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 so that candidate subsets can be screened before full validation. 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; the cores observed in practice are far smaller. 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 (adapter-block bootstrap 95% interval $[0.75, 0.98]$); under linear merging, one fit on fifty screens 1,096 pair and triple merges at AUROC 0.95 (0.94 on the off-design triples alone). 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 (retention labels; Section 5.3) 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 (parameter-efficient fine-tuning) (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; each update is trimmed to its largest entries and sign conflicts are resolved before averaging) (Yadav et al., 2023), and variants of DARE (drop and rescale; random entries of each update are dropped and the survivors rescaled) (Yu et al., 2024) (Section 6), and all of them assume the subset has already been chosen. We ask which subsets are worth validating.

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. To answer such queries without validating every candidate, we fit, once for each task, a simple quadratic model of how that task's validation loss changes as the merge weights vary. The fit uses a small number of merged models scored on the task's own prompts, and afterward no further merged model needs to be evaluated. A later query specifies which adapters may enter the merge and which tasks must be kept; answering it is a

small convex optimization over the fitted models. When a query is accepted, the answer lists the best merge weights and the few adapters that carry them; when it is rejected, the answer lists a small group of tasks that cannot all be kept at once. Both answers carry a *surrogate certificate*, a pair of matching bounds for the fitted models that can be checked without re-solving the problem. Throughout, a surrogate certificate is a guarantee about the fitted convex models; upgrading it to a *deployed-loss certificate*, a guarantee about the model that is actually served, additionally requires uniform approximation envelopes over the whole query chart (Theorem 2.3), which we measure on finite designs but do not establish uniformly. The per-library record of these verdicts is the *Compatibility Atlas* (Section 4; the full pipeline is Algorithm 1).

The number of tasks and adapters named in these answers depends on the dimension d of the merge-weight domain rather than on the size of the library. For the fitted models, rejection is always forced by at most $d{+}1$ of the protected tasks, and an accepted merge can place its free weight on at most $d{+}1$ of the adapters. For $d \geq 2$ no thresholded pairwise compatibility graph decides every three-way case (Sections 2 and 3). On the subset-specific charts of the public runs, more than ten fifty-adapter Qwen triples under TIES merging have all three pairs compatible yet no evaluated weight setting retains all three tasks (Table 6), and on TinyLlama-v0.3 under linear merging two triples pass held-out validation while two of their pairs do not; the fixed-chart obstruction itself is realized analytically in Table 4 and observed once in the 146-triple shared-chart probe (Section 5.3).

When every task's fitted loss is flat along some directions of the weight region, the bound $d{+}1$ drops to $r{+}1$, where r counts the directions that matter (Theorems 3.1 and 3.3).

The loss can be scored on every prompt token or only on a task's answer tokens, and it follows task accuracy only in the second case (Pearson correlation -0.72, against $+0.37$ on the prompt). With this scoring, one shared fit on a sixteen-task Mistral library scores every triple and predicts measured loss on 120 independently selected off-design triples, still screening subset queries at AUROC 0.95 when the library grows to fifty adapters, while on the separate fifty-adapter Qwen library a score refit for each query recovers nearly half of the passing merges at about a quarter of the exhaustive validation cost (Section 5).

Concurrent work predicts whether a supplied set will merge (Zhou et al., 2026; Wang et al., 2026; Rahamim et al., 2026). The atlas returns the merge weights, the few adapters that carry the free weight, and the tasks that block a rejected query.

The experiments cover six base panels; the largest uses Qwen2.5-0.5B-Instruct with fifty adapters and 4000 pair and triple queries across four merge operators (linear, TIES, DARE-TIES, magnitude pruning). Five smaller panels, on TinyLlama-v0.3 (Zhang et al., 2024) at $m{=}12$ and on Mistral-7B-Instruct (Jiang et al., 2023), Phi-2 (Javaheripi and Bubeck, 2023), Pythia-410M, and Pythia-1B (Biderman et al., 2023) at $m{=}10$ each, enumerate every pair and triple. A separate TinyLlama sweep covers 5,720 queries across thresholds and seeds. We release uncertainty estimates, operating points, baseline comparisons, and per-query certificate records.

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 collects notation.

Let f_{θ_0} be a base model and let $\Delta_i \in \mathbb{R}^p$ be LoRA updates trained from the same checkpoint (Hu et al., 2022). Allowed merges are parameterized by a compact convex set $C \subset \mathbb{R}^d$, the *chart*; each point $c \in C$ specifies 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 $\|\ell\|_1 \leq 1$, a minimum weight per selected adapter. In the introduction's example, U holds the math, SQL, and summarization adapters, T the math and SQL tasks, and $\ell_i = 0.2$ on each adapter in U. 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 $s_\ell = 0$ the chart is the single point ℓ. The *Helly dimension* that enters the core bounds is this intrinsic dimension rather than the ambient m.

Table 1: Core notation. Symbols used only in the appendices are defined there.

