-3.2-1B/3B pools screen reliably with no within-stratum sign reversal; pooled associations remain predominantly between-rate effects (Appendix E.4).

4.2 The evaluation budget decides what is decidable

Per-document scores let us vary only the budget while holding everything else fixed. Documents are split into draw and reference halves; budgets are subsampled from the draw half (400 draws each) and scored against the reference half (Appendix A.7).

We plot the prediction of Theorem 2 against measured regret in Figure 2a: both fall with budget and the bound holds with slack. The sharper empirical test is pairwise: using the plug-in variance estimate, the predicted relation, $\Pr[\text{inversion}] \le e^{-\nu_{ij}}$ is violated 0 **times in** $11{,}734$ **pair-by-budget checks**, and it separates what it should: comparisons the index calls resolvable at $\delta = 0.05$ invert 0.005% of the time against 22.8% for those it calls unresolvable (Figure 2b). Empirical-Bernstein and distribution-free Hoeffding variants also yield zero violations, although Hoeffding certifies only 2 of 11,734 pair-budgets (Table 19).

Out-of-sample regret on the retrospective development suite falls from 0.0268 at 125 documents to 0.0084 at $3{,}883$ against a chance regret of 0.0347, reliability rises $0.046 \to 0.707$, and rank agreement with the reference half rises $0.238 \to 0.592$; the prospective 0.8B pool behaves the same way ($0.0106 \to 0.0032$). A single-budget study would have reported any one point on that curve as a property of the metrics.

We find out-of-sample regret is monotone in only 3 of the 6 curves; the exceptions are the pools with the smallest viable spread (the 3B ladder's viable candidates span 0.023 in accuracy) where more documents recover the ranking of the half they were drawn from without improving choice on a disjoint half. Even at the largest budget its resolvable-pair fraction is 0.152. That is "unresolved at this budget" operationally, and by Corollary 2 it is a property of the pool's gap geometry rather than a defect of the metrics.

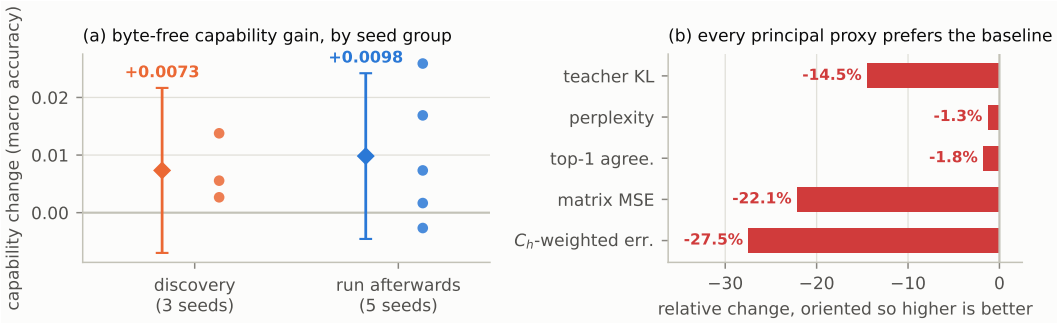

Figure 3: **A byte-free improvement that every proxy ranks below the baseline.** (a) Paired seed deltas, with the three seeds available when the arm was chosen separated from the five run afterwards; diamonds are group means with 95% paired-t intervals. (b) All five fidelity changes, oriented so that positive means "better". *Takeaway:* fidelity and capability can point in opposite directions even within a fixed byte budget.

4.3 What screening is worth: a fixed-cost comparison

Proposition 2 exposes the tradeoff between shrinking the survivor set and discarding the optimum, and both terms are measurable: we fix a total budget of candidate-documents, screen to the top m candidates by teacher KL, split the budget equally over the survivors, and score the choice on the disjoint reference half. The screen costs teacher forwards over a fixed corpus and zero benchmark documents, so this is an equal-document comparison, not a wall-clock one; separately, on Qwen2.5-0.5B CPU, matrix MSE costs 0.197 s/candidate versus 126.3 s for capability evaluation (642× cheaper; Appendix F).

Screening wins in all six pool/suite combinations, by factors of 1.4–9.3× in regret at the smallest budget, and the margin is largest where the candidate pool is largest (Appendix F). On the retrospective pool at 4,559 candidate-documents, evaluating all 47 candidates gives regret 0.0199 while screening to 5 and spending the same budget gives 0.0021, with $\epsilon_{\text{screen}} = 0$ at every m we tried because the screen never discards the best candidate. The second term of Equation 8 is visible too: on P1's development suite an aggressive screen to 6 has $\epsilon_{\text{screen}} = 0.0256$ and is no longer the best column, so both regimes are observed.

The proxy need not identify the winner, only reduce K to m safely, after which capability evaluation solves a much easier problem on the same budget.

4.4 At local resolution, fidelity points the wrong way

We ask the deployment question directly via the local matrix: with representation family, recipe, steps, and served size fixed, which single change should be kept? Every structural arm fits inside the baseline's 29.729 MB. We find that restoring each row's native norm (which costs no bytes) improves macro-average capability by $+0.0089$ across eight paired training seeds (95% paired-t CI $[+0.0011, +0.0167]$; seven of eight deltas positive), yet it worsens every principal proxy: teacher KL $0.1628 \to 0.1864$, perplexity $16.61 \to 16.83$, top-1 agreement $0.841 \to 0.826$, matrix MSE $1.41 \to 1.72 \times 10^{-4}$, and $\mathbf{C}_h$-weighted error $9.67 \to 12.32$ (Figure 3). This is stronger than a low correlation: all five proxies make the opposite binary decision.

Because norm restoration was selected from an exploratory screen, we preregistered non-selected A0-versus-A6 replications before training. SmolLM2-1.7B confirms the reversal: $+0.0050$ over 12 paired seeds (95% CI $[+0.0029, +0.0071]$, $p = 1.6 \times 10^{-4}$; 11/12 positive), while 4/5 fidelity metrics prefer the baseline. Llama-3.2-1B is unresolved ($+0.0036$, CI $[-0.0040, +0.0112]$), and a fresh-Qwen redraw fails the registered directional test (-0.0035, CI $[-0.0149, +0.0079]$; Appendix C.2). We therefore treat Qwen as the selected discovery and SmolLM2 as the preregistered independent-architecture confirmation.

We separate document-level from seed-level resolution: the same comparison has $\nu = 13.6$ at the document level ($\Delta = 0.0079$ over 7,500 documents, $\sigma_{\text{eff}} = 0.0015$) but only $\nu = 3.6$ at the seed

level: the benchmark resolves this effect comfortably and training-seed variance is what makes it marginal. Separating the two is a reporting requirement, since a document-only interval would have called several arms of this matrix winners.

The reverse direction fails too: we find that protecting the top eight $\mathbf{C}_h$ directions cuts activation-weighted error $3.6\times$ to 2.72 with no established gain at three seeds, and caching $2{,}048$ frequent rows gives the matrix's best KL, perplexity, and agreement while its paired-seed capability interval contains zero. Fidelity still catches the clearly damaged top-16-KL arm; it screens, but misses the positive local choice (Appendix B.9).

4.5 Mechanism checks, and a target that is not scalar

Two prospective mechanism checks (full curves in Appendix E.3): Matched-norm perturbations placed in the bottom, isotropic, or top eigenspaces of $\mathbf{C}_h$ have bit-identical matrix MSE within each model, yet teacher KL spans $19{,}429\times$ at 0.8B and $12{,}939\times$ at 3B – far past the pre-registered $3\times$ prediction – while frozen-suite capability moves from 0.4590 to 0.3215 and from 0.6370 to 0.3765 (Corollary 1(ii); single draws, so reachability rather than prevalence). We test Corollary 1(ii) on a controlled equal-MSE grid spanning seven language models, with 120 perturbations per model. At identical matrix MSE, top- versus bottom-eigenspace teacher KL differs by $127\times$–$2.6 \times 10^4\times$, while variation across six random directions within an eigenspace is at most $1.45\times$ (Table 25). Tightening the fidelity geometry does not consistently improve ranking: moving from table to activation/logit geometry changes the best viable-pool correlation by +0.001, -0.003, 0.000, -0.035 across the four pools, with a substantial gain appearing only for distribution geometry in the heterogeneous pool (+0.356; Appendix E.1); the activation bound's slack relative to KL varies $143\times$ across candidates. The direct quadratic explains only 22.6–25.7% of three compositional codes' realized logit MSE against 0.66–$1.34\times$ elsewhere, so Equation 2's second term is missing from the metric (Appendix E.2). A matched pathway intervention confirms that the local effect is output-driven: for top-eigenspace perturbations, the head-only path accounts for 97.8–98.1% of tied KL, and head-only versus tied capability rankings agree at $\rho = 0.958$ on dev and 0.902 held-out.

Finally, Proposition 3 matters only if candidates are actually Pareto-incomparable in our pools, and they are: among viable candidates 72–93% of pairs are Pareto-incomparable across tasks, 3–5 candidates are best on some single task, and Dirichlet-sampled deployment weightings produce 5–8 distinct winners per pool. No scalar proxy, and no scalar benchmark average, can be right for all of them at once. The scope extends beyond language: on ImageNet-1K (Deng et al., 2009) compression pools for ResNet50 and ViT-B/16 (Dosovitskiy, 2020), teacher KL screens broken candidates with AUROC $1.000/0.963$ and orders every resolvable pair correctly (Appendix E.5).

5 A protocol for compressed-model selection

The four limits above motivate the following five-step protocol. We state it explicitly because the required diagnostics are cheap to add to an existing compression paper and have measurable consequences in our experiments.

1. **Screen with fidelity, and report it separately from ranking.** A one-sided certificate is the right tool for rejecting damaged candidates (Theorem 1), and we find it works reliably (AUROC 0.97–1.00). Conflating screening with ranking can inflate apparent proxy skill by mixing the between-stratum signal with the within-stratum question.
2. **Control for compression rate and family.** Carry serialized bytes, form rate/family strata, and do not interpret a ranking that bytes reproduce as evidence of design-level selection without within-stratum support. (Proposition 1). Without this control, a metric that merely recovers how many bytes were spent will look like a capable ranker.
3. **Specify the deployment target and measure its resolution before trusting a correlation.** When candidates are Pareto-incomparable, first specify the deployment weighting that defines the target. Then report suite reliability and the resolvable-pair fraction (Equation 7) alongside every ρ. A ranking over comparisons that the evaluation budget cannot resolve should not be interpreted as evidence of proxy quality.

4. **Spend capability evaluation on the survivors, and separate noise sources.** Equation 8 quantifies the tradeoff between screening regret and evaluation error, and Section 4.4 shows that document-level and seed-level resolution can diverge ($\nu = 13.6$ vs. $\nu = 3.6$). Reporting only a document-level interval can declare wins that disappear when additional seeds are run. Which source binds is itself model-dependent: document noise dominates the SmolLM2 comparison, whereas seed-level variation dominates the Qwen and Llama-3.2-1B comparisons (Table 24).
5. **Allow "unresolved at this budget" as an answer.** We find it is the correct answer in 3 of our 6 curves at every affordable budget; this is better than a ranking that does not survive a disjoint half of the same suite. Forcing a verdict when the candidates are too close to separate can produce conclusions that do not survive an independent evaluation split.

We do not claim that fidelity is cheaper in wall-clock or FLOPs (all comparisons are at equal document counts), nor that any allocation rule is optimal. Making stage 4 a racing algorithm with an anytime guarantee is the natural next step; we do not evaluate it here.

5.1 Scope and limitations

Our main evidence is Qwen output-head compression; SmolLM2/Llama and ResNet50/ViT are scope checks, while whole-model compression is retrospective or simulated. The norm-restoration result was selected exploratorily and then tested in a pre-registered SmolLM2 replication. We measure teacher-forced multiple-choice accuracy against serialized bytes, not latency, energy, runtime memory, free-running generation, or safety. Further details are in Appendices B-C.

6 Related work

Evaluating compressed models. LLM-KICK shows that similar perplexity can hide different losses in knowledge, reasoning, generation, and retrieval (Jaiswal et al., 2024), and ACBench extends compressed-model testing to agentic capabilities (Dong et al., 2025). Both ask whether compression *preserves* capability; we ask whether a cheap measurement can *choose* among candidates that work; we show that a metric can detect damage well and still be a bad decision rule (Theorem 1, Section 4.4).

