The Compatibility Atlas: A Convex-Geometric Index for LoRA Libraries

Sushaan Kandukoori, Aarit Atreja, Devom Brahmbhatt, Avaneesh Parvathareddy, Vladimir Filkov

sushaankandukoori@gmail.com, aaritatreja@gmail.com, devom.hb@gmail.com,

avaneesh.parvathareddy@gmail.com, vfilkov@ucdavis.edu

Abstract

LoRA adapters merge into one model at low cost, but a fifty-adapter library already yields more than two million five-way subsets, too many for exhaustive held-out validation. We fix the base model, merge rule, allowed merge weights, evaluation prompts, and the tokens on which the loss is scored, then fit a quadratic model of each task's loss in the merge weights. Each query names adapters to combine and tasks to retain; a small convex optimization over the fitted models returns the verdict and a surrogate certificate for it, with no further evaluation of the merged model; certifying the deployed loss itself additionally requires uniform approximation envelopes, which we measure on finite designs rather than establish over the whole chart. The per-library record of these verdicts is the Compatibility Atlas. By Helly's theorem, any rejection of the fitted models is forced by at most $d{+}1$ tasks; by Carathéodory's theorem, an accepted merge places its free weight on at most $d{+}1$ adapters, where d is the number of free merge weights; and for $d \geq 2$, checking every pair at a fixed threshold is not enough to decide every triple. One fit on sixteen adapters, using the loss on answer tokens, predicts the measured loss of 120 off-design three-adapter merges at Pearson correlation 0.96 (bootstrap interval $[0.75, 0.98]$); under linear merging, one fit on fifty screens 1,096 pair and triple merges at AUROC 0.95. On a separate library of fifty adapters and four merge methods, the locally fitted score recovers 45.5% of the merges that pass held-out retention validation at 78% precision, using 27% of the exhaustive validation cost.

1 Introduction

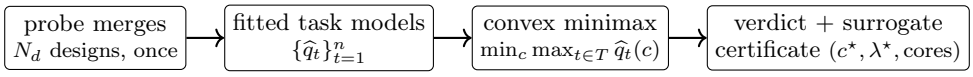
Given a library of LoRA adapters trained on the same base, exhaustively validating every candidate merge on held-out data is expensive, since a fifty-adapter library already has $\binom{50}{5} = 2{,}118{,}760$ five-way candidates. Our running example is a Qwen2.5-0.5B-Instruct base [Qwen Team, 2024] with a Hugging Face PEFT [Mangrulkar et al., 2022] library of fifty LoRA adapters spanning instruction following, code generation, GSM8K-style math [Cobbe et al., 2021], SQL writing, summarization, refusal training, and multi-turn chat. Many merge operators exist, among them weight averaging [Wortsman et al., 2022], task arithmetic [Ilharco et al., 2023], TIES (trim, elect sign, and merge) [Yadav et al., 2023], and variants of DARE (drop and rescale) [Yu et al., 2024], and all of them assume the subset has already been chosen. We ask which subsets are worth validating.

Compatibility Atlas. Given calibration losses for a small design of merge coefficients, fit one convex quadratic $\widehat{q}_t(c)$ per task, once. A query (T, U, ℓ) names the tasks to retain, the adapters allowed into the merge, and a per-adapter weight floor; answer it by solving

$$\min_{c \in C(U,\ell)} \max_{t \in T} \widehat{q}_t(c).$$

Return merge weights $c^\star$, a surrogate accept/reject certificate, a small blocking task core if rejected, and a sparse adapter core if accepted. No merged model is evaluated at query time.

A concrete query from this library asks whether the math, SQL, and summarization adapters can be merged, each with weight at least 0.2, so that the math and SQL tasks keep their single-adapter performance.



Reusable atlas	Result
Mistral index: 16 adapters, 136 calibration merges	0.96 Pearson on 120 off-design triples
Mistral index: 50 adapters, 1,275 calibration merges	0.95 AUROC on 1,096 pair/triple queries
Local Qwen screen, 20% validation budget	45.5% recall, 77.6% precision, 73% compute saved

Figure 1: The atlas pipeline and its main results: probe merges are scored once per task, and every later subset query is a convex minimax over the fitted models, returning a surrogate-certified verdict with task and adapter cores of order at most $d{+}1$.

Every certificate in this paper is a *surrogate certificate*, a guarantee about the fitted convex models; upgrading it to a *deployed-loss certificate*, a guarantee about the model that is actually served, requires uniform approximation envelopes over the query chart (Theorem 2.3), which we measure on finite designs but do not establish uniformly. The per-library record of verdicts is the *Compatibility Atlas* (Figure 1; the full pipeline is Algorithm 1).

Our contributions:

- **A surrogate-certified convex screen.** Sion duality gives every verdict a checkable primal–dual certificate (Theorem 2.1), and supplying approximation envelopes upgrades it to a sound accept/reject/inconclusive test for the deployed loss (Theorem 2.3).
- **Two-sided finite cores.** Rejection is forced by at most $d{+}1$ protected tasks (Helly) and an accepted merge places its free weight on at most $d{+}1$ adapters (Carathéodory), the two cores forming a discrete saddle (Theorem 2.4); a shared rank r lowers both orders to $r{+}1$ (Theorem C.2).
- **Pairwise impossibility.** For $d \geq 2$ no thresholded pairwise compatibility graph decides every triple (Theorem 2.5); the disagreements occur in practice (Table 4).
- **Evidence at library scale.** We validate atlas indexes fitted once at sixteen to fifty adapters, a four-operator screening run, and deployed coefficient interventions (Figure 1; Section 4).

Complete proofs and the full experimental record appear in the extended version [Kandukoori et al., 2026].

2 Compatibility queries over tasks and adapters

A query names the tasks the merge must retain, the adapters allowed into it, and a minimum weight for each selected adapter; Table 1 in Appendix B collects notation.

Let f_{θ_0} be a base model and $\Delta_i \in \mathbb{R}^p$ LoRA updates trained from the same checkpoint [Hu et al., 2022]; allowed merges are parameterized by a compact convex *chart* $C \subset \mathbb{R}^d$, each point $c \in C$ specifying merge weights and a merged update $\Delta(c)$.

A query fixes the adapters $U \subseteq [m]$ allowed in the merge, the tasks $T \subseteq [n]$ it must retain, and a coefficient floor $\ell \in \mathbb{R}^m_{\geq 0}$ supported on U with $\|\ell\|_1 \leq 1$; the free mass is $s_\ell := 1 - \|\ell\|_1$. The merge chart is the simplex on U raised above the floor, $C(U, \ell) = \{\ell + s_\ell z : z \in \Delta_U\} \subseteq \mathbb{R}^m$, a compact convex chart of intrinsic affine dimension $|U|{-}1$ when $s_\ell > 0$; when $s_\ell = 0$ it is the single point ℓ. The *Helly dimension* that enters the core bounds is this intrinsic dimension rather than the ambient m; for a set $V \subseteq [m]$, $C_\ell(V) = \{\ell + s_\ell z : z \in \Delta_V\}$ spreads the free mass over V alone.

For task t, let $J_t : \mathbb{R}^p \to \mathcal{Y}_t$ be the linear map sending a weight update to its effect on task t's outputs, with $\mathcal{Y}_t$ a Hilbert space. With anchor $a_t \in C$, the squared change in task t's output as the merge moves from a_t to c is

$$q_t(c) = \|J_t(\Delta(c) - \Delta(a_t))\|^2_{\mathcal{Y}_t}\,, \tag{1}$$

a convex quadratic in the coefficients when $\Delta(c)$ is linear in c, as for the exact weighted sum $\sum_i c_i \Delta W_i$. The factor-space operator used in the public screening panels (Section 4.3; Appendix B) is not of that form; for that operator, Eq. (1) motivates rather than derives the quadratic, which we treat as a fitted surrogate whose fidelity is measured directly. For validation loss claims we fit a second-order approximation of the

retained loss f_t, scored on every prompt token or only on the tokens of a task's answer, the *answer loss*. The true second derivative need not be PSD for nonlinear losses [Garipov et al., 2018, Frankle et al., 2020, Ortiz-Jiménez et al., 2023], so we fit an unprojected quadratic model $\widetilde{q}_t$ with curvature $\widetilde{H}_t$, take the positive semidefinite projection $H_t := \Pi_{\mathrm{PSD}}(\widetilde{H}_t)$, and run every certificate computation on the surrogate

$$\widehat{q}_t(c) = \alpha_t + b_t^\top (c - a_t) + \tfrac{1}{2}(c - a_t)^\top H_t (c - a_t), \tag{2}$$

with $H_t \succeq 0$, logging the projection size per query. The theory below is stated for generic convex task models q_t; in the experiments q_t is instantiated by the fitted $\widehat{q}_t$, so every certificate is a statement about the fitted models rather than the deployed loss.

For a protected-task subset $S \subseteq [n]$, define the minimax retention score

$$\rho_S(C) = \min_{c \in C} \max_{t \in S} q_t(c). \tag{3}$$

The subset is compatible at threshold τ iff $\rho_S(C) \le \tau$. Eq. (3) is a quasiconvex program [Eppstein, 2005, Amenta, 1994] and the diagonal case of a two-sided score that separates the protected tasks from the merge adapters and their floor,

$$\rho(T; U, \ell) = \min_{c \in C(U,\ell)} \max_{t \in T} q_t(c), \tag{4}$$

with $\rho(A; V, \ell)$ the same minimax on $C_\ell(V)$, while a general chart C (a box, or a router whose gating weights live in a lower-dimensional convex set) keeps the duality and Helly certificates. A set of tasks is jointly retainable at level r when their sublevel sets $F_t(r) = \{c \in C : q_t(c) \le r\}$ share a point (Figure 3).

Theorem 2.1 (Primal-dual certificate)**.** *For compact convex C and continuous convex quadratic task models q_t,*

$$\rho_S(C) = \max_{\lambda \in \Delta_S} \min_{c \in C} \sum_{t \in S} \lambda_t q_t(c), \tag{5}$$

where Δ_S is the probability simplex on S [Sion, 1958, Rockafellar, 1970]. Thus any primal $c \in C$ and dual $\lambda \in \Delta_S$ give $L(\lambda) := \min_{z \in C} \sum_t \lambda_t q_t(z) \le \rho_S(C) \le U(c) := \max_t q_t(c)$. For any threshold τ, $U(c) \le \tau$ certifies compatibility at level τ, $L(\lambda) > \tau$ certifies incompatibility at level τ, and $U(c) - L(\lambda)$ certifies optimality of $\rho_S(C)$.

When the two bounds meet, the pair (c, λ) certifies the exact value.