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
$F_t(r) = \{c \in C : q_t(c) \le r\}$	task-t sublevel set at threshold r
Δ_S	probability simplex on S (dual to the inner max)
K_r	threshold complex at level r: which task subsets have a common feasible point
$N_d = 1 + d + d(d{+}1)/2$	number of coefficients of a quadratic model on a d-chart
$\sigma\text{-}\operatorname{adim}_\epsilon(\mathcal{L}) \le r$	spectral effective dimension: rank-r projector onto the top eigenspace with strict max-task residual $\le \epsilon$
$\bar{H}$, H_t, L_F	mean Hessian over the queried tasks, per-task PSD Hessian, $L_F := \max_t \|H_t\|_F$
τ	retention threshold (used $\{0.55, 0.65, 0.70, 0.75, 0.85\}$)

For a set $V \subseteq [m]$ over which the free mass may be spread, not required to contain $\operatorname{supp}\ell$, write $C_\ell(V) = \{\ell + s_\ell z : z \in \Delta_V\}$ for the chart that fixes the floor ℓ and spreads the free mass s_ℓ over V alone. A coefficient there has complete support $\operatorname{supp}\ell \cup V$. Taking $V = U$ recovers the merge chart, $C(U,\ell) = C_\ell(U)$, and at zero floor $C_0(V) = \Delta_V$. The score of a residual set is

$$\rho(A; V, \ell) = \min_{c\in C_\ell(V)} \max_{t\in A} q_t(c).$$

In the *atlas index* the chart (U,ℓ) is held fixed, one fit serves every later query, and only the protected set T varies; the canonical case is one mixture weight simplex in a shared coefficient space (Figure 1a), and a router or a rule with one weight per layer group is a general compact convex chart $C \subset \mathbb{R}^d$ that the duality and Helly certificates (Theorems 2.1, 2.4) cover as well. Only in the atlas index is the query downward-closed, so that every subset of a compatible task set is itself compatible.

In the *restricted atlas* the task models are fitted once on the full ambient simplex, and (T,U,ℓ) then vary by restriction with no further evaluation of the merged model.

The *local quadratic screen* instead refits the models after (T,U,ℓ) is chosen; the public PEFT runs use this regime on the diagonal $T{=}U{=}S$ with $\ell_i{=}0.2$ on $i \in U$, scoring a pair on $C(\{i,j\},\ell) \subset \mathbb{R}^2$ and a triple on $C(\{i,j,k\},\ell) \subset \mathbb{R}^3$.

For a fixed floor ℓ the score is nondecreasing in the protected set T and nonincreasing in the residual adapter set U (Eq. (4)). With the chart held fixed, compatibility is therefore downward-closed in T (Theorem 2.4), and the only possible pair/triple disagreement is the classical one, in which every pair is feasible but the triple is not.

On the diagonal, however, both T and U vary with S; enlarging S adds a retention constraint while enlarging the coefficient domain, and no general monotonicity remains. With a query-dependent floor the charts themselves stop nesting. Once $\ell \neq 0$ the pair chart $C(\{i,j\},\ell)$ is not a face of the triple chart $C(\{i,j,k\},\ell)$, and a triple can then be feasible while a pair sub-query is not; the change of chart produces this second disagreement (Figure 1b). Table 6 counts both disagreements, the same score evaluated at different (U,ℓ), on the Qwen $m{=}50$ panel.

For task t, let $J_t : \mathbb{R}^p \to \mathcal{H}_t$ be the linear map sending a weight update to its effect on task t's outputs, with $\mathcal{H}_t$ 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

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

When $\Delta(c)$ is linear in c, as for the exact weighted sum $\sum_i c_i \Delta W_i$, this is a convex quadratic in the coefficients, and the derivation is exact. The factor-space operator used in the PEFT runs is not of that form, since it