GPTQ, Fisher-weighted SVD, ASVD, and SVD-LLM establish that plain reconstruction error is misaligned with activation or loss geometry (Frantar et al., 2022; Hsu et al., 2022; Yuan et al., 2023; Wang et al., 2025); we claim no novelty for activation weighting and our equal-MSE controls confirm its value, but even the *fidelity* geometry is one-sided and need not be the *selection* geometry.

Best-arm identification, and benchmark precision. Choosing the best of K candidates from noisy evaluations is a pure-exploration problem. Theorem 2 and Proposition 2 are elementary members of that family (Maron & Moore, 1993; Even-Dar et al., 2006; Audibert & Bubeck, 2010; Karnin et al., 2013), which we import rather than extend. On the measurement side, split-half agreement and the Spearman–Brown correction are classical (Spearman, 1961; Brown, 1910), seed variance is a known source of false conclusions (Reimers & Gurevych, 2017), statistical power in NLP has been analyzed directly (Card et al., 2020; Dror et al., 2018), and tinyBenchmarks studies efficient *estimation* of benchmark scores (Polo et al., 2024) where we ask about *ordering* a specific candidate set. Our contribution is to connect these ideas to compressed-model selection: candidate-pool geometry determines resolution, while fidelity can serve as an effective pre-elimination screen. Extended comparisons are in Appendix D.

7 Conclusion

Fidelity can screen damaged models without reliably ranking viable ones. Compressed-model selection faces four limits: pool structure turns compression rate into apparent proxy skill; one-sided fidelity geometry need not preserve local order; finite budgets resolve only sufficiently separated candidates; and Pareto-incomparable candidates admit no universal scalar ranking. Empirically, fidelity screens reliably (AUROC 0.97–1.00), the resolution bound holds in 11,734 checks, and a byte-free fidelity–capability reversal replicates on an independent architecture in one of three pre-registered tests. We therefore advocate a simple protocol: screen with fidelity, control for rate and family, evaluate the survivors, and report "unresolved at this budget" when necessary.

AI USE STATEMENT

Generative AI tools were used to provide feedback on the experimental methodology. They were not used to generate synthetic datasets, develop theoretical or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, translate research material, process datasets, perform qualitative or thematic data analysis, or interpret experimental results.

Additionally, generative AI tools were used to identify and search for relevant literature, brainstorm and review research ideas, assist with writing and debugging experimental code, assist with the preparation of scientific figures, draft portions of the manuscript, and improve the clarity and readability of the writing. All AI-assisted code, figures, literature suggestions, and manuscript text were reviewed by the authors, and AI-assisted experimental code was tested before use. The authors take responsibility for the final content of this work.

REFERENCES

Mélisande Albert. Concentration inequalities for randomly permuted sums. In *High Dimensional Probability VIII: The Oaxaca Volume*, pp. 341–383. Springer, 2019.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pp. 13–p, 2010.

William Brown. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, 3(3):296–322, 1910.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pp. 7750–7774. PMLR, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.

Peijie Dong, Zhenheng Tang, Xiang Liu, Lujun Li, Xiaowen Chu, and Bo Li. Can compressed llms truly act? an empirical evaluation of agentic capabilities in llm compression. *arXiv preprint arXiv:2505.19433*, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. *arXiv preprint arXiv:2401.06118*, 2024.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35–48, 2024.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. *arXiv preprint arXiv:2207.00112*, 2022.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*, 2016.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. Compressing llms: The truth is rarely pure and never simple. In *International Conference on Learning Representations*, volume 2024, pp. 36973–36993, 2024.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 12186–12215, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pp. 1238–1246. PMLR, 2013.

Maurice G Kendall et al. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 785–794, 2017.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Minchong Li, Feng Zhou, and Xiaohui Song. Bild: Bi-directional logits difference loss for large language model distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1168–1182, 2025.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Computing and Communications*, 28(4):12–17, 2025.

Anemily Machina and Robert Mercer. Anisotropy is not inherent to transformers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4892–4907, 2024.

Oded Maron and Andrew Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. *Advances in neural information processing systems*, 6, 1993.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Paolo Paruolo. Multivariate reduced rank regression, theory and applications, 2000.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, 2017.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 338–348, 2017.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Charles Spearman. The proof and measurement of association between two things. 1961.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Kohki Tamura, Naoki Yoshinaga, and Masato Neishi. Further compressing distilled language models via frequency-aware partial sparse coding of embeddings. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pp. 388–399, 2024.

Dan Tito Svenstrup, Jonas Hansen, and Ole Winther. Hash embeddings for efficient word representations. *Advances in neural information processing systems*, 30, 2017.

Haoyu Wang, Ruirui Li, Haoming Jiang, Zhengyang Wang, Xianfeng Tang, Bin Bi, Monica Cheng, Bing Yin, Yaqing Wang, Tuo Zhao, et al. Lighttoken: A task and model-agnostic lightweight token embedding framework for pre-trained language models. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2302–2313, 2023.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pp. 19299–19319, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577, 2024.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

A Full protocol

A.1 The five interface settings

(S-1) A tied sparse lexical interface. Replace $\mathbf{E}^0 \in \mathbb{R}^{V\times h}$ by $u_t = z_t\mathbf{B}$ with z_t k-sparse over D learned atoms, $\mathbf{B} \in \mathbb{R}^{D\times h}$ shared. Because the head is tied, the output softmax factorizes exactly as $\boldsymbol{\ell} = \mathbf{Z}(\mathbf{B}h)$, at cost $O(Dh + Vk)$ against dense $O(Vh)$. Codes are $z_t = \mathrm{TopK}_k(\mathrm{ReLU}((x_t - b)\mathbf{W}_{\mathrm{enc}}^\top))$, solved offline and cached; the transformer is **frozen in every run** and only $\mathbf{B}$ (7.34M parameters at $D = 8192$, $h = 896$) is in the optimizer by default. The deployed artifact is the cached $\mathbf{Z}$ (indices + coefficients) plus $\mathbf{B}$; $\mathbf{W}_{\mathrm{enc}}$ and $\mathbf{E}^0$ are discarded. Loss $(1-\lambda)\,\mathrm{CE} + \lambda\,\mathrm{KL}(p_{\text{base}} \| p_{\text{student}})$ with the frozen native model as teacher (Hinton et al., 2015), Adam at 10^{-3}, 8000 steps unless stated. Operating point $D = 8192$, $\bar{k} = 64$.

(S-2) An output-only low-rank functional repair. Keep the sparse input side and add $+(\mathbf{E}^0 - U)\mathbf{Q}_r\mathbf{Q}_r^\top$ to the head. Storage is $V \cdot r$ values for any $\mathbf{Q}_r$, which is what isolates *which* subspace a fixed budget should repair (Table 8).

(S-3) A cross-state functional repair. Given the compressed model's own hidden states, fit the smallest rank-r operator mapping them to the teacher's logits — reduced-rank regression, which has a closed-form global optimum and therefore an exact rate–distortion curve inside its family.

(S-4) An input-only side memory. Compress an untied *input* table and patch the input side only, head native. This setting has no tied head and therefore no head-input covariance, so it tests whether the mechanisms of Section 4.5 are properties of tying or of embedding interfaces generally ($\lambda_{\max}$ share 0.052 against 0.59).

(S-5) Whole-model heterogeneous compression. Quantize the backbone sequentially with GPTQ and allocate a fixed whole-model byte budget between backbone and interface. Native bf16 Qwen2.5-0.5B is 988.07 MB (715.65 backbone + 272.27 interface); budgets are computed from the uniform-quantizer totals, not chosen: $139.07/200.81/262.56$ MB.

Baselines. Group-RTN, AWQ and GPTQ on the same table at 2–6 bits; post-hoc and LM-trained low-rank factorization; product and additive quantization; sparse PCA; magnitude pruning; a random-dictionary control. All are implemented in-repo and priced by the same accountant.

A.2 Proofs

Proof of Theorem 1. Fix a hidden state h, write $\ell = Eh$, $\delta = \Delta h$, and let $A(\ell) = \log \sum_j \exp(\ell_j)$. Since $\nabla A(\ell) = p_\ell$, teacher KL is the Bregman divergence of log-sum-exp,

$$D_{\mathrm{KL}}(p_\ell \| p_{\ell+\delta}) = A(\ell+\delta) - A(\ell) - \nabla A(\ell)^\top \delta.$$

Taylor's theorem with integral remainder gives the exact identity

$$D_{\mathrm{KL}}(p_\ell \| p_{\ell+\delta}) = \int_0^1 (1-t)\, \delta^\top \left[\mathrm{Diag}(p_{\ell+t\delta}) - p_{\ell+t\delta} p_{\ell+t\delta}^\top\right] \delta \, dt.$$

The bracketed matrix is the covariance of a categorical one-hot vector. Its operator norm is at most $1/2$, with equality for a two-point distribution of masses $1/2, 1/2$. Hence

$$D_{\mathrm{KL}}(p_\ell \| p_{\ell+\delta}) \le \int_0^1 (1-t) \tfrac{1}{2} \|\delta\|_2^2 dt = \tfrac{1}{4} \|\Delta h\|_2^2.$$

Taking expectation over h yields $\mathbb{E}\|\Delta h\|_2^2 = \mathrm{Tr}(\Delta \mathbf{C}_h \Delta^\top)$. Finally, $\mathrm{Tr}(\Delta \mathbf{C}_h \Delta^\top) \le \lambda_{\max}(\mathbf{C}_h) \|\Delta\|_F^2$.

For the first reverse direction, choose $\Delta = \mathbf{1} a^\top$. Then $\Delta h = (a^\top h)\mathbf{1}$ is a common logit shift, so KL is identically zero while the activation-weighted quadratic is positive whenever $a^\top h$ is nonzero with positive probability. For the second, no lower constant can hold uniformly over covariances: place Δ in a null direction, or in an eigenvector whose eigenvalue tends to zero. If a particular $\mathbf{C}_h$ is positive definite, the exact fixed-covariance lower bound is $\mathrm{Tr}(\Delta \mathbf{C}_h \Delta^\top) \ge \lambda_{\min}(\mathbf{C}_h) \|\Delta\|_F^2$; its condition number is precisely the possible slack. □

Proof of Corollary 1. Let $\Delta_i = \sqrt{c}\, u q_i^\top$, where $\|u\|_2 = \|q_i\|_2 = 1$ and $\mathbf{C}_h q_i = \lambda_i q_i$. Then $\|\Delta_i\|_F^2 = c$, while

$$\mathrm{Tr}(\Delta_i \mathbf{C}_h \Delta_i^\top) = c\ \mathrm{Tr}(u q_i^\top \mathbf{C}_h q_i u^\top) = c\lambda_i.$$

Thus equal-Frobenius perturbations span a ratio $\lambda_{\max}/\lambda_{\min} = \kappa(\mathbf{C}_h)$ when $\mathbf{C}_h$ is positive definite, and an infinite ratio when it has a nullspace. A strict ordering reversal follows by choosing eigenvalues $\lambda_i > \lambda_j$ and norms $c_1 < c_2$ such that $1 < c_2/c_1 < \lambda_i/\lambda_j$: then the smaller-MSE perturbation has the larger activation-weighted cost. □

Proof of Proposition 1. Apply iterated expectation to the centered product:

$$\mathrm{Cov}(U, V) = \mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V].$$

Adding and subtracting $\mathbb{E}_S[\mathbb{E}(U \mid S)\mathbb{E}(V \mid S)]$ gives Equation 3. Because U and V are the global rank transforms, their Pearson correlation is pooled Spearman correlation. To see that the between-stratum term can dominate, form an increasing number of ordered, nonoverlapping strata and permute proxy and capability ranks independently inside each one. Every conditional covariance is zero. The within-stratum rank variation remains bounded while the variance of the common stratum rank grows quadratically in the number of strata, so pooled Spearman correlation approaches one. □

