MEASURE THE MERGE IN THE CURRENCY YOU DEPLOY: CALIBRATED PATH DECISIONS ON A BUDGET

Anonymous authors
Paper under double-blind review

ABSTRACT

Merging two fine-tuned models by weight averaging can quietly destroy the task: averaging two capable, officially released Qwen2.5 specialists can leave a model that solves nothing, while another pair from the same family survives — and nothing readable from the endpoints, from task labels to weight distance to training logs, says which is which. Part of that cost is exact rather than empirical: the conventional 50/50 average forfeits half the endpoint gap on any asymmetric pair, because net gain over the better endpoint decomposes pointwise as dip depth minus an endpoint-asymmetry wedge. The rest must be measured along the interpolation path, but path evaluations cost GPU time: how many to buy, and what to read them in? How many: a midpoint accept/reject decision is a three-evaluation problem — three points plus one calibrated safety margin, which transfers unchanged to three external populations and is measured wider on the fourth; a denser grid raises coverage on paper but admits no more pairs at the same calibrated risk. For finding a coefficient, density buys real structure at a minimax-optimal k^{-2} rate that flattens past about six queries. In what currency: at matched wall-clock budget, rules that read the proxy-loss curve repeatedly committed coefficients below the better endpoint — the worst landing -17.0 points below it — while rules that measured the task metric directly did so far less often; where paths are flat, nothing separates the two currencies. The coefficient, not the recipe, is the operative variable: TIES and DARE collapse at full strength and recover at the measured coefficient where the path allows it. Read the path's depth once, then measure the merge in the currency you deploy.

1 INTRODUCTION

On the officially released Qwen2.5 specialist models, averaging two fine-tuned checkpoints at the conventional weight 0.5 can quietly destroy the task. On the 7B Instruct×Coder pair, generation accuracy falls from 78–82$\%$ at the endpoints to 40% at the midpoint. The pair passes every inexpensive check: same base family, byte-identical architecture and tokenizer, both released Instruct-line checkpoints. And the failure is not a property of the family — a 3B sibling pair survives averaging at 55%. Nothing readable from the endpoints — task labels, weight-space distance, training logs — separates the pair that collapses from the pair that survives. What separates them is the shape of the interpolation path, and reading that shape costs evaluations.

This is the practitioner's real problem. Merging is never free to decide: endpoint attributes and fixed recipes are cheap but indirect, and evaluating every candidate coefficient is direct but expensive. Between the two sits a measurement-design question — how many points of the path to query, and in which metric to read the answers — and it has been answered mostly by folklore. This paper answers it with measurements, and the answer is not a number. *The value of a path evaluation is set by the decision it serves and the currency it is read in, not by the query count.*

Three findings carry that claim, and each comes with something a practitioner can use.

A midpoint decision is a three-evaluation problem. Querying the two endpoints and the midpoint, and adding one calibrated safety margin, yields a screen that controls false-safe decisions on every sealed pool we test; a pair-level conformal argument (Theorem 1) extends the same cover to a

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.

Selection strategy	evals	sel. n	acc. (%)	Δ (pt)	95% CI	$<$ better ep.	worst
Task metric (generation accuracy) measured directly							
Direct-2 (endpoints only)	2	254–500	42.8	—	—	0/8	$+0.0$
Direct-3	3	169–333	43.1	$+0.4$	$[-0.5, +1.7]$	1/8	-1.4
Direct-5	5	101–200	42.9	$+0.1$	$[-1.1, +1.5]$	4/8	-2.4
Proxy loss curve read for its minimum							
CE-3	3	500	37.6	-5.1	$[-10.8, -1.7]$	7/8	-19.6
CE-5	5	500	40.1	-2.7	$[-6.5, -0.2]$	6/8	-19.6
CE-11	11	500	40.0	-2.8	$[-6.4, -0.5]$	6/8	-17.0

Table 1: Six selection strategies at one matched wall-clock budget. 8 LoRA adapter units (Qwen2.5-1.5B and -3B, four pair templates $\times$ two scales), GSM8K generation accuracy, budget $b^* = 1140.8\,\mathrm{s}$ per unit charged identically to every strategy, coefficient switching and weight instantiation included. Each strategy picks a coefficient on the selection half and is scored on the sealed half; **sel.** n is the selection-half sample size per coefficient. Δ is against Direct-2, which queries only $\alpha \in \{0, 1\}$ and keeps the endpoint winning on the selection half — a deployable rule, not a sealed-test oracle. $<$ **better ep.** counts units scoring worse than the sealed better endpoint. Intervals are exact 4^4 pair-template cluster enumerations (Appendix C). One unit contributes 74% of the Direct-5 mean gap (median unit -0.5 points): eight measurements on one task, not a ranking of merging methods.

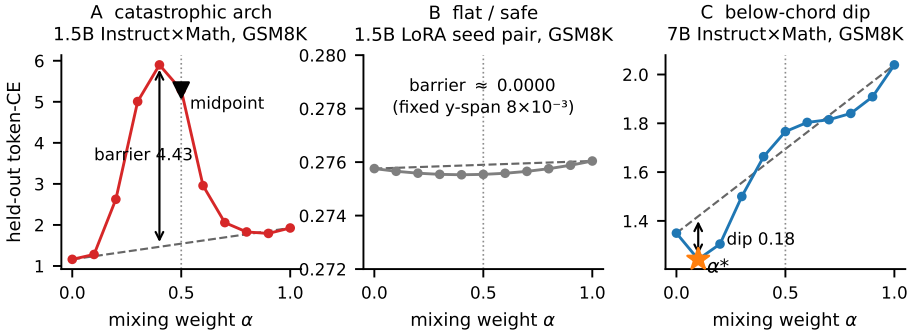

Figure 1: Three path shapes, three different decisions. Held-out loss along the interpolation between two fine-tuned models (dashed: endpoint chord). **A:** a catastrophic arch, where a single midpoint query detects severe damage (generation accuracy $45.0\% \to 1.0\%$). **B:** a flat path, where averaging is effectively free. **C:** a below-chord dip with an off-center candidate that takes resolution to localize.

future pair under exchangeability — today a worst-of-nineteen statement at the tested level (§3). Densifying the path raises coverage on paper but buys no additional admission at the same calibrated risk (§4) — the extra points cannot say more about a midpoint that has already been queried.

Density buys a map, and the map is priced in loss. For finding an unknown off-center coefficient, a uniform grid attains the minimax-optimal k^{-2} resolution rate (Proposition 1), and a decomposition of held-out regret shows why the returns flatten: past about six queries, additional points fight sample mismatch, which they cannot reduce, rather than resolution, which they can (§5).

The currency decides. All of the structure above is measured in loss, and loss improvements need not move the deployment metric. At matched wall-clock budget on a pool whose paths carry cliffs, the three rules that read the proxy-loss curve committed coefficients below the better endpoint on 7, 6 and 6 of 8 units, against 0, 1 and 4 for the three that measured the task metric directly; the worst proxy pick committed $\alpha = 0.9$ and landed -17.0 points below the better endpoint, and three-point direct selection is worth 3.2 accuracy points at the same budget (Table 1); on a second pool with flat paths, nothing separates the two currencies (§6). Terrain — the depth of the loss path itself, readable from one cheap scan before any generation run — tells the practitioner which regime a pair is in, and therefore which currency the budget should be spent in (§7).

2 Related Work

Connectivity describes a path; we study how to measure it. A substantial literature asks whether independently trained solutions can be aligned into one low-loss basin by removing permutation or

other symmetries (Entezari et al., 2022; Ainsworth et al., 2023; Qin et al., 2022; Theus et al., 2025; Frankle et al., 2020; Draxler et al., 2018; Garipov et al., 2018). Shared initialization removes a major alignment ambiguity for fine-tuned descendants, but it does not guarantee a low-loss segment. Our question starts where that literature ends: with the parameterization fixed, which features of the segment does a finite query budget resolve, and which decisions remain valid on fresh evaluation samples and new model pairs?

Merge recipes and path measurements answer complementary questions. Soups, TIES, DARE, and task-arithmetic variants specify how weights are combined (Wortsman et al., 2022; Yadav et al., 2023; Yu et al., 2024a; Ilharco et al., 2023; Matena and Raffel, 2022); AdaMerging learns per-layer coefficients by entropy minimization on unlabeled data (Yang et al., 2024). These methods need a coefficient, and most ship with a default one. Our readout supplies that coefficient from measurement: it embeds into TIES and DARE unchanged, and on official pairs the coefficient — not the recipe — turns out to be the operative variable (§6). The closest evaluation-side precedent is MAP (Li et al., 2025), which fits a quadratic surrogate of task metrics over sampled coefficient vectors; its one-dimensional analogue is one of our budget-matched comparators (Appendix E). Searching the coefficient space directly (Akiba et al., 2025) is the complementary optimization view — and, read in the deployment metric, it is the strategy class that wins our budget-matched study (Table 1).

Calibration and global optimization address different halves of the problem. Split conformal turns exchangeability plus a calibration sample into marginal finite-sample validity (Angelopoulos and Bates, 2023; Angelopoulos et al., 2024); deterministic global optimization prices the query complexity of an unknown extremum under Lipschitz derivatives (Lera and Sergeyev, 2013), with optimal adaptive rates in the continuum-armed bandit literature (Kleinberg, 2004; Auer et al., 2007; Bubeck et al., 2011; Audibert and Bubeck, 2010). Theorem 1 maps a pair-unit conformal score to a merge decision, and Proposition 1 prices the path extremum — keeping statistical transfer across pairs and approximation error along a path visibly separate, because they have different remedies. Structural features can also predict mergeability outright (Zhou et al., 2026; Bolton et al., 2026; Marczak et al., 2025; Kim et al., 2025); we test the inexpensive versions of those cues and find them wanting (§7), which is what leaves measurement as the reliable option.

3 The Measurement Problem, and One Safety Margin

The measured object. For two fine-tuned models with weights θ_0, θ_1, write $L(\alpha) = \mathcal{L}((1-\alpha)\theta_0 + \alpha\theta_1)$ for the loss along the interpolation, under a specified dataset, metric, and data split. Change any of those choices and the object changes, so we treat a path as an evaluation-conditioned function, not an intrinsic property of two checkpoints.

Three decisions, two baselines. A practitioner faces three distinct decisions: *screen* a precommitted midpoint (safe to average, yes/no), *choose* a coefficient, and *certify* the choice against damage on sealed data. Two baselines must be kept apart throughout: damage against the endpoint chord, $\Delta(\frac{1}{2}) = L(\frac{1}{2}) - \frac{1}{2}[L(0) + L(1)]$, and damage against the *better endpoint*, $R_{\text{best}} = L(\frac{1}{2}) - \min\{L(0), L(1)\}$. A merge can beat the chord while losing to the stronger endpoint, by exactly the endpoint gap $A = |L(0) - L(1)|$; much of what follows prices that difference.

Protocol. Selection side A and sealed side B are disjoint halves of question chains; every coefficient and every margin is fixed before side B is scored. The validation pool contains 810 questions not used for training, with losses computed from 2,000 stratified step records per evaluation condition; the same question bank is reused at every grid point, which correlates per-question noise across points and matters for what density can buy (§5). All path values come from forward evaluations of interpolated weights.

The safety margin, defined once. A measured envelope becomes a certificate by adding a width q: accept a midpoint only if the selection-half envelope plus q stays under the tolerance. We calibrate q once on a held-out pool, freeze it at $+0.05$, and carry it unchanged to every external population (§4); Appendix D reports the fold-level variants, which differ from the frozen point but are never used to screen. The pair-level version of the same idea carries the prospective guarantee:

Theorem 1 (Pair-unit calibrated midpoint decision)**.** *For pair i with prespecified halves A and B, let $B_3 = \max\{A/2, \Delta(\frac{1}{2})\}$ be the three-point envelope and define the pair score $R_i =$*

$\max\{\Delta_{i,B}(\frac{1}{2}) - B_{3,i,A},\ \Delta_{i,A}(\frac{1}{2}) - B_{3,i,B}\}$. *Let n calibration pairs and one future pair be exchangeable units, put $k = \lceil (n+1)(1-\delta) \rceil$, and let q_δ be the kth order statistic of the n calibration scores ($q_\delta = +\infty$ if $k = n+1$). Then, simultaneously for both half-to-half directions of the future pair,*

$$\Pr\big\{\Delta_{n+1,B}(\tfrac{1}{2}) \le B_{3,n+1,A} + q_\delta,\ \Delta_{n+1,A}(\tfrac{1}{2}) \le B_{3,n+1,B} + q_\delta\big\} \ge 1 - \delta.$$

Accepting from side A only when $B_{3,A} + q_\delta \le \tau$ therefore keeps the probability of accepting while sealed midpoint damage exceeds τ at most δ.