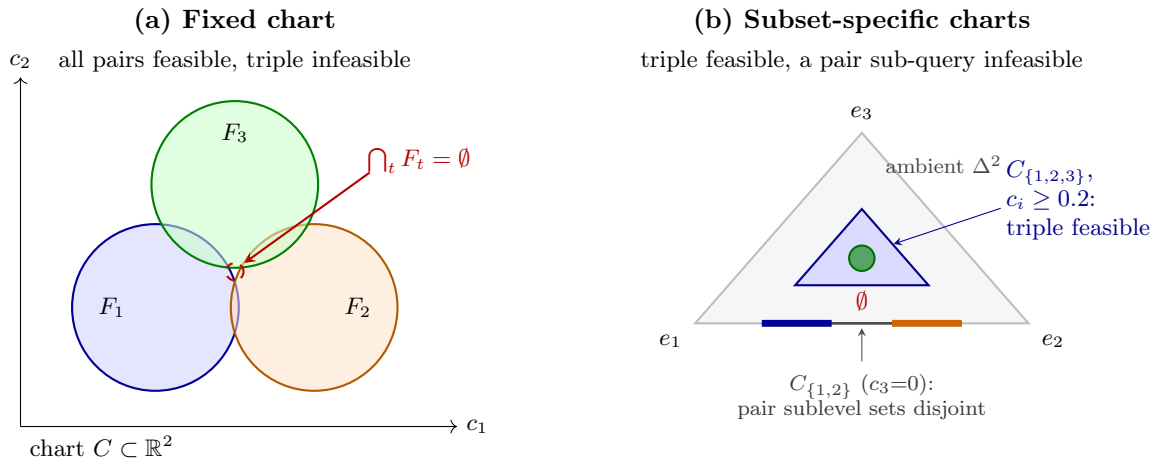

Figure 1: Two ways that pair and triple checks can disagree. **(a) Fixed chart**, where only the protected set varies: three convex sublevel sets in $C \subset \mathbb{R}^2$ can intersect pairwise while having no common triple intersection (Theorem 2.8). **(b) Subset-specific charts** with floor ℓ_i=0.2: the pair chart $C_{\{1,2\}}$ (on the c_3=0 edge) and the triple chart $C_{\{1,2,3\}}$ (the interior sub-simplex) 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. Table 6 separates the two cases on the PEFT panels.

scales each factor pair by the square root of its coefficient and so contributes cross terms (Appendix B); there 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, which may be scored on every prompt token or only on the tokens of a task's answer; we call the latter choice the *answer loss*. The true second derivative $\widetilde{H}_t = \nabla^2 f_t$ need not be PSD for nonlinear losses, and whether the loss is convex along weight interpolations is the subject of the mode-connectivity literature (Garipov et al., 2018; Frankle et al., 2020), with linearized fine-tuning (Ortiz-Jiménez et al., 2023) the regime where quadratic-in-coefficient structure is most plausible; we therefore fit the unprojected model

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

then take the positive semidefinite (PSD) projection $H_t := \Pi_{\mathrm{PSD}}(\widetilde{H}_t)$ (replace negative eigenvalues by zero) 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 $H_t \succeq 0$; we log the projection size $\|\widetilde{H}_t - H_t\|_F$ per query. The fitted convex quadratic $\widehat{q}_t$ is the *quadratic task-loss model*, or task model for short; it stores a value α_t, a slope b_t, and a curvature H_t of the loss in the merge weights, the data of a second-order Taylor expansion around the anchor. The theory below is stated for generic convex task models q_t; in the experiments, q_t is instantiated by the fitted model $\widehat{q}_t$, 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 $\rho_S(C) \leq \tau$; the normalized default takes τ=1. Eq. (3) is a quasiconvex program (Eppstein, 2005; Amenta, 1994): the finite-core structure of Theorem 2.4 requires only that the sublevel sets $F_t(r)$ be convex, while the duality of Theorem 2.1 uses convexity of the task models themselves. Eq. (3) is the diagonal case of a two-sided score that separates the protected tasks T from the merge adapters U and their floor,

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

so the diagonal $\rho(S;S,0)$ is Eq. (3) on the simplex chart Δ_S, while a general chart C (a router or a box) keeps the duality and Helly certificates. Define the sublevel-set family and the threshold complex by

$$\begin{aligned} F_t(r) &= \{c \in C : q_t(c) \le r\}, \\ K_r &= \{S \subseteq [n] : \textstyle\bigcap_{t\in S} F_t(r) \neq \emptyset\}. \end{aligned} \tag{5}$$

Each $F_t(r)$ collects the merge weights at which task t's predicted loss stays below r, so a set of tasks is jointly retainable at level r when their regions share a point; as r grows the regions grow, and K_r records which subsets have become jointly feasible.

A concrete $d = 2$ instance shows directly how pairs can pass while a triple fails. Three disks in $\mathbb{R}^2$ whose centers form an equilateral triangle can pairwise overlap yet share no common point, the $(d{+}1){=}3$ obstruction core that pairwise screening accepts but the atlas rejects (Proposition D.4; Figure 1).

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

$$\rho_S(C) = \max_{\lambda\in\Delta_S} \min_{c\in C} \sum_{t\in S} \lambda_t q_t(c), \tag{6}$$

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

$$\begin{aligned} &L(\lambda) := \min_{z\in C} \textstyle\sum_t \lambda_t q_t(z), \quad U(c) := \max_t q_t(c), \\ &L(\lambda) \le \rho_S(C) \le U(c). \end{aligned}$$

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 $\rho_S(C)$.

Any single coefficient thus gives an upper bound on the score and any weighting of the tasks a lower bound; when the two meet, the pair (c, λ) certifies the exact value.