Proposition 2.2 (Projection is accept-sound)**.** *Since $H_t = \Pi_{\mathrm{PSD}}(\widetilde{H}_t) \succeq \widetilde{H}_t$, we have $\widehat{q}_t \ge \widetilde{q}_t$ on C, so any primal c with $\max_t \widehat{q}_t(c) \le \tau$ certifies compatibility for the unprojected fitted models as well: projection can remove an accept but cannot introduce one. Errors relative to the deployed loss still depend on the Taylor remainder and the estimation error in the fitted coefficients, not on the projection.*

Theorem 2.3 (Target certificate for the deployed loss)**.** *Fix a query (T, U, ℓ) and let f_t be the true retained loss of task t, with $\rho^f(T; U, \ell)$ the two-sided score under the deployed loss. Suppose envelopes $\epsilon_t^-, \epsilon_t^+ \ge 0$ bracket the task model against this loss on the query chart, $\widehat{q}_t(c) - \epsilon_t^- \le f_t(c) \le \widehat{q}_t(c) + \epsilon_t^+$ for all $t \in T$ and $c \in C(U, \ell)$. For a primal c and a dual $\lambda \in \Delta_T$, set $U_{\mathrm{true}}(c) := \max_t(\widehat{q}_t(c) + \epsilon_t^+)$ and $L_{\mathrm{true}}(\lambda) := \min_c \sum_t \lambda_t \widehat{q}_t(c) - \sum_t \lambda_t \epsilon_t^-$. Then $L_{\mathrm{true}}(\lambda) \le \rho^f(T; U, \ell) \le U_{\mathrm{true}}(c)$: $U_{\mathrm{true}}(c) \le \tau$ accepts the query, $L_{\mathrm{true}}(\lambda) > \tau$ rejects it, and otherwise the query is inconclusive; both verdicts are sound for the true retained loss, and they cannot disagree. (Proof in Appendix D.2.)*

The envelopes collect the calibration sampling error, the design interpolation error, and the nonlinear remainder; setting every envelope to zero recovers the surrogate certificate, and Section 4.1 measures the envelopes on dense three-adapter charts.

Helly's theorem says that if every $d{+}1$ members of a finite family of convex sets in $\mathbb{R}^d$ have a common point then the whole family does; Carathéodory's, that any point in the convex hull of a set in $\mathbb{R}^d$ is a convex combination of at most $d{+}1$ of its points.

When every task model depends on the coefficient only through a shared linear map, $q_t(c) = \tilde{q}_t(Rc)$ with $\tilde{q}_t$ convex, write $k(C) = \dim \operatorname{aff} R(C)$ for the latent-image dimension, $a_i = Rv_i$ for the images of the adapter atoms v_i (the pure single-adapter settings), and $k_U = \dim \operatorname{aff}\{a_i : i \in U\}$, so $k_U \le \min(|U|{-}1, \operatorname{rank} R)$ and $k_U = k(C(U, \ell))$ when $s_\ell > 0$.

Theorem 2.4 (Finite cores, one- and two-sided)**.** *Let $C \subset \mathbb{R}^d$ be compact and convex. (i)* Fixed chart. *If S is incompatible at threshold r, some $T \subseteq S$ with $|T| \le d{+}1$ is already incompatible at r [Helly, 1923]; the full score is realized by a sparse core,*

$$\rho_S(C) = \max_{T \subseteq S,\ |T| \le d+1} \rho_T(C), \tag{6}$$

and the order $d{+}1$ is tight. Suppose further that $q_t(c) = \tilde{q}_t(Rc)$ with $\tilde{q}_t$ convex, and consider the two-sided score $\rho(T; U, \ell)$ of Eq. (4). (ii) Task core. *Applying Helly in the latent image gives $A^\star \subseteq T$, $|A^\star| \le k(C(U,\ell))+1$, with $\rho(A^\star; U, \ell) = \rho(T; U, \ell)$; this needs only a compact convex chart, so it holds for a router or box as well. (iii)* Adapter core. *On the chart $C(U,\ell)$ with $s_\ell > 0$, some minimizer $c^\star$ has residual support $V^\star = \{i : c_i^\star > \ell_i\}$ of size at most $k_U{+}1$, and restricting the free mass to $V^\star$ preserves the score; at zero floor the merge therefore uses at most $k_U{+}1$ adapters. (iv)* Discrete saddle. *For $s_\ell > 0$, over nonempty $A \subseteq T$ and $V \subseteq U$ of size at most $k_U{+}1$, $\max_A \min_V \rho(A; V, \ell) = \min_V \max_A \rho(A; V, \ell) = \rho(T; U, \ell)$. Since $k_U \le d$ every core has order at most $d{+}1$. (Proofs in Appendix D.)*

Every fixed-chart threshold decision therefore reduces to a max-core lookup, so the $(d{+}1)$-core score table is a sufficient statistic for exact fixed-chart threshold decisions (formal statement and two-sided generalization in the extended version [Kandukoori et al., 2026]).

An accepted query never needs more than $k_U{+}1$ adapters above its floor, while rejection is already forced by some $k(C(U,\ell))+1$ of the protected tasks; on a simplex chart with every adapter free the adapter bound is vacuous, and it binds when the models share a lower-rank map or the chart is a router of dimension below $|U| - 1$. Both orders are simultaneously tight: a symmetric instance with $m = n = r{+}1$ and adapter atoms at the vertices of a regular r-simplex forces each core to use all $r{+}1$ indices (Theorem D.1, Appendix D.5).

Theorem 2.5 (No pairwise surrogate on a fixed chart)**.** *There is no universally exact representation of multiway adapter compatibility by a pairwise mergeability graph. For $d \ge 2$, two libraries can have identical pairwise mergeability graphs on the same adapters but different three-way compatibility. Hence clique completion of a pairwise graph is incomplete for local LoRA-library merging. (Proof in Appendix D.8.)*

Beyond this fixed-chart obstruction, query-dependent floors create a second, chart-mismatch disagreement, in which a triple is feasible on its own chart while a pair sub-query fails on the pair chart; Figure 3 in Appendix B illustrates both, and Table 4 counts them on the Qwen $m{=}50$ panel.

For update-space merging the atlas is invariant to LoRA factor reparameterizations, so no raw-factor-coordinate or cosine rule can be universally exact (Theorem C.3, Appendix C); the deployed factor-space operator falls outside this class, and Section 4.1 compares it with the exact update sum.

3 Effective dimension and the atlas as an index

If the task models share an r-dimensional representation, both core bounds improve from $d{+}1$ to $r{+}1$, and the order is tight; the minimal such r is exactly the rank $r_{\text{exact}}(C)$ of an intrinsic curvature matrix computed in the chart's own coordinates (Theorems C.1 and C.2, Appendix C), and a factorization holding only within a uniform error δ moves every score by at most 2δ (Appendix C).

The rank equals d unless the family omits directions, so exact compression needs a low-rank library; measured across our panels r_{exact} matches d, and the small cores there come from sparse optima, not low dimension. Three statements are distinct: the worst-case order $d{+}1$ always holds, a certified shared rank r lowers it to $r{+}1$, and an optimum can lie on a low-dimensional face even when $r_{\text{exact}} = d$; the experiments exhibit this last kind of sparsity. As a computable diagnostic, the mean-Hessian projector residual certifies an upper bound on the spectral effective dimension $\sigma\text{-}\operatorname{adim}_\epsilon$, the smallest rank at which any projector's strict per-task residual falls below ϵ (Appendix C).

Implementation modes. In the primary regime, the *atlas index*, task models are fitted once on a shared coefficient chart and reused for every later query, the only regime in which compatibility is downward-closed in the protected set; two variants, the *restricted atlas* (fit on the ambient simplex, answer by restriction) and the *local quadratic screen* (refit per query, the regime of the public screening panels), are specified in Appendix B.

The atlas index. An atlas is a chart $C \subset \mathbb{R}^d$ with one atom per adapter, one fitted task model per protected task, and one verdict per queried subset, all relative to a fixed base model, merge operator, ambient chart, and retention loss; a new operator or a new choice of scored tokens needs a new atlas. Each task has a PSD-projected channel for certificates and an unprojected channel that ranks better where projection was active (+0.04 pooled AUROC, Appendix A), and a verdict record stores the coefficient, cores, dual weights, certificate gap, and fit provenance; Algorithm 1 (Appendix B) collects the pipeline.

On a shared m-adapter simplex with design step $\frac{1}{2}$, the poised quadratic design (poised: the losses measured at its points determine the quadratic's coefficients uniquely) consists of all single adapters and all equal-weight pairs; this is Scheffé's quadratic simplex-lattice design $\{m, 2\}$ [Scheffé, 1958], the classical saturated design for quadratic response surfaces on a mixture simplex [Box and Wilson, 1951]. At $d = m-1$ this gives $N_d = m + \binom{m}{2}$ design merges, so building the fifty-adapter atlas takes one pass over 50 singles and 1,225 equal-weight pairs. Every later query is a convex minimax over the fitted quadratics; on a laptop CPU core the released solver answers one in 13 to 46 milliseconds on the m=8 charts and at a median of 0.46 seconds for a five-way query at d=49 (zero-floor timings; at the default floor 0.2 a five-way query reduces to the single equal-floor point). At the measured effective ranks the far smaller $(r+1)$-core tables reproduce the full scores on the Qwen and TinyLlama m=8 shared charts (exactly at r=3 on Qwen, to below 10^{-3} at r=2 on TinyLlama), and the gap admits a per-library interval certificate (Proposition C.5) that settles 23% of 211 subset queries on the Qwen chart and none on the lossier TinyLlama chart; no certified surrogate decision disagreed at any threshold in a 201-point sweep.

A verdict is valid for the base and adapter versions recorded with it, and the minimax score is 1-Lipschitz in the task models, so a stored verdict, as a statement about the fitted models, stays valid while the accumulated drift bound stays below its certificate margin γ; measured insertion and consolidation drifts sit at the scale of typical margins (Appendix B.5), so the rule triggers often, and its present value is the soundness of every served surrogate verdict, not refit avoidance.

4 Experiments

The experiments cover six public panels evaluating the local quadratic screen, with task models refitted per queried subset: the largest uses Qwen2.5-0.5B-Instruct with fifty adapters and 4,000 pair and triple queries across four merge operators, and five smaller panels on TinyLlama-v0.3 [Zhang et al., 2024] at m=12 and Mistral-7B-Instruct [Jiang et al., 2023], Phi-2 [Javaheripi and Bubeck, 2023], Pythia-410M, and Pythia-1B [Biderman et al., 2023] at m=10 enumerate every pair and triple. The answer-loss panels, on Mistral-7B at m=8 to 50 and Qwen2.5-1.5B at m=8, instead fit once on the ambient simplex and answer later queries on its faces. All evaluations share one configuration fixing the feasibility rule, splits, the four operators (linear; TIES; DARE-TIES; and *magnitude pruning*, TIES with the trim step but without sign-conflict resolution), the retention thresholds, and the floor ℓ_i=0.2. Feasibility labels are finite-design labels: a positive query has at least one evaluated coefficient that passes the retention rule on held-out prompts, and a negative label does not rule out an unevaluated coefficient elsewhere on the chart (Appendix B).

4.1 Prediction and coefficient selection from a single fit

The retention labels throughout are loss-based, and whether loss screening transfers to task accuracy depends on where the loss is scored. On five classification tasks from the released Mistral-7B panel (45 equal-weight records, adapter gains +0.20 to +0.97 accuracy), the answer loss tracks measured accuracy (Pearson −0.72 pooled) while the prompt-scored loss does not (+0.37); as a ranker at the top-30% budget the answer loss recovers 48% of accuracy-retaining records at 71% precision against 24% at 36% for the prompt loss. Quadratic models fitted once on the answer loss retain the association, at Pearson 0.81 against measured answer loss and AUROC 0.78 for accuracy retention (Appendix A).