The guarantee is marginal over the named pair population: it covers a future pair drawn from the same mechanism, not any particular pair, and shared endpoints between pairs must be designed around (Appendix B). On a 19-pair endpoint-disjoint subpool of our measurements, the quantile is finite at both tested levels: $q_{0.10} = 0.0495$ and $q_{0.05} = 0.0903$. These are extremal ranks — at nineteen pairs the 0.05 quantile *is* the sample maximum — so the certificate currently says no more than "no worse than the worst of nineteen pairs." That is a real but modest statement; what makes it more than vacuous is that the margin it prices is the same $+0.05$ class of width that survives every external population we test (§4).

4 Three Evaluations Screen a Midpoint

Both envelopes contain the midpoint, so on a single fixed split the screen holds by construction; the content of this section is what survives fresh data. On the 98-side cross-half pool (Llama-3.1-8B, Qwen3-8B, Pythia-1.4B; PRM-BCE readouts), pair-grouped cross-fold calibration puts both the three-point and the eleven-point envelope at $95/98$ held-out sides. On the pooled measurements the raw eleven-point envelope covers $87/98$ sides, and the frozen $+0.05$ margin lifts this to $98/98$; the three-point envelope reaches $97/98$ at the same margin. The dense scan's entire increment over its own three-point nesting, at the operating point a deployment would actually use, is one side.

The frozen margin travels. Carried unchanged to four external populations, $+0.05$ holds everywhere it is tested, and it is tested in exactly one corner (Appendix D). On generation-preserving LoRA adapters under token-CE, the raw envelope already covers every unit, so the margin is never stressed; on a Pythia-2.8B stack held out of calibration entirely, the same happens — the relation travels out of family, with the margin's width untested there. On wild Mistral-7B specialists the raw envelope breaks shallowly (excesses $\le +0.014$) and $+0.05$ covers unrecalibrated. Only the officially released Qwen2.5 pairs reach the margin: $+0.05$ covers $29/32$ units (90.6%, below the nominal 95%), all three misses in Instruct×Coder external evaluations, and 95% coverage needs $+0.085$. One corner of one population is where the margin must be widened; everywhere else it holds as shipped.

Coverage is cheap; admission is the test. A margin can always buy coverage with width, so the screen's value is decided by how much it admits. Calibrating the midpoint probe, B_3, and B_{11} to the same folded coverage and sweeping the tolerance (Figure 2), the two envelopes never separate by more than 11.2 points, and at the upper-quartile tolerance their difference is -1.0 points with an interval crossing zero: density shows no admission advantage. The calibrated midpoint probe admits substantially more at low tolerances — $+7.1$ points over B_{11} at the median — because its score prices the midpoint directly rather than the path maximum, and it is the only screen that reaches $\tau = 0$ under the chord target, admitting $8/98$ sides, all confirmed non-damaging on sealed data. Against the better-endpoint baseline the picture is starker: no screen admits a single side at $\tau = 0$, because no selection-half dip on this pool covers both the margin and the wedge. A zero-tolerance pass against the better endpoint is not something this pool sells; what the margin sells is priced admission at a stated tolerance, and on this pool the cheapest way to buy it is three measurements plus calibration.

5 What More Measurements Buy: A Map of the Path

Finding an unknown off-center coefficient is a resolution problem, and resolution has a price with a known rate.

Proposition 1 (Minimax path-resolution law)**.** *Let $S \subset [0,1]$ contain both endpoints, let $\rho(S)$ be its fill distance, and suppose the path derivative is C-Lipschitz. Then the continuous maximum or*

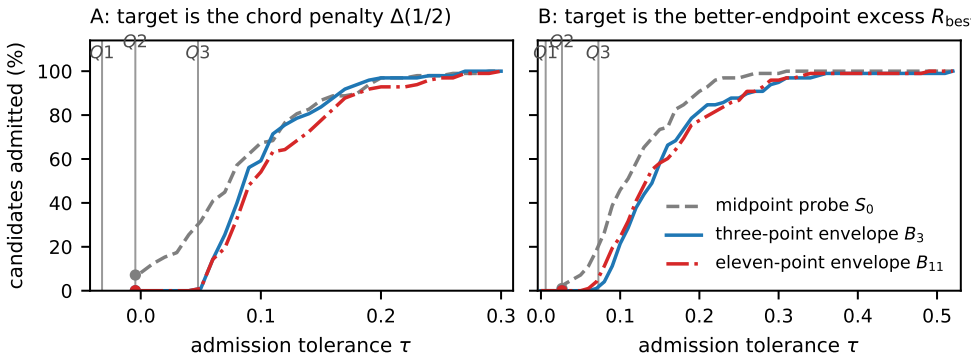

Figure 2: Admission, not coverage, separates the screens. Three screeners calibrated to the same folded coverage, tolerance swept on two baselines. Gray verticals mark the quartiles of the pooled sealed-half penalty distribution. **A:** chord target — the midpoint probe alone reaches $\tau \leq 0$. **B:** better-endpoint target — nothing is admitted at $\tau = 0$, and at usable tolerances the probe admits more than either envelope.

minimum differs from its sampled counterpart by at most $C\rho(S)^2/2$ — at most $C/[8(k-1)^2]$ on a uniform k-point grid — and no k-query design, deterministic or randomized, adaptive or not, improves the worst-case k^{-2} rate over this smoothness class (proof and the noisy-query extension in Appendix B).

Corollary 1 (Held-out coefficient regret)**.** *Let $\widehat{\alpha}_A$ minimize L_A over S, let L'_B be C_B-Lipschitz, and let $\Gamma_{A,B} = \sup_\alpha |L_A(\alpha) - L_B(\alpha)|$ be the split discrepancy. Then $L_B(\widehat{\alpha}_A) - \min_\alpha L_B(\alpha) \leq 2\Gamma_{A,B} + C_B\rho(S)^2/2$.*

The two terms have different remedies, and the decomposition is visible in the measurements. Replaying nested subgrids of the stored curves against the eleven-point reference: from three to six queries the discretization term contracts $7.2\times$; from six to eleven the two terms offset — discretization falls, split mismatch rises, and the pool cannot resolve any remaining movement in the total. Past six queries, additional points fight the term they cannot reduce. This is also why adaptive search cannot dominate the grid here: on twelve freshly measured sealed sides, five uniform and five adaptive queries tie on mean regret (0.0145 vs. 0.0142), with the eleven-point grid at 0.0087 on more than twice the budget — while both five-query designs underprice the real sealed-half barrier, by up to 0.1901. What the grid keeps and optimizers lose is the path's shape.

How much shape survives a coarser grid is itself measurable: across 461 stored curves, five points retain a median 86% of the eleven-point barrier, and 17- and 33-point rescans of two pairs bound the remaining movement at 3.6%. The design guidance that falls out: three evaluations for a screen (§4), about eleven for a map, and more only if the map itself is the deliverable. On arch-shaped official-pair paths the map's low point often *is* an endpoint — the rule falls back on the stronger endpoint instead of inventing an interior optimum, a useful degeneracy that the generative results of §6 confirm from the other side. On the 49-side split-half pool, the curve's low point carries mean held-out regret 0.0106 against 0.021 for choosing among the endpoints and midpoint and 0.0646 for the naive 0.5 convention.

6 The Currency Decides

Everything measured so far is loss. Whether loss improvements move the deployment metric is a separate question, and Table 1 answers it at matched wall-clock budget: the proxy-loss designs lose to the direct ones, and the losses do not order by density. The worst proxy pick commits $\alpha = 0.9$ on a shallow loss dip that turns out to be an accuracy cliff, landing -17.0 points below the better endpoint. Four controls fix what the table means. The ordering is not a budget artifact: no budget in a sixteenfold range reverses the point estimate (Appendix C). It is not grid density: within the proxy currency, eleven points versus five moves the mean by 0.1 points, while at the same five-coefficient grid the direct metric beats the proxy by 2.7 points — the gap is metric mismatch, not resolution. It is not information volume: the proxy arms score the most selection questions per coefficient and select worst. And it is not one unit: recomputing the pool with each unit dropped in turn keeps the CE-11 vs. Direct-5 point estimate negative in every replicate (-3.4 to -0.9 points),

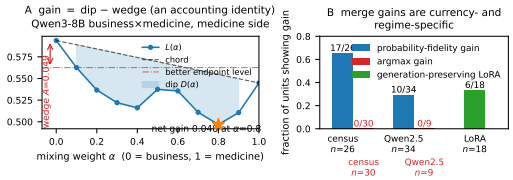

Figure 3: Where loss-path gains reach the deployment metric, and where they do not. A: net gain decomposes as dip depth minus the asymmetry wedge (Eq. 1). **B:** probability-fidelity gains (blue) appear in the controlled census and on official pairs while argmax gains (red) do not ($0/30$, $0/9$); the generation-preserving LoRA arm (green) supplies the measured unit-level gains.

and dropping the unit that carries 74% of the mean gap shrinks it to -0.9 points without flipping its sign (Appendix C.4).

Why the proxy fails: the argmax blind spot. Mean cross-entropy prices probability changes that a thresholded decision never registers (Gneiting and Raftery, 2007; Hinton et al., 2015; Guo et al., 2017; Vlaar and Frankle, 2022). In the controlled full-weight census, captured dip gains improve BCE, calibration, and ensemble correlation, yet none of the thirty ranking sides passes the argmax criterion (one-sided 95% upper bound 9.5%), and the loss- and accuracy-optimal coefficients coincide on only $11/43$ sides. The mechanism is exact: a margin-band argument (Appendix B) shows harmful argmax flips are bounded by the endpoint's band mass plus a loss-increment term — so small loss gains bound *harm*, while sub-threshold probability improvements never cross the argmax at all. The blind spot is the same thresholding read in the two directions.

The one identity a practitioner needs. Net gain over the better endpoint decomposes exactly as dip depth minus an endpoint-asymmetry wedge, pointwise along the path:

$$G(\alpha) \;=\; D(\alpha) \;-\; A\,\delta(\alpha), \qquad D(\alpha) = \mathrm{chord}(\alpha) - L(\alpha), \;\; A = |L(0) - L(1)|, \tag{1}$$

with $\delta(\alpha)$ the distance from the better endpoint. At $\alpha = 0.5$ the wedge tax is exactly $A/2$: naive fifty-fifty averaging forfeits half the endpoint gap on asymmetric pairs, before any question of terrain. The net-gain peak shifts from the dip bottom toward the better endpoint, which motivates the gain-seeking coefficient $\alpha^{\dagger}$ (the curve minimum shrunk toward the better endpoint); across the 40-side Pythia pool its hit rate is monotone in the selection-half gain signal and it degenerates harmlessly to an endpoint where that signal is absent (Appendix F).

Where gains do cross the argmax — and the limit of that evidence. When endpoints retain language-modeling competence and the task has headroom, merged models beat *both* endpoints on $6/18$ generation units, by up to $+6.5$ points, with the wins concentrated where the dips live (Appendix F). This establishes possibility and scope: loss gains can become accuracy gains in this regime. Two cautions are measured, not rhetorical: the count is an any-of-five-candidates rate, and the one cross-scale pattern we re-tested prospectively — GSM8K-side gains on GSM8K×MBPP pairs, held-out seeds at five times the original sample size — came back exactly null (pooled paired McNemar $p = 1.0$). Individual wins are real; the pattern is not yet a rule.

On released specialists, the margin and the currency agree. Under true generation on the official Qwen2.5 pairs, midpoint averaging takes the 1.5B Instruct×Coder pair from MBPP pass@1 endpoints reaching 64% to 0% — the three other 1.5B pair-sides, each involving the Math specialist, land at 0–1% as well — and roughly halves accuracy at 7B, every observed collapse lies in the set the loss curve flags for endpoint selection ($6/7$ pairs), and no measured point beats both endpoints on the 9 generation cells. The one a-priori confound is visible and priced: the Math specialist ships a different RoPE base, worth 17 GSM8K points on its own (Appendix F). The curve coefficient also slots directly into sparsifying recipes: TIES and DARE at full strength collapse on every cell, and re-armed at the curve-selected coefficient they recover to LERP-at-α^{*} levels on exactly the large low-barrier cells — the coefficient, not the recipe, is the operative variable.