Proposition 2.2 (Projection is accept-sound)**.** *Because $H_t = \Pi_{\mathrm{PSD}}(\widetilde{H}_t)$ replaces the negative eigenvalues of $\widetilde{H}_t$ by zero, $H_t \succeq \widetilde{H}_t$, so $\widehat{q}_t(c) \ge \widetilde{q}_t(c)$ for every t and every $c \in C$, with equality for all $c \in C$ whenever $\widetilde{H}_t \succeq 0$. (For a chart of full affine dimension the converse also holds; a lower-dimensional chart can avoid the negative eigenspace of $\widetilde{H}_t$, so equality on C does not force $\widetilde{H}_t \succeq 0$.) Hence $\max_{t\in S} \widehat{q}_t(c) \ge \max_{t\in S} \widetilde{q}_t(c)$, and any primal c with $U(c) = \max_t \widehat{q}_t(c) \le \tau$ certifies compatibility for the unprojected fitted models as well. Relative to the unprojected model, projection can remove an accept but cannot introduce one. Errors relative to the deployed loss still depend on the Taylor remainder $f_t - \widetilde{q}_t$ and the estimation error in the fitted coefficients, not on the projection.*

Theorem 2.3 (Target certificate)**.** *Fix a query (T, U, ℓ) and let f_t be the true retained loss of task t, so that*

$$\rho^f(T; U, \ell) = \min_{c\in C(U,\ell)} \max_{t\in T} f_t(c)$$

is the two-sided score under the deployed loss. Suppose two-sided envelopes $\epsilon_t^-, \epsilon_t^+ \ge 0$ bracket the task model $\widehat{q}_t$ of Eq. (2) against this loss on the query chart,

$$\widehat{q}_t(c) - \epsilon_t^- \;\le\; f_t(c) \;\le\; \widehat{q}_t(c) + \epsilon_t^+ \qquad \text{for all } t \in T,\ c \in C(U,\ell).$$

For a primal $c \in C(U,\ell)$ and a dual weight λ in the probability simplex Δ_T, set

$$U_{\mathrm{true}}(c) := \max_{t\in T} \big(\widehat{q}_t(c) + \epsilon_t^+\big), \qquad L_{\mathrm{true}}(\lambda) := \min_{c\in C(U,\ell)} \sum_{t\in T} \lambda_t \widehat{q}_t(c) \;-\; \sum_{t\in T} \lambda_t \epsilon_t^-.$$

Then $L_{\mathrm{true}}(\lambda) \le \rho^f(T; U, \ell) \le U_{\mathrm{true}}(c)$. Hence $U_{\mathrm{true}}(c) \le \tau$ 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. 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 the left-hand side over the chart gives $\rho^f(T;U,\ell) \le U_{\text{true}}(c)$, so $U_{\text{true}}(c) \le \tau$ certifies $\rho^f(T;U,\ell) \le \tau$. For the lower bound, $\sum_{t\in T}\lambda_t f_t(c) \ge \sum_{t\in T}\lambda_t \widehat{q}_t(c) - \sum_{t\in T}\lambda_t\epsilon_t^-$ for every c, so $\min_{c\in C(U,\ell)}\sum_{t\in T}\lambda_t f_t(c) \ge L_{\text{true}}(\lambda)$; weak duality, which holds for f_t with no convexity assumed, gives $\rho^f(T;U,\ell) = \min_c \max_{t\in T} f_t(c) \ge \max_{\lambda'}\min_c \sum_{t\in T}\lambda'_t f_t(c) \ge L_{\text{true}}(\lambda)$, so $L_{\text{true}}(\lambda) > \tau$ certifies $\rho^f(T;U,\ell) > \tau$. The sandwich $L_{\text{true}}(\lambda) \le \rho^f(T;U,\ell) \le U_{\text{true}}(c)$ rules out returning both verdicts. □

Concretely, if the fitted quadratic is known to lie within ϵ_t^+ below and ϵ_t^- above the true loss everywhere on the chart, then padding the accept test upward and the reject test downward by those amounts makes both verdicts valid for the deployed model, and the two tests can never both fire. The envelopes $\epsilon_t^\pm$ collect the pointwise gap between the task model $\widehat{q}_t$ and the deployed loss f_t on the chart, in three parts, the calibration sampling error in the fitted coefficients, the interpolation error of the poised design (both bounded in Appendix D.9), and the nonlinear remainder $f_t - \widetilde{q}_t$ of the quadratic approximation. The PSD-projection correction $\widehat{q}_t - \widetilde{q}_t \ge 0$ of Proposition 2.2 is one-sided and so enters ϵ_t^- alone. Two further error sources stay outside the envelopes. The solver gap on the surrogate minimax lands in the primal-dual bracket of Theorem 2.1, since $U_{\text{true}}(c)$ and $L_{\text{true}}(\lambda)$ bound $\rho^f(T;U,\ell)$ for any primal c and dual λ. A library update moves the true-loss score by at most the largest per-task change, by the 1-Lipschitz dependence of the minimax on the task models (Section 4), so a verdict whose band clears that change stays valid. Setting every envelope to zero recovers the surrogate certificate of Theorem 2.1. Section 5.1 measures these 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 theorem says 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.