Proof of Equation 4. Write $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$. The assumed sub-Gaussian moment bound and the Chernoff method give

$$\Pr(\bar{X}_n \le 0) = \Pr(\bar{X}_n - \Delta_{ij} \le -\Delta_{ij}) \le \exp\left(-\frac{n\Delta_{ij}^2}{2\sigma_{ij}^2}\right) = e^{-\nu_{ij}(n)},$$

which is Equation 4 (Hoeffding, 1963; Albert, 2019). Solving $e^{-\nu_{ij}(n)} \le \delta$ for n gives $n \ge 2\sigma_{ij}^2\Delta_{ij}^{-2}\log(1/\delta)$, the Δ^{-2} law and the threshold in Equation 7. For a union bound over K candidates, apply the pairwise statement to each comparison of the unique best candidate with one competitor at failure probability $\delta/(K-1)$; simultaneous correct selection then holds with probability at least $1-\delta$ once $n \ge \max_{i\neq\star} 2\sigma_{\star i}^2\Delta_i^{-2}\log\frac{K-1}{\delta}$. □

Proof of Theorem 2. The event $\{\hat{c}_n = c_i\}$ implies $\widehat{A}_n(c_i) \ge \widehat{A}_n(c^\star)$, i.e. the paired difference $\bar{X}_n^{(i)} \le 0$, so $\Pr(\hat{c}_n = c_i) \le e^{-\nu_i(n)}$ by Equation 4. Because $\mathrm{Reg}(\hat{c}_n) = \sum_{i\neq\star}\Delta_i\,\mathbf{1}\{\hat{c}_n = c_i\}$ (the events are disjoint and exhaust the complement of a correct choice), linearity of expectation gives $\mathbb{E}[\mathrm{Reg}] = \sum_{i\neq\star}\Delta_i\Pr(\hat{c}_n = c_i) \le \sum_{i\neq\star}\Delta_i e^{-\nu_i(n)}$. For the tail statement, $\{\mathrm{Reg} \ge \epsilon\} = \bigcup_{i:\Delta_i\ge\epsilon}\{\hat{c}_n = c_i\}$, so a union bound gives $\Pr[\mathrm{Reg} \ge \epsilon] \le \sum_{i:\Delta_i\ge\epsilon} e^{-\nu_i(n)}$. Ties in $\widehat{A}_n$ are counted against the selector, which the non-strict inequality $\widehat{A}_n(c_i) \ge \widehat{A}_n(c^\star)$ already does. □

Proof of Corollary 2. With no true ties, Kendall's τ_a over K items is $1 - 2D_n/\binom{K}{2}$, where D_n is the number of pairs whose estimated order disagrees with the true order, counting an estimated tie as a disagreement (Kendall et al., 1938). Each pair contributes to D_n only if $\widehat{A}_n(i) \le \widehat{A}_n(j)$ for the truly better i, an event of probability at most $e^{-\nu_{ij}(n)}$. Hence $\mathbb{E}[D_n] \le \sum_{i<j} e^{-\nu_{ij}(n)}$ and $\mathbb{E}[\tau_n] \ge 1 - 2\binom{K}{2}^{-1}\sum_{i<j} e^{-\nu_{ij}(n)} = 1 - \frac{4}{K(K-1)}\sum_{i<j} e^{-\nu_{ij}(n)}$. No independence between pairs is needed, because only expectations are combined. □

Macro-averaged targets (Section 3.3). Let the macro estimator be

$$\widehat{A} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{u=1}^{n_t} Y_{tu},$$

where the independent per-document paired differences Y_{tu} are σ_t^2-sub-Gaussian. A weighted sum of independent sub-Gaussian variables with weights $w_{tu} = 1/(Tn_t)$ is sub-Gaussian with parameter

$$\sigma_{\text{eff}}^2 = \sum_{t=1}^{T}\frac{\sigma_t^2}{T^2 n_t}.$$

Substituting σ_{eff}^2 for the variance term in Equation 4 gives

$$\nu = \frac{\Delta^2}{2\sigma_{\text{eff}}^2}.$$

In the balanced case with n total evaluation documents, $n_t = n/T$ and $\sigma_t \equiv \sigma$, so $\sigma_{\text{eff}}^2 = \sigma^2/n$ and therefore

$$\nu = \frac{n\Delta^2}{2\sigma^2},$$

which reduces exactly to Equation 4. □

Proof of Proposition 2. The screen is a function of quantities independent of the evaluation examples, so S may be treated as fixed. For each $c \in S$, a two-sided sub-Gaussian bound gives $\Pr(|\widehat{A}_n(c) - A(c)| > t) \le 2e^{-nt^2/(2\sigma^2)}$; a union bound over the m survivors makes $\max_{c\in S}|\widehat{A}_n(c) - A(c)| \le t$ hold with probability at least $1 - 2me^{-nt^2/(2\sigma^2)}$. Setting that to δ gives $t = \sigma\sqrt{2\log(2m/\delta)/n}$. On that event, writing $c_S = \arg\max_{c\in S} A(c)$,

$$A(c_S) - A(\hat{c}) \le \widehat{A}_n(c_S) + t - \big(\widehat{A}_n(\hat{c}) - t\big) \le 2t,$$

because $\widehat{A}_n(\hat{c}) \geq \widehat{A}_n(c_S)$ by definition of $\hat{c}$. Adding $A(c^\star) - A(c_S) = \epsilon_{\text{screen}}$ gives Equation 8. At a fixed total budget of $B = mn$ candidate-examples, $n = B/m$ and the statistical term is $2\sigma\sqrt{2m\log(2m/\delta)/B}$, which is increasing in m while ϵ_{screen} is non-increasing in m for a nested screen. □

Proof of Proposition 3. Pareto-incomparability gives tasks t, s with $A_t(c_i) > A_t(c_j)$ and $A_s(c_i) < A_s(c_j)$. Take $w = e_t$ and $w' = e_s$. For strictly positive weights, perturb: $w = e_t + \varepsilon\mathbf{1}$ and $w' = e_s + \varepsilon\mathbf{1}$ preserve both strict inequalities for small enough $\varepsilon > 0$ by continuity. Hence no single ordering of c_i, c_j is correct for all weightings, and a scalar proxy induces exactly one ordering. □

A.3 The tied chain rule, and why the two settings are not interchangeable

For tied embeddings $f_E(x) = E\,h_E(x)$, perturbing $E \to E + \Delta$ gives

$$\delta f = \underbrace{\Delta\,h_E}_{\text{direct output path}} + \underbrace{E\,Dh_E[\Delta]}_{\text{input/state path}} + O(\|\Delta\|^2). \tag{9}$$

This is the chain rule and we claim no novelty for it, but it is what makes Section 3 exact for one setting and first-order for another: the output-only repair (S-2) and the head-only prospective pools exercise the direct term alone, the input-only side memory (S-4) the indirect term alone, and a tied compression both. **Theorem 1 is exact for the direct path at fixed hidden states**; this is why the prospective challenges are head-only, and it also lets every candidate share one set of cached teacher hidden states, which is what makes 48 candidates affordable. A tied-subset analysis must recompute both fidelity (the student's hidden states are no longer the teacher's) and capability. We do not use that subset in any headline claim (Section 5.1).

A.4 Pre-registration mechanics

For both prospective challenges the pipeline is staged: `validate` → `download` → `corpus` → `fidelity` → `capability` → `analyse`. The driver **refuses to run the capability stage without a complete fidelity artifact**, and fidelity is written atomically after every candidate, so the order in the pre-registration is enforced by the code rather than by intention. The candidate pool, metric battery, task suites, above-chance rule, viability threshold and sensitivity grid, statistics and predictions are all in the pre-registration file (Appendix C) with a timestamp preceding the first capability score.

Fidelity is measured on 64 deterministic WikiText-103 test windows (Merity et al., 2016) of length 256, tokenized by each model's own tokenizer (16,320 scored positions), with $\mathbf{C}_h$ estimated on those same native hidden states.

The two prospective pools use the nine-task suite fixed before capability evaluation: ARC, HellaSwag, OpenBookQA, WinoGrande, BoolQ, MMLU, RACE, CommonsenseQA and COPA (Clark et al., 2018; Zellers et al., 2019; Mihaylov et al., 2018; Sakaguchi et al., 2021; Clark et al., 2019; Hendrycks et al., 2020; Lai et al., 2017; Talmor et al., 2019; Roemmele et al., 2011), with at most 500 documents per task and per-example scores retained.

For the resolution intervention, its evaluation limit is increased to 2,000 documents. Retaining per-example scores is what makes split-half self-consistency and honest regret computable: without them, "the metrics do not rank" is indistinguishable from "the measurement cannot resolve anything".

A.5 Byte accounting

One function prices every row: structured interfaces, GPTQ/AWQ/RTN, compositional codes, and native bf16. We use decimal MB (10^6 bytes) throughout, with a bf16 denominator of 272.27 MB for the Qwen2.5-0.5B table; MiB is not used in any reported compression ratio or budget. Sizes are recomputed rather than quoted at a flat ratio: ragged codes are charged $\sum_t k_t$; pinned rows drop their code slots and add $n \cdot h$; an untied head counts shared indices once and both value sets; a selective residual over n_r rows costs $\lceil n_r h b/8 \rceil + 2n_r + 4n_r$ bytes. Whole-model accounting sums backbone linears at their quantized width, norms and biases at fp16, and the interface's own byte dictionary

passed in rather than recomputed; the embedding is asserted not to appear among the backbone linears; mean bits equals nominal plus $32/\text{group}$ exactly.

For the byte-matched improvement matrix (P4), the accountant *solves* for the largest k' that fits the baseline's budget and ties are broken **downward**, so every non-baseline arm is handed fewer bytes than the baseline and no win can be a byte artifact. Twenty-seven verification checks run before the batch is submitted, including the byte match, the softmax-gauge identity, the $\mathbf{B}_\perp Q$ projection, the free-ness of the norm restore, and the two changed loss terms against a direct reimplementation.

A.6 Whole-model quantization protocol

Sequential GPTQ, not the cheap variant: each decoder layer's Hessians are measured with layers $0..i-1$ *already quantized*, one calibration pass per layer (64 windows $\times$ 512 tokens). The cheap "collect every Hessian from the fp32 model once" version would weaken our own baseline. Calibration length is decoupled from evaluation length so the baseline's calibration set cannot be silently shrunk to whatever the evaluation uses. The teacher is the *whole* native model: a shared-backbone fidelity function (correct only while the table moves) would score a quantized-backbone student through the native transformer, and we verified the failure mode directly (on a backbone-only-quantized student it reports KL 0.0000 where the real whole-model KL is 0.0023), so two real models are run instead. Backbone per-layer bit allocation is floored at 3 bits.

A.7 Statistics

Paired bootstrap at the example level. Pairing over nine task means gives $n = 9$, too wide to resolve a 0.004 delta. We record per-example scores and resample *within* each task with the same indices for both arms. Verified: intervals $3.7\times$ tighter on truly paired arms (0.0208 against 0.0775), and a true null returns delta exactly 0.0000. For the improvement matrix, each arm has at least three seeds. Seeds are averaged *per document* first, then documents are resampled within task and task means are macro-averaged over $4{,}000$ bootstrap draws (Efron, 1992). Norm restoration was selected after the initial three-seed screen. A power analysis then fixed the total at eight before the additional seeds were inspected; baseline and selected arm use the same seed IDs. Their reported analysis is the paired task-macro seed difference with a two-sided paired-t interval. The remaining arms stay exploratory at three seeds, for which seed-level intervals are necessarily wide.

Screening AUROC is over the broken/viable split with variant-level bootstrap intervals.

Within-viable ranking reports Spearman ρ with Fisher-z 90% intervals and a bootstrap 95% interval, Kendall τ, all-pair and resolvable-pair accuracy, and the pre-declared equivalence margin $|\rho| < 0.5$.

Suite ranking reliability is a split-half Spearman over 200 random pairs of disjoint halves of the examples, Spearman–Brown corrected (Spearman, 1961; Brown, 1910) for the full suite. Attenuation-corrected metric ρ divides by $\sqrt{r_{\text{SB}}}$.