7 Terrain Decides When the Currency Matters

The proxy deficit of Table 1 has a measurable boundary. Re-running the same six-strategy comparison with both swappable factors changed — the deployment task is now PubMedQA, and the pool

is eight freshly trained LoRA units sharing no adapter with the first pool — no budget separates the eleven-point proxy curve from direct five-point selection: every interval includes zero, and at b^* all six strategies sit within 0.53 points of one another (Appendix C). The two pools differ in exactly the quantity the theory says should matter. The median per-unit accuracy range along the path is 2.2 points on the flat pool (max 5.8) against 7.7 (max 21.4) on the cliff pool, and the worst shortfall below the better endpoint across all strategies is -1.6 points there against -17.0 here. The unit that carried 74% of the cliff pool's mean gap has no counterpart on flat terrain. On flat paths the currencies tie because there is little left to decide; on cliff paths the proxy's blind spot is expensive because that is where the decision lives. Terrain — the depth of the loss path itself, read from one cheap scan before any generation run — tells the practitioner which regime a pair is in, and therefore which currency the budget should be spent in.

The per-unit deficit ranks with the per-unit accuracy-path range inside each pool separately (Spearman $+0.33$ on the cliff pool, $+0.59$ on the flat one, exact-permutation $p = 0.43$ and 0.14; $+0.43$ pooled over 16 units), and — on the loss side, from the same eleven-point chain scans the screen already runs — the deficit ranks with the loss-path amplitude on the cliff pool ($\rho = 0.40$, exact $p = 0.33$, 8 units; Appendix C.4).

One cheap scan, before any generation run, turns the two studies into a decision procedure:

1. **Scan once in the cheap currency.** An eleven-point loss curve prices the terrain: barrier, dips, endpoint asymmetry.
2. **Flat path: stop scanning.** Flat rules out averaging damage, not the endpoint gap: on an asymmetric pair the wedge (Eq. 1) still taxes the naive 0.5 by $A/2$, so keep the better endpoint's side and spend nothing more. Where the deployment target is itself loss-shaped, proxy loss remains the right currency.
3. **Cliff-bearing path: select in the deployment currency.** Measure the task metric at three to five coefficients; never commit a blind midpoint on an asymmetric pair, where the wedge alone forfeits half the endpoint gap.
4. **Price the margin.** Envelope plus the frozen $+0.05$; against the better-endpoint baseline, state the tolerance bought rather than expecting a zero-tolerance pass.

One cue helps before any path measurement, at stack level only: the endpoints' token-CE multiple over the base model orders the merge cost across our four stacks ($5\times$ to $23\times$; median accuracy cost $+0.4$ to -1.5 points at the curve-selected coefficient; Appendix F). The cheaper endpoint cues do not survive contact: task labels fail to order the path (within-domain pairs can be the less mergeable ones, and the ordering flips across stacks), weight-space distance is statistically indistinguishable across groups whose barriers differ by orders of magnitude, a layer-level mechanism reversed under fp32 re-measurement, and the RoPE confound above is worth more than most barriers (Appendix G). Learned multi-feature prediction remains the open alternative to measurement; nothing simpler currently is.

8 Limitations

The currency experiments are two tasks, sixteen units, one model family, and the terrain reading of the contrast is post-hoc: within-pool trends point the same way on the accuracy side ($+0.43$ pooled over sixteen units) and on the loss side ($\rho = 0.40$, eight units), but the prospective prediction — loss-path depth should forecast which pairs repay direct selection — is stated, not yet run. The pair-level certificate at nineteen disjoint pairs is a worst-of-nineteen statement at $\delta = 0.05$ (§3). Accuracy evidence on the self-trained full-weight stacks uses PRM-stratified readouts, and the generative arms put the regime boundary at 1.5B–7B. The resolution law's constant inherits grid-estimated curvature, and convergence to the continuous optimum is bounded by two pairs' 33-point rescans, not established. Multi-model mixtures face dimensional resolution decay $\Theta(k^{-2/(d-1)})$ (Appendix B), and our three-model probe is consistent with it. Rates are reported per pool without cross-pool multiplicity correction.

9 CONCLUSION

How much of a model-merging path must we measure? Less than folklore suggests, and in a different currency than folklore uses. Three evaluations plus a calibrated margin settle a midpoint decision on every pool we test — with a pair-level guarantee that is today a worst-of-nineteen statement, survives four external populations, and fails gracefully in one named corner. About eleven evaluations map the path at the minimax rate, and past six they mostly fight the wrong term. And when the decision is a deployment metric, the measurements that decide it should be read in that metric — a rule whose cost is one cheap scan to learn whether your pool is the one where it matters. Merging is already how the ecosystem reuses fine-tuning; these results make it a measured operation with a priced risk instead of an act of faith, and they tell practitioners exactly where the measurement budget changes the answer.

Reproducibility Statement

All experiments run offline on cached open-weight checkpoints and one open dataset (MMLU-Pro-CoT-Train-Labeled). The accompanying anonymized repository contains every scan and validation ledger, the scripts that regenerate each table and figure from those ledgers, versioned evaluation protocols with exact split definitions, and the training and scanning pipeline (8-bit AdamW, batch 128, one epoch; eleven-point interpolation grids in fp32). Checkpoints and per-record dumps exceed the package size limit and are regenerated end-to-end by the included pipeline; one 8B stack's fine-tuning plus census scans costs on the order of 250 GPU-hours. Model revisions, data filtering, evaluation pools, and compute budgets are specified in Appendix H.

AI Use Statement

Generative AI tools worked on this study as engineering and writing assistants under the authors' direction. They cleaned and reformatted the training data (the malformed-row filtering of the chain set), implemented the experiment pipeline (training queues, interpolation scans, validation harnesses, and the analysis scripts that regenerate the tables), helped schedule the experiment arms, searched and summarized the literature, prepared the figures, and drafted and edited manuscript text. The research question, the evaluation protocols, and every threshold and decision rule were set by the authors. All experimental data are real model evaluations; no AI-generated data were used. Every reported number is produced by a script from a named ledger, and the authors checked the headline results against the raw artifacts. The authors reviewed all AI-assisted work and take responsibility for the full contents of this submission, including text, code, and claims produced with AI assistance.

References

L. Adilova, M. Andriushchenko, M. Kamp, A. Fischer, and M. Jaggi. *Layer-wise Linear Mode Connectivity.* In *ICLR*, 2024. arXiv:2307.06966.

S. Ainsworth, J. Hayase, and S. Srinivasa. *Git Re-Basin: Merging Models modulo Permutation Symmetries.* In *ICLR*, 2023.

T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha. *Evolutionary Optimization of Model Merging Recipes. Nature Machine Intelligence*, 7(2):195–204, 2025. arXiv:2403.13187.

A. N. Angelopoulos and S. Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. arXiv:2107.07511.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. *Conformal Risk Control.* In *ICLR*, 2024. arXiv:2208.02814.

J.-Y. Audibert and S. Bubeck. *Best Arm Identification in Multi-Armed Bandits.* In *COLT*, 2010.

P. Auer, R. Ortner, and C. Szepesvári. *Improved Rates for the Stochastic Continuum-Armed Bandit Problem.* In *COLT*, pp. 454–468, 2007.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. *Convexity, Classification, and Risk Bounds. Journal of the American Statistical Association*, 101(473):138–156, 2006.

S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skripnik, P. Liskovich, Q. G. Liao, N. Piel, A. Patel, A. Jones, L. Gao, and S. Goldfarb. *Pythia: A Suite for Analyzing Large Language Models.* In *NeurIPS*, 2023.

O. Bolton, Aakanksha, A. Ahmadian, S. Hooker, M. Fadaee, and B. Ermis. *SimMerge: Learning to Select Merge Operators from Similarity Signals.* In *ICLR*, 2026. arXiv:2601.09473.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. *$\mathcal{X}$-Armed Bandits. Journal of Machine Learning Research*, 12:1655–1695, 2011.

Y. Cao, D. Ran, Y. Guo, M. Wu, S. Chen, L. Li, W. Yang, and T. Xie. *An Empirical Study and Theoretical Explanation on Task-Level Model-Merging Collapse.* arXiv:2603.09463, 2026.

F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. *Essentially No Barriers in Neural Network Energy Landscapes.* In *ICML*, 2018.

R. Entezari, H. Sedghi, O. Saukh, and B. Neyshabur. *The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks.* In *ICLR*, 2022.

J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. *Linear Mode Connectivity and the Lottery Ticket Hypothesis.* In *ICML*, 2020.

J.-C. Gagnon-Audet, R. P. Monti, and D. J. Schwab. *AWE: Adaptive Weight-space Ensembling for Few-shot Fine-tuning.* In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)*, 2023.

T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson. *Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs.* In *NeurIPS*, 2018.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. *Measuring Massive Multitask Language Understanding.* In *ICLR*, 2021.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

W. Hoeffding. *Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association*, 58(301):13–30, 1963.

G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. *Editing Models with Task Arithmetic.* In *ICLR*, 2023.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed. *Mistral 7B. arXiv:2310.06825*, 2023.

J. Juneja, R. Bansal, K. Cho, J. Sedoc, and N. Saphra. *Linear Connectivity Reveals Generalization Strategies.* In *ICLR*, 2023. arXiv:2205.12411.

T. Kim, H. Gouk, M. Kim, and T. Hospedales. *Model Merging is Secretly Certifiable: Non-Vacuous Generalisation Bounds for Low-Shot Learning. arXiv:2505.15798*, 2025.

R. D. Kleinberg. *Nearly Tight Bounds for the Continuum-Armed Bandit Problem.* In *NIPS*, 2004.

Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. *BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains.* In *Findings of ACL*, 2024.

D. Lera and Y. D. Sergeyev. *Acceleration of Univariate Global Optimization Algorithms Working with Lipschitz Functions and Lipschitz First Derivatives. SIAM Journal on Optimization*, 23(1):508–529, 2013.

L. Li, T. Zhang, Z. Bu, S. Wang, H. He, J. Fu, Y. Wu, J. Bian, Y. Chen, and Y. Bengio. *MAP: Low-compute Model Merging with Amortized Pareto Fronts via Quadratic Approximation.* In *ICLR*, 2025. arXiv:2406.07529.

D. Marczak, S. Magistri, S. Cygert, B. Twardowski, A. D. Bagdanov, and J. van de Weijer. *No Task Left Behind: Isotropic Model Merging with Common and Task-Specific Subspaces.* In *ICML*, 2025.

M. Matena and C. Raffel. *Merging Models with Fisher-Weighted Averaging.* In *NeurIPS*, 2022.

C. Oh, Y. Li, K. Song, S. Yun, and D. Han. *DaWin: Training-free Dynamic Weight Interpolation for Robust Adaptation.* In *ICLR*, 2025.

Y. Qin, C. Qian, J. Yi, W. Chen, Y. Lin, X. Han, Z. Liu, M. Sun, and J. Zhou. *Exploring Mode Connectivity for Pre-trained Language Models.* In *EMNLP*, pp. 6726–6746, 2022. arXiv:2210.14102.

S. P. Singh, L. Adilova, M. Kamp, A. Fischer, B. Schölkopf, and T. Hofmann. *Landscaping Linear Mode Connectivity.* In *ICML 2024 Workshop on High-dimensional Learning Dynamics (HiLD)*, 2024. arXiv:2406.16300.

Teknium. *OpenHermes-2.5-Mistral-7B.* Hugging Face repository, 2023. https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B.

A. Theus, A. Cabodi, S. Anagnostidis, A. Orvieto, S. P. Singh, and V. Boeva. *Generalized Linear Mode Connectivity for Transformers.* In *NeurIPS*, 2025. arXiv:2506.22712.

A. B. Tsybakov. *Optimal Aggregation of Classifiers in Statistical Learning. The Annals of Statistics*, 32(1):135–166, 2004.

T. J. Vlaar and J. Frankle. *What Can Linear Interpolation of Neural Network Loss Landscapes Tell Us?* In *ICML*, 2022.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World. Springer*, second edition, 2022.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark.* In *NeurIPS*, 2024. arXiv:2406.01574.

H. Wendland. *Scattered Data Approximation. Cambridge University Press*, 2005.

M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt. *Model Soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.* In *ICML*, 2022.

P. Yadav, D. Tam, L. Choshen, C. Raffel, and M. Bansal. *TIES-Merging: Resolving Interference When Merging Models.* In *NeurIPS*, 2023.

E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao. *AdaMerging: Adaptive Model Merging for Multi-Task Learning.* In *ICLR*, 2024. arXiv:2310.02575.

V. G. Vizing. *On an estimate of the chromatic class of a p-graph. Diskretnyi Analiz*, 3:25–30, 1964.

L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li. *Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch.* In *ICML*, 2024. arXiv:2311.03099.

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. *MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models.* In *ICLR*, 2024.

L. Zhou, B. Zhao, R. Yu, and E. Rodolà. *Demystifying Mergeability: Interpretable Properties to Predict Model Merging Success.* In *ICML*, 2026. arXiv:2601.22285.