Theorem 2.4 (Finite core theorem)**.** *Let $C \subset \mathbb{R}^d$ be compact and convex, and let S be any subset. (i) If S is incompatible at threshold r, then some $T \subseteq S$ with $|T| \le d+1$ is incompatible at threshold r (Helly, 1923). (ii) The full score is realized by a sparse core:*

$$\rho_S(C) = \max_{T\subseteq S,\ |T|\le d+1} \rho_T(C). \tag{7}$$

(iii) The order $d+1$ is tight for families of compact convex sets in $\mathbb{R}^d$.

Theorem 2.4 is one half of a two-sided statement. Suppose 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. For a compact convex chart C, write $k(C) = \dim \operatorname{aff} R(C)$ for the dimension of the *latent image* $R(C)$, which alone bounds the task core. An *atomic* chart is generated by adapter atoms v_i, the pure single-adapter settings ($v_i = e_i$ on the ambient simplex). Write $a_i = Rv_i$ for their latent images and $k_U = \dim \operatorname{aff}\{a_i : i \in U\}$ for the residual adapter dimension. In general $k_U \le \min(|U|{-}1, \operatorname{rank} R)$, and $k_U = k(C(U,\ell))$ when $s_\ell > 0$.

Theorem 2.5 (Two-sided cores and discrete saddle)**.** *For the two-sided score $\rho(T;U,\ell)$ of Eq.* (4) *with $q_t(c) = \tilde{q}_t(Rc)$ and $\tilde{q}_t$ convex, alongside the primal-dual saddle (Theorem 2.1):*

(i) Task core (Helly, any compact convex chart). *Applying Helly in the latent image of dimension $k(C(U,\ell)) \le k_U$ (Theorem 2.4) gives $A^\star \subseteq T$, $1 \le |A^\star| \le k(C(U,\ell))+1$, with $\rho(A^\star;U,\ell) = \rho(T;U,\ell)$. This needs only that $C(U,\ell)$ is compact convex, so it holds for a router or box chart as well.*

(ii) Adapter core (Carathéodory, atomic simplex chart). *On the atomic chart $C(U,\ell) = \{\ell + s_\ell z : z \in \Delta_U\}$ with $s_\ell > 0$, some minimizer $c^\star$ can be chosen with 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, $\rho(T;V^\star,\ell) = \rho(T;U,\ell)$ on $C_\ell(V^\star)$; the complete support is $\operatorname{supp}\ell \cup V^\star$, and at zero floor $V^\star$ is the whole support, so the merge uses at most $k_U{+}1$ adapters. When $s_\ell = 0$ the chart is the single point ℓ, the residual core is empty, and $\rho(T;U,\ell) = \max_{t\in T} q_t(\ell)$.*

(iii) Discrete saddle. *For $s_\ell > 0$, writing $\beta_\ell(A,V) = \rho(A;V,\ell)$ and ranging over $1 \le |A|, |V| \le k_U{+}1$, the two cores form a saddle point,*

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

so $\max_A \min_V \beta_\ell(A,V) = \min_V \max_A \beta_\ell(A,V) = \rho(T;U,\ell)$.

Since $k_U \le d$ both cores have order at most $d{+}1$; when the task models depend on the coefficient only through a common rank-r linear map the order drops to $r{+}1$ (Section 3).

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 (proof in Appendix D.3). Note that on a simplex chart with every adapter free, $k_U = |U| - 1$ and the adapter bound is vacuous; it binds when the models share a lower-rank linear map or when the chart is a router of dimension below $|U| - 1$. The adapter-core clause is stated for the simplex atoms $v_i{=}e_i$ used throughout, and the same Carathéodory reduction holds on a general atomic chart once the floor is replaced by its coefficient-space image $b_\ell = \sum_i \ell_i v_i$.

A single symmetric instance forces both cores to their maximal order $r{+}1$ at the same time. For $r{=}2$ it is three adapters at the corners of an equilateral triangle with the three matching tasks. The barycenter is the only mixture that retains all three, dropping any one task lets the score improve, and dropping any one adapter puts some task out of reach.

Theorem 2.6 (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, so $\|v_i\| = 1$, $\sum_i v_i = 0$, and $\langle v_i, v_j \rangle = -1/r$ for $i \neq j$. Take these $r{+}1$ points as the task centers and as the latent images of the adapter atoms: the chart is the coefficient simplex $C = C([m], 0) = \Delta_{[m]}$ with $m = n = r{+}1$, the linear map R sends each coefficient atom e_i to v_i (an affine bijection onto $\operatorname{conv}\{v_0, \ldots, v_r\}$, so $a_i = Re_i = v_i$), and $q_t(c) = \|Rc - v_t\|^2$. Then $k_{[m]} = r$, and with $T = [n]$, $U = [m]$, $\ell = 0$ the two orders of Theorem 2.5 are tight together. (i)* Full score: *the unique minimizer of $\rho([n]; [m], 0)$ is the barycenter $\bar{v} = 0$, with $\rho([n]; [m], 0) = 1$. (ii)* Task core tight: *dropping any single task lowers the score, $\rho([n] \setminus \{j\}; [m], 0) = 1 - 1/r^2 < 1$ for every j, so no task core of size r attains the full score. (iii)* Adapter core tight: *dropping any single adapter raises the score, $\rho([n]; [m] \setminus \{j\}, 0) \geq 1 + 2/r > 1$ for every j, so no adapter core of size r carries the merge. Hence the task core and the adapter core of Theorem 2.5 each require all $r{+}1 = k_{[m]}{+}1$ indices (proved in Appendix D.4).*