Table 3 (Appendix B) and Figure 2 summarize six panels scored on the answer loss, across two bases, three operators, and libraries of four to sixteen adapters (the sixteen from the Lots-of-LoRAs collection [Brüel-Gabrielsson et al., 2024]); every panel has positive adapter gains and its measured answer loss moves with task accuracy (capability Pearson −0.54 to −0.80). Fidelity depends on how much off-design variation the chart contains: the exactly poised m=8 linear panel sits near the fit noise floor (0.47 on triples), TIES and DARE-TIES span wider loss ranges and reach 0.89 and 0.79, and the denser m=16 chart reaches 0.96 on

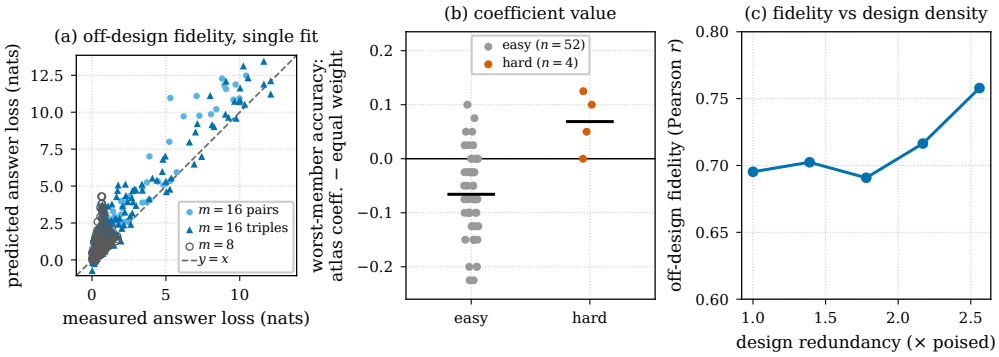


Figure 2: The index fitted once on the answer loss. (a) Predicted versus measured answer loss on off-design merges: one 136-point build at m=16 predicts all 120 off-design triples (triangles) at Pearson 0.96, pairs on the design (circles) at 0.98. (b) Worst-member accuracy at the atlas coefficient minus equal weights: on three of the four triples where equal weights leave a member below 0.6 accuracy the coefficient improves that member (+0.07 mean), and it changes little elsewhere. (c) Off-design fidelity rises with design redundancy.

triples alone. The Qwen2.5-1.5B panel marks the limit of the local approximation: merges that leave the calibration region reach answer losses near 25 nats and triple fidelity falls to 0.04, yet threshold screening stays informative (AUROC 0.76) because the fitted score still separates near-threshold merges.

Exhaustive validation is expensive at m=16: the $\binom{16}{3}$=560 triples cost hours to validate while one 136-point build scores them all and every higher order, with fidelity 0.96 on 120 independently evaluated off-design triples and screening AUROC 0.99. Because Theorem C.3 constrains gauge-invariant merge maps, we also reran this panel under the exact weighted sum of updates, which is in that class: the exact sum reaches fidelity 0.938 on off-design triples against 0.912 for the factor-space operator, a paired difference of +0.026 whose bootstrap interval $[-0.04, +0.16]$ contains zero, in the direction Eq. (1) predicts but below what sixteen adapters resolve. The rerun also puts the factor-space arm at 0.912 against the 0.957 of the original run on different hardware, so fidelity values carry at least that much variation between environments and should be read to one decimal place rather than three.

The same construction extends to thirty and fifty adapters on the same Mistral-7B base. Each build fits the singles-and-pairs design once (465 and 1,275 design merges) and answers every subset query on its faces; validation covers every pair plus sampled triples, 443 and 1,096 subsets with every member scored. At the 70% retention threshold the screen gives AUROC 0.94 at m=30 and 0.95 at m=50 (bootstrap intervals $[0.87, 0.99]$ and $[0.89, 0.98]$; triples alone 0.93 and 0.94), and pooled fidelity on off-design triples is 0.86 and 0.85. Two scored tasks have negative single-adapter gains; excluding them and the 106 subsets containing them leaves the pooled AUROC at 0.94. The cost of scale appears in the per-task statistics: between the two builds the within-task median of triple fidelity falls from 0.69 to 0.54, and at the strictest threshold the triples-only AUROC drops from 0.88 to 0.79, so the pooled screen holds at fifty adapters while the per-task rankings coarsen.

Six three-adapter charts are sampled densely (91 points each); on the original three, the quadratic model fit on only the six poised design points reproduces the full answer-loss surface at median R^2=0.91, and all six grids yield empirical finite-design envelopes for Theorem 2.3. On the original three charts the fitted models lie within ϵ_t^+ of 0.07 to 1.6 nats below and ϵ_t^- of 0.08 to 2.5 nats above the measured answer loss; padding the retention thresholds by these envelopes certifies 41 of 42 task-subset verdicts for the deployed score over the evaluated design, every bracket contains the deployed score, and the one remaining query returns inconclusive rather than a wrong verdict. Three further charts over tasks new at fifty adapters repeat the measurement with wider envelopes: 32 of 42 verdicts certified, all correct, the other ten inconclusive, all brackets containing the deployed score.

4.2 Coefficient and adapter-core intervention

Constructing the cores of Theorem 2.4 on three libraries of fitted task models gives task cores of 1 to 5 tasks and adapter cores of 3 to 5 adapters, each reconstructing the full minimax score exactly (Appendix B).

On one eight-way Mistral-7B query (linear merge, zero floor, 40 held-out items per task disjoint from the fit split) we test whether the fitted score transfers to the deployed model (Table 2, Appendix B). The score-optimal merge is naturally sparse, using 5 of the 8 adapters because its optimum lies on a face; it cuts the worst task's held-out answer loss from 3.65 under equal weights to 1.59 and raises mean accuracy from 0.63 to 0.72, and its fitted worst score (1.58) matches the deployed value to within 0.01. A fifteen-query replication on the sixteen-adapter chart removes the single-query scope: five queries each at orders four, five, and six give the atlas coefficient a lower deployed worst-member answer loss than equal weights on thirteen of fifteen (median improvement 2.25 nats); the fitted models predict an improvement on all fifteen, and no failed prediction reverses by more than 0.21 nats.

Restricting to three adapters and reoptimizing and deploying all $\binom{8}{3}$=56 subsets, the fitted worst score ranks them against the deployed worst at Spearman 0.88 and the best-fitted subset is the deployed optimum, so the fitted score orders supports without a new model evaluation, while dropping to three adapters costs worst-task loss (2.03 against the five-adapter 1.59; per-support detail in Table 2). This is one query; whether two coefficients with the same fitted latent image also behave identically on the deployed model is untested.

4.3 Screening across public libraries

Pair queries run on $C_{\{i,j\}} \subset \mathbb{R}^2$ and triples on $C_{\{i,j,k\}} \subset \mathbb{R}^3$; for every query we evaluate a coefficient design for held-out labels, fit retention models on calibration prompts, and record the projection size. The released public panel totals 7,784 evaluated pair/triple records: Qwen2.5 m=50 balanced/circulant slices plus full pair/triple enumeration for TinyLlama m=12 and Mistral, Phi-2, Pythia-410M, and Pythia-1B m=10 panels.

A direct test of the fixed-chart Helly bound fixes one shared Δ^2 chart over three adapters, where each pair sub-query is the c_k=0 face, and runs 146 unique triples from the m=50 linear panel (Appendix B.2); it records a single Helly witness, a lower bound on the rate since pair feasibility attained only in the chart interior is not tested. Synthetic libraries verify the d+1 obstruction order analytically at d=2, 3, 4, with the obstructing core size exactly d+1 in every case (analytic and numerical values in the extended version [Kandukoori et al., 2026]).

On the Qwen m=50 run, whose calibration prompts are generic instructions, the atlas AUROC is linear 0.89, magnitude pruning 0.85, TIES 0.79, and DARE-TIES 0.76 (DARE-Linear sensitivity 0.87; pooled AUPRC 0.74), and every per-operator bootstrap and clustered interval excludes 0.5, as does an adapter-resampling bootstrap [Owen, 2007] over the fifty-adapter pool (pooled $[0.75, 0.88]$, Appendix B.4). We therefore interpret the Qwen result as retention screening relative to the base, not as an end-task capability guarantee; the capability claims rest on the answer-loss panels above. Linear and magnitude pruning are the two strongest operators on three of the five smaller panels, while the two Pythia panels deviate and AUPRC is the more informative metric at their low base rates; the Phi-2 cells at AUROC 1.00 rest on 68 feasible events and should be read as point estimates (per-panel tables in the extended version [Kandukoori et al., 2026]). Screening quality is lower on the events whose fitted curvature required projection (Appendix A). Figure 4 (Appendix B.3) plots the per-operator AUROC, the operating points, and the low-cost baselines.

At a 20% validation budget (full validation $\approx$ 33 MI300X-hours), validating the atlas's top-k recovers 45.5% of the subsets that pass held-out validation at 77.6% precision and saves 73% of the GPU-hours including the $\sim$2 s per-query screening cost (adapter-resampling bootstrap: recall $[34.2, 59.6]$%, precision $[60.8, 91.3]$%; full budget curve in Table 5, Appendix B.3). Because the per-query cost scales linearly, screening all $\binom{50}{5}$ five-way candidates at this rate would take roughly 1,180 GPU-hours, so exhaustive screening of large subset spaces needs the atlas index, whose build cost is amortized across queries, rather than the local quadratic screen used in these runs. On the TinyLlama-12 library, selecting the lowest-score 20% of candidates clears the retention threshold on every constituent task in 30% of pairs and 12% of triples, versus 7% and 4% under random selection.

Atlas screening outperforms the label-free low-cost screens on Qwen, at per-operator AUROC 0.76 to 0.89 versus 0.64 to 0.73 for a single-adapter calibration-gain floor and 0.63 to 0.81 for leave-one-slice-out logistic regression, and cosine baselines on raw (B, A) and on ΔW=BA trail it by at least 25 AUROC points (per-operator table in the extended version [Kandukoori et al., 2026]). Clique screening over a fully validated

pair table ranks triples at 0.77 to 0.87, near the atlas, but presumes the $\binom{50}{2}$ held-out pair evaluations the other screens do not receive: a pair table costs about 10 GPU-hours and answers only pair queries, whereas the atlas build scores 1,275 design merges at probe cost and answers every subset query at every order. A hybrid ranking triples by the atlas's own predicted pair feasibility improves every budgeted operating point at no extra cost (Appendix B.1).

Supervised baselines characterize the break-even: gradient-boosted trees and a small MLP lose to the atlas on new or pair-screened libraries (cold-start gaps +0.263 to +0.392 AUROC; +0.24 on the TinyLlama sweep) and overtake it only in the transductive full-history regime, after receiving adapter identities and feasibility labels from the target library itself (+0.025 to +0.107; full table in the extended version [Kandukoori et al., 2026]). The cold-start comparison is asymmetric, since the atlas consumes calibration evaluations of the actual merged model; granting a GBM and an MLP the atlas's probe-derived outputs shifts per-operator AUROC by −0.02 to +0.02 relative to the atlas rule, so the fixed convex rule extracts essentially all the screening signal its calibration evaluations carry.