Honest selection regret selects on one random half of the examples and scores on the disjoint half, 400 splits, reported beside capability selection with the same number of evaluation examples, random selection and the worst arm. This is an equal-document statistical comparison, not matched FLOPs, teacher-forward cost, or wall-clock time. All intervals are percentile bootstraps with a fixed generator seed, reproducible without third-party numerics.

Budget curves and bound calibration. Each task's documents are split once into a draw half H_1 and a disjoint reference half H_2. For every budget we take 400 subsamples of H_1 without replacement, using the same documents for every candidate, and compute the task-macro score. Two sets of statistics are then reported, because they answer different questions. *Out-of-sample* columns (reliability, τ, regret, probability of choosing the best) score the budget-n ranking against H_2, so nothing is evaluated on the data that selected it. *Calibration* columns score the same draws against the H_1 population mean, which is the estimand a subsample of H_1 is actually estimating; gaps Δ_{ij} and variances σ_t^2 come from the same half, so a bound violation cannot be manufactured by half-to-half

noise. Both are given in Appendix E for all six pool/suite combinations. Documents are resampled with the trained artifact held fixed, so these curves isolate target noise from training-seed noise by construction.

Node matching. Every whole-model comparison runs on one CPU node family; Appendix E gives the measured cross-family scope and how it is handled.

B FULL TABLES

B.1 THE COMPLETE SELECTOR BATTERY ON THE RETROSPECTIVE POOL

Table 1 is the full within-viable battery on the cleanest slice of R: the 19 viable arms whose capability run stored per-example scores, which is what makes split-half regret and the suite ranking self-consistency computable. Byte-matched at 30.3–50.5 MB, 18 of 19 in the sparse-interface family, MEAN_ACC 0.5362–0.5562 (a 0.0200 spread, 3.3× the seed s.d.).

Table 1: R, 19 viable arms with per-example scores. The final column shows the standard attenuation adjustment, $\rho/\sqrt{0.758}$, where 0.758 is the suite's Spearman–Brown reliability estimate. We treat it as a descriptive noise adjustment, not a selector ceiling. Naive regret against the empirical maximum is biased for every selector; honest split-half regret is reported in the machine-readable battery. Hidden cosine's apparent lead is only 3.5×10^{-5} and its top arms are the largest int4 interfaces, so it is not evidence beyond the deployed-rate axis.

selector	ρ	90% CI	pairwise	naive R_M	vs chance	ρ corrected
hidden cosine	**+0.717**	[+0.45, +0.86]	75.6%	0.0000	0.00×	**+0.823**
teacher-argmax NLL	+0.699	[+0.43, +0.86]	78.0%	0.0000	0.00×	+0.803
top-1 agreement	+0.672	[+0.38, +0.84]	74.9%	0.0000	0.00×	+0.772
EAR ratio	+0.661	[+0.36, +0.84]	74.4%	0.0000	0.00×	+0.759
JSD	+0.658	[+0.36, +0.83]	75.0%	0.0000	0.00×	+0.756
teacher KL	+0.653	[+0.35, +0.83]	74.4%	0.0000	0.00×	+0.750
top-5 overlap	+0.573	[+0.24, +0.79]	70.8%	0.0000	0.00×	+0.658
reconstruction MSE	+0.473	[+0.10, +0.73]	66.1%	0.0072	0.75×	+0.544
perplexity	+0.401	[+0.01, +0.68]	65.5%	0.0102	1.06×	+0.460
logit MSE	+0.249	[−0.16, +0.58]	58.9%	0.0116	1.19×	+0.286
$\mathbf{C}_h$-weighted error	+0.246	[−0.16, +0.58]	60.1%	0.0116	1.19×	+0.283
random choice	*0.000*	—	*50.0%*	*0.0097*	*1.00×*	—
suite split-half	**+0.610**	±0.118	—	—	—	*(S–B +0.758)*

B.2 SCREENING AUROC PER METRIC

Table 2: Screening AUROC with 90% bootstrap intervals (broken vs. viable; viability declared in advance at a 0.05 capability drop). Every principal metric screens strongly on both prospective pools. The only point estimates below 0.5 are perplexity on the retrospective pool; both corresponding intervals include chance, so we do not claim systematic inversion.

metric	P1 dev	P1 held out	P2 dev	P2 held out	R dev	R held out
teacher KL	0.990 [0.966, 1.000]	0.986 [0.955, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.987 [0.949, 1.000]
JSD	0.993 [0.974, 1.000]	0.990 [0.963, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.982 [0.939, 1.000]	0.974 [0.924, 1.000]
teacher-argmax NLL	0.991 [0.970, 1.000]	0.986 [0.955, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.984 [0.939, 1.000]
top-1 agreement	0.990 [0.965, 1.000]	0.988 [0.958, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.947 [0.870, 1.000]	0.939 [0.857, 0.994]
top-5 overlap	0.990 [0.965, 1.000]	0.986 [0.955, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.982 [0.939, 1.000]	0.968 [0.905, 1.000]
EAR ratio	0.990 [0.965, 1.000]	0.984 [0.950, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.930 [0.839, 0.989]	0.929 [0.838, 0.989]
logit cosine	0.995 [0.979, 1.000]	0.981 [0.944, 1.000]	0.990 [0.946, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.994 [0.971, 1.000]
logit MSE	0.981 [0.937, 1.000]	0.970 [0.917, 1.000]	0.990 [0.946, 1.000]	1.000 [1.000, 1.000]	0.962 [0.881, 1.000]	0.946 [0.848, 1.000]
reconstruction MSE	0.983 [0.946, 1.000]	0.976 [0.931, 1.000]	0.990 [0.946, 1.000]	1.000 [1.000, 1.000]	0.845 [0.638, 1.000]	0.821 [0.586, 1.000]
matrix cosine	0.993 [0.973, 1.000]	0.981 [0.945, 1.000]	0.990 [0.946, 1.000]	1.000 [1.000, 1.000]	0.956 [0.883, 1.000]	0.965 [0.895, 1.000]
$\mathbf{C}_h$-weighted error	0.981 [0.937, 1.000]	0.970 [0.917, 1.000]	0.990 [0.946, 1.000]	1.000 [1.000, 1.000]	0.728 [0.442, 1.000]	0.692 [0.374, 1.000]
perplexity	0.990 [0.965, 1.000]	0.986 [0.955, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.333 [0.000, 0.667]	0.381 [0.008, 0.753]

B.3 THRESHOLD SENSITIVITY OF THE VIABILITY BOUNDARY

The 0.05 capability-drop boundary is declared in advance. We vary the boundary after the fact only as a sensitivity analysis. The prospective pools are stable across the tested range; the retrospective pool is more sensitive because the number and identity of broken candidates change sharply.

Table 3: Viability-threshold sensitivity. Entries are within-viable Spearman correlations, oriented so that larger is better. P2 is especially stable: on its held-out suite the threshold moves fourfold without changing which candidate lies on either side of the boundary.

pool / suite	drop	broken	viable	teacher KL	perplexity	reconstruction MSE
P1 dev	0.03	26	22	$+0.395$	$+0.400$	$+0.366$
P1 dev	0.05	25	23	$+0.467$	$+0.475$	$+0.441$
P1 dev	0.08	23	25	$+0.579$	$+0.583$	$+0.547$
P1 dev	0.12	19	29	$+0.716$	$+0.724$	$+0.692$
P1 held out	0.03	29	19	$+0.561$	$+0.619$	$+0.589$
P1 held out	0.05	26	22	$+0.717$	$+0.755$	$+0.732$
P1 held out	0.08	23	25	$+0.787$	$+0.811$	$+0.794$
P1 held out	0.12	16	32	$+0.895$	$+0.907$	$+0.890$
P2 dev	0.03	7	15	$+0.828$	$+0.815$	$+0.828$
P2 dev	0.05	7	15	$+0.828$	$+0.815$	$+0.828$
P2 dev	0.08	7	15	$+0.828$	$+0.815$	$+0.828$
P2 dev	0.12	6	16	$+0.858$	$+0.848$	$+0.852$
P2 held out	0.03	6	16	$+0.756$	$+0.738$	$+0.758$
P2 held out	0.05	6	16	$+0.756$	$+0.738$	$+0.758$
P2 held out	0.08	6	16	$+0.756$	$+0.738$	$+0.758$
P2 held out	0.12	6	16	$+0.756$	$+0.738$	$+0.758$
R dev	0.03	11	36	$+0.757$	$+0.419$	$+0.661$
R dev	0.04	9	38	$+0.791$	$+0.504$	$+0.471$
R dev	**0.05**	9	38	$\mathbf{+0.791}$	$\mathbf{+0.504}$	$\mathbf{+0.471}$
R dev	0.06	7	40	$+0.821$	$+0.290$	$+0.546$
R dev	0.08	3	44	$+0.864$	-0.029	$+0.658$
R dev	0.10	3	44	$+0.864$	-0.029	$+0.658$

At the loosest retrospective thresholds only three candidates are labelled broken, so screening AUROC itself becomes too poorly determined to interpret. We therefore use this sweep as a ranking-sensitivity check rather than as evidence for or against screening in that regime.

B.4 WHOLE-MODEL SCREENING SCOPE CHECK

We additionally test a simulated whole-model pool of 15 Qwen2.5-1.5B candidates spanning GPTQ, AWQ, and group-RTN at 2–4 bits. Five candidates remain viable, all at 4 bits. This is a scope check rather than prospective evidence.

B.5 POOL SURGERY AND RANDOM SUBSETS

Pooling across byte budgets inflates teacher KL on P3 from $+0.567$ (Fisher-pooled within byte class) to $+0.791$ (pooled over all viable), with $+0.673$ within family; the corresponding numbers are top-1 $+0.672/+0.737/+0.505$, reconstruction MSE $+0.500/+0.593/+0.471$, perplexity $+0.270/+0.384/+0.504$ and $\mathbf{C}_h$-weighted $+0.284/+0.542/+0.422$. On P1 dev the single stratum with enough arms (RTN, $n = 18$) has *every* metric at $\rho \approx -0.065$ against $+0.468$ across strata, with a within-stratum accuracy spread of 0.034 and a split-half self-consistency of $+0.297$. **Pooled selector correlations are partly a rate artifact and the size of the artifact is now measured.**

Table 4: Whole-model screening scope check. Fidelity cleanly removes the damaged 2–3 bit candidates, but the surviving 4-bit candidates remain too close for reliable ranking at the available evaluation budget.

quantity	value
candidates	15
viable candidates	5 (all 4-bit)
screening AUROC, all seven proxies	$\mathbf{1.000}$
regret, evaluate all 15	0.0049
regret, screen to $m = 3$	$\mathbf{0.0016}$
regret reduction	$3.0\times$
broken candidates retained after screening	0
resolvable-pair fraction at full budget	0.30
best-vs-selected viable gap	0.0020
resolution index for that pair	$\nu = 0.08$

B.6 Rate-controlled correlations

The effect of pooling is not uniform across metrics or pools. On the retrospective development pool, teacher KL falls from $\rho = +0.791$ pooled to $+0.567$ after conditioning on byte class, while several other metrics move in the opposite direction. The relevant quantity is therefore the decomposition itself, not an assumption that pooling always inflates correlation.

Table 5: Retrospective development pool: pooled rank correlation versus Fisher-pooled within-byte and within-family correlation. Pooling contributes substantial between-rate signal for teacher KL, but not identically for every metric.

selector	pooled	within byte class	within family
teacher KL	$+\mathbf{0.791}$	$+0.567$	$+0.673$
EAR ratio	$+0.684$	$+0.604$	$+0.621$
top-1 agreement	$+0.505$	$+0.672$	$+0.737$
perplexity	$+0.504$	$+0.270$	$+0.384$
reconstruction MSE	$+0.471$	$+0.500$	$+0.593$
$\mathbf{C}_h$-weighted error	$+0.422$	$+0.284$	$+0.542$

The cleanest prospective instance is the largest homogeneous P1 stratum. On development, all principal selectors are near zero inside the RTN stratum even though their across-stratum correlations are positive; on held-out data, the same stratum retains substantial within-stratum association.