Corollary 2.7 (Finite score table)**.** *Let $r = r_{\text{exact}} = \operatorname{rank}(M_C)$ be the exact shared dimension (Theorem 3.1), through which every task model factors exactly, so that $k_U \leq r$ for every U; when the models factor only within a uniform error δ through a smaller effective rank (Section 3), the same table holds approximately, with the maximin and minimax bracket of Theorem 3.4.*

At zero floor the two-sided core table

$$\mathcal{B}_r = \big\{\beta(A,V) : 1 \leq |A| \leq r{+}1,\ 1 \leq |V| \leq r{+}1\big\}, \qquad \beta(A,V) = \rho(A; V, 0),$$

over task subsets A and adapter subsets V, determines every query $\rho(T; U, 0)$ through the discrete saddle of Theorem 2.5,

$$\rho(T; U, 0) = \max_{A \subseteq T,\ 1 \leq |A| \leq r+1} \ \min_{V \subseteq U,\ 1 \leq |V| \leq r+1} \beta(A, V).$$

The restriction to order $\leq r{+}1$ is lossless, since for each A the inner minimum equals $\rho(A; U, 0)$ and is attained by an adapter core of at most $k_U{+}1 \leq r{+}1$ atoms, while the outer maximum equals $\rho(T; U, 0)$ and is attained by a Helly task core of at most $k_U{+}1 \leq r{+}1$ tasks. A single table then determines every pair (T, U) at a fixed zero floor.

For a fixed nonzero floor ℓ the same reading holds on the residual-support charts $C_\ell(V)$, with entries $\beta_\ell(A, V) = \rho(A; V, \ell)$ over residual sets V of size at most $r{+}1$; the residual core $V^\star$ gives the free-mass support above ℓ, and the complete adapter support is $\operatorname{supp} \ell \cup V^\star$.

Under floors that vary with the query the fitted models stay reusable without new evaluation of the merged model (Section 4), but the numeric table depends on the floor, since each floor induces its own $\mathcal{B}_r$. The cross-table is far larger than the one-sided task-core table, with worst-case size $\big(\sum_{a=1}^{r+1} \binom{n}{a}\big)\big(\sum_{v=1}^{r+1} \binom{m}{v}\big)$, so in practice the index stores the fitted models and answers each query by a direct convex optimization, and the full cross-table is precomputed only for small m, n, r; the one-sided fixed-chart $(d{+}1)$-core table (Corollary D.1) is far cheaper and is precomputed when U is fixed and its core order is small.

The fixed-chart score is the upper envelope of its low-order table,

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

so every threshold decision $\rho_S(C) \leq \tau$ reduces to a max-core lookup over the $\leq (d{+}1)$-cores of S, and the $(d{+}1)$-core score table is a sufficient statistic for exact fixed-chart threshold decisions (Theorem D.2;

Corollary D.1 in Appendix D.6 reads ρ_S as a birth time and bounds the cost of recovering an accepting coefficient). When the task Hessians share a low-rank structure, the core order shrinks from $d{+}1$ to $r{+}1$ (Theorem 3.2), with a computable candidate projector and a certified residual (Appendix C.1).

Theorem 2.8 (No pairwise surrogate on a fixed chart)**.** (Fixed chart C shared across subset queries.) *There is no universally exact representation of multiway adapter compatibility by a pairwise mergeability graph. For $d \geq 2$, 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.*

The obstruction order, the size of the largest minimal blocking task set, determines what pairwise screening can detect. With one global weight the order is two and the pairwise graph decides compatibility; a two-dimensional chart already needs order three, and a router chart of dimension d needs order $d{+}1$, beyond the reach of any pairwise screen. The numeric pair scores do not help either; two libraries can share every pair score while a triple score differs (Proposition D.3).

LoRA factors carry a reparameterization, since the pairs (B, A) and $(BR, R^{-1}A)$ encode the same weight update for any invertible R; a rule that reads raw factor coordinates can therefore change its answer while the model does not change.