Performance is stable over the tested ranges of the coefficient floor and the retention threshold, with AUROC 0.92 to 0.97 across a five-point floor sweep and pooled AUROC in $[0.79, 0.87]$ as the retention threshold sweeps the base rate from 9% to 67%; a TinyLlama sweep over 20 threshold-and-seed cells (5720 queries) gives mean AUROC 0.938, and in its four τ=0.55 cells 17 accepted triples are non-cliques in the pair graph, false negatives of pairwise screening on subset-specific charts (sweep detail in Appendix B.4). Measured spectral effective dimensions across all charts are in Appendix B: certified rank 2 on the homogeneous m=8 chart, rising with task heterogeneity to 23 on the Qwen m=50 chart, flat at 5 to 6 on the task-homogeneous Mistral substrate from sixteen through fifty adapters.

5 Discussion and related work

Several methods improve a merge after its subset has been chosen: KnOTS [Stoica et al., 2025] aligns update subspaces with an SVD, Core Space [Panariello et al., 2025] and NSC [Lee et al., 2026] compress the updates, Pico [Tang and Yang, 2026] and OSRM [Zhang and Zhou, 2025] calibrate the factor spaces, LoRA Soups [Prabhakar et al., 2025] composes checkpoints, LoraHub [Huang et al., 2024] searches coefficients on a fixed subset, and Fisher merging [Matena and Raffel, 2022] and RegMean [Jin et al., 2023] compute merged weights for a supplied subset; full-model fusers [Wan et al., 2025, Jang et al., 2024, Huang et al., 2024, Lu et al., 2024] likewise target the merge itself and produce no feasibility score. The atlas screens the subsets, and its coefficient comes from one convex optimization over the fitted models, with no black-box search; a practitioner can screen with the atlas first, then merge the survivors with any of these operators. Task grouping [Fifty et al., 2021] predicts which tasks should train together, a different problem from screening already-trained adapters.

An alignment-based screener built from KnOTS's SVD step is at chance on 999 linear Qwen events (AUROC 0.500, versus 0.894 for the atlas on the same events), and per-adapter gradient alignment is also at chance on the eight-task answer-loss panel where the atlas scores 0.95 to 0.98; a different alignment estimator could behave otherwise. Recent work studies pre-merge signals and causes of mergeability for a supplied set: Zhou et al. [2026] regress merge success on interpretable gradient and subspace properties, Wang et al. [2026] score merging compatibility from limited unlabeled data, and Rahamim et al. [2026] analyze the role of base-model knowledge, with related set-level analyses by Cao et al. [2026] and Tang et al. [2026]. The atlas instead uses task-conditioned calibration loss to answer a query over protected tasks and eligible adapters, returning optimized coefficients and sparse witnesses for the fitted problem; Theorem 2.5 limits rules determined only by pair information, not set-level learned predictors, and we make no matched-information comparison with predictors whose calibration access and outputs differ from ours.

6 Limitations

Validity for deployed losses. Every certificate we report is a surrogate certificate: exact for the fitted task models, and silent about the deployed loss until envelopes are supplied. Decisions about the deployed loss additionally require uniform approximation envelopes over the continuous chart (Theorem 2.3); these remain unestimated. Section 4.1 evaluates finite-design versions on six dense charts, while the larger panels'

certificates remain statements about the fitted models, and the deployed tests of coefficient selection cover sixteen queries in total.

Weak calibration targets on Qwen. The generic prompts in the m=50 Qwen run give weak single-adapter gains, so those results measure retention relative to the base; the capability claims rest on the Mistral panels scored on the answer loss, and because the atlas records where each channel's loss is scored, moving to tasks with recoverable answers requires only a refit.

Dependent events. Pair and triple records share adapters and tasks, so the adapter-resampling and leave-one-adapter-out analyses measure stability within the observed libraries and make no claim of population-level coverage.

Index scale. The atlas index is validated to fifty adapters on one base and one merge operator; the four-operator evidence at fifty still comes from the local quadratic screen on Qwen, within-task fidelity medians fall as the library grows, and six of the fifty tasks are unscoreable because their prompts exceed the context limit.

Rank compression is per-core. Small ranks occur on the local and m=8 charts and the Mistral ambient charts hold certified rank 5 to 6 through fifty adapters, but the heterogeneous Qwen m=50 chart needs rank 23, so direct convex queries remain the default there.

Conclusion

Compatibility in a LoRA library depends on the tasks to protect and on the adapters available to the merge. The Compatibility Atlas models this relation with reusable task models; each answer names the merge weights and the small task and adapter cores that explain it, under a surrogate certificate for the fitted models rather than an established uniform guarantee for the deployed loss, with the strongest current evidence collected in Figure 1. The dual certificates of rejected queries are constructive minimal blocking task cores, supporting sub-library synthesis and library covering, which we develop in a companion paper; the main open problem is dynamic maintenance after accepted merges change the base model.

Ethical considerations. The method reduces validation waste by screening candidate merges before GPU time is spent on every one. Cheaper merging could lower the cost of assembling specialized models with undesirable capabilities; we mitigate this by not releasing model weights or adapters. Every repository we use reports an Apache-2.0 or MIT-compatible license, and no human subjects or private data are involved.

Code and data

The screening runner, the analysis scripts, and every released result are at `https://github.com/sushaan-k/lora-compatibility-atlas`. Running `verify.py` there recomputes the paper's headline numbers from the released records and checks them against the values printed here.

Author contributions

S.K. conceived the project, designed the experimental program, and led the research and the writing. A.A. contributed to the experiments and the codebase. D.B. developed the mathematical derivations and sourced the compute. A.P. carried out the data analysis and curation. V.F. supervised the work and advised on methodology and presentation.

References

Sushaan Kandukoori, Aarit Atreja, Devom Brahmbhatt, Avaneesh Parvathareddy, and Vladimir Filkov. The Compatibility Atlas: A convex-geometric index for LoRA libraries (extended version). Ancillary file `extended_paper.pdf` accompanying this submission, 2026.

Nina Amenta. Helly-type theorems and generalized linear programming. *Discrete & Computational Geometry*, 12:241–261, 1994.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

George E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B*, 13(1):1–45, 1951.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Rickard Brüel-Gabrielsson, Jiacheng Zhu, Onkar Bhardwaj, Leshem Choshen, Kristjan Greenewald, Mikhail Yurochkin, and Justin Solomon. Compress then serve: Serving thousands of LoRA adapters with little overhead. *arXiv preprint arXiv:2407.00066*, 2024.

Yuan Cao, Dezhi Ran, Yuzhe Guo, Mengzhou Wu, Simin Chen, Linyi Li, Wei Yang, and Tao Xie. An empirical study and theoretical explanation on task-level model-merging collapse. *arXiv preprint arXiv:2603.09463*, 2026.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016.

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

David Eppstein. Quasiconvex programming. In *Combinatorial and Computational Geometry*, volume 52 of *MSRI Publications*. Cambridge University Press, 2005.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Eduard Helly. Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *First Conference on Language Modeling (COLM)*, 2024.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. EMR-merging: Tuning-free high-performance model merging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *European Conference on Computer Vision (ECCV)*, 2024.

Mojan Javaheripi and Sébastien Bubeck. Phi-2: The surprising power of small language models. Microsoft Research Blog, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Wonyoung Lee, Wooseong Jeong, and Kuk-Jin Yoon. Label-free cross-task LoRA merging with null-space compression. *arXiv preprint arXiv:2603.26317*, 2026.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Michael S. Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007.

Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew D. Bagdanov, and Simone Calderara. Core space merging for efficient LoRA fusion. *arXiv preprint arXiv:2509.17786*, 2025.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. LoRA soups: Merging LoRAs for practical skill composition tasks. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Adir Rahamim, Asaf Yehudai, Boaz Carmeli, Leshem Choshen, Yosi Mass, and Yonatan Belinkov. Will it merge? On the causes of model mergeability. *arXiv preprint arXiv:2601.06672*, 2026.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Henry Scheffé. Experiments with mixtures. *Journal of the Royal Statistical Society: Series B*, 20(2):344–360, 1958.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

George Stoica, Pratik Ramesh, Boglárka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the KnOTS. In *International Conference on Learning Representations (ICLR)*, 2025.

Yixuan Tang and Yi Yang. Crowded in B-Space: Calibrating shared directions for LoRA merging. *arXiv preprint arXiv:2604.16826*, 2026.

Lin Tang, Wei Zhang, Jing Li, Hongyu Chen, Ming Zhao, and Yuxuan Wang. Predicting mergeability of parameter-efficient fine-tuning updates. *arXiv preprint arXiv:2606.19549*, 2026.

Fanqi Wan, Longguang Zhong, Ziyi Yang, Ruijun Chen, and Xiaojun Quan. FuseChat: Knowledge fusion of chat models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21618–21642, 2025.

Tiantong Wang, Yiyang Duan, Haoyu Chen, Tiantong Wu, and Wei Yang Bryan Lim. M-Loss: Quantifying model merging compatibility with limited unlabeled data. *arXiv preprint arXiv:2602.08564*, 2026.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Mitchell Wortsman, Gabriel Ilharco, Samir Y. Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Haobo Zhang and Jiayu Zhou. Unraveling LoRA interference: Orthogonal subspaces for robust model merging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers*, pages 26459–26472, 2025.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Luca Zhou, Bo Zhao, Rose Yu, and Emanuele Rodolà. Demystifying mergeability: Interpretable properties to predict model merging success. *arXiv preprint arXiv:2601.22285*, 2026.

A From fitted task models to deployed loss

PSD-projection stability. The PSD projection is frequently active, with projection rate 0.63 to 0.93 across the six base panels, and on Qwen m=50 the pooled AUROC drops from 0.886 on naturally PSD events to 0.783 on projected ones. The linear cell barely moves (0.897 unprojected versus 0.887 projected) and DARE-TIES is most sensitive (0.879 versus 0.718), and dropping the largest-correction quartile leaves the operator ranking intact.

Debiasing ablation. Projection is accept-sound but not order-preserving, so it can cost ranking on the projected subset. Across two Qwen m=50 slices (637 projected events), re-scoring each projected event by the unprojected model at the same certificate solution raises pooled AUROC from 0.839 to 0.879 (+0.040). The gain grows with the size of the projection correction: DARE-TIES 0.756→0.842 and magnitude pruning +0.012, against a flat linear (−0.008) and TIES (−0.010). The certificate retains the soundness guarantee, while the unprojected model is a better statistic for prioritizing projected events.