Table 6: P1 RTN stratum. The development suite exhibits the strongest pool effect: the same selectors that correlate positively across strata are near zero inside the homogeneous stratum. The held-out suite is reported as a control showing that the effect is pool- and suite-specific rather than a necessary consequence of stratification.

suite	selector	within-stratum ρ	across-stratum ρ
dev	teacher KL	-0.065	$+0.467$
dev	perplexity	-0.044	$+0.475$
dev	reconstruction MSE	-0.065	$+0.441$
dev	$\mathbf{C}_h$-weighted error	-0.065	$+0.437$
dev	top-1 agreement	-0.065	$+0.465$
held out	teacher KL	$+0.683$	$+0.717$
held out	perplexity	$+0.727$	$+0.755$
held out	reconstruction MSE	$+0.683$	$+0.732$
held out	$\mathbf{C}_h$-weighted error	$+0.683$	$+0.729$
held out	top-1 agreement	$+0.683$	$+0.733$

The development RTN stratum contains 18 candidates, has a capability spread of 0.034, and has split-half ranking reliability $+0.309$.

B.7 ANISOTROPY, AND THE RANK-8 SUBSPACE SWEEP IN FULL

Table 7: Head-input second-moment anisotropy in the original geometry pass. We report the quantities measured for all four models in this pass rather than mixing measurements from different corpus/window configurations.

model	energy in top-1 direction	effective rank
Qwen2.5-0.5B	59.0%	2.84/896
Qwen3.5-0.8B	36.0%	2.78/1024
Qwen2.5-3B	64.2%	1.56/2048
GPT-2	98.9%	$\approx 1.0/768$

For Qwen2.5-0.5B, the same pass gives 74.2% energy in the top eight directions and $\kappa(\mathbf{C}_h) = 1.1 \times 10^8$; for GPT-2, the top-eight share is 99.8%.

Table 8: Four rank-8 output corrections at **identical** cost (+1.22 MB), plus rank ablations. Reconstruction MSE spans 2.5% across the four and cannot distinguish them; teacher KL spans 2.2×. The *Frobenius-optimal* subspace — the rank-8 SVD of the residual, by construction the best rank-8 approximation of the error matrix — is the only arm that makes fidelity *worse*, and it has the best MSE of the four. Single artifact, reconstruction-only, CPU; ordering and sign, not magnitudes.

r	subspace	teacher KL	top-1 agree.	matrix MSE	extra MB
—	*no correction*	3.876	0.144	9.16e−5	0
8	**top-8 eigenvectors of $\mathbf{C}_h$**	**1.859**	**0.314**	9.10e−5	+1.22
8	bottom-8 eigenvectors of $\mathbf{C}_h$	3.881	0.145	9.08e−5	+1.22
8	random orthonormal	3.858	0.143	9.09e−5	+1.22
8	residual-SVD (*Frobenius optimum*)	**4.018**	0.136	8.94e−5	+1.22
1	top-1 eigenvector of $\mathbf{C}_h$	2.316	0.280	9.16e−5	+0.15
16	top-16 eigenvectors of $\mathbf{C}_h$	1.771	0.321	9.04e−5	+2.43
64	top-64 eigenvectors of $\mathbf{C}_h$	1.534	0.409	8.61e−5	+9.72
—	*ceiling: fully native output head*	0.993	0.579	0	+272

A second instance of the same dissociation, on allocation rather than repair: at identical bytes and identical mean $k = 8.00$ on GPT-2, a damped functional allocator reaches teacher KL 3.6438 while a Euclidean allocator reaches 37.2856 — 10× worse — while having the **best matrix MSE of all six arms tested** (1.476e−2 against the winner's 1.544e−2). And decomposing two byte-matched error matrices on GPT-2: an additive code has *more* total MSE (1.168e−2 against 1.052e−2) and 95–103× *less* error in the top-$\mathbf{C}_h$ direction, i.e. 21× less $\mathbf{C}_h$-weighted error — and its teacher KL is 4.2× better.

B.8 FREE-RUNNING FIDELITY

Figure 4 plots the same rows.

Table 9: Generation-level fidelity, 500 untouched prompts, greedy. **Every** compressor loses ≈0.71 of its agreement when the model consumes its own output — including GPTQ-int3, which is benchmark-lossless (ΔMEAN_ACC $= -0.001$). Single-position teacher-forced fidelity was over-reporting for all of them, not differentially for one.

interface	prefix match	exact full	free-running	teacher-forced	**gap**
native	1.0000	1.0000	1.0000	1.0000	0.0000
sparse interface	0.1332	0.0300	0.1976	0.9074	**0.7098**
GPTQ int3	0.0747	0.0100	0.1414	0.8649	**0.7234**
AWQ int4	0.0959	0.0180	0.1513	0.8886	**0.7373**

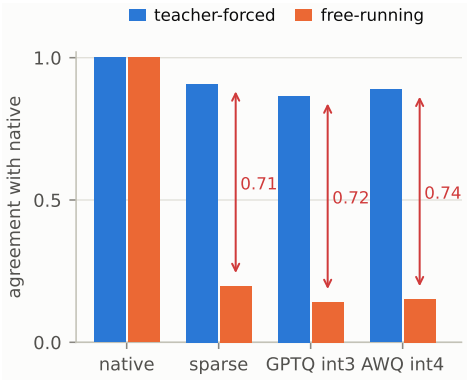

Figure 4: Agreement with the native model on 500 untouched prompts. Because the ≈ 0.71 collapse is nearly identical for the structured interface and both quantizers — one of which is benchmark-lossless — it is not evidence that any of them is worse than the others.

B.9 The completed local-control matrix

The local matrix is evaluated at the task-macro level using matched seed IDs. Three-seed arms are exploratory and necessarily have wide intervals; only A0 and the selected norm-restoration arm A6 were extended to eight matched seeds. Intervals below are two-sided paired-t intervals over seed-level task-macro differences and are unadjusted across the exploratory matrix.

Table 10: Seed-level audit of the local intervention matrix. "Unresolved" means insufficient evidence, not equivalence. A6 is the only positive arm whose paired-seed interval excludes zero; A2 is the only established loss.

arm	seeds	Δ capability	paired-t 95% CI	reading
A1	3	$+0.0030$	$[-0.0116, +0.0176]$	unresolved
A2	3	$\mathbf{-0.0352}$	$[\mathbf{-0.0595}, \mathbf{-0.0109}]$	loss
A3	3	-0.0134	$[-0.0376, +0.0108]$	unresolved
A4-E	3	$+0.0005$	$[-0.0205, +0.0215]$	unresolved
A4-F	3	$+0.0031$	$[-0.0096, +0.0159]$	unresolved
A4-O	3	$+0.0077$	$[-0.0239, +0.0393]$	unresolved
A5	3	-0.0042	$[-0.0252, +0.0169]$	unresolved
A6	**8**	$\mathbf{+0.0089}$	$[\mathbf{+0.0011}, \mathbf{+0.0167}]$	**win**
A7	3	$+0.0004$	$[-0.0049, +0.0057]$	unresolved
A0-CE	3	-0.0065	$[-0.0168, +0.0038]$	unresolved
A2-CE	3	-0.0139	$[-0.0385, +0.0108]$	unresolved
A3-CE	3	-0.0041	$[-0.0244, +0.0161]$	unresolved
A4-S	3	-0.0060	$[-0.0269, +0.0148]$	unresolved
A7-F	3	-0.0007	$[-0.0150, +0.0136]$	unresolved

The matched CE control isolates the top-16 objective itself: A2-CE versus A0-CE has $\Delta = -0.0073$ with 95% CI $[-0.0282, +0.0135]$. We therefore do not claim that top-16 supervision itself is harmful; the severe A2 failure arises when the remaining vocabulary is left unconstrained.

Table 11: Capability and fidelity for the controls needed by the main-text mechanism argument. A6 improves capability while *all five* principal fidelity proxies prefer A0. A5 sharply improves the activation-weighted quadratic without an established capability gain, while the frequent-row cache A7-F gives the best KL, perplexity, and agreement among these controls without an established capability gain.

arm	intervention	MEAN_ACC	teacher KL	PPL	top-1	matrix MSE	$\mathbf{C}_h$-weighted
A0	baseline	0.5418	0.1628	16.614	0.8411	1.41×10^{-4}	9.669
A2	top-16 KL, no CE	0.5073	1.8798	65.010	0.7179	2.32×10^{-4}	8.613
A5	exact top-8 $\mathbf{C}_h$ channel	0.5383	0.1510	16.419	0.8424	1.48×10^{-4}	**2.721**
A6	**norm restoration**	**0.5507**	0.1864	16.828	0.8262	1.72×10^{-4}	12.324
A7-F	2,048 frequent-row cache	0.5418	**0.1346**	**16.216**	**0.8855**	1.56×10^{-4}	4.806

Thus the selected A6 intervention improves capability by $+0.0089$ across eight matched seeds while worsening teacher KL, perplexity, top-1 agreement, matrix MSE, and $\mathbf{C}_h$-weighted error relative to A0. The benchmark resolves that capability effect, whereas the A5 and A7-F capability comparisons remain unresolved at three seeds.

B.10 THE TWO SECOND MOMENTS

A hazard worth stating separately because we conflated the two objects once. The *uncentered* head-input second moment $\mathbb{E}[hh^\top]$ has effective rank $\mathbf{2.84}$ with 59% of its energy in one direction — this is the object in Equation 1. The *centered* covariance $\mathbb{E}[hh^\top] - \mu\mu^\top$ has effective rank $\mathbf{23.31}$ with 15.5% in one direction, an $8\times$ difference. Any argument quoting 2.84 while using a centered covariance is quoting the wrong number. An independent measurement on a different corpus slice and split reproduces the uncentered figures to two significant digits (effective rank 2.868, top-1 energy 58.6%, top-8 74.3%).

C PRE-REGISTRATIONS AND SCORECARD

The two prospective pool registrations were timestamped before their corresponding capability stages. Their staged driver refuses to run capability before a complete fidelity artifact, so the order is enforced in code (Appendix A). After the local norm-restoration arm was selected, we separately preregistered non-selected replications before any replication training run was launched. We report all completed registered outcomes, including refutations and unresolved tests, and explicitly identify the preregistered analysis that was not carried out in this paper.

C.1 WHAT WAS FIXED BEFORE CAPABILITY

For P1 (Qwen3.5-0.8B), we fixed 48 deployable candidates and three mechanism controls by name, the 16-metric battery, 64 WikiText-103 windows, development and frozen task lists, the above-chance keep rule, the 0.05 viability drop with a sensitivity grid, and all selector statistics. The three mechanism controls were declared non-deployable and excluded from selector statistics.

P2 (Qwen2.5-3B) fixed the same objects for 22 deployable candidates concentrated in the high-rate region. Its registration explicitly calls it a scale replication, not an architecture replication. The tied subset was to be chosen mechanically from family medians and fidelity extremes.

For the later local replication, the two arms were fixed in advance as A0 (flat $k = 64$ codes) and A6 (norm-restored initialization). The registered primary statistic was a one-sided paired t-test on the nine-task task-macro capability difference, in the direction A6 $>$ A0. The seed lists and five fidelity diagnostics—teacher KL, perplexity, top-1 agreement, matrix MSE, and $\mathbf{C}_h$-weighted error—were fixed before training.

C.2 PREREGISTERED LOCAL REPLICATION

Because norm restoration was selected from an exploratory Qwen screen, its eight-seed result is treated as selected discovery rather than a confirmatory test. We therefore preregistered fresh A0-versus-A6 comparisons before training on SmolLM2-1.7B, Llama-3.2-1B, and fresh Qwen2.5-0.5B seeds.

The SmolLM2 result is positive on $11/12$ seed pairs (one-sided $p = 1.6 \times 10^{-4}$), while four of the five registered fidelity metrics prefer A0; $\mathbf{C}_h$-weighted error is the exception. The fresh-Qwen registered test is reported as failed. A post-hoc analysis attributes much of its variance to BoolQ, but we do not replace the registered statistic with the BoolQ-free result.