A Notation and Evaluation Pools

$L(\alpha)$, chord	loss at mixing weight α; segment from $L(0)$ to $L(1)$
B_3, B_{11}	three-point midpoint envelope; eleven-point path envelope
$D(\alpha)$, A	dip depth $\mathrm{chord}(\alpha) - L(\alpha)$; endpoint gap $\lvert L(0) - L(1)\rvert$
α^*, $\alpha^\dagger$	curve low point; gain-seeking variant (shrunk toward the better endpoint)
q, q_δ	safety margin; pair-level conformal quantile (Theorem 1)
$\Gamma_{A,B}$	split discrepancy $\sup_\alpha \lvert L_A(\alpha) - L_B(\alpha)\rvert$

Table 2: Path quantities. Each is tied to the dataset, metric, and split on which $L(\alpha)$ is evaluated.

Every number in the paper names its pool. The main ones: the penalty census (58 Gemma-2-2B + 147 Llama-3.1-8B sides); the 98-side cross-half pool (Llama, Qwen3-8B, Pythia-1.4B); the 49-side selector pool (43 Llama + 6 Pythia); the Pythia gain-rule pools (15 pairs, then a 40-side expansion); the wild arm (Mistral-7B and Qwen2.5 families); the LoRA arms (Gemma-2-2B and Qwen3-8B, 15 pairs each; Qwen2.5-1.5B/3B with headroom); and the two currency pools of Appendix C (eight units each). Pools share model families, endpoints, and evaluation items, so pool rates summarize scope rather than independent draws, and are reported without cross-pool multiplicity correction.

B PROOFS

B.1 FINITE-SAMPLE VALIDITY ACROSS PAIR UNITS

Proof of Theorem 1. The $n+1$ scores $R_1, \dots, R_{n+1}$ are exchangeable, each being the same measurable function of one pair unit and its two halves. With conservative tie handling, the rank of R_{n+1} is no larger than k with probability at least $k/(n+1) \geq 1-\delta$. If $k \leq n$ this event implies $R_{n+1} \leq q_\delta$; if $k = n+1$ the convention $q_\delta = +\infty$ makes the statement immediate. Since R_{n+1} is the maximum of the two directional residuals, $R_{n+1} \leq q_\delta$ gives both displayed inequalities, and $\{B_{3,A} + q_\delta \leq \tau,\ \Delta_B(\frac{1}{2}) > \tau\} \subseteq \{R_{n+1} > q_\delta\}$, so the false-safe probability is at most δ. This is the standard split-conformal rank argument Angelopoulos and Bates (2023); no parametric loss model is required. □

The unit of exchangeability is the substantive assumption. If one endpoint appears in several pair edges, those edges are not automatically exchangeable calibration units; a sufficient construction is to partition the comparison graph into endpoint-disjoint classes and conformalize the per-class maximum scores. By Vizing's theorem Vizing (1964) the pair graph's edges partition into at most $\Delta+1$ such matchings, so shared endpoints degrade the effective calibration size, not the guarantee's logic: use the largest class, draw one class at random, or calibrate each class and combine levels by Bonferroni.

The 19-pair instantiation. A deterministic exact matching over our pair graph — selection reads the topology only, never a score — extracts a maximum endpoint-disjoint subpool of 18 pairs, extended by one designated disjoint pair (chosen score-blind) to $n = 19$. Both split directions of every score are computable from the stored scans, so no pair enters with a missing direction. With $n = 19$ the rank $k = \lceil (n+1)(1-\delta) \rceil$ is 18 at $\delta = 0.10$ and 19 at $\delta = 0.05$: the quantiles are the second-largest and the largest of the nineteen scores, $q_{0.10} = 0.0495$ and $q_{0.05} = 0.0903$. The requirement curve is steep — the 95% quantile remains the sample maximum up to $n = 38$ and reaches a decile-deep rank only at $n = 199$ — which is why we read the certificate as a worst-of-nineteen statement. The subpool spans three stacks; the pooled statement assumes the stack mixture is exchangeable, and per-stack instantiation is underpowered at this size.

B.2 RESOLUTION: PROOF OF THE MINIMAX LAW

Proof of Proposition 1. Let f be the path loss and $\alpha_{\max}$ a maximizer. If the maximizer is an endpoint it is sampled; otherwise $f'(\alpha_{\max}) = 0$, and for $s \in S$ with $|s - \alpha_{\max}| \leq \rho(S)$ the smoothness inequality gives $f(s) \geq f(\alpha_{\max}) - \frac{C}{2}|s - \alpha_{\max}|^2$. Hence $0 \leq \max_{[0,1]} f - \max_S f \leq C\rho(S)^2/2$; applying the same argument to $-f$ proves the minimum bound, and a uniform grid has fill distance $1/[2(k-1)]$.

For sharpness, suppose k queries including the endpoints have all returned zero; some adjacent pair encloses a gap of length $g \geq 1/(k-1)$. On that gap place the $C^{1,1}$ bump $f(\alpha) = \frac{Cg^2}{2}\psi((\alpha - a)/g)$ with $\psi(t) = t^2(1-t)^2$ and $f = 0$ elsewhere: the function and derivative match zero at the boundaries, f' is C-Lipschitz since $\sup|\psi''| = 2$, and the hidden maximum is $Cg^2/32 \geq C/[32(k-1)^2]$. An adversary can place this bump after the zero responses even against an adaptive deterministic design. Randomization does not change the rate: by Yao's principle, hide the bump in one of $m = 3k$ equal intervals chosen uniformly; a k-query design misses it with probability at least $1 - k/m$, giving expected extremum error $\Omega(C/k^2)$. □

For $f(\alpha) = \mathcal{L}(\theta_0 + \alpha\Delta\theta)$ with a Hessian, $f''(\alpha) = \Delta\theta^\mathsf{T}\nabla^2\mathcal{L}(\theta_0 + \alpha\Delta\theta)\Delta\theta$, so the smoothness constant is a bound on directional curvature along the segment, not on the full Hessian norm.

Proof of Corollary 1. With $s_B \in \arg\min_{s \in S} L_B(s)$: $L_B(\widehat{\alpha}_A) \le L_A(\widehat{\alpha}_A) + \Gamma_{A,B} \le L_A(s_B) + \Gamma_{A,B} \le L_B(s_B) + 2\Gamma_{A,B}$, and Proposition 1 bounds $L_B(s_B) - \min_\alpha L_B(\alpha)$ by $C_B \rho(S)^2/2$. □

Noisy queries. The measured curves are empirical means. If each of k grid points is estimated from n fresh independent units with per-unit losses in $[0, M]$ — total budget $B = kn$ — a Hoeffding radius Hoeffding (1963) at each point plus a union bound gives, with probability $1 - \delta$, extremum error at most

$$\frac{C}{8(k-1)^2} \;+\; M\sqrt{\frac{k \log(2k/\delta)}{2B}},$$

whose minimizer scales as $k^* \asymp B^{1/5}$. Two conditions are load-bearing: bounded per-unit losses (enforced by clipping at $M = 8$ nats; the clip never activated on measured data) and independence (the operative independent unit is the question chain, since steps within a chain correlate). Priced at the full range M the bound is vacuous at our unit counts; at the measured dispersions and grid-estimated curvatures the plug-in objective minimizes at $k^* = 9$–13 — the grid density we use, not an order of magnitude below it. The curvature constants are second-difference estimates on the same eleven-point grid, so this is a consistency check, not an independent derivation; the resolution-independent anchor is the $17/33$-point rescan movement of at most 3.6% on two pairs. Under the actual protocol the same question bank is reused at every grid point, so the simultaneous estimation error grows only as $\sqrt{\log k}$ while resolution improves as k^{-2}: density is compute-limited, not statistics-limited.

Multi-model mixtures. The exponent is dimensional. Mixtures of d models live on the $(d-1)$-simplex, where k points have fill distance $\Theta(k^{-1/(d-1)})$ Wendland (2005), giving extremum error $\Theta(k^{-2/(d-1)})$ provided the faces are sampled as well; the bump lower bound lifts unchanged. A three-model soup already degrades to $\Theta(k^{-1})$ — the backdrop against which our three-model probe's inconsistent dominance ordering should be read (Appendix F).

B.3 Path geometry: the wedge and the quadratic regime

The exact statements need no model of the path; the quadratic-regime statements discard $O(\|\Delta\theta\|^3)$ terms and their measured remainders are the tightness deficit and mixed-morphology sides.

Chord-excess lemma. Expanding the side loss about the midpoint, $E(\alpha) = L(\alpha) - [(1-\alpha)L(0) + \alpha L(1)] = -\frac{1}{2}\alpha(1-\alpha)Q + O(\|\Delta\theta\|^3)$ with $Q = \Delta\theta^\mathsf{T}\mathbf{H}(\bar{\theta})\Delta\theta$ — the lemma of Singh et al. (2024), validated here at LLM fine-tuning scale. The sign of Q carries the morphology (arch $\iff Q < 0$, excess peak $-Q/8$; dip $\iff Q > 0$), and magnitude factors out: direction, not displacement size, decides the path shape — the reason Gaussian perturbation at matched weight displacement produces $\sim$800$\times$ smaller barriers than training-induced displacement. With signed endpoint slope $s = L(1) - L(0)$, the endpoint-mean barrier is piecewise

$$\mathcal{B} \;=\; \begin{cases} A/2, & Q \ge 0 \text{ or } |s| > |Q|/2, \\ -Q/8 \;-\; s^2/(2Q), & Q < 0,\ |s| \le |Q|/2, \end{cases} \tag{2}$$

peaked at $\alpha_{\text{peak}} = \frac{1}{2} - s/Q$ in the interior branch. The two conventions coincide at balanced endpoints; otherwise the endpoint-mean convention adds a slope term, which over-reads arches on sloped sides (the label flips are always arch-under-mean to dip-under-chord, never the reverse).

The wedge vertex. $A = \left|\int_0^1 \nabla\mathcal{L}(\theta(\alpha))^\mathsf{T}\Delta\theta\, d\alpha\right|$ exactly (fundamental theorem of calculus), and in the quadratic regime the net-gain vertex sits at $\alpha_{\text{vertex}} = \frac{1}{2} - s/Q$ — pulled *toward* the better endpoint, the sign being load-bearing. An interior gain point exists iff $A/Q < \frac{1}{2}$. Empirically all 44 grid-resolvable vertex drifts point toward the better endpoint (sign test $p < 2^{-43}$), which is what motivates the shrunk coefficient $\alpha^\dagger$.

B.4 From loss bounds to argmax flips

The risk bounds price a loss, not an argmax ranking; the standard margin argument links them. Work in the binary case: $p(\theta; x) \in [0, 1]$ the probability of the correct label, $\ell = -\log p$, and $\Delta\ell(x)$ the

increment between a reference endpoint θ_E and the merge θ_M. A record is *harmfully flipped* if the endpoint's argmax is correct and the merge's is wrong. On a harmful flip whose endpoint confidence clears the band, $p(\theta_E; x) \geq (1+m)/2$, the crossing forces $\Delta\ell(x) \geq \log(1+m)$; inside the band the flip is charged to the band mass. Taking expectations of the pointwise inequality gives the exact bound

$$\Pr[\text{harmful flip}] \;\leq\; \eta(m) \;+\; \frac{\mathbb{E}\big[(\Delta\ell)^+\big]}{\log(1+m)}, \qquad \eta(m) = \Pr\Big[\tfrac{1}{2} < p(\theta_E) < \tfrac{1+m}{2}\Big], \tag{3}$$

where the band mass $\eta(m)$ is estimable from the endpoint alone. Converting the one-sided mean to the certified two-sided one requires one bounded-tail assumption: $\mathbb{E}[(\Delta\ell)^-] \leq \tau$. Two readings matter. Against a confident endpoint (thin band), the measured loss increments (median regret 0.004) contribute $\sim$0.01 at $m = \frac{1}{2}$ — the loss bound becomes an argmax-stability bound of the same order. And the bound is one-directional: small loss increments bound *harmful* flips, while sub-threshold improvements never cross the argmax — the blind spot of §6 is the same thresholding read in the other direction. The band mass is necessary, not proof slack: for any m and target flip mass one can construct distributions approaching the right-hand side arbitrarily closely, so no bound depending on $\mathbb{E}[\Delta\ell]$ alone can exist.

B.5 Transfer to decisions

The generic conversion from surrogate-loss risk to classification risk is classical Bartlett et al. (2006), and the margin-mass condition is that of Tsybakov (2004); what is specific to this paper is the coupling of grid density, split discrepancy, and the calibrated margin into one bound.