Proposition A.1 (Projection acts only through an active task core)**.** *Let $\widehat{\lambda}$ be an optimal dual for the projected surrogate on query S, chosen with support $A = \{t : \widehat{\lambda}_t > 0\}$ of size $|A| \le d{+}1$ (a sparse optimal dual always exists, by Carathéodory applied to the dual stationarity condition), and let $\widehat{c}$ be a matched primal. If $\widetilde{H}_t \succeq 0$ for every $t \in A$, then projection changes nothing on this query: $L(\widehat{\lambda})$ and $U(\widehat{c})$ coincide for the unprojected and projected models, and the two scores agree at the same primal-dual pair. A query's certificate therefore moves only when one of its at most $d{+}1$ core tasks is indefinite, so the per-task projection rate overstates how often projection reaches the certificate. (Proof in the extended version [Kandukoori et al., 2026].)*

Certificate-margin reliability. Binning the 4,000 Qwen events by certificate margin quintile gives feasibility rates $\{8.6, 13.9, 22.5, 48.0, 77.6\}$% from least- to most-feasible quintile, with Wilson intervals [Wilson, 1927]; the per-panel reliability curves are in the extended version [Kandukoori et al., 2026]. Under leave-one-slice-out Platt scaling [Platt, 2000] the score calibrates to a Brier score [Brier, 1950] of 0.164, versus 0.225 for a constant base-rate predictor, with 10-bin expected calibration error [Guo et al., 2017] of 0.068.

Capability probe: loss versus accuracy. On Mistral-7B-Instruct with the SNLI adapter [Bowman et al., 2015] (base at chance 0.35, adapter 0.95), across 19 equal-weight merges, the input-prompt loss does not track accuracy (Pearson +0.43) while the same loss restricted to the answer token does (−0.96). The five-task panel replicates the pattern (pooled −0.72 versus +0.37; per task −0.96 to −0.15), and the correlation weakens where merges barely change accuracy or the adapter gain is small, that is, where there is little signal to track.

B Protocols, algorithm, and additional data

Chart mismatch. For a fixed floor the score is nondecreasing in T and nonincreasing in U, so with the chart held fixed compatibility is downward-closed in the protected set and the only possible pair/triple disagreement is the classical Helly one, every pair feasible but the triple not. With a query-dependent floor the charts stop nesting: once $\ell \neq 0$ the pair chart $C(\{i,j\},\ell)$ is not a face of the triple chart $C(\{i,j,k\},\ell)$, so a triple can be feasible while a pair sub-query is not (Figure 3b); Table 4 counts both disagreements on the Qwen m=50 panel.

Operator semantics and item splits. The deployed linear operator is PEFT's `add_weighted_adapter` with `combination_type=linear`, which scales each adapter's factor pair by the square root of its coefficient before summing: the resulting update is $\sum_{i,j} \sqrt{c_i c_j}\, s\, B_i A_j$, whose diagonal terms recover $c_i\, \Delta W_i$ and whose cross terms mix distinct adapters' factors (the exact weighted sum is `combination_type=cat`). Every fit and every validation merge in a panel use the same operator, so all numbers are statements about the operator actually deployed. On the Mistral answer-token panels, off-design refers to merge configurations absent from the fit design; their evaluation items are the task's calibration items, so those numbers measure generalization across coefficient settings, while the deployed intervention runs evaluate on held-out instances disjoint from

Table 1: Core notation.

m, n, d	adapters, protected tasks ($n{=}m$ in experiments), chart dimension
$C \subseteq \mathbb{R}^d$	compact convex coefficient chart (the allowed merge-weight settings)
f_t, q_t, $\widetilde{q}_t$, $\widehat{q}_t$	deployed task loss; generic convex task model in the theory; unprojected quadratic fit; PSD-projected model used in optimization
$\rho_S(C) = \min_{c\in C} \max_{t\in S} q_t(c)$	minimax retention score for subset S on C
$T \subseteq [n]$, $U \subseteq [m]$, ℓ	protected tasks, merge adapters, coefficient floor ($s_\ell{=}1{-}\|\ell\|_1$); the diagonal sets $T{=}U{=}S$
$C(U,\ell){=}\{\ell{+}s_\ell z : z \in \Delta_U\}$	merge chart: the simplex on U raised above the floor
$\rho(T;U,\ell)$	two-sided score $\min_{c\in C(U,\ell)} \max_{t\in T} q_t(c)$; diagonal $\rho(S;S,0){=}\rho_S(\Delta_S)$
R, $a_i{=}Rv_i$, k_U	shared linear map into the latent space, adapter atoms ($v_i{=}e_i$ on the ambient simplex), their affine dimension
Δ_S, Δ_U	probability simplex on the index set S (or U)
$F_t(r) = \{c \in C : q_t(c) \le r\}$	task-t sublevel set at threshold r
$N_d = 1 + d + d(d{+}1)/2$	number of coefficients of a quadratic model on a d-chart
τ	retention threshold (used $\{0.55, 0.65, 0.70, 0.75, 0.85\}$)

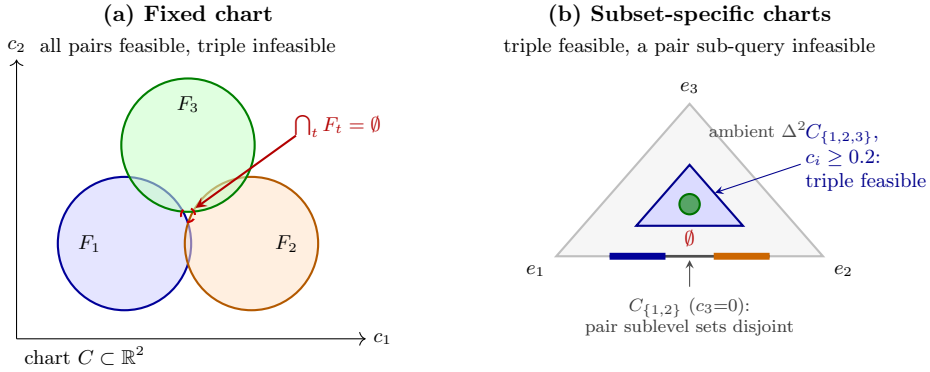


Figure 3: Two ways that pair and triple checks can disagree. **(a) Fixed chart**, where only the protected set varies: three convex sublevel sets in $C \subset \mathbb{R}^2$ can intersect pairwise while having no common triple intersection (Theorem 2.5). **(b) Subset-specific charts** with floor $\ell_i{=}0.2$: the pair chart and the triple chart are different subsets of the ambient Δ^2, neither a face of the other, so a triple can be feasible on its chart while a pair fails on the pair chart.

the fit split. The Qwen panels' held-out scores come from validation prompts separate from the calibration probes throughout.

Regimes. The *atlas index* holds the chart fixed and fits once, so one fit serves every later query. The *restricted atlas* fits once on the full ambient simplex and answers each query by restriction, with no further evaluation of any merged model. The *local quadratic screen* refits the task models after (T, U, ℓ) is chosen; the public PEFT panels use this regime on the diagonal $T{=}U{=}S$ with $\ell_i{=}0.2$, scoring a pair on $C(\{i,j\},\ell) \subset \mathbb{R}^2$ and a triple on $C(\{i,j,k\},\ell) \subset \mathbb{R}^3$.

Feasibility rule. For task t, let $L_0(t)$ be the base loss, $L_1(t)$ the single-adapter loss, and $L_c(t)$ the merged loss at coefficient c. A query is feasible at retention threshold τ iff every selected task satisfies

$$L_0(t) - L_c(t) \ge \tau\big(L_0(t) - L_1(t)\big).$$

The denominator is the single-adapter gain, clamped below at 0.005 nats, so a task whose adapter does not measurably improve on the base contributes the criterion that the merge must not degrade the base loss. The clamp is often active on the Qwen panel, whose calibration probes are generic instruction prompts (median gain -0.02 nats, 43 of 50 below the floor), so most Qwen labels reduce to holding the merge near the

Algorithm 1 Atlas pipeline: task models are fitted once on a fixed chart (the atlas index), fitted once on the ambient simplex and reused by restriction (the restricted atlas), or refitted per query on subset-specific charts (the local quadratic screen).

Input: Protected tasks $T \subseteq [n]$, merge adapters $U \subseteq [m]$, coefficient floor ℓ supported on U with $\|\ell\|_1 \leq 1$, threshold τ, margin η; linear map R with adapter atoms $a_i{=}Re_i$; per-task calibration prompts $\{\mathcal{D}_t\}_{t\in T}$.
Output: Verdict $\in \{\text{accept}, \text{reject}, \text{inconclusive}\}$ with primal $c^\star$, adapter core $V^\star$, task core $A^\star$, dual $\lambda^\star$, gap, and a robust primal–dual interval.
Fit or reuse task models. Regime 1 (atlas index): fit once on the shared chart, amortized cost nN_d scalar probes. Regime 2 (restricted atlas): fit once on the ambient simplex, answer by restriction. Regime 3 (local quadratic screen): refit for the current chart at cost $|T|N_d$. The anchored model is $\widehat{q}_t$ of Eq. (2). *Exact mode* solves on the full models; *reduced mode* first restricts them to the top-r eigenspace of $\bar{H} = |T|^{-1}\sum_{t\in T} H_t$ when its strict residual check passes, and its score ρ_r satisfies $|\rho - \rho_r| \leq \delta$ for the uniform reduction error $\delta = \sup_{t,c} |\widehat{q}_t(c) - \widehat{q}_t^{(r)}(c)|$ (Proposition C.5).
Solve the minimax. Compute $\rho^\circ = \min_{c\in C(U,\ell)} \max_{t\in T} q_t^\circ(c)$ for the models in force, with primal minimizer $c^\star$ and optimal dual $\lambda^\star$ (Theorem 2.1).
Carathéodory elimination. Write $x^\star{=}Rc^\star$ and split off the floor along the free mass: the residual $w^\star{=}(x^\star - R\ell)/s_\ell$ lies in $\text{conv}\{a_i : i \in U\}$ (when $s_\ell{=}0$ the chart is the single point ℓ and this step is skipped). While more than $k_U{+}1$ atoms carry positive weight, solve for an affine dependence among the active a_i and shift weights along it until one vanishes, holding $w^\star$ fixed. This returns the residual core $V^\star$ with $|V^\star| \leq k_U{+}1$, which reproduces the score on $C_\ell(V^\star)$ (Theorem 2.4).
Task core. Extract a Helly task core by greedily deleting tasks from the active set while the score of the remainder stays within tolerance of ρ°, then re-solving the dual on the remainder. This returns $A^\star$ with $|A^\star| \leq k_U{+}1$ whose restricted score equals ρ°; a blocking core when $\rho^\circ > \tau$.
Verify the cores. Recompute the score on the residual-support chart and confirm it equals ρ°, so the task and adapter cores jointly reconstruct the solved score.
Robust intervals and verdict. Form $U_{\text{true}}(c^\star)$ and $L_{\text{true}}(\lambda^\star)$ from the per-task envelopes $\epsilon_t^\pm$ (Theorem 2.3); in reduced mode the final verdict is checked against the full fitted models. With empty envelopes this is the surrogate-ranking mode of the reported experiments. Accept if $U_{\text{true}}(c^\star) \leq \tau{-}\eta/2$; reject if $L_{\text{true}}(\lambda^\star) > \tau{+}\eta/2$; otherwise return inconclusive.

base loss; the task-aligned Mistral panel has a stronger target (median gain +0.05). We fit PSD-projected quadratic models to the normalized retention deficit on calibration prompts, solve the minimax certificate, and label feasibility on held-out prompts from the evaluated design; every query returns a coefficient, primal and dual values, dual weights, certificate gap, per-task margins, and, when rejected, a blocking task core.