C.3 HOW THE THESIS CHANGED

The original broad claim—that standard fidelity metrics fail to select viable candidates—did not survive. Distribution metrics select usefully over broad pools, and direct capability selection becomes useful when the evaluation has enough documents. The surviving claim is narrower: screening, broad ranking, and local design selection are different measurement regimes. Fidelity is excellent in the

Table 12: Preregistered replication of the norm-restoration effect. The original Qwen row is shown only as the selected discovery that motivated the registered tests. SmolLM2 provides the independent-architecture confirmation; Llama-3.2-1B is unresolved, and the fresh Qwen redraw fails the registered directional test.

run	seeds	Δ(A6−A0)	95% CI	verdict
SmolLM2-1.7B	12	$\mathbf{+0.0050}$	$[\mathbf{+0.0029}, \mathbf{+0.0071}]$	**confirmed**
Llama-3.2-1B	12	$+0.0036$	$[-0.0040, +0.0112]$	underpowered
Qwen2.5-0.5B, fresh seeds	8	-0.0035	$[-0.0149, +0.0079]$	failed
original Qwen, selected discovery	8	$+0.0089$	$[+0.0011, +0.0167]$	discovery

first, can inherit compression rate in the second, and can reverse a resolved local choice in the third. The selected Qwen reversal is independently confirmed by the preregistered SmolLM2 result.

C.4 Completed follow-ups and remaining open work

The 3B frozen suite is complete for all 22 candidates, the retrospective evaluation is complete for all 47 arms at both document budgets, and all local-matrix arms landed. The original norm-restoration result was selected from the three-seed screen and then extended to eight matched seeds after a power calculation. The preregistered replication is also complete: SmolLM2 confirms the effect, Llama-3.2-1B is underpowered, and the fresh-Qwen redraw fails the registered directional test.

Remaining open work includes the mechanically specified tied-subset analysis, further independent replication of the local effect, and population-level estimates for the exploratory three-seed arms.

D Related work in full

Beyond perplexity in compressed-model evaluation. LLM-KICK (Jaiswal et al., 2024) shows that similar perplexities can conceal different losses in knowledge, reasoning, generation, retrieval, and summarization. ACBench (Dong et al., 2025) extends compressed-model evaluation to 12 agentic tasks spanning four capability categories and 15 models. Related work also studies differences between quantization and pruning strategies (Dettmers & Zettlemoyer, 2023). These works broaden the capability evaluation used to judge a compressed model. Our question is complementary: given several compressed candidates, under what conditions can a cheap measurement reliably *select* among them?

Our contribution is to make that selection problem explicit. Three distinctions are central.

1. **We measure the target's resolution as well as the proxy.** A small $\rho(M, A)$ is difficult to interpret when the capability ranking A itself is unstable over the same candidate pool. We therefore report split-half ranking reliability and its descriptive Spearman–Brown correction, together with a document-budget intervention. This follows classical reliability-attenuation analysis (Spearman, 1961; Brown, 1910), but we use reliability only as a diagnostic, not as a selector ceiling. tinyBenchmarks (Polo et al., 2024) also studies reducing evaluation examples while preserving aggregate benchmark estimates; our target is instead the candidate-selection decision and its regret.
2. **We separate screening from selection among viable candidates.** Prior compression studies establish that fidelity measurements can miss capability degradation or that richer downstream evaluation can reveal it. We separately measure screening AUROC, rate-controlled within-viable ranking, target resolution, and selection regret against a capability selector using the same number of evaluation documents. The prospective registrations and their complete scorecard are reported in Appendix C, including predictions that were confirmed, qualified, refuted, or left unresolved.
3. **We separate geometric non-identifiability from statistical non-resolution.** For reconstruction error, the limitation can be structural: the one-sided geometry of Section 3 does not imply an ordering of downstream behavior, and equal-MSE controls realize this failure directly. For distribution-side metrics, the empirical picture is different: they are strong screens and often useful broad-pool rankers, but can still select the wrong local design at fixed bytes. Treating

these two mechanisms as a single statement that "fidelity metrics are unreliable" would therefore obscure the distinction the experiments are designed to measure.

Activation-aware compression, and why $\mathbf{C}_h$-weighting is not our contribution. The mismatch between ordinary matrix reconstruction and model behavior is well established. GPTQ uses approximate second-order information (Frantar et al., 2022); Fisher-weighted SVD weights reconstruction by task sensitivity (Hsu et al., 2022); and activation-aware low-rank methods such as ASVD and SVD-LLM use activation statistics and whitening (Yuan et al., 2023; Wang et al., 2025). Our contribution is not the observation that activation geometry matters. Rather, Equation 1 is *one-sided*: even the corresponding activation-weighted quadratic does not supply the converse required to rank candidates by capability.

The local controls make the distinction concrete (Appendix B.9). Protecting the top-$\mathbf{C}_h$ subspace sharply reduces the activation-weighted error without an established capability gain, while norm restoration improves capability even though its $\mathbf{C}_h$-weighted error worsens. Thus a better value of the quadratic surrogate need not imply a better local deployment choice. The Frobenius-optimal rank-r correction is given by Eckart–Young (Eckart & Young, 1936), while activation-weighted and cross-state forms are related to reduced-rank regression (Izenman, 1975; Paruolo, 2000); the rank-8 controls in Table 8 illustrate why optimizing one such quadratic is not itself a guarantee about functional behavior.

Anisotropy. Transformer representations are often strongly anisotropic (Ethayarajh, 2019; Gao et al., 2019), although the extent and universality of this phenomenon remain model-dependent: Godey et al. (2024) argue that anisotropy may arise intrinsically from self-attention, while Machina & Mercer (2024) report substantially more isotropic regimes in large Pythia models. Our result is therefore conditional, not universal: when $\mathbf{C}_h$ is ill-conditioned, equal Frobenius errors can have widely different activation-weighted costs. We measure the relevant second-moment geometry directly in our language-model controls (Table 7) rather than assuming anisotropy.

Top-k and margin-based distillation. BiLD (Li et al., 2025) studies task-specific LLM distillation using only top-k teacher and student logits, motivated by noise in the long logit tail, and reports strong results with $k = 8$. Our top-16 experiments answer a different question. The unrestricted-support A2 arm is an established loss, but A3 is unresolved at the seed level (Appendix B.9). Adding a matched full-vocabulary cross-entropy term reduces A2's perplexity from 65.0 to 14.7, and the isolating A2-CE versus A0-CE capability comparison is unresolved. We therefore do not conclude that restricting supervision to top logits is intrinsically harmful. Instead, the original failure is explained by leaving the remaining $151{,}920$ logits unconstrained while they share the same softmax normalizer. Likewise, the preregistered margin-based *measurement* reduces algebraically to $1-$top-1 agreement, and its lower-margin variants do not provide a superior selector. These observations concern selection and measurement, not whether top-k targets are useful for distillation.

Compressing the lexical interface. Large lexical tables have been compressed using low-rank factorization (Lan et al., 2019), hashing and shared parameter pools (Tito Svenstrup et al., 2017), low-rank-plus-binary residual coding (Wang et al., 2023), frequency-aware sparse coding (Tamura et al., 2024), and post-training quantization (Frantar et al., 2022; Lin et al., 2025; Dettmers et al., 2022; 2023). One-shot pruning methods such as SparseGPT and Wanda (Frantar & Alistarh, 2023; Sun et al., 2024) are an adjacent family evaluated with similar reconstruction and activation statistics. Weight tying (Press & Wolf, 2017; Inan et al., 2016) makes the lexical table simultaneously an input representation and an output classifier, which introduces the additional input/state pathway analyzed in Section 3. Product and additive quantization (Jegou et al., 2010; Johnson et al., 2019) provide further byte-axis baselines. We treat these approaches as candidate families whose *selection* must be evaluated; we propose no new compressor.

Evaluation practice. We use `lm-evaluation-harness` (Gao et al., 2021) and retain per-example and per-task scores rather than only an aggregate benchmark mean. This enables paired uncertainty estimates, task-macro evaluation, split-half ranking reliability, and disjoint-half selection regret. Our treatment of training variation follows standard warnings about seed sensitivity (Reimers & Gurevych, 2017); we additionally distinguish uncertainty from evaluation documents from uncertainty

across independently trained seeds, because either can be the binding source of resolution for a model-selection claim.

E ADDITIONAL EMPIRICAL ANALYSES

This appendix reports additional mechanism and scope analyses supporting the main results: stage-by-stage selector behavior, tied-path failures of the direct quadratic, controlled equal-MSE constructions, and cross-architecture language-model and vision extensions.

E.1 STAGE-BY-STAGE SELECTOR VALUE

Table 13: Within-viable ρ of the best metric at each stage of the chain in Equation 1. **The activation-aware step — table geometry to logit geometry, the correction GPTQ, Fisher-weighted SVD and SVD-LLM adopt — buys nothing as a selector in any pool**, and is worse on the only pool that discriminates. The whole gain lives in the softmax step, and only where the candidate pool mixes compressor families. P2's viable region is a single RTN rate ladder, so every stage ties there by construction and the column is uninformative.

stage of the chain	P1 dev	P1 held out	P2 dev	**R dev**
table geometry ($\|\Delta\|_F^2$, matrix cos)	$+0.461$	**+0.732**	$+0.828$	$+0.471$
logit geometry ($\mathrm{Tr}(\Delta \mathbf{C}_h \Delta^\top)$, logit MSE/cos)	$+0.462$	$+0.729$	$+0.828$	$+0.436$
distribution geometry (KL, JSD, top-5, argmax NLL)	**+0.486**	$+0.717$	$+0.828$	**+0.792**
benchmark self-consistency (split-half)	*+0.493*	*+0.546*	*+0.365*	*+0.610*

The activation-aware step changes the best within-viable correlation by $+0.001$, -0.003, 0.000, and -0.035 across P1 dev, P1 held out, P2 dev, and R dev. Split-half regret on R dev is 0.0014 $[0.0000, 0.0053]$ for distribution geometry, 0.0084 $[0.0029, 0.0138]$ for table geometry, and 0.0130 $[0.0080, 0.0180]$ for direct-logit geometry; the distribution and logit intervals are disjoint. The direct-logit point estimate is not distinguishable from random choice (0.0113 $[0.0002, 0.0220]$).

Decision equivalence between fidelity metrics. Rank correlations can exaggerate differences between metrics that ultimately make the same deployment choice. We therefore measure a stronger notion of decision equivalence: for each pair of fidelity metrics, sample random viable candidate subsets and ask whether both metrics select the same candidate on *every* sampled subset.

The result is strongly pool-dependent. Most metric pairs are decision equivalent on P1, and several remain equivalent on the single-family P2 ladder. None of the ten tested pairs is decision equivalent on the more heterogeneous retrospective pool. Thus the apparent redundancy of different fidelity geometries is itself conditional on the candidate pool: metrics that collapse to the same choice on a rate-dominated pool can disagree once the pool contains qualitatively different trained and post-hoc constructions.

E.2 TIED COMPOSITIONAL-CODE FAILURE

For tied embeddings, $f_E(x) = Eh_E(x)$ and $\delta f = \Delta h_E + EDh_E[\Delta] + O(\|\Delta\|^2)$. The activation-weighted quadratic measures the first term only. The three compositional-code arms are exactly the outliers in Table 15; every other family has a direct-to-realized logit-MSE ratio between 0.66 and 1.34. On the head-only prospective pools, where hidden states are fixed by construction, $\mathbf{C}_h$-weighted error is close to teacher KL. The observed failure is therefore specific to tied artifacts whose input-side error is structured across rows.

E.3 EQUAL RECONSTRUCTION ERROR, UNEQUAL FUNCTIONAL DAMAGE

The top-, isotropic-, and bottom-eigenspace perturbations use identical Frobenius error. On P1, the bottom and isotropic controls are both viable yet differ by 0.0260 capability and $478\times$ teacher KL at bit-identical MSE. The top control is behaviorally destroyed and supports a screening, not a

Table 14: Decision equivalence of fidelity metrics. The middle column counts metric pairs that select the identical candidate on 100% of sampled viable subsets.