Coefficient side. Let α_L minimize the target path loss and $\widehat{\alpha}$ the grid-selected coefficient, with regret $\varepsilon_k = 2\Gamma_{A,B} + C_B\rho(S)^2/2$. Under (i) quadratic growth $L(\alpha) - L(\alpha_L) \geq \frac{\mu}{2}|\alpha - \alpha_L|^2$, (ii) margin stability $|m_\alpha - m_{\alpha'}| \leq K|\alpha - \alpha'|$, and (iii) margin mass $\Pr(|m_{\alpha_L}| \leq t) \leq Ht^\beta$:

Proposition 2 (Accuracy transfer)**.** $|\operatorname{Acc}(\widehat{\alpha}) - \operatorname{Acc}(\alpha_L)| \leq \min\{1,\, H(K\sqrt{2\varepsilon_k/\mu})^\beta\}$.

Proof. The regret bound and (i) give $|\widehat{\alpha} - \alpha_L| \leq \sqrt{2\varepsilon_k/\mu}$. The two accuracies differ only on records where margin signs disagree, and there $|m_{\alpha_L}| \leq K|\widehat{\alpha} - \alpha_L|$ by (ii); (iii) bounds that band's mass. □

On a uniform grid with $\Gamma_{A,B} = O(k^{-2})$ the bound is $O(k^{-\beta})$: the loss-resolution law becomes an accuracy-approach bound, with a floor $H(2K\sqrt{\Gamma_{A,B}/\mu})^\beta$ that no densification removes — the decision-level instance of Corollary 1's two terms. The proposition does not assert that minimizing loss maximizes accuracy; no such assertion is available (mean cross-entropy can fall while accuracy falls), and margin information is precisely what transfer needs beyond the mean loss.

Penalty side. Fix a reference endpoint θ_E, loss penalty $d(\alpha) = L(\alpha) - L(\theta_E)$, accuracy penalty $r(\alpha) = \operatorname{Acc}(\theta_E) - \operatorname{Acc}(\alpha)$, and a calibration event on which grid residuals are bounded by q (same-sample choice $q = 0$ exact).

Proposition 3 (Path-wide accuracy-risk bound)**.** *On the calibration event, with $\widetilde{d}_k$ the piecewise-linear reconstruction from the k-point grid, $\eta_E(m)$ the endpoint band mass, and $T_\alpha = \mathbb{E}[(\ell_E - \ell_\alpha)^+]$ the improvement mass,*

$$\sup_{\alpha\in[0,1]} r(\alpha) \;\leq\; \frac{\widehat{U}_k + q}{\log(1+m)} + \frac{C\rho(S)^2}{2\log(1+m)} + \eta_E(m) + \frac{T_{\max}}{\log(1+m)},$$

where $\widehat{U}_k = \max_{s\in S}\widehat{d}(s)$; at measured points the resolution term vanishes.

The proof chains Eq. 3 with the positive-part identity $\mathbb{E}[(\Delta\ell)^+] = d(\alpha) + T_\alpha$ and the smoothness inequality between grid points. With the lower-loss endpoint as reference, $U_k = \mathcal{B}_k + A/2$: the measured barrier plus half the endpoint gap prices the whole path's accuracy penalty, up to the resolution and conversion terms. Priced end-to-end on four stored per-record dumps, the bound is non-vacuous on three (minimized right-hand side 0.59–0.84) and vacuous on the fourth, where the band mass dominates — a usable tool on three of four measured paths, honestly vacuous on the fourth.

B.6 WHY DENSITY CANNOT IMPROVE THE SCREEN

Two statements formalize §4. First, estimating the midpoint better by smoothing neighboring readings is capped by the bank-reuse correlation:

Proposition 4 (Smoothing-gain ceiling)**.** *Let $Y_\alpha = f(\alpha) + e_\alpha$ with f' C-Lipschitz, $\mathrm{Var}(e_\alpha) = \sigma^2$, midpoint-neighbor correlation r_1. Among symmetric linear estimators $\hat{f}_w = (1-2w)Y_{1/2} + w(Y_{1/2-\rho} + Y_{1/2+\rho})$, the worst-case MSE gain over the single midpoint reading is at most $4\sigma^2(1-r_1)^2/(C\rho^2)^2$ — exactly zero when $r_1 = 1$.*

Proof. $\mathrm{Var}(\hat{f}_w) = \sigma^2[1 - 4(1-r_1)w + \kappa w^2]$ with $\kappa = 6 - 8r_1 + 2r_2 \geq 4(1-r_1)^2$; the bias is w times the second difference, at most $C\rho^2$ under the smoothness class. Minimizing the quadratic in w and dropping $\sigma^2\kappa \geq 0$ from the denominator gives the ceiling. □

Measured on four per-record dumps, the midpoint-shell correlation averages 0.86–0.98 and the gain ceiling is 1.70%–0.22%; the all-shell extension caps at 2.5%–0.3%. The statement is falsifiable: resampling a fresh question bank at every grid point drives r_1 toward zero and lifts the ceiling by one to two orders of magnitude — three-point sufficiency is a property of this protocol, not of midpoint estimation in general.

Second, admission-set comparisons of nested screens reduce to an identity:

Proposition 5 (Nested-screen admission identity)**.** *Let $S \leq S'$ be nested envelope scores on the same units, calibrated on the same folds at the same level with margins q, q', and write $G(x) = S'(x) - S(x)$, $g^* = q - q'$. Then the certified scores differ by exactly $U'(x) - U(x) = G(x) - g^*$ for every unit: if G is constant the screens admit identically at every tolerance, and the denser screen is strictly better at x exactly when $G(x) < g^*$.*

The proof is one line. On the 98-side pool $G = B_{11} - B_3$ is exactly zero on 69 sides, so unit-varying separation is confined to the other 29 — consistent with the measured admission coincidence of §4, which is an empirical fact, not a corollary of the identity.

B.7 FLOORS THAT PRICE THE NEGATIVE RESULTS

Proposition 6 (Hidden-gain floor)**.** *For any deterministic, possibly adaptive k-query design there is a path with C-Lipschitz derivative on which an available better-endpoint gain of at least $C/[32(k-1)^2]$ hides between query points while every queried value agrees with the flat profile; for randomized designs the same order holds in expectation.*

Proof. Answer every query with the flat profile, which fixes the adaptive query set. Some gap has width $g \geq 1/(k-1)$; place $f(\alpha) = -\frac{C}{2g^2}[(\alpha - s)(s + g - \alpha)]^2$ there: $|f''| \leq C$, all queried values match, and the hidden dip is $Cg^2/32$. □

At the grid-estimated curvatures the floor reads 0.0354–0.1066 at $k = 3$ against 0.0014–0.0043 at $k = 11$: a coarse design can step over an off-center gain, and no query pattern within budget can rule one out.

Proposition 7 (Transfer-floor tightness)**.** *For any $\Gamma > 0$ and any query density there are selection and target curves with $\sup_\alpha |L_A - L_B| = \Gamma$ on which the coefficient selected on A incurs better-endpoint regret arbitrarily close to 2Γ on B.*

Proof. $L_A(\alpha) = \varepsilon\alpha^2$, $L_B = L_A - \Gamma(2\alpha - 1)$: every budget on A selects $\hat{\alpha} = 0$, and $R_{\mathrm{best}}(0) = 2\Gamma - \varepsilon$. □

A consequence already visible in the admissions: the certified score is a cross-half residual, negative on most sides, so $q_\delta > 0$ whenever the calibration pool carries enough violators — as every screen calibrated here does — and certification at $\tau = 0$ then requires a raw score of at most $-q_\delta$, which no side reaches. The zero strict-tolerance admissions of §4 are structural, not a sample-size limitation: they persist at any pool size while $q_\delta > 0$.

C THE CURRENCY EXPERIMENTS

Three studies share one harness: LoRA adapter pairs on Qwen2.5-1.5B and -3B, a selection half S and a sealed test half T (disjoint rows of the task's evaluation slice), and a cost model that charges every strategy identically for coefficient switching, weight instantiation, and generation. The strategies: CE-11/CE-5/CE-3 read the proxy cross-entropy curve on S at eleven, five, or three coefficients and commit its minimum; Direct-5/Direct-3/Direct-2 measure generation accuracy on S at five, three, or two coefficients. Budgets are $\{0.25, 0.5, 1, 2, 4\} \times b^*$, with b^* defined as Direct-5 at 200 questions per arm on the 1.5B stack; budgets nest by prefix sampling, and a candidate arm below $n_{\min} = 25$ questions is infeasible.

C.1 CLIFF TERRAIN: GSM8K

Eight units — four LoRA pair templates (GSM8K/MBPP/MedQA adapters, two seed pairings) crossed with the two scales — fixed before outcomes were seen; GSM8K rows 0–499 form S and rows 500–999 form T. Table 3 gives the per-unit ledger at b^*.

unit	best ep. (oracle)	CE-11		Direct-5		Direct-3		Direct-2		D (pt)
		α	acc.	α	acc.	α	acc.	α	acc.	
1.5B g s0×g s1	30.0	0	29.8	1	30.0	1	30.0	1	30.0	-0.2
1.5B g s0×c s0	40.4	0.6	41.2	0.75	41.2	1	40.4	1	40.4	0.0
1.5B g s0×m s0	29.8	0.8	30.6	0.5	34.2	0.5	34.2	0	29.8	-3.6
1.5B g s1×c s1	43.6	0.7	42.6	0.75	43.0	1	43.6	1	43.6	-0.4
3B g s0×g s1	46.0	1	42.4	0.5	44.6	0.5	44.6	0	46.0	-2.2
3B g s0×c s0	54.6	0.7	53.8	0.75	54.4	1	54.6	1	54.6	-0.6
3B g s0×m s0	46.0	0.9	29.0	0	46.0	0	46.0	0	46.0	-17.0
3B g s1×c s1	51.8	0.7	50.4	0.5	49.4	1	51.8	1	51.8	1.0

Table 3: Per-unit ledger at the reference budget b^*. Test-half accuracy (%) of each strategy's committed coefficient; $D = \mathrm{Acc}_T(\text{CE-11}) - \mathrm{Acc}_T(\text{Direct-5})$ in points. *best ep. (oracle)* is the better endpoint read on the sealed test half — not a deployable selection. *Direct-2* keeps the selection-half better endpoint; it matches the oracle on 8/8 units. Unit labels: stack then adapter pair (g/c/m = GSM8K/MBPP/MedQA, s = seed).

The primary difference is $D(b^*) = -2.9$ points (95% CI $[-7.2, -0.2]$, pair-cluster bootstrap over the eight units), favoring Direct-5; Direct-3 gives -3.2 (CI $[-7.4, -0.5]$). Degradation below the better endpoint: CE-11 on 6/8 units, Direct-3 on 1. The budget sweep (Table 4) reverses nothing; the two lowest budgets' intervals include zero. The worst proxy pick (3B g s0×m s0, $\alpha = 0.9$) lands -17.0 points below the better endpoint — a shallow proxy dip that is a generation cliff.

budget	units	$D(b)$ (pt)	95% CI
$0.25 \times b^*$	4	-0.7	$[-5.3, 3.3]$
$0.5 \times b^*$	8	-2.6	$[-7.7, 0.6]$
b^*	8	$\mathbf{-2.9}$	$[-7.2, -0.2]$
$2 \times b^*$	8	-3.1	$[-7.3, -0.4]$
$4 \times b^*$	8	-3.3	$[-7.4, -0.5]$

Table 4: Proxy-loss versus direct-metric selection at matched wall-clock budget. Negative favors direct evaluation. Direct-5 is infeasible at $0.25 \times b^*$ on 3B, so that row pools 4 units. Units share adapters (four templates × two scales): exact enumeration of the 4^4 pair-template cluster resamples widens the b^* interval to $[-7.8, 0.0]$. $b^* \approx 19$ min on this hardware.