Core construction on fitted libraries. We construct the cores of Theorem 2.4 directly on three libraries of fitted task models, the homogeneous Qwen chart ($m{=}8$) and the Mistral charts at $m{=}8$ and $m{=}16$. On all three the exact shared dimension r_{exact} equals the chart dimension $(7, 7, 15)$, so no exact rank reduction is available; the cores are small regardless. The Helly task core reduces greedily to 1, 5, and 2 tasks and the Carathéodory adapter core to 3, 5, and 5 adapters; each reconstructs the full minimax score exactly. The task core is the binding set: dropping one member relaxes ρ by up to 0.41 ($m{=}8$) and 5.85 ($m{=}16$), and equal-size random task subsets fall below ρ in 88 to 100% of draws.

Panels and manifest. The public panel uses the four Qwen2.5 $m{=}50$ slices plus the $m{=}10$ to $m{=}12$ panels for TinyLlama, Mistral, Phi-2, Pythia-410M, and Pythia-1B; the non-Qwen manifests enumerate every pair and triple per operator with floor 0.2. Balanced slices draw triples so that each adapter appears equally often; circulant slices take the pairs $(i, i{+}k)$ at fixed offsets. All runs use a single fixed random seed for the calibration/held-out split, taken before any feasibility label is computed; the released runs used a single AMD MI300X with ROCm 7.0, and every number in the tables is reproduced by the archive's entry point with the corresponding panel configuration. A separate TinyLlama sweep covers 5,720 queries across five thresholds and four seeds, and a DARE-Linear re-run on the same Qwen slices (1000 events) reaches atlas AUROC 0.870.

Table 2: Deployed adapter-core intervention on one eight-way Mistral-7B m=8 query (linear merge, zero floor, 40 held-out items per task). *Fitted worst* is the surrogate $\max_t \widehat{q}_t$ at the coefficient; *deployed worst* and *mean accuracy* are measured on held-out answer tokens. The last two rows evaluate all $\binom{8}{3}$=56 three-adapter supports, each reoptimized on its own chart; the final row reports the median deployed worst loss with interquartile range and the median support-level mean accuracy.

Coefficient	Support	Fitted worst (nats)	Deployed worst (nats)	Mean acc.
Equal weights	8	6.40	3.65	0.63
Full atlas minimizer	5	1.58	1.59	0.72
Top-3 weight truncation	3	2.76	2.90	0.65
Rank-2 atlas core	3	3.45	3.46	0.67
Best-fitted 3-subset	3	2.12	2.03	0.74
All 56 reoptimized 3-subsets	3	–	5.84 [3.30, 8.28]	0.57

Table 3: Capability panels scored on the answer loss, at the 70% retention threshold. Fidelity is the Pearson correlation, on triples alone, of predicted against measured answer loss on off-design triples; the m=16 value 0.96 has adapter-block bootstrap 95% interval $[0.75, 0.98]$ and leave-one-adapter-out range $[0.86, 0.97]$. AUROC (adapter-block bootstrap 95% intervals) and AUPRC screen subsets by the worst member's predicted deficit, and capability is the pooled Pearson of answer loss against task accuracy. †: the m=8 linear panel has all-feasible pairs and four infeasible triples, all sharing one adapter, so the 0.98 is a point estimate rather than a stable AUROC.

Base	m	Operator	Queries	Feasible	Fidelity	AUROC (95% interval)	AUPRC	Capability
Mistral-7B	8	linear	84	0.95	0.47	0.98†	–	−0.67
Mistral-7B	8	TIES	84	0.29	0.89	0.71 [0.62, 0.91]	0.62	−0.61
Mistral-7B	8	DARE-TIES	84	0.55	0.79	0.55 [0.31, 1.00]	0.71	−0.71
Mistral-7B	16	linear	240	0.83	0.96	0.99 [0.96, 1.00]	1.00	−0.54
Qwen2.5-1.5B	8	linear	84	0.63	0.04	0.76 [0.54, 0.99]	0.85	−0.68
Mistral-7B	4	linear	10	0.90	–	–	–	−0.80

Measured effective dimension. The spectral diagnostic certifies $\sigma\text{-}\operatorname{adim}_{0.20} \leq 2$ on the homogeneous m=8 Qwen shared chart and median 2 across twenty random five-adapter cores; the certified rank rises with task heterogeneity, to 4 to 5 on the diverse TinyLlama charts, 6 on the sixteen-task Mistral chart, and 23 on the heterogeneous Qwen m=50 chart, where no small shared rank is certified. Library size alone does not drive the growth, since the task-homogeneous Mistral substrate certifies rank 5 to 6 flat from sixteen through fifty adapters, and clustering the adapters by curvature affinity does not lower the shared rank (spectra and per-chart values in the extended version [Kandukoori et al., 2026]).

B.1 Qwen baselines and hybrid screening

A hybrid of clique screening with the atlas. A triple can be ranked by how many of its three pair sub-queries are feasible, breaking ties by the triple atlas score. Using the atlas's own *predictions* costs no extra validation, and the resulting hybrid improves on the atlas score alone at every budget on the 4,400 pooled public triples: recall/precision at the 10/20/30/50% budgets is 0.32/0.79, 0.56/0.69, 0.75/0.61, 0.90/0.44, versus 0.29/0.71, 0.54/0.67, 0.66/0.54, 0.82/0.40 for the atlas alone. Using held-out pair labels instead does slightly better still, but presumes a fully validated pair table; both hybrids keep the atlas's coefficients and certificates for every candidate they forward.

B.2 Fixed-chart Helly probe on Qwen

On the Qwen runs, pairs are screened on the chart $C_{\{i,j\}}$ while triples use the larger chart $C_{\{i,j,k\}}$, so the pairwise graph screen can wrongly accept a triple; the probe tests whether those triples also realize the Helly obstruction on one fixed chart at the correct intrinsic dimension. We run a shared Δ^2 chart probe over 146 unique triples from the m=50 linear panel with no coefficient floor, so each pair sub-query is the c_k=0 face of the same chart. The probe is one-sided: pair feasibility certified on a face carries over to the full

Table 4: Pair-vs-triple disagreement on balanced Qwen m=50. The closure row joins the balanced-slice run (which leaves 400 pair faces unevaluated per operator) with a separate face-pair run that evaluates exactly those missing pair sub-queries; after the join the missing-face count drops to 0. The released events contain 70 pair keys whose held-out labels conflict across slices, and the original run's join of those conflicts is not recoverable; under the join variants in the release the counts move by a few in each cell, always with the same conclusion, that chart-mismatch disagreements dominate fixed-chart Helly events.

	Linear	TIES	DARE-TIES	Mag. pruning
All 3 pair sub-queries feasible, triple infeasible	1	14	13	3
Triple feasible, ≥ 1 pair sub-query infeasible	102	78	110	99
Triple feasible, exactly 2 pair sub-queries infeasible	27	21	34	28
Triple feasible, all 3 pair sub-queries infeasible	8	5	16	11
Triples lacking evaluated pair faces after closure	0	0	0	0

chart, so each recorded event is a valid fixed-chart Helly witness, but pair feasibility attained only in the interior is not tested, and the count is a lower bound. The deduplicated breakdown: 48 triples with all four sub-queries feasible, 66 with all four infeasible, 31 mixed, 1 Helly event; the per-triple records are in the extended version [Kandukoori et al., 2026].

B.3 Operating points and budget curve

At a 30% budget, the atlas recovers 55 to 73% of feasible events depending on the operator, compared with 30% for random screening; the extra cost is about 2 s of screening versus 30 s of full validation per query, and the savings in Table 5 include it. The full per-operator operating-point table and the low-cost baseline table are in the extended version [Kandukoori et al., 2026].

Table 5: Compute savings from atlas pre-screening on the Qwen m=50 balanced panel. Full validation of all 4000 events takes approximately 33.3 MI300X-hours. Screening with the atlas costs $\sim$2 s per query; the top-k atlas-ranked candidates then receive full validation. k chosen at each validation budget; cost in MI300X-hours; savings vs full validation.

budget	k	feasible recall	precision	cost (GPU-h)	savings
5%	200	14.2%	97.0%	3.89	88%
10%	400	24.8%	84.8%	5.56	83%
20%	800	45.5%	77.6%	8.89	73%
30%	1200	63.3%	72.0%	12.22	63%
50%	2000	80.8%	55.1%	18.89	43%

B.4 Stability diagnostics

Recomputing AUROC on the events that avoid each of the 50 Qwen adapters in turn, the leave-one-out median agrees with the full-data estimate to within 0.002 and the per-operator range spans 0.021 to 0.043, so no single adapter accounts for the screening signal. The pigeonhole bootstrap of Owen [2007], resampling adapters instead of events (B=2,000), gives 95% AUROC intervals of linear $[0.821, 0.946]$, magnitude pruning $[0.758, 0.918]$, TIES $[0.691, 0.883]$, DARE-TIES $[0.673, 0.835]$, pooled $[0.748, 0.884]$, every interval excluding 0.5. Owen's coverage guarantees are for crossed random-effects designs, so these are a stability analysis over the adapter population rather than calibrated coverage for the pair/triple hypergraph. In the parameter sweeps, the α=0 panel keeps AUROC 0.80 to 0.90 across operators, and the TinyLlama threshold-and-seed sweep spans AUROC 0.843 to 0.976 with top-quartile enrichment 3.858 over a 2.41% base rate. Per-operator tables, clustered intervals, the floor and threshold sweeps, and the per-panel certificate diagnostics are in the extended version [Kandukoori et al., 2026].

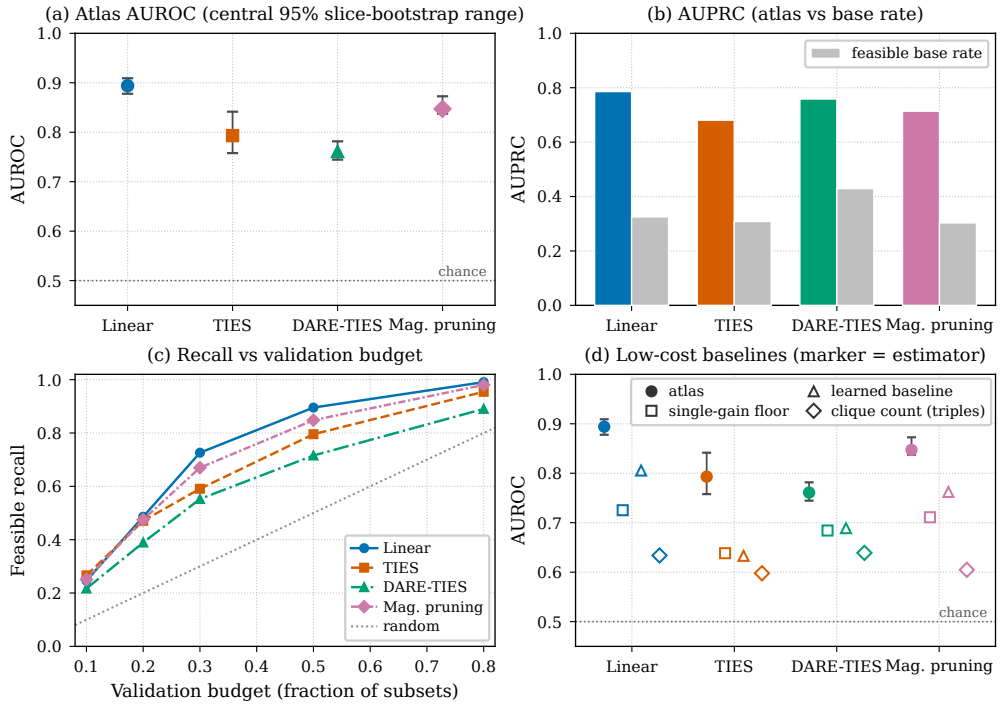


Figure 4: The Qwen m=50 panel (4,000 pair and triple events, four predeclared slices). (a) Atlas AUROC per operator with cluster-bootstrap stability ranges. (b) AUPRC per operator against the feasible base rate. (c) Recall of top-k screening across validation budgets. (d) The atlas versus the low-cost baselines on triples.