Pool / suite	Identical metric pairs	Subsets
P1 / dev	10/15	4,000
P1 / held out	10/15	4,000
P2 / dev	6/15	2,008
P2 / held out	6/15	2,982
R / dev	0/10	4,000
R / held out	0/10	4,000

Table 15: The three compositional-code arms of R (group size G, 256 centroids), against the median of the 38 viable arms in the same pool. They *out-fit* most viable arms on both matrix-side metrics while destroying the output distribution. The last column is the diagnosis: the ratio of the direct output-path quadratic to the *realized* logit error. For every other family in the pool that ratio is 0.66–1.34; for this one it is 0.23–0.26, so the metric is seeing about a quarter of the damage.

arm	matrix MSE	$\mathbf{C}_h$-w err	logit MSE	teacher KL	top-1	$\mathbf{C}_h$-w / logit MSE
compose-$G8$	1.686e−4	**2.925**	12.94	**4.493**	0.054	**0.226**
compose-$G16$	1.550e−4	**2.424**	10.74	**3.744**	0.102	**0.226**
compose-$G32$	1.364e−4	**1.932**	7.51	**2.487**	0.250	**0.257**
median of the 38 viable arms	1.508e−4	*6.354*	*5.894*	*0.132*	*0.867*	≈ 1.05

viable-ranking, claim. At 3B the corresponding viable pair differs by 0.0048, below the measured seed standard deviation, so it is not counted as a ranking replication. These controls use one noise seed and establish reachability of the theorem's worst case, not prevalence.

E.4 Cross-architecture scope extension

The main prospective candidate pools use Qwen models. Separately from the preregistered local SmolLM2 replication in Appendix C.2, we repeated the *pool-level* analysis post hoc on SmolLM2-1.7B and Llama-3.2 at 1B and 3B scale. These experiments are scope extensions rather than preregistered pool tests: they ask whether the screening, rate-confounding, and resolution phenomena are specific to the Qwen lineage.

Table 17 summarizes the resulting pools. Screening transfers cleanly. Across all complete architecture pools, the fidelity metrics separate broken from viable candidates with AUROC between .996 and 1.000. Among viable candidates, teacher KL also orders the comparisons that the capability evaluation can statistically resolve: resolvable-pair accuracy is .925 on Llama-3.2-1B and 1.000 on the complete SmolLM2 pools. The Llama-3.2-3B development result is reported only as provisional because the pool contains 16 of the planned 48 candidates.

The architecture extension also clarifies what does and does not generalize about the pool limit. In the original analysis, the middle Llama-3.2-1B rate stratum appeared to show a small negative within-stratum association, $\rho = -.029$. That model's native BoolQ score, however, is close to the majority-label baseline, so BoolQ acts largely as a label-prior readout. When the pool is rebuilt without BoolQ—including recomputing the viability boundary rather than merely deleting the task after selection—the same stratum changes to $\rho = .657$. Its low/middle/high rate-stratum correlations become $.643/.657/.771$, and all 18 pairs declared resolvable are ordered correctly by teacher KL.

The other architectures do not produce this artifact. The provisional Llama-3.2-3B strata have correlations $.800/.600/.200$, while the SmolLM2 strata have $.900/.800/.700$. Thus none of these architecture extensions provides evidence for a within-stratum sign reversal once the BoolQ issue is handled.

They do, however, reproduce the weaker and more general pool-limit phenomenon: pooled association can be dominated by between-rate variation. Only .179 of the rank covariance in the Llama-3.2-3B pool and .110 in the SmolLM2 pool is within rate strata, corresponding to approximately 82% and 89% of the pooled covariance arising between strata. Screening remains useful under the same

Table 16: The designed controls. Prediction H3 registered a teacher-KL top/bottom ratio ≥ 3; measured $19{,}429\times$ and $12{,}939\times$. **Read the two rows above each rule separately from the third.** `bot` and `iso` are both *viable* on P1 dev and differ by 0.0260 accuracy at bit-identical MSE — a within-viable ordering failure. `top` is a destroyed model (top-1 agreement 0.002, PPL $2.2\text{e}21$), so the $0.0864/0.1380$ spans it creates are a *screening* failure of matrix MSE, not a ranking one. Single noise seed; these are constructions, not replicated measurements.

model (native MEAN_ACC)	control	matrix MSE	$\mathbf{C}_h$-w	teacher KL	dev	held
Qwen3.5-0.8B	**bot-$\mathbf{C}_h$** *(viable)*	1.9604e−4	0.00756	**0.0024**	**0.5212**	**0.4590**
0.5220 / 0.4625	**iso** *(viable)*	1.9604e−4	3.235	1.147	0.4952	0.3990
	top-$\mathbf{C}_h$ *(broken)*	1.9604e−4	209.98	**45.93**	**0.4348**	**0.3215**
	ratio top/bot	1.000×	27,790×	**19,429×**	Δ.0864	Δ.1380
Qwen2.5-3B	bot-$\mathbf{C}_h$ *(viable)*	7.9326e−6	0.00309	**0.0009**	**0.6468**	—
0.6488	iso *(viable)*	7.9326e−6	0.2122	0.0679	0.6420	—
	top-$\mathbf{C}_h$ *(broken)*	7.9326e−6	41.07	**11.02**	**0.5268**	—
	ratio top/bot	1.000×	13,297×	**12,939×**	Δ.1200	—

Table 17: Post-hoc cross-architecture scope extension. AUROC ranges over the fidelity metrics. "Resolvable accuracy" is the teacher-KL pairwise accuracy restricted to pairs declared resolvable by the resolution index. The Llama-3.2-3B row is provisional.

Pool	Cand./viable	Native	AUROC	ρ_{KL}	Resolvable acc.	Regret	RPF	Violations
Llama-3.2-1B / dev	48/18	.5588	.998–1.000	.754	.925 (37/40)	.0000	.261	0/918
Llama-3.2-3B / dev (provisional)	16/14	.6464	1.000	.591	1.000	.0023	.286	0/546
SmolLM2-1.7B / dev	48/14	.6381	.996–1.000	.916	1.000	.0000	.462	0/546
SmolLM2-1.7B / held out	48/15	.5691	1.000	.825	1.000	.0052	.190	0/630

matched-document accounting: the best observed screen-then-select improvement is $20.6\times$ on Llama-3.2-1B, $15.8\times$ on SmolLM2-1.7B, and $3.3\times$ on the provisional Llama-3.2-3B pool.

These experiments therefore broaden the scope of the screening and rate-confounding conclusions, but we do not use them to claim that local fidelity/capability reversals occur universally across architectures.

E.5 Outside-language replication

As an outside-language scope check, we evaluate ImageNet-1K (20 images per class; $20{,}000$ total). ResNet50 and ViT-B/16 pools contain 20/10 and 21/12 total/viable candidates, respectively. Teacher-KL screening AUROC is $1.000/.963$, viable-pool Spearman ρ is $.966/.846$, and resolvable-pair accuracy is $1.000/1.000$; selection regret is $.0000/.0009$. The resolution bound has $0/315$ and $1/462$ violations, respectively.

F All resolution curves and screening economics

The protocol (draw half, disjoint reference half, 400 subsamples per budget, gaps and variances from the reference or draw half as appropriate) is in Appendix A. "RPF" is the resolvable-pair fraction of Equation 7 at $\delta = .05$; "inv. unres." and "inv. res." are measured pairwise inversion rates among the pairs the index calls unresolvable and resolvable. At the largest budget the subsample is the whole draw half, so there is no sampling variation left and both inversion rates are zero by construction.

F.1 Budget curves, all six pool/suite combinations

pool	suite	docs	rel.	τ	regret	RPF	inv. unres.	inv. res.	bound
P1 (0.8B)	dev	125	+0.019	0.236	0.0106	0.000	0.385	–	0.256
		250	+0.239	0.292	0.0075	0.000	0.304	–	0.204
		500	+0.519	0.354	0.0056	0.028	0.213	0.0000	0.152

pool	suite	docs	rel.	τ	regret	RPF	inv. unres.	inv. res.	bound
		1,250	+0.821	0.466	0.0032	0.182	0.019	0.0000	0.097
P1 (0.8B)	frozen	100	-0.204	0.162	0.0254	0.000	0.377	–	0.376
		200	-0.345	0.242	0.0211	0.000	0.283	–	0.302
		350	-0.263	0.321	0.0148	0.000	0.194	–	0.246
		800	-0.026	0.418	0.0070	0.000	0.000	–	0.190
P2 (3B)	dev	125	-0.085	0.091	0.0137	0.000	0.385	–	0.172
		250	+0.045	0.113	0.0155	0.000	0.300	–	0.129
		500	+0.234	0.152	0.0176	0.000	0.199	–	0.080
		1,250	+0.545	0.109	0.0232	0.152	0.000	0.0000	0.028
P2 (3B)	frozen	100	-0.301	0.083	0.0168	0.000	0.451	–	0.145
		200	-0.418	0.133	0.0141	0.000	0.361	–	0.128
		350	-0.355	0.162	0.0116	0.000	0.258	–	0.108
		800	+0.152	0.197	0.0130	0.000	0.000	–	0.082
R (0.5B)	dev	125	+0.046	0.238	0.0268	0.000	0.393	–	0.706
		250	+0.113	0.302	0.0242	0.000	0.335	–	0.544
		500	+0.291	0.390	0.0195	0.000	0.263	–	0.347
		1,250	+0.534	0.507	0.0139	0.062	0.172	0.0006	0.130
		2,250	+0.670	0.568	0.0102	0.192	0.112	0.0000	0.056
		3,883	+0.707	0.592	0.0084	0.299	0.000	0.0000	0.028
R (0.5B)	frozen	100	-0.201	0.071	0.0172	0.000	0.416	–	0.675
		200	-0.147	0.104	0.0132	0.000	0.343	–	0.493
		350	-0.238	0.112	0.0098	0.000	0.272	–	0.329
		800	-0.299	0.128	0.0064	0.013	0.180	0.0000	0.175
		1,550	-0.415	0.128	0.0063	0.028	0.108	0.0000	0.115
		2,660	-0.500	0.104	0.0077	0.047	0.000	0.0000	0.093

Table 18: Budget curves. Reliability, τ and regret are scored against a disjoint reference half; RPF, inversion rates and the bound are the calibration columns of Appendix A. Across all six curves the bound is violated 0 times in 11,734 pair-by-budget checks.

Alternative resolution indices and calibration versus validation. The main analysis uses the plug-in paired variance in the resolution index. To test whether its apparent calibration depends on that choice, we repeat all 11,734 pair-by-budget checks with an empirical-Bernstein construction and with a distribution-free Hoeffding bound. Table 19 reports the result.

The zero-violation conclusion is therefore not specific to the plug-in index. The difference is statistical efficiency. The paired per-document variance in these experiments is approximately .05–.15, whereas the distribution-free Hoeffding construction must use the full bounded range. Its exponent is consequently roughly 7–19$\times$ smaller, and it certifies only two pair-budget cells. This is the expected price of a distribution-free guarantee rather than evidence that the tighter indices are miscalibrated.

We separately distinguish calibration against the draw-half population from generalization to a disjoint half. Across the disjoint-half comparison, 20% of candidate pairs change sign between halves. Every observed sign change is an *unresolvable* comparison according to the index; sign agreement on the resolvable pairs is 1.000. Thus the resolution label predicts not only subsampling instability within the calibration half but also which pairwise conclusions survive an independent evaluation half.

F.2 Screening economics, all six pools

At a fixed total budget of candidate-documents, we keep the top-m candidates by teacher KL and divide the budget equally among them; $m = K$ is no screening. "sat." marks a budget larger than the documents available per candidate. ϵ_{screen} is the screening regret of the survivor set, and "broken kept" counts survivors that the pre-declared viability rule labels broken.

Table 19: Calibration under three resolution indices, aggregated over the six committed pool/suite budget curves. "Resolvable" is the number of pair-by-budget cells certified at the fixed inversion tolerance.