Concentration and the cheapest direct design. The mean is carried by one unit: dropping the worst leaves -0.86 points, the median across units is -0.5, and that unit contributes 74% of the signed total. The deployable two-endpoint baseline prices the same question without an oracle (Table 5): Direct-2 evaluates two existing checkpoints and instantiates no merged weights, so the same wall-clock buys 500–254 selection questions per arm; it selects the sealed-better endpoint on 8/8 units and degrades on zero. The proxy curve loses to it by -2.8 points, and Direct-3's edge over it is 0.4 point. Read against §5's useful degeneracy: on this pool the endpoint answer is not the instrument tolerating a failure mode — it is the correct selection. That record rides on a pool where sealed gains over the better endpoint are rare; it does not make endpoint selection the better strategy in general.

unit	Direct-2 $\hat{\alpha}$ / acc.	CE-11 acc.	Direct-3 acc.	best ep. (oracle)
1.5B g s0×g s1	1 / 30.0	29.8	30.0	30.0
1.5B g s0×c s0	1 / 40.4	41.2	40.4	40.4
1.5B g s0×m s0	0 / 29.8	30.6	34.2	29.8
1.5B g s1×c s1	1 / 43.6	42.6	43.6	43.6
3B g s0×g s1	0 / 46.0	42.4	44.6	46.0
3B g s0×c s0	1 / 54.6	53.8	54.6	54.6
3B g s0×m s0	0 / 46.0	29.0	46.0	46.0
3B g s1×c s1	1 / 51.8	50.4	51.8	51.8

Table 5: The deployable two-endpoint baseline, per unit (test-half accuracy in points at b^*). Direct-2 lands on the sealed-test better endpoint on all 8 units and never falls below it, while the eleven-point proxy curve averages -2.8 pt against it.

C.2 Density controls within the proxy currency

Same harness, same eight units, same split and budgets; CE-3 and CE-5 query the same proxy curve at $\{0, \frac{1}{2}, 1\}$ and $\{0, 0.25, 0.5, 0.75, 1\}$, committing the minimum exactly as CE-11 does. The three designs differ only in how many coefficients of the same curve are queried at the same budget — isolating the density factor that the proxy-versus-direct comparison confounds with the metric change. Coarser grids buy per-coefficient samples (at $0.25 \times b^*$ on 3B: $n = 500/338/147$ for CE-3/CE-5/CE-11).

Eleven to five is free: $D_5(b^*) = 0.1$ points (CI $[-0.8, 0.9]$), and at half budget the coarser grid wins (0.5, CI $[0.1, 1.0]$) — five coefficients at the full 500 samples against eleven at 307–367. Eleven to three is not free: $D_3(b^*) = -2.3$ (CI $[-4.4, -0.8]$), at every budget. The mechanism is visible in the ledger (Table 6): the proxy minima sit off-center (0.6–0.9 on six of eight units), so the three-point grid mis-localizes and falls back to an endpoint on 4 units — and endpoint commitments are where the proxy's blind spot is expensive.

unit	CE-3		CE-5		CE-11		D_3 (pt)	D_5 (pt)
	α	acc.	α	acc.	α	acc.		
1.5B g s0×g s1	0	29.8	0	29.8	0	29.8	0.0	0.0
1.5B g s0×c s0	0.5	40.6	0.75	41.2	0.6	41.2	-0.6	0.0
1.5B g s0×m s0	1	21.8	0.75	32.6	0.8	30.6	-8.8	2.0
1.5B g s1×c s1	0.5	39.2	0.75	43.0	0.7	42.6	-3.4	0.4
3B g s0×g s1	1	42.4	1	42.4	1	42.4	0.0	0.0
3B g s0×c s0	0.5	51.6	0.75	54.4	0.7	53.8	-2.2	0.6
3B g s0×m s0	1	26.4	1	26.4	0.9	29.0	-2.6	-2.6
3B g s1×c s1	0.5	49.4	0.75	51.2	0.7	50.4	-1.0	0.8

Table 6: Per-unit ledger of the same-currency density control at b^*. $D_3 = \mathrm{Acc}_T(\text{CE-3}) - \mathrm{Acc}_T(\text{CE-11})$, D_5 likewise, in points.

budget	$D_5(b)$ (pt)	95% CI	$D_3(b)$ (pt)	95% CI
$0.25 \times b^*$	0.6	$[-0.8, 2.2]$	-1.9	$[-3.2, -0.8]$
$0.5 \times b^*$	0.5	$[0.1, 1.0]$	-2.0	$[-4.2, -0.4]$
b^*	**0.1**	$[-0.8, 0.9]$	$\mathbf{-2.3}$	$[-4.4, -0.8]$
$2 \times b^*$	0.1	$[-0.8, 0.9]$	-2.3	$[-4.5, -0.8]$
$4 \times b^*$	0.1	$[-0.8, 0.9]$	-2.3	$[-4.4, -0.8]$

Table 7: Grid density within the proxy currency at matched wall-clock budget. Positive favors the coarser grid. At b^* and above, all strategies saturate the 500-row selection half, so delivered coefficients coincide across those tiers.

The same-grid contrasts isolate the metric: at five coefficients, generation selection beats proxy selection by -2.7 points; at three, by -5.5. What the direct designs buy is a measurement of the deployment metric, not a coarser grid. The three-point row carries the mirror lesson: the grid strongest under the deployment metric (Direct-3) is weakest under the proxy, because on three points the proxy commits endpoints — precisely where its blind spot costs most.

C.3 FLAT TERRAIN: PUBMEDQA WITH FRESH ADAPTERS

The same six-strategy comparison with both swappable factors changed: the deployment task is PubMedQA (rows 0–499 / 500–999 of the frozen local copy), and the pool is eight freshly trained LoRA units — four pair templates (MedQA×GSM8K and MedQA×MBPP, two seeds each) crossed with the two scales — sharing no adapter with the GSM8K pool. The cost model, budget grid, and strategies are inherited unchanged; $b^* = 257.2$ s on this task against 1140.8 s on GSM8K.

No budget separates the proxy curve from direct selection (Table 8): at b^*, $D = 0.0$ points (CI $[-0.6, 0.7]$), and every budget's interval includes zero. All six strategies sit within 0.53 points of one another at b^* (Table 9); the Direct-2 baseline picks the sealed-better endpoint on 6 of 8 units. Nothing wins, nothing loses.

budget	$D(b)$ **(pt)**	**95% CI (unit bootstrap)**	**95% CI (cluster enum.)**
$0.25 \times b^*$		infeasible under the frozen $n_{\min}$ rule	
$0.5 \times b^*$	0.1	$[-0.6, 0.9]$	$[-0.2, 0.4]$
b^*	**0.0**	$[-0.6, 0.7]$	$[-0.4, 0.3]$
$2 \times b^*$	0.2	$[-0.4, 0.8]$	$[-0.2, 0.5]$
$4 \times b^*$	-0.1	$[-0.6, 0.3]$	$[-0.5, 0.1]$

Table 8: The proxy–direct gap does not replicate on PubMedQA. $D(b) = \mathrm{Acc}_T(\hat{\alpha}_{\text{CE-11}}) - \mathrm{Acc}_T(\hat{\alpha}_{\text{Direct-5}})$ on the sealed test half, in points; positive favors the proxy. Every interval includes zero.

strategy	**mean acc. (%)**	**< better ep.**	**mean margin (pt)**
Direct-2	60.6	2/8	−0.28
Direct-3	60.3	4/8	−0.55
Direct-5	60.9	2/8	−0.03
CE-3	60.7	4/8	−0.25
CE-5	60.8	5/8	−0.10
CE-11	60.9	4/8	−0.03

Table 9: Six selection strategies at b^* on the PubMedQA pool. Sealed test-half accuracy averaged over the 8 units; all six means lie within 0.53 points.

Terrain. The median per-unit selection-half accuracy range along the path is 2.2 points here (max 5.8), against 7.7 (max 21.4) on the GSM8K pool; the worst sealed-half shortfall below the better endpoint across all six strategies is -1.6 points here against -17.0 there. On flat paths the currency choice has little left to decide. What survives both pools is the unsigned statement: at matched budget, direct task-metric selection was never worse than the proxy curve. The checkable prediction — loss-path depth, computable before any generation run, should predict which pairs repay direct selection — is stated in §7 and remains open; the within-pool reads it motivates are in §C.4.

C.4 ROBUSTNESS OF THE ORDERING, AND THE WITHIN-POOL TERRAIN READS

Leave-one-unit-out on the cliff pool. One unit carries 74% of the mean gap, so we recompute all six strategies' sealed-half means at b^* with each of the eight units dropped in turn (`t360_loo_robustness.py`; no new evaluations — the committed coefficients and sealed accuracies are read from the same three ledgers that produce Table 1). $\mathrm{Acc}_T(\text{CE-11}) - \mathrm{Acc}_T(\text{Direct-5})$ stays negative in all eight replicates, spanning -3.4 to -0.9 points, and the replicate dropping the dominant unit is the least negative at -0.9 points. This is a statement about the sign under unit removal, not about significance: at seven remaining units a -0.9-point mean sits well inside the full pool's cluster-bootstrap interval, so what survives is the direction of the ordering, not a resolved effect of that size.

Loss-side within-pool read (post-hoc). §7 reads terrain off the loss scan, while both the pool contrast and the preregistered within-pool check measure it in accuracy. We therefore compute the loss-side version on the cliff pool (`t362_terrain_loss_withinpool.py`): per-unit loss-path amplitude from the eleven-point token-CE curve on selection half A — the same scan the screen of §4 already pays for — against the per-unit proxy shortfall $-[\mathrm{Acc}_T(\text{CE-11}) - \mathrm{Acc}_T(\text{Direct-5})]$. Spearman $\rho = 0.40$, exact two-sided permutation $p = 0.33$ over all $8!$ relabelings. The flat pool is not tested on this side because its per-unit deficits carry no variation to explain (§C.3). The statistic was not preregistered — the frozen analysis plan registered the accuracy-side range as the terrain

statistic and no loss-side variant — and we report it as an exploratory trend at $n = 8$, not as a test of the prediction that preregistration leaves open.

Where the accuracy-side figure comes from. That preregistered check, frozen before any correlation was computed, resolves in neither pool, and leave-one-pair-template-out on the cliff pool locates its positive sign: dropping any of the other three templates leaves the correlation between $+0.49$ and $+0.60$, while dropping the template holding the cliff unit leaves no positive correlation at all ($n = 6$ each). We do not read the sign of that residual in either direction: with the cliff unit removed, the remaining six units' path ranges span only 1.0–9.2 points and their deficits -1.0 to $+2.2$, a spread below twice the sealed half's own sampling scale (the unpaired binomial standard error at $n = 500$ is 2.2 points per arm, itself an upper bound on the paired noise in the deficit). The defensible reading is the frozen one: the cliff pool's $+0.33$ is consistent with the pool contrast but not separable from that one template at $n = 8$.

D Calibration Details

Risk-envelope comparison. Table 10 scores four selection-half signals against the sealed-half midpoint penalty on the 98-side pool. The eleven-point barrier's increment over the nested three-point envelope is 9 sides raw and 1 side at $+0.05$; the envelopes coincide on $69/98$ sides. The midpoint penalty itself is the best point predictor (CV $R^2 = 0.6608$ against the barrier's -0.0084) but under-covers as an envelope — predict the penalty with the midpoint, certify it with the barrier.

A-half risk signal	raw coverage	$+0.05$ coverage	margin for 95/98
endpoint asymmetry $A/2$	$61/98$	$82/98$	$+0.0913$
midpoint penalty $\Delta(\frac{1}{2})$	$46/98$	$89/98$	$+0.0717$
three-point envelope B_3	$78/98$	$97/98$	$+0.0434$
eleven-point barrier B_{11}	$87/98$	$98/98$	$+0.0410$

Table 10: Risk-envelope comparison on the 98-side cross-half pool. The last column is the minimal in-pool margin covering $95/98$ sides — smaller than the fold-recomputed cross-fit medians ($+0.046$ for B_3, $+0.0445$ for B_{11}), both below the frozen $+0.05$ operating point.

Margin portability. Table 11 collects the external readouts, sorted by whether the frozen margin is testable at all. Re-running the admission sweep at the wider official-pair margin ($+0.085$ instead of $+0.05$) leaves median-tolerance admission unchanged and drops one unit at the upper-quartile tolerance: the wider margin prices admission rather than collapsing it.

Unit pool (metric)	units	B_{11} raw	B_{11} @$+0.05$	in-pool margin 95%
sealed-half sides (PRM-BCE; calibration pool)	98	$87/98$	$98/98$	$+0.041$
LoRA adapters (token-CE)	36	$36/36$	$36/36$	raw (untestable)
Pythia-2.8B held-out stack (PRM-BCE)	12	$12/12$	$12/12$	raw (untestable)
official Qwen2.5 pairs (token-CE)	32	$23/32$	$29/32$	$+0.085$

Table 11: Portability of the frozen margin, sorted by testability. The wild-specialist arm (raw failures with excesses $\leq +0.014$; covered at $+0.05$) is a fourth, shallow-break family not tabulated here.