B.5 Maintenance of the index

If maintenance changes each task model by at most d_t on the chart, then the score moves by at most $\max_{t \in T} d_t$, by the 1-Lipschitz dependence of the minimax on the models. Insertion has three cases: a new adapter adds a coordinate, with each existing model acquiring $m{+}1$ new coefficients fixed by the same $m{+}1$ insertion evaluations (replaying the build one adapter at a time moves the shared models by a median 0.16 nats per added coordinate); a new protected task adds no coordinate and fits one new model on the N_d existing design points; and the combined case does both. Consolidation writes an accepted merge into the base weights and changes every task model: consolidating an accepted pair moves the remaining adapters' models by a median 0.09 and at most 0.41 nats in single-adapter answer loss, small relative to the multi-nat gains but of the same size as typical certificate margins, which supplies the index's refit criterion.

C Additional theory

Theorem C.1 (Effective dimension finite core)**.** *Let $C \subset \mathbb{R}^d$ be compact convex. Suppose there is a linear $\Pi : \mathbb{R}^d \to \mathbb{R}^r$ and convex $\tilde{q}_t : \mathbb{R}^r \to \mathbb{R}$ with $q_t(c) = \tilde{q}_t(\Pi c)$ for every $t \in [n], c \in C$. Then for every $S \subseteq [n]$, $\rho_S(C) = \max_{T \subseteq S,\, |T| \le r+1} \rho_T(C)$, and the order $r{+}1$ is tight. (Proof in Appendix D.7.)*

Exact shared dimension. The fitted models may vary along fewer than d directions of a given chart. That number is a matrix rank computed in the chart's own coordinates: writing $\mathrm{aff}(C) = c_0 + L$ with Q an orthonormal basis of L, the restricted model has slope g_t and curvature $\widehat{H}_t = Q^\top H_t Q$, and the intrinsic curvature matrix is $M_C = \sum_t (\widehat{H}_t^2 + g_t g_t^\top)$ with $r_{\mathrm{exact}}(C) := \mathrm{rank}(M_C)$; its kernel is exactly the set of intrinsic directions along which every task model is constant on C, while the ambient matrix $\sum_t (H_t^2 + b_t b_t^\top)$ overstates the dimension.

Theorem C.2 (Exact shared dimension)**.** *Let C have non-empty relative interior in* $\mathrm{aff}(C)$*, and let* Π_{exact} *be the projection onto the top* $r_{\mathrm{exact}}(C)$ *eigenvectors of* M_C*. Each restricted model factors as* $\widehat{q}_t(c_0 + Qz) = \tilde{q}_t(\Pi_{\mathrm{exact}} z)$ *for a convex* $\tilde{q}_t$*, so for every* S*,*

$$\rho_S(C) = \max_{T \subseteq S,\, |T| \le r_{\mathrm{exact}}(C)+1} \rho_T(C)$$

holds exactly, the induced map realizes the shared linear map R of Theorem 2.4 with both cores of order at most $r_{\mathrm{exact}}(C)+1$*, and no linear map of rank below* $r_{\mathrm{exact}}(C)$ *preserves every fitted model on C without error. (Proof in Appendix D.6.)*

Gauge invariance. LoRA factors carry a reparameterization, since (B, A) and $(BR, R^{-1}A)$ encode the same weight update for any invertible R, so a rule that reads raw factor coordinates can change its answer while the model does not change. A merge map is *gauge invariant* when $\Delta(c)$ depends on the adapters only through the represented updates $\{\Delta W_i\}$, as the exact weighted sum does; factor-space operators need not lie in this class, since PEFT's `linear` combination produces cross terms $B_i A_j$ and the merged model itself can then depend on the factorization. The ellipsoidal sublevel sets also depend on the task metric, which already rules out raw task-vector cosine as a universal rule; Theorem C.3 shows the sublevel-set family is exactly the gauge-invariant content of the fitted local models.

Theorem C.3 (Sublevel-set identifiability and raw-geometry impossibility)**.** *The quadratic task models determine the sublevel-set family, all scores, finite cores, dual witnesses, and optimal coefficient sets, and conversely the sublevel-set family recovers each q_t on C. If the merge map is gauge invariant, then LoRA gauge transformations preserve every task model and every answer, so no non-gauge-invariant factor-coordinate rule and no raw-parameter cosine rule can be universally exact for adapter compatibility. (Proof in Appendix D.9.)*

Spectral effective dimension. Let $\mathcal{L} = \{H_1, \ldots, H_n\}$ be the per-task PSD Hessians. The *spectral effective dimension at residual* ϵ is

$$\sigma\text{-}\operatorname{adim}_\epsilon(\mathcal{L}) := \min\big\{r : \ \exists \text{ rank-}r \text{ orthogonal projection } \Pi \text{ with } \max_t \|H_t - \Pi H_t \Pi\|_F \le \epsilon \max_t \|H_t\|_F\big\}.$$

Remark C.4 (Mean-Hessian projector diagnostic). The top-r eigenspace Π_r of the mean Hessian $\bar{H} = \frac{1}{n}\sum_t H_t$ is a computable candidate projector, obtained in $O(d^3 + nd^2)$ time. Certifying its strict per-task residual gives a computable upper bound on $\sigma\text{-}\operatorname{adim}_\epsilon(\mathcal{L})$ and an obstruction order at most $r+1$ with additive certificate slack $\epsilon L_F D^2$ for $L_F = \max_t \|H_t\|_F$ and $D = \mathrm{diam}(C)$ (a non-zero linear term adds $2\max_t \|(I - \Pi_r) b_t\|_2 D$). This certifies the residual of one specific projector; it is not a solver for the minimum-rank projector.

Approximate reduction. When the models factor through a rank-r map only within a uniform error δ, every score moves by at most 2δ: $|\rho_S(C) - \max_{|T| \le r+1} \rho_T(C)| \le 2\delta$, and the two-sided restricted maximin and minimax scores agree with the true score to within the same 2δ (statements and proofs in the extended version [Kandukoori et al., 2026]).

Proposition C.5 (Interval certificate for rank-reduced models)**.** *Let $C \subseteq \{c \in \mathbb{R}^d : c \ge 0,\ \sum_i c_i \le 1\}$ be compact convex, with models in the uncentered parametrization. For an orthogonal projector Π of rank r, let $\widehat{q}_t^{(r)}$ be the reduced model with coefficients $(\alpha_t, \Pi b_t, \Pi H_t \Pi)$ and $\epsilon_S := \max_{t \in S} \sup_{c \in C} |\widehat{q}_t(c) - \widehat{q}_t^{(r)}(c)|$, computable exactly by face enumeration when C is a polytope. Then $|\rho_S(C) - \rho_S^{(r)}(C)| \le \epsilon_S$ for every S; each reduced model is convex and factors through Π, so the $(r+1)$-core identity holds exactly for the reduced models; and whenever the reduced core-table score differs from τ by more than ϵ_S, the full-model decision coincides with the reduced core-table decision. (Proof in the extended version [Kandukoori et al., 2026].)*

Estimation. Estimated models with uniform error η move every score by at most η; reaching sup-norm precision η under poised quadratic interpolation needs $O(\kappa^2 B^2 \eta^{-2} \log(nN_d/\delta_0))$ example evaluations per task and design point, for per-prompt deficits bounded by B, poised-design conditioning constant κ, and failure probability δ_0 (proof via Hoeffding [Hoeffding, 1963] and interleaving stability [Chazal et al., 2016] in the extended version [Kandukoori et al., 2026]).

D Proofs

D.1 Proof of Theorem 2.1

Set $F(c,\lambda) = \sum_{t\in S} \lambda_t q_t(c)$. Since C and Δ_S are compact convex sets and F is continuous, convex in c, and affine in λ, Sion's minimax theorem [Sion, 1958] applies:

$$\min_{c\in C} \max_{\lambda\in\Delta_S} F(c,\lambda) = \max_{\lambda\in\Delta_S} \min_{c\in C} F(c,\lambda).$$

The inner max over λ collapses to $\max_{t\in S} q_t(c)$, which gives the stated equality. The certificate inequalities are weak duality. For any threshold τ, $U(c) \le \tau$ puts c inside every sublevel set at level τ, and $L(\lambda) > \tau$ forces every c above level τ on at least one task.

D.2 Proof of Theorem 2.3

For every $c \in C(U,\ell)$ and every $t \in T$, $f_t(c) \le \widehat{q}_t(c) + \epsilon_t^+ \le U_{\text{true}}(c)$, so $\max_{t\in T} f_t(c) \le U_{\text{true}}(c)$; minimizing over the chart gives $\rho^f(T;U,\ell) \le U_{\text{true}}(c)$. For the lower bound, $\sum_t \lambda_t f_t(c) \ge \sum_t \lambda_t \widehat{q}_t(c) - \sum_t \lambda_t \epsilon_t^-$ for every c, so $\min_c \sum_t \lambda_t f_t(c) \ge L_{\text{true}}(\lambda)$; weak duality, which holds for f_t with no convexity assumed, gives $\rho^f(T;U,\ell) \ge \max_{\lambda'} \min_c \sum_t \lambda'_t f_t(c) \ge L_{\text{true}}(\lambda)$. The sandwich rules out returning both verdicts. □

D.3 Proof of Theorem 2.4, part (i)

Monotonicity (≤). For any $T \subseteq S$, $\max_{t\in T} q_t \le \max_{t\in S} q_t$ pointwise on C, so $\rho_T \le \rho_S$ and hence $\max_{|T|\le d+1,\, T\subseteq S} \rho_T \le \rho_S$.