Theorem 2.9 (Sublevel-set identifiability and raw-geometry impossibility)**.** *For the centered local geometry of output changes, the quadratic task models determine the sublevel-set family $\{F_t(r)\}_{t,r}$, the nested complexes $\{K_r\}_r$, and all scores, finite cores, dual witnesses, and optimal coefficient sets; conversely the sublevel-set family recovers each q_t on C through $q_t(c) = \inf\{r : c \in F_t(r)\}$, while the complexes alone fix the compatibility topology but not the models themselves. Suppose further that the merge map is* gauge invariant*, meaning $\Delta(c)$ depends on the adapters only through the represented updates $\{\Delta W_i\}$, as the exact weighted sum $\Delta(c) = \sum_i c_i \Delta W_i$ does. 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.10). Factor-space operators need not lie in this class: PEFT's* `linear` *combination mixes the factors themselves and produces cross terms $B_i A_j$, and reparameterizing adapter i as $(B_i R_i, R_i^{-1} A_i)$ leaves ΔW_i fixed while moving $B_i R_i R_j^{-1} A_j$, so the merged model itself can depend on the factorization (Appendix B). The theorem therefore constrains rules for update-space merging, and the panels merged with the factor-space operator fall outside its scope; Section 5.1 measures a panel merged with the exact weighted sum for comparison.*

3 Effective dimension

The $d{+}1$ Helly bound of Theorem 2.4 is worst-case over arbitrary convex sublevel sets, but the fitted models of adapters trained from the same base may vary along fewer than d directions of a given chart. That number is a matrix rank, and it must be computed in the chart's own coordinates rather than the ambient ones. Write the affine hull of the chart as $\mathrm{aff}(C) = c_0 + L$ and let Q have orthonormal columns spanning L, so a chart point is $c = c_0 + Qz$ in the intrinsic coordinate z. Substituting into the anchored model of Eq. (2) and collecting terms in z gives, up to an additive constant,

$$\widehat{q}_t(c_0 + Qz) = \mathrm{const}_t + g_t^\top z + \tfrac{1}{2} z^\top \widehat{H}_t z, \qquad \widehat{H}_t = Q^\top H_t Q, \quad g_t = Q^\top \big(b_t + H_t(c_0 - a_t)\big),$$

with $\widehat{H}_t \succeq 0$; here g_t is the slope and $\widehat{H}_t$ the curvature of task t's model along the chart. Form the intrinsic curvature matrix

$$M_C = \sum_{t \in [n]} \big(\widehat{H}_t^{\,2} + g_t g_t^\top\big) \succeq 0, \qquad r_{\mathrm{exact}}(C) := \mathrm{rank}(M_C), \tag{8}$$

and assume C has non-empty relative interior in $\mathrm{aff}(C)$. Because M_C is a sum of PSD terms, $\ker M_C = \bigcap_t (\ker \widehat{H}_t \cap \ker g_t^\top)$ is exactly the set of intrinsic directions along which every task model is constant on C, so each model depends only on the projection Π_{exact} of the intrinsic coordinate onto $(\ker M_C)^\perp$, a subspace of dimension $r_{\mathrm{exact}}(C)$ and no smaller. The ambient matrix $\sum_t (H_t^2 + b_t b_t^\top)$ instead counts directions outside the chart's affine hull and overstates the dimension; on the segment $C = \{(x, 0) : |x| \leq 1\}$ the model $q(x, y) = y^2$ is constant, so $r_{\mathrm{exact}}(C) = 0$, while the ambient rank is 1.

Theorem 3.1 (Exact shared dimension)**.** *Let C have non-empty relative interior in* $\mathrm{aff}(C) = c_0 + L$*, and let Q, $\widehat{H}_t$, g_t, M_C and $r_{\mathrm{exact}}(C) = \mathrm{rank}(M_C)$ be as above, with $\Pi_{\mathrm{exact}} = WW^\top$, where W collects the top $r_{\mathrm{exact}}(C)$ eigenvectors of M_C; the induced map on the ambient coefficient space is $R_C := W^\top Q^\top$. Each restricted model factors as $\widehat{q}_t(c_0 + Qz) = \tilde{q}_t(\Pi_{\mathrm{exact}} z)$ for a convex $\tilde{q}_t$, and Theorem 3.2 applies with $r = r_{\mathrm{exact}}(C)$ at zero slack: for every S,*

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

holds exactly. The shared linear map R of Theorem 2.5 is realized here by R_C, whose adapter atoms $a_i = R_C v_i = R_C e_i$ (with $v_i = e_i$ the coefficient-space atom of adapter i, and the common anchor term $R_C c_0$ absorbed) span an affine set of dimension $k_U \le r_{\mathrm{exact}}(C)$, so the two-sided cores of Theorem 2.5 have order $k_U{+}1 \le r_{\mathrm{exact}}(C){+}1$ for every U. No linear map of rank below $r_{\mathrm{exact}}(C)$ preserves every fitted model on C without error (proved in Appendix D.5).

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 (Section 5), and the small cores there come from sparse optima rather than low dimension. Three statements are therefore 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 $r_{\mathrm{exact}} = d$; the experiments exhibit exactly this last kind of sparsity. When the models share only an approximate low-rank subspace, a rank-r projector trades a controlled slack for a smaller core (proofs in Appendix D).