Group-conditional margins. The pooled margin pools heterogeneous sub-populations. Group-conditional (Mondrian) calibration Vovk et al. (2022) within named groups at the same nominal level: cross-domain sides ($n = 88$) need only $+0.0221$, the Pythia stack is fully covered by $+0.0215$, while Qwen3-8B requires its maximum $+0.0495$ — the pooled margin is simultaneously loose on the easiest groups and tightest on the hardest. The within-domain group ($n = 10$) is too small to certify 95% at all, which is the operative tradeoff of group-conditional calibration. The frozen $+0.05$ covers every listed group on this pool. The stack and domain labels that fail as *predictors* of path shape (Appendix G) re-enter here as *grouping* variables: they cannot replace measurement, but they do describe the population a margin is pooled over.

Pair-level coverage. Re-aggregating per pair (a pair counts covered only if its worst side is covered), the eleven-point envelope covers $44/54$ pairs raw and $54/54$ at $+0.05$; the $90\%/95\%$ pair-max margins are $+0.021/+0.043$. Pair-level exchangeability is approximate in this pool (seed grids share endpoints), so this is a sensitivity read; the prospective pair-level statement is Theorem 1's.

Direct three-point baselines. The strongest curve-free comparator directly evaluates the two endpoints and the midpoint on the selection half. On the 49-side pool: deploying the argmin of the three direct measurements carries mean regret 0.021 against the curve low point's 0.011; a margin-screened variant is worse still (0.037); and the curve captures $19/49$ of available gains against the three-point rule's $6/49$. Direct midpoint evaluation answers "is the midpoint damaged" cheaply; it cannot locate a coefficient, classify the path, or price the peak.

Margin variants. Two calibration widths appear in this paper beyond the frozen operating point: the fold-recomputed cross-fit margin (a median over folds, quoted as a diagnostic; no single unit is screened at it) and the in-pool minimal margin (the smallest width attaining stated coverage on a pool — descriptive, never an external guarantee). Theorem 1's q_δ is the only one carrying a prospective guarantee. A number from one layer is never quoted as if it came from another.

E Resolution Details

Coefficient selectors head-to-head. On the 49-side split-half pool (A-side selection, sealed B-side scoring, loss), the curve low point is the regret-first readout and the naive convention is not competitive:

Selector	median regret	mean regret	capture rate	harmed sides
grid ceiling (in-sample upper bound)	0	0	1.00	0/49
curve low point (**regret-first**)	**0.0022**	**0.0106**	0.68	**11**/49
$\alpha^\dagger$ wedge-shrunk (gain-seeking)	0.0047	0.0137	**0.79**	**11**/49
naive $\alpha = 0.5$ (convention)	0.0540	0.0646	0.25	38/49

Table 12: Coefficient-selector head-to-head (49 held-out pair-sides; loss metric). The 0.5 default carries $6.1\times$ the curve rule's mean regret and significantly harms $38/49$ sides.

Point optimizers versus path measurement. On the same pool we simulate four search families at budgets of 5, 7, and 11 selection-half evaluations — golden-section, successive parabolic interpolation, coarse-to-fine, uniform grids, plus a 20-seed random-search control — each method's structure estimates computed from its own samples against the full-grid readouts. Adaptive searches match or beat the grid on coefficient regret alone (parabolic 0.0084 at eleven evaluations against the grid's 0.0106), but the same query sets lose the path's shape: golden-section underprices the barrier on $16/16$ arch sides, arch regimes are missed outright, off-center peaks localized at 14%–85% accuracy. An independent 162-side pool reproduces both halves of the dissociation, and the golden-section failure tail concentrates exactly on the wild pairs where global coverage matters. Table 13 gives the full bake-off.

Real-evaluation check. The bake-off queries piecewise-linear reconstructions of stored curves; we re-ran the comparison with real evaluations at matched budget (decide on half A, score on sealed half B against a sealed eleven-point grid oracle). Twelve sides — three pairs each on Llama-8B and Qwen3-8B, freshly measured — compare the eleven-point grid, the uniform five-point grid, and five-query parabolic search:

Design	queries	mean regret (12 real sides)
uniform grid	5	0.0145
adaptive parabolic search	5	0.0142
uniform grid (high-budget reference)	11	0.0087

Table 14: Real-evaluation coefficient regret at matched budget. The paired adaptive-minus-uniform difference is opposite-signed across the two stacks, so the surrogate's regret ordering does not transfer as a law; both five-query designs underprice the real sealed-half barrier, by up to 0.0882 and 0.1901 respectively, which does transfer.

Where the regret sits: an exact decomposition. Replaying the nested subgrids $S_3 = \{0, \frac{1}{2}, 1\}$ and $S_6 = \{0, 0.2, \ldots, 1\}$ of the stored curves against their common S_{11} reference, sealed-half regret decomposes pointwise as $\mathrm{tot}_k = \mathrm{sel}_k + \mathrm{disc}_k$ (split mismatch plus discretization). Pooled means: $\mathrm{tot} = 0.0275/0.008/0.0087$ at $k = 3/6/11$. Discretization contracts $7.2\times$ from three to six queries (consistent with h^2 scaling) and accounts for 78.1% of the three-query regret; from six to eleven

Method	B	Mean regret	Arch det.	Barrier und.	Peak err	Dip det.	Gain acc
uniform grid	5	0.0138	16/16	0.0080	0.050	39/42	0.857
golden-section	5	0.0116	14/16	0.0087	0.082	39/42	0.918
parabolic	5	0.0099	13/16	0.0245	0.100	40/42	0.939
coarse-to-fine	5	0.0105	16/16	0.0093	0.075	40/42	0.918
uniform grid	7	0.0127	16/16	0.0061	0.033	41/42	0.959
golden-section	7	0.0106	14/16	0.0087	0.082	41/42	0.980
parabolic	7	0.0093	13/16	0.0245	0.100	41/42	0.939
coarse-to-fine	7	0.0090	16/16	0.0082	0.038	40/42	0.939
uniform grid	11	0.0106	16/16	0.0000	0.000	42/42	1.000
golden-section	11	0.0106	14/16	0.0087	0.082	42/42	0.980
parabolic	11	0.0084	13/16	0.0245	0.100	41/42	0.959
coarse-to-fine	11	0.0085	16/16	0.0082	0.038	41/42	0.939
random*	5	0.0146	0.66	0.0389	0.126	0.87	0.800
random*	7	0.0126	0.81	0.0220	0.085	0.94	0.905
random*	11	0.0112	0.89	0.0106	0.048	0.96	0.944

Table 13: Fixed-budget bake-off: point optimizers vs. path measurement on the 49-side sealed-half pool. Regret is sealed-half loss against the sealed-grid minimum. Structure columns report what each method's own query set recovers of the full-grid readouts. *20-seed means.

the two terms offset — the paired difference is 0.0007 with CI $[-0.0023, 0.0046]$ crossing zero — and at eleven queries the regret is entirely split mismatch by construction. Two controls pin the reading: on same-half data, where split mismatch is identically zero, the minima gaps contract $8.9\times$; and supersampled 33-point rescans agree with the eleven-point minimum to within 1.46×10^{-4} on every side column. The decomposition, not any ratio of means, is the evidence that the two terms move against each other.

Grid-density audit. Subsampling all 461 census curves: the barrier is median-lossless at 7–9 points, while off-center peaks lose up to 45% at 9 points and exact coefficient agreement is 87% (99% within one grid step); the eleventh point re-homes α^* on $62/461$ curves and closes all remaining peak-miss cases. These are barrier under-reads, never midpoint-bound violations, since every sub-grid contains $\alpha = \frac{1}{2}$.

F The Generative Arms

F.1 Endpoint language geometry: the triage cue

Full-weight specialist endpoints are not language models: under each stack's official language head, their token-CE on a frozen general slice runs 5–23× the base model's (70/70 audited endpoints, zero exceptions), and merge midpoints sit at or above the better endpoint's CE on all five audited pairs — weight-space interpolation does not restore LM geometry. The multiples order the merge cost:

Stack	**endpoint CE ×base**	**median Δacc (pt)**	**acc ≥ better ep.**	**within 1 pt**	**worst side**
Pythia-1.4B	5.0×	+0.4	69%	97%	−3.9 pt
Llama-3.1-8B	12.0×	−0.1	35%	67%	−12.5 pt
Qwen3-8B	12.9×	0.0	50%	77%	−4.4 pt
Gemma-2-2B	22.9×	−1.5	2%	45%	−11.4 pt

Table 15: The accuracy-level cost of merging at the curve-selected coefficient, ordered by endpoint LM degradation (held-out B-side arms; Δacc = acc(merge) − acc(better endpoint), in points). The medians order with the endpoints' token-CE multiple over the base model — a stack-level triage cue, not a pair-level predictor.

F.2 Official specialist pairs under true generation

Setup. Officially released Qwen2.5 specialists (Instruct/Coder/Math) at 1.5B/3B/7B: pairs Instruct×Math, Instruct×Coder, Math×Coder at 1.5B and 7B, Instruct×Coder at 3B (7 pairs), plus a 14B Instruct×Coder token-CE sanity arm. A configuration gate verified weight-architecture and tokenizer identity within each size family; MBPP training rows pass a two-tier decontamination

gate against the evaluation slice. Curve sides use evaluation rows disjoint from the generation slices; generation uses GSM8K exact-match (n=200) and MBPP pass@1 (n=120), evaluated at $\{0, 0.5, 1, \alpha^*, \alpha^\dagger\}$ under the pair's shared chat template.

Pair	side/bench	α^*	barrier	dip	mid pen.	epA	epB	mid	α^*-merge
1.5b i×c	mbpp	0.00	0.806	0.019	0.806	61.7	64.2	0.0	61.7
1.5b i×m	gsm8k	0.00	4.338	0.400	3.705	45.0	27.5	1.0	45.0
1.5b m×c	gsm8k	1.00	3.098	0.322	1.530	27.5	31.5	1.0	31.5
1.5b m×c	mbpp	1.00	5.509	0.564	1.192	16.7	64.2	0.0	64.2
3b i×c	mbpp	0.00	0.221	0.044	0.221	72.5	71.7	55.0	72.5
7b i×c	mbpp	0.00	0.761	0.108	0.761	78.3	81.7	40.0	78.3
7b i×m	gsm8k	0.00	0.355	0.449	0.080	54.0	38.5	43.5	54.0
7b m×c	gsm8k	1.00	0.898	0.449	0.895	38.5	45.5	36.5	45.5
7b m×c	mbpp	1.00	0.946	0.597	0.785	23.3	81.7	1.7	81.7

Table 16: Official-pair generative cells. CE readouts (barrier, dip, midpoint penalty; selection half) alongside true-generation accuracies. All observed midpoint collapses occur in the flagged set: the curve chooses an endpoint on 6 of 7 pairs, and barrier magnitude tracks collapse severity across scales. On 2 of 9 cells the CE-selected endpoint is the generation-*worse* one — the two currencies disagree even at endpoint selection.

RoPE subgroup. The Qwen2.5-Math endpoints ship RoPE θ=10^4 against their siblings' 10^6 — a runtime geometry difference, not a weight difference. Math-involving pairs carry the largest barriers of the arm, and dual-scoring the Math endpoint under its native θ versus the merged shared θ=10^6 prices the RoPE-shift component at 17 points of GSM8K accuracy, separate from the merge penalty proper.

Recipes at the curve coefficient. TIES and DARE at full strength (λ=1) collapse on all cells (median accuracy 0%) — including cells where plain midpoint LERP retains 40–55%: sparsification at full dose is strictly more destructive than naive averaging here. Re-armed at the curve-selected coefficient, both recipes recover to 61.5–74.2% on exactly the three large low-barrier cells, statistically indistinguishable from LERP at the same point, while the high-barrier Math×Coder pairs stay at 0% under every configuration (Table 17).

Pair	bench	LERP mid	LERP α^*	TIES λ=1	TIES λ=α^*	DARE λ=1	DARE λ=α^*
1.5b i×c	mbpp	0.0	61.7	0.0	0.8	0.0	0.8
1.5b i×m	gsm8k	1.0	45.0	0.0	9.5	1.5	9.5
1.5b m×c	gsm8k	1.0	31.5	0.5	0.5	0.0	0.0
1.5b m×c	mbpp	0.0	64.2	0.0	0.0	0.0	0.0
3b i×c	mbpp	55.0	72.5	0.0	71.7	0.0	71.7
7b i×c	mbpp	40.0	78.3	0.0	74.2	0.0	74.2
7b i×m	gsm8k	43.5	54.0	0.0	61.5	0.5	61.5
7b m×c	gsm8k	36.5	45.5	0.0	0.0	0.0	0.0
7b m×c	mbpp	1.7	81.7	0.0	0.0	0.0	0.0

Table 17: Sparsifying merge recipes under the curve coefficient. TIES/DARE at full strength and re-armed at λ=α^*, against plain LERP. The two recipes agree exactly on 14/18 cells: the coefficient, not the recipe, is the operative variable.