Attainment (≥) and clause (i). Fix $r < \rho_S$. The compact convex sets $\{F_t(r)\}_{t\in S}$ have empty intersection in C. By Helly's theorem in $\mathbb{R}^d$ [Helly, 1923], some $T \subseteq S$ with $|T| \le d{+}1$ already has $\bigcap_{t\in T} F_t(r) \cap C = \emptyset$, that is $\rho_T > r$. Applying this at every $r < \rho_S$ yields

$$\rho_S = \max_{T\subseteq S,\, |T|\le d+1} \rho_T.$$

Tightness. Place $d{+}1$ vertices $v_0,\dots,v_d$ of a regular d-simplex on the unit sphere in $\mathbb{R}^d$, take any compact convex chart containing the unit ball, and set $F_t = C \cap \{c : \|c - v_t\|_2^2 \le R_0^2\}$ with $\sqrt{1-1/d^2} < R_0 < 1$. Every d-subset has non-empty intersection (the face circumcenter is within distance R_0 of each face vertex), but no point is within distance $R_0 < 1$ of all $d{+}1$ vertices. Pairwise insufficiency is the $d{=}2$ instance, embedded into higher dimensions by product with a compact ball. □

D.4 Proof of Theorem 2.4, parts (ii)–(iv)

Write $c^\star$ for a minimizer of $\rho(T;U,\ell)$, so $c^\star = \ell + s_\ell z^\star$ with $z^\star \in \Delta_U$, and let $w^\star = \sum_{i\in U} z_i^\star a_i$.

Task core. Write $k_C = \dim \operatorname{aff} R\,C(U,\ell)$. Each sublevel set $\{c \in C(U,\ell) : q_t(c) \le r\}$ is the preimage under $c \mapsto Rc$ of a convex set inside the k_C-dimensional affine image. Applying Helly in that image (part (i) with ambient dimension k_C) gives $A^\star \subseteq T$, $|A^\star| \le k_C{+}1$, with $\rho(A^\star;U,\ell) = \rho(T;U,\ell)$. The same argument bounds any inclusion-minimal such set, justifying the greedy extraction of Algorithm 1.

Adapter core. The hull $\operatorname{conv}\{a_i : i \in U\}$ lies in an affine space of dimension k_U, so Carathéodory writes $w^\star = \sum_{i\in V^\star} z_i' a_i$ with $|V^\star| \le k_U{+}1$. Put $c' = \ell + s_\ell \sum_{i\in V^\star} z_i' v_i \in C_\ell(V^\star)$. Then $Rc' = Rc^\star$, so $q_t(c') = q_t(c^\star)$ for every t; hence c' is also a minimizer with residual support in $V^\star$. Since $C_\ell(V^\star) \subseteq C(U,\ell)$, $\rho(T;V^\star,\ell) \ge \rho(T;U,\ell)$, and c' attains equality.

Discrete saddle. Write $\beta_\ell(A,V) = \rho(A;V,\ell)$ and $\rho = \rho(T;U,\ell)$. Task monotonicity and adapter antitonicity, together with $\rho(A^\star;U,\ell) = \rho = \rho(T;V^\star,\ell)$, give

$$\beta_\ell(A,V^\star) \le \rho(T;V^\star,\ell) = \rho = \rho(A^\star;U,\ell) \le \beta_\ell(A^\star,V),$$

and $\rho \le \rho(A^\star;V^\star,\ell) \le \rho(T;V^\star,\ell) = \rho$ forces $\beta_\ell(A^\star,V^\star) = \rho$. Restricting each extremum to order at most $k_U{+}1$ leaves it unchanged, since the cores already attain it. □

D.5 Two-sided tightness

Theorem D.1 (Simultaneous two-sided tightness)**.** *Fix $r \geq 1$ and let $v_0, \ldots, v_r$ be the vertices of a regular r-simplex in $\mathbb{R}^r$ with unit circumradius and circumcenter at the origin. Take these points as the task centers and as the latent images of the adapter atoms: $C = \Delta_{[m]}$ with $m = n = r{+}1$, R sends e_i to v_i (an affine bijection onto* $\mathrm{conv}\{v_0, \ldots, v_r\}$*, so $k_{[m]} = r$), and $q_t(c) = \|Rc - v_t\|^2$. Then with $T = [n]$, $U = [m]$, $\ell = 0$: (i) the unique minimizer is the barycenter, with $\rho([n]; [m], 0) = 1$; (ii) $\rho([n] \setminus \{j\}; [m], 0) = 1 - 1/r^2 < 1$ for every j; (iii) $\rho([n]; [m] \setminus \{j\}, 0) \geq 1 + 2/r > 1$ for every j. Hence the task core and the adapter core each require all $r{+}1$ indices.*

Proof. Unit vectors summing to zero satisfy $\langle v_i, v_j \rangle = -1/r$ for $i \neq j$. *(i)* For $x = Rc$, averaging the centers gives $\max_t \|x - v_t\|^2 \geq \frac{1}{r+1} \sum_t \|x - v_t\|^2 = \|x\|^2 + 1 \geq 1$, with equality forcing $x = 0$; the coefficient barycenter attains it. *(ii)* The r centers $\{v_i : i \neq j\}$ have barycenter $-v_j/r$, the latent image of a chart point. At $x = -v_j/r$ every $\|x - v_i\|^2 = 1/r^2 - 2/r^2 + 1 = 1 - 1/r^2$; the averaging bound over $\{v_i : i \neq j\}$, using $\sum_{i \neq j} v_i = -v_j$, shows no chart point does better. *(iii)* Every x in the face $\mathrm{conv}\{v_i : i \neq j\}$ has $\langle x, v_j \rangle = -1/r$, so $\|x - v_j\|^2 = \|x\|^2 + 2/r + 1 \geq 1 + 2/r$; since task j stays in the protected set, the score on the face is at least $1 + 2/r$. No convex hull of only r latent atoms contains the circumcenter, whose barycentric representation uses all $r{+}1$ vertices. □

D.6 Proof of Theorem C.2

Work in the intrinsic coordinate z, where $c = c_0 + Qz$. Substituting into the anchored model gives $\widehat{q}_t(c_0 + Qz) = \kappa_t + g_t^\top z + \frac{1}{2} z^\top \widehat{H}_t z$ with $\widehat{H}_t = Q^\top H_t Q \succeq 0$, a convex quadratic on the compact convex set $K = \{z : c_0 + Qz \in C\}$ with non-empty interior.

Constant directions. For a unit u, constancy of $\widehat{q}_t$ along u forces $u^\top \widehat{H}_t u = 0$, hence $\widehat{H}_t u = 0$ since $\widehat{H}_t \succeq 0$, and then $g_t^\top u = 0$; conversely these two conditions make $\widehat{q}_t$ invariant under translation by u. The constant directions of task t are exactly $\ker \widehat{H}_t \cap \ker g_t^\top$, and by the interior hypothesis this local characterization is global.

Kernel of M_C. $u^\top M_C u = \sum_t (\|\widehat{H}_t u\|^2 + (g_t^\top u)^2)$, so $u \in \ker M_C$ iff u is a common constant direction of the whole family. Hence each $\widehat{q}_t$ depends on z only through the projection onto $(\ker M_C)^\perp$, and $\tilde{q}_t(\Pi_{\mathrm{exact}} z) := \widehat{q}_t(c_0 + Qz)$ is well defined and convex.

Minimality. If every model factors through a linear map Π of rank r', every $u \in \ker \Pi$ is a common constant direction, so $\ker \Pi \subseteq \ker M_C$ and $r' \geq r_{\mathrm{exact}}(C)$.

Finite core. Writing $\Pi_{\mathrm{exact}} = WW^\top$, Theorem C.1 applies to K with the map $z \mapsto W^\top z$ and zero slack, and the composite $R_C = W^\top Q^\top$ realizes the shared linear map of Theorem 2.4, with $k_U \leq r_{\mathrm{exact}}(C)$ for every U. □

D.7 Proof of Theorem C.1

Let $\tilde{F}_t(\tau) = \{\tilde{c} : \tilde{q}_t(\tilde{c}) \leq \tau\}$ and $K = \Pi(C)$, compact convex in $\mathbb{R}^r$. For every τ and S: some $\tilde{c} \in K \cap \bigcap_t \tilde{F}_t(\tau)$ lifts to $c \in C$ with $q_t(c) = \tilde{q}_t(\tilde{c}) \leq \tau$ for all $t \in S$, and conversely any feasible c projects to a feasible $\tilde{c}$; so the two intersections are empty together. By Helly's theorem in $\mathbb{R}^r$, if the full intersection is empty then some T with $|T| \leq r{+}1$ has empty intersection, that is $\rho_T(C) > \tau$. Applying this at every $\tau < \rho_S(C)$ yields the equality. For tightness, place r-simplex vertices in $\mathbb{R}^r$, set $\tilde{q}_t(\tilde{c}) = \|\tilde{c} - v_t\|^2$, and lift cylindrically to $\mathbb{R}^d$; between the thresholds at which every r-subset and the full $(r{+}1)$-subset become compatible, the bound binds. □

D.8 Pairwise insufficiency

Proposition D.2 (Compact ellipsoidal triple obstruction)**.** *In $d = 2$ there is a compact convex chart C and three centered quadratic sublevel sets $F_i = C \cap \{c : \|c - v_i\|_2^2 \leq R_0^2\}$ such that every pair intersects but $F_1 \cap F_2 \cap F_3 = \emptyset$.*

Proof. Place v_1, v_2, v_3 at the vertices of an equilateral triangle of side length 2 with circumcenter at the origin, take $C = [-2, 2]^2$, and pick R_0 with $1 < R_0 < 2/\sqrt{3}$. Each pairwise center distance is $2 < 2R_0$ and

the pairwise lens lies inside C, but the smallest enclosing disk of the three centers has radius $2/\sqrt{3} > R_0$, so the triple intersection is empty. At $R_0 = 1.1$ the triple score is $4/3$, with primal witness at the circumcenter and dual $\lambda^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, so pairwise screening accepts the triple while the minimax score rejects it. □

Proof of Theorem 2.5. Proposition D.2 gives a library $\mathcal{L}_1$ with a complete pair graph and an empty triple. Take $\mathcal{L}_2$ to be three identical balls at the same vertices with radius $R_0' > 2/\sqrt{3}$, so the triple intersects and the pair graph is still complete. The two libraries are pairwise-graph-indistinguishable yet disagree on the triple decision, so any rule that depends on the library only through the thresholded pairwise graph is non-exact on this pair. □

D.9 Proof of Theorem C.3

The models determine every sublevel set, score, optimizer, core, and dual certificate by definition, and the sublevel-set family recovers $q_t(c) = \inf\{r : c \in F_t(r)\}$. Let the merge map be gauge invariant, so $\Delta(c)$ is a function of the represented updates alone. A LoRA gauge transformation leaves each represented update fixed, hence leaves $\Delta(c)$ and the tangent action $J_t\Delta(c)$ unchanged, so every task model and every answer is preserved, while a non-gauge-invariant rule can move. The hypothesis is needed: for a factor-space operator the merged update carries cross terms B_iA_j, and the gauge $(B_i, A_i) \mapsto (B_iR_i, R_i^{-1}A_i)$ moves $B_iR_iR_j^{-1}A_j$ while fixing every ΔW_i, so the argument does not apply. Finally, holding ambient updates fixed and varying the task metric moves the ellipsoids in coefficient space without changing raw adapter cosine, which rules out any universally exact rule built only on raw parameter geometry. □