Index	Checks	Violations	Resolvable	Inv. resolvable	Inv. unresolvable
Plug-in variance	11,734	0	562	4.7×10^{-5}	.228
Empirical Bernstein	11,734	0	399	3.3×10^{-5}	.224
Hoeffding	11,734	0	2	0	.217

pool	suite	total docs	regret at m survivors			
P1 (0.8B)	dev	$m =$	6	12	24	48
		1,488	0.0270	0.0244	0.0235	0.0361
		2,976	0.0268	0.0245	0.0236	0.0262
		6,000	0.0265	0.0239	0.0204	0.0244
		14,976	sat.	0.0280	0.0232	0.0227
		30,000	sat.	sat.	0.0280	0.0224
		60,000	sat.	sat.	sat.	0.0280
		ϵ_{screen}	0.0256	0.0168	0.0000	0.0000
		broken kept	0	0	2	25
P1 (0.8B)	frozen	$m =$	6	12	24	48
		960	0.0162	0.0164	0.0185	0.0235
		1,920	0.0161	0.0163	0.0162	0.0265
		3,840	0.0157	0.0153	0.0142	0.0182
		9,600	sat.	0.0160	0.0102	0.0149
		19,200	sat.	sat.	0.0160	0.0119
		38,400	sat.	sat.	sat.	0.0160
		ϵ_{screen}	0.0060	0.0050	0.0000	0.0000
		broken kept	0	0	3	26
P2 (3B)	dev	$m =$	2	5	11	22
		682	0.0069	0.0047	0.0073	0.0218
		1,364	0.0071	0.0039	0.0075	0.0146
		2,750	sat.	0.0046	0.0070	0.0123
		6,864	sat.	sat.	0.0078	0.0123
		13,750	sat.	sat.	0.0120	0.0092
		27,500	sat.	sat.	sat.	0.0120
		ϵ_{screen}	0.0056	0.0000	0.0000	0.0000
		broken kept	0	0	0	7
P2 (3B)	frozen	$m =$	2	5	11	22
		440	0.0111	0.0102	0.0071	0.0176
		880	0.0118	0.0109	0.0073	0.0137
		1,760	sat.	0.0114	0.0078	0.0128
		4,400	sat.	sat.	0.0097	0.0131
		8,800	sat.	sat.	0.0120	0.0107
		17,600	sat.	sat.	sat.	0.0120
		ϵ_{screen}	0.0080	0.0080	0.0000	0.0000
		broken kept	0	0	0	6
R (0.5B)	dev	$m =$	5	11	23	47
		4,559	0.0021	0.0078	0.0129	0.0199
		9,118	0.0019	0.0050	0.0113	0.0175
		18,236	0.0020	0.0017	0.0070	0.0141
		45,590	sat.	sat.	0.0044	0.0069
		91,227	sat.	sat.	sat.	0.0047
		182,501	sat.	sat.	sat.	0.0028
		ϵ_{screen}	0.0000	0.0000	0.0000	0.0000

Table 21: Median total scored documents required to reach regret $\leq .005$ as the candidate-set size grows. Screened runs first retain approximately $K/4$ candidates by fidelity and then allocate capability evaluation to the survivors.

Pool / suite	K	Unscreened	Screened to $\approx K/4$
R / dev	4	2,000	250
R / dev	8	8,000	500
R / dev	16	29,600	2,000
R / dev	32	82,128	28,664
R / held out	16	28,080	800
P2 / held out	8	5,200	200
Whole-model pool	4	1,800	450

pool	suite	total docs	regret at m survivors			
		broken kept	0	0	0	7
R (0.5B)	frozen	$m =$	5	11	23	47
		3,102	0.0266	0.0185	0.0208	0.0252
		6,251	0.0279	0.0167	0.0204	0.0249
		12,502	0.0352	0.0155	0.0201	0.0243
		31,255	sat.	sat.	0.0158	0.0171
		62,510	sat.	sat.	sat.	0.0133
		125,020	sat.	sat.	sat.	0.0081
		ϵ_{screen}	0.0106	0.0000	0.0000	0.0000
		broken kept	0	0	0	9

Table 20: Screening economics at equal document cost.

Scaling with candidate-set size. The fixed-pool comparisons above leave open how selection cost grows as the number of candidates K increases. We therefore subsample candidate sets of different sizes and measure the median total number of scored documents required to reach selection regret at most .005. Table 21 compares uniform capability evaluation with screening first to approximately $K/4$ candidates.

The cost increase with K is substantial because larger candidate sets introduce more near competitors while dividing a fixed evaluation budget more thinly. Screening can change that scaling dramatically when it safely removes distant candidates: at $K = 16$ on R/held-out, the median requirement falls from 28,080 to 800 scored documents, and at $K = 8$ on P2/held-out it falls from 5,200 to 200. The advantage is not guaranteed. On P1/dev at $K = 16$, the screened experiment remains censored without reaching the .005 target, consistent with Proposition 2: screening helps only when the reduction in candidate count outweighs its screening regret.

Deployment tolerance and ϵ-optimal selection. Exact identification of the empirical best candidate can be unnecessarily strict when several deployments are practically interchangeable. We therefore also report whether the selected candidate lies within a deployment tolerance ϵ of the reference-half optimum. Table 22 uses the largest available evaluation budget for each committed pool/suite.

The tolerance view separates "identify the unique best" from "make a deployment decision whose loss is negligible." Four of the six pool/suite combinations are within .01 of the reference optimum at the largest budget, despite several having low resolvable-pair fractions. Conversely, P2/dev remains outside even a .02 tolerance.

A second consequence is that spending more benchmark documents does not mechanically dominate a proxy selector when the ranking itself is poorly resolved. On three of the six pool/suite combinations, the zero-document proxy has lower point-estimate regret than capability selection on the full reference-half budget. We do not interpret this as universal superiority of the proxy; it is another manifestation of finite-resolution best-arm selection, where noisy attempts to distinguish near ties can be less stable than a coarse but deterministic rule.

Table 22: Selection under deployment tolerance. $P(\epsilon\text{-optimal})$ is evaluated at the largest available capability budget. The final column is the regret of the zero-benchmark-document proxy selector.

Pool / suite	Docs	Mean regret	$P(\epsilon$-opt.), $\epsilon = .005$	$\epsilon = .01$	$\epsilon = .02$	Proxy-only regret
P1 / dev	1,250	.0032	1	1	1	.011
P1 / held out	800	.0070	0	1	1	.011
P2 / dev	1,250	.0232	0	0	0	.002
P2 / held out	800	.0130	0	0	1	.001
R / dev	3,883	.0084	0	1	1	.000
R / held out	2,660	.0077	0	1	1	.010

F.3 THE LOCAL MATRIX BY SEED GROUP

Table 23: Norm restoration versus baseline at matched served bytes. The discovery seeds are the three that existed when the arm was chosen; the confirmation seeds were run afterwards, once a power calculation had fixed the total at eight.

seed group	n	task-macro Δ	95% paired-t CI	one-sided p	positive seeds
all eight	8	$+0.0089$	$[+0.0011, +0.0167]$	0.016	7/8
discovery	3	$+0.0073$	$[-0.0070, +0.0217]$	0.079	3/3
run afterwards	5	$+0.0098$	$[-0.0046, +0.0242]$	0.065	4/5

Document-level resolution index of the same comparison: $\nu = 13.6$ ($\Delta = 0.0079$ over 7,500 documents, $\sigma_{\text{eff}} = 0.0015$). Seed-level index: $\nu = 3.6$. The benchmark resolves this effect; the training-seed variance is what makes it marginal.

Which source of resolution binds is model-dependent. The local Qwen example has two distinct sources of uncertainty: finite evaluation documents and variation over trained seeds. To test whether their relative importance is itself stable across architectures, we decompose the variance of the paired A6–A0 effect for the preregistered SmolLM2, Llama-3.2-1B, and fresh-Qwen runs, alongside the original Qwen experiment. Within each trained seed we estimate the document-level standard error from paired per-example differences. We then compare it with the across-seed standard deviation of the task-macro effect.

The binding source of uncertainty changes across models. For SmolLM2, the across-seed variation is no larger than the average document-level standard error, and the residual seed component is zero at the displayed precision: more evaluation documents are therefore the relevant way to sharpen the comparison. On Llama-3.2-1B and both Qwen experiments, by contrast, the estimated seed component is much larger than the document standard error. Once that regime is reached, scoring more documents from the same trained artifacts cannot resolve the population-level training claim.

This is not explained by obviously unstable language-model training: Llama-3.2-1B and SmolLM2 have similar perplexity coefficients of variation (.15% and .14%), yet very different paired capability variance. At the observed Llama effect and seed standard deviation, approximately 67 paired seeds would be required for 80% one-sided power at $\alpha = .05$.

We also vary the number of Qwen seed pairs and evaluation documents jointly, taking the sign of the all-seed paired mean as the target decision. Across the tested document budgets, the probability of selecting the correct sign is .75–.93 with two seeds, .81–.97 with three, .91–1.00 with five, and .94–1.00 with eight. Thus benchmark size and training replication are not interchangeable resources: once seed variation dominates, additional independent training runs are the effective way to increase decision resolution.

Table 24: Document- versus seed-level uncertainty for norm restoration. "Seed component" is the estimated seed-level standard deviation remaining after accounting for the mean within-seed document standard error. A value of 0 means zero at the displayed precision, not that trained models are deterministic.

Experiment	Across-seed SD	Mean document SE	Seed component	A0/A6 item churn
SmolLM2-1.7B	.0034	.0035	**.0000**	10.2%
Llama-3.2-1B	.0120	.0040	**.0113**	13.5%
Qwen original	.0094	.0036	.0087	10.8%
Qwen fresh seeds	.0137	.0035	.0132	10.7%

Table 25: Equal-MSE anisotropy across seven language models. Perturbations are matched in matrix MSE and placed in different eigenspaces of C_h. "Top/bottom" is the ratio of median teacher KL in the top versus bottom eigenvalue bands; "within" is the largest variation across six random directions within a band, and "across" is the smallest separation between bands.

Model	$\mathrm{rank}_{\mathrm{eff}}(C_h)$	Top/bottom	Within	Across
GPT-2	1.03	2.24–2.63×10^4	$1.23\times$	2.47×10^4
Qwen3-0.6B	1.55	282–283$\times$	$1.45\times$	$357\times$
Qwen2.5-0.5B	1.69	270–272$\times$	$1.37\times$	$307\times$
SmolLM2-1.7B	1.85	820–827$\times$	$1.26\times$	$924\times$
Llama-3.2-1B	2.53	127–128$\times$	$1.21\times$	$137\times$
Qwen3.5-0.8B	2.74	235–236$\times$	$1.22\times$	$259\times$
Llama-3.2-3B	3.59	$134\times$	$1.18\times$	$153\times$

F.4 Equal-MSE controls and task-level crossings

F.5 Measured selector cost

On Qwen2.5-0.5B CPU, matrix MSE costs $0.197\,\mathrm{s}$ per candidate versus $126.3\,\mathrm{s}$ for capability evaluation, a $642\times$ reduction in marginal selector cost. Including amortized setup, the measured pool-level reduction is 966–$1{,}678\times$.

Table 26: Prospective equal-MSE controls. Within each model the three perturbations have bit-identical reconstruction error; only the eigenspace of $\mathbf{C}_h$ they occupy differs.

model	perturbation	matrix MSE	teacher KL	dev acc.	frozen acc.
Qwen3.5-0.8B	bottom-8 eigenspace	1.960×10^{-4}	0.0024	.5212	.4590
	isotropic	1.960×10^{-4}	1.1469	.4952	.3990
	top-8 eigenspace	1.960×10^{-4}	45.93	.4348	.3215
Qwen2.5-3B	bottom-8 eigenspace	7.933×10^{-6}	0.0009	.6468	.6370
	isotropic	7.933×10^{-6}	0.0679	.6420	.6335
	top-8 eigenspace	7.933×10^{-6}	11.02	.5268	.3765

Table 27: Task-level crossings among viable candidates. "distinct winners" counts candidates that are $\arg\max_c w^\top a(c)$ for at least one of 2,000 Dirichlet($\mathbf{1}$) task weightings.

pool / suite	viable	tasks	Pareto-incomparable pairs	single-task winners	distinct winners
P1 / dev	23	5	.925	3	8
P1 / frozen	22	4	.792	4	6
P2 / dev	15	5	.924	5	8
P2 / frozen	16	4	.725	4	5
R / dev	40	5	.855	5	7
R / frozen	38	4	.863	4	6

Table 28: Measured CPU selector cost.

Selector	s / candidate	Relative cost
Matrix MSE	0.197	1/642
Capability	126.3	1