Three-model simplex probe. A 66-point lattice over the Qwen2.5-1.5B Instruct×Math×Coder simplex: the centroid mixture carries excess beyond every pairwise edge barrier on two of three sides (pairwise curves do not price the soup), the worst MBPP point is strictly interior — invisible to any edge scan — and gain is essentially absent from the interior. Pairwise envelopes are necessary, not sufficient, for multi-model soups; the dimensional decay of Appendix B is the backdrop.

side (half)	centroid excess	max edge barrier	PS1	worst point	interior beats chord
gsm8k (A)	4.05	4.34	N	4.32 @ (0.6,0.4,0.0) edge	1/36
gsm8k (B)	4.05	4.37	N	4.36 @ (0.6,0.4,0.0) edge	0/36
mbpp (A)	6.83	6.44	**Y**	6.91 @ (0.2,0.4,0.4) interior	0/36
mbpp (B)	6.57	6.36	**Y**	6.57 @ (0.3,0.3,0.4) interior	0/36
wikitext (A)	3.91	3.89	**Y**	4.79 @ (0.0,0.2,0.8) edge	0/36
wikitext (B)	3.81	3.73	**Y**	4.62 @ (0.0,0.2,0.8) edge	0/36

Table 18: Three-model simplex probe (66-point λ-lattice, excess over the barycentric chord). Edge points reproduce the dedicated pairwise scans; the global worst mixture is strictly interior on MBPP.

F.3 GENERATION-PRESERVING LORA ARMS

Construction. Specialists fine-tuned as LoRA rank 16 (attention+MLP only; embeddings and LM head never updated, so generation is preserved by design) on the math/code/medical domains, two seeds each, on Gemma-2-2B and Qwen3-8B. Under a shared base, merging is base + tensor-wise average of the LoRA deltas — exactly the full-weight averaging the instrument assumes. All 15 pairs per stack scanned on the frozen eleven-point split-half protocol under token-CE, plus generative accuracy on GSM8K ($n = 200$) and PubMedQA ($n = 500$) at $\{0, 0.5, 1, \alpha^\dagger\}$.

Penalty side: carries over unchanged. Zero cross-half violations on both stacks ($0/54$ side-directions each). The selected coefficient beats the 0.5 convention on $17/27$ sealed-half sides at 2B and $18/27$ at 8B (median held-out regret 0.0031 vs. 0.0126; significant harm on $10/27$ vs. the convention's $20/27$), the wins concentrating on cross-domain sides. The convention's harm rate sits at the full-weight level but its magnitude shrinks fivefold: smaller displacements tax less, not less often.

Accuracy side: the blind spot persists as a power statement. Median Δacc($\alpha^\dagger$ vs. 0.5) is within the noise floor on both stacks, and nothing collapses — contrast the 5–23$\times$ CE blow-up of full-weight endpoints. But dips of this magnitude do not cross the argmax at these sample sizes and floor-adjacent endpoint accuracies.

Replication with headroom. Re-trained with endpoints that answer (Qwen2.5-1.5B/3B LoRA; GSM8K/MBPP/MedQA; in-domain endpoint accuracies 33–48% GSM8K): the gain side changes qualitatively. The merged model beats *both* endpoints on $6/18$ generation units, by up to $+6.5$ points, the wins concentrating on cross-domain pairs scored on a strength side — where the dips live. With endpoints that answer, task-accuracy gains appear where the loss geography says they should. The count is an any-of-five-candidates rate (per-candidate: $3/18$ at the midpoint, $4/18$ at α^*, $2/18$ at $\alpha^\dagger$), and individual wins, while real on their items, are under-powered at these benchmark sizes — every item-bootstrapped interval reaches zero — which is why the unit-level rate carries the claim.

The prospective retest. We re-tested the one pattern shared across both scales — GSM8K-side gains on GSM8K$\times$MBPP pairs — on held-out seeds at five times the original sample size ($n = 1000$). The retest does not confirm it: $+1.1$ points at 1.5B against -1.1 at 3B, per-stack item bootstraps spanning zero on both sides, and a pooled paired McNemar exactly null ($p = 1.0$; 126 versus 126 discordant items). The combined result is consistent with no net accuracy gain; the original cross-scale observation stands as unreplicated, and a single prespecified selection rule on a fresh, larger unit pool is the designated continuation.

F.4 WILD SPECIALISTS AND THE GAIN-SIGNAL EXPANSION

On seven wild Mistral-7B specialist pairs Jiang et al. (2023); Yu et al. (2024b); Labrak et al. (2024); Teknium (2023) (GSM8K $n = 200$, PubMedQA $n = 500$), the curve readout embeds into TIES unchanged: replacing the symmetric mixing weights by $\alpha^\dagger$ improves default TIES in 3 of 6 two-specialist cells, and LERP+$\alpha^\dagger$ improves on the 50-50 convention in $9/14$ cells. The penalty bound holds on the distribution-level metric as well ($0/7$ violations), below-chord dips are universal, and the BCE direction between merge@0.5 and merge@α^* agrees with the generative-accuracy direction on $7/7$ pairs. A sealed-half expansion over thirteen cells (the seven pairs on PubMedQA and MMLU Hendrycks et al. (2021), plus six Qwen2.5-7B crosses) freezes the coefficient on one half and scores the other: the raw bound fails shallowly (excesses $\leq +0.014$) while $+0.05$ covers, and the frozen α^* matches or beats merge@0.5 on 25/26 cells in BCE.

On the Pythia gain-rule pool expanded to forty sides, the hit rate of $\alpha^\dagger$ is monotone in the census-gain signal ($0/6 \rightarrow 13/16$ across bands; 69% above gain 0.002 vs. 9% below), domain pairing predicts nothing once signal strength is held fixed, null sides degenerate harmlessly to an endpoint ($6/6$ no interior harm), and $11/40$ sides lose >0.005 to the better endpoint (worst -0.124). The gain readout prices the probability of transfer with a measured resolution floor; it is not a guarantee. The readout also composes with operators learned from auxiliary data (Gagnon-Audet et al., 2023; Oh et al., 2025). A three-model soup probe (ten triplets) aggregates pairwise curves into soup coefficients that beat uniform $\frac{1}{3}$ on $22/29$ sides by loss — positive on average, unordered by the pairwise gain signal, and unpriced, consistent with the dimensional decay of Appendix B.

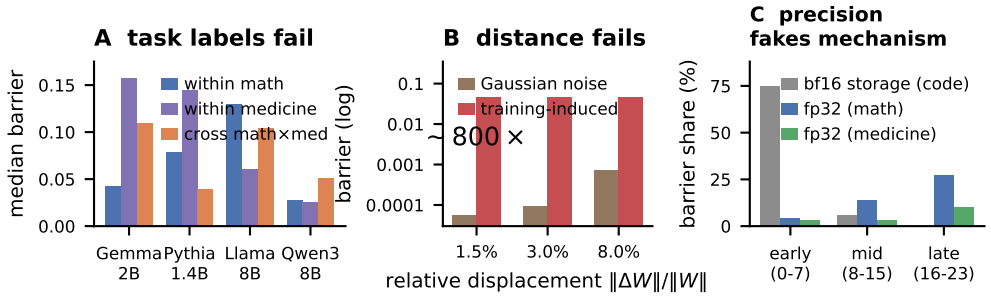

Figure 4: Why endpoint cues fail. A: task labels do not order the path — the unstable pole changes across stacks. **B:** distance does not price the path — Gaussian perturbation at matched displacement yields $\sim$800$\times$ smaller barriers than training-induced displacement. **C:** numerical precision can fake the mechanism — the bf16-era early-layer concentration reverses under fp32 re-measurement.

κ sensitivity. The shrinkage $\kappa = \frac{1}{2}$ was fixed as part of the geometric construction, not selected from outcomes. Across five side pools (146 sides), no currency makes it all-pool optimal: pooled median regret is minimized at $\kappa = 0$ (the unshrunk minimum), while the gain-side hit rate is a flat plateau over $\kappa \in [0, 0.75]$. The per-pool minimizer disagrees across pools, so §5 keeps the regret-first low point as the default and §6 uses $\alpha^\dagger$ only where gain is the objective.

G What Endpoint Cues Do Not Tell You

The inexpensive alternatives to measurement fail in measured ways (Figure 4).

Task labels do not order the path. Within-domain seed pairs can be less mergeable than cross-domain pairs — medical seed pairs on Pythia-1.4B show 3$\times$ higher within- than cross-domain barriers — yet the ordering is stack-dependent (math leads at 8B), pair-specific (a five-seed medical analysis spans a 15$\times$ barrier range by partner Juneja et al. (2023)), and it replicates with sharpened separation on three disjoint corpus families (GSM8K/MBPP/MedQA, 27 pair-sides: every within-domain side above every cross-domain side on chord deviation, cross-domain curves dipping $\sim$7$\times$ deeper). Under matched label granularity the pattern holds and the arch magnitude itself is regime-dependent.

Weight-space distance does not price the path. Within- and cross-domain pairs are statistically indistinguishable in weight space (median relative ΔW 0.0325 vs. 0.0322, MWU $p = 0.785$; a 210-pair all-pairs census), and Gaussian perturbation at matched displacement produces $\sim$800$\times$ smaller barriers ($\sim$63$\times$ at fivefold-inflated displacement). The barrier prices the directional structure of training updates, not their magnitude — the chord-excess lemma's Rayleigh-quotient reading (Appendix B), consistent with the weak parameter-space signal reported by Cao et al. (2026).

Layer-level readings do not survive storage precision. A bf16-era early-layer concentration of the barrier reverses under fp32 re-measurement (late blocks carry the largest share; layer-wise protocol of Adilova et al. (2024)), and a precision-only control reproduces the reversal with no training at all: sub-quantum coordinates contribute a rounding floor that reads a true early-layer share of 6.6% as 50.8% under bf16. A matched-step training audit shows the bf16 arm reaching nearly identical training loss while moving only about three-quarters as far — loss equivalence is not geometric equivalence.

Endpoint configuration is a confound, not a weight property. The Math specialist's RoPE base differs from its siblings' and inflates its pairs' barriers; the dual-geometry control of Appendix F.2 prices that component apart from the merge penalty. Even a shared checkpoint is not the operative boundary: cross-pretrain pairs (Pythia-1.4B versus its deduplicated variant, same-domain fine-tuned) reconverge into a shared basin; only fully independent initializations remain the outer boundary. Training-log stability proxies rank with the within-domain barrier on some stacks and reverse sign under a caliber change on others — they do not replace the scan.

H EVALUATION SETUP

Data. MMLU-Pro-CoT-Train-Labeled (derived from MMLU-Pro Wang et al. (2024)) filtered to 79,204 well-formed chains; a 7-domain fine grid derived by label rewriting only, preserving the question-level train/validation partition (Table 19). Evaluation: stratified per-category sampling with whole-chain truncation, $n = 2{,}000$ step records per side; selection/test split by question-ID hash rank, with the complement half verified disjoint per run.

Grid (stacks)	Label	Source categories	chains
4-domain (Pythia, Gemma)	math	math	5,952
	code	computer science + engineering	5,677
	medical	health + biology	11,184
	general	the nine remaining categories	44,456
fine-7 (Llama, Qwen3, Pythia)	math, physics, chemistry, engineering, business	each a single native category	4.3k–6.0k
	social_fine	economics + psychology	13,079
	medicine	health + biology	11,184

Table 19: Domain registry. Labels name datasets, not abstractions: each label's exact category composition in the canonical set. The two grids share the medicine composition.

Models. Pythia-1.4B Biderman et al. (2023), Gemma-2-2B, Llama-3.1-8B, Qwen3-8B-Base, Qwen2.5-1.5B/3B/7B/14B (Instruct/Coder/Math release lines), Pythia-2.8B, and the Mistral-7B family (base, Instruct, MetaMath, BioMistral, OpenHermes-2.5) — all open-weight checkpoints cached at fixed revisions. Fine-tuning: 8-bit AdamW, learning rate 10^{-4}, effective batch 128, one epoch, fp32 master weights.

Interpolation. All parameters interpolated (normalization gains and biases included, no special-casing); eleven-point grid $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$ computed in fp32 on CPU and loaded to GPU per point.

Compute. Fine-tuning runs ≈45 minutes at 1.4B–2B and ≈36 minutes at 8B; each eleven-point scan ≈25–50 minutes at the small scale and ≈3 hours at 8B (two domains). The census covers 84 pairs (Llama), 69 (Gemma), 51 (Pythia re-scan), and 84 (Qwen3). All experiments ran offline; regenerating one 8B stack end-to-end costs on the order of 250 GPU-hours.