Theorem 3.2 (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. Thus the $(d{+}1)$-core table of Corollary D.1 can be replaced by an $(r{+}1)$-core table whenever such a Π exists.*

Theorem 3.3 (Approximate effective dimension finite core)**.** *Under the same setup but with $\sup_{c \in C} |q_t(c) - \tilde{q}_t(\Pi c)| \le \delta$, for every S,*

$$\left|\rho_S(C) - \max_{|T| \le r+1} \rho_T(C)\right| \le 2\delta.$$

Theorem 3.4 (Approximate two-sided finite core)**.** *Fix $T \subseteq [n]$, $U \subseteq [m]$ and write $\rho := \rho(T; U, 0)$ for the two-sided score of Eq.* (4) *at zero floor. Suppose each task model is uniformly close to a function of a shared linear image, $\sup_{t \in [n],\, c \in C} |q_t(c) - \tilde{q}_t(Rc)| \le \delta$ with every $\tilde{q}_t$ convex and $\mathrm{rank}\, R = r$. With $\beta(A, V) := \rho(A; V, 0)$ define the restricted maximin and minimax scores*

$$\underline{\rho}_r(T, U) := \max_{A \subseteq T,\, 1 \le |A| \le r+1} \; \min_{V \subseteq U,\, 1 \le |V| \le r+1} \beta(A, V), \quad \overline{\rho}_r(T, U) := \min_{V \subseteq U,\, 1 \le |V| \le r+1} \; \max_{A \subseteq T,\, 1 \le |A| \le r+1} \beta(A, V).$$

Then $|\rho - \underline{\rho}_r(T, U)| \le 2\delta$ and $|\rho - \overline{\rho}_r(T, U)| \le 2\delta$, and their finite-game gap is bounded,

$$0 \le \overline{\rho}_r(T, U) - \underline{\rho}_r(T, U) \le 2\delta.$$

When $\delta = 0$ this is the discrete saddle of Theorem 2.5 on a rank-r linear map; with slack the accept side (an adapter core) and the reject side (a task core), both of order at most $r{+}1$, still agree to within 2δ (proof in Appendix D.13).

When one rank-r projector serves a whole shared chart, the $(r{+}1)$-core table replaces the $(d{+}1)$-core table. Whether such a projector exists is an empirical question; Section 5 finds median rank $r{=}2$ on five-adapter Qwen charts. Write $\mathcal{L} = \{H_1, \ldots, H_n\}$ for the per-task PSD Hessians. We take the top-r eigenspace of the mean Hessian $\bar{H} = \frac{1}{n} \sum_t H_t$ as a computable candidate subspace, in $O(d^3{+}nd^2)$ time, and certify its strict per-task residual; the spectral effective dimension $\sigma\text{-}\,\mathrm{adim}_\epsilon(\mathcal{L})$ (Appendix C.1) is the smallest rank whose residual falls below ϵ, with the projector diagnostic given as Remark C.1.

4 The atlas as an index

The local quadratic screen refits the task models for every query. On a shared ambient chart the task models depend only on the library, so one quadratic task-loss model per protected task, fitted once, answers every

later subset query by a small convex program over the fitted quadratics. Section 5 measures this index at m=8 on Qwen and TinyLlama shared charts and on Mistral ambient charts at sixteen, thirty, and fifty adapters, and evaluates the locally refit score on the six public panels.

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 choice of retention loss; a new operator or a new choice of scored tokens needs a new atlas. Each task has two fitted channels, a PSD-projected model for certificate computation (Proposition 2.2) and an unprojected model that ranks better on the queries whose fitted curvature required projection (+0.04 pooled AUROC, Appendix A). Either channel can be fit on the prompt-scored loss the released panels use or on the answer loss, which correlates with task accuracy (Section 5). A verdict record stores the optimized coefficient, its residual adapter core (the at most d+1 adapters above the floor; the complete support adds $\operatorname{supp}\ell$, and at zero floor the core is the whole support), the critical task core, dual weights, certificate gap, and PSD diagnostics. It also records the base model identity and fit provenance.

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 fitting quadratic response surfaces on a mixture simplex, in the response-surface tradition of Box and Wilson (1951). At $d = m-1$ this gives

$$N_d = m + \binom{m}{2} = 1 + d + \tfrac{d(d+1)}{2}$$

design merges, and scoring each on the n protected tasks costs nN_d scalar loss evaluations (n=m here). Building the fifty-adapter atlas therefore takes one pass over 50 singles and 1,225 equal-weight pairs; both m=50 builds in Section 5 run exactly this design. 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 (ninetieth percentile 0.54) for a five-way query on the fifty-adapter chart (d=49),[1] timings reproduced by the released query-timing scripts. A precomputed core table answers threshold queries by lookup in microseconds (Corollary D.1 gives the (d+1) table). Algorithm 1 collects the pipeline: fit one quadratic model per task, solve a minimax over the merge weights, and return either accepting weights or a small blocking set of tasks.

At the measured effective ranks the far smaller (r+1)-core tables reproduce the full scores on both shared charts, exactly at r=3 on Qwen and to below 10^{-3} at r=2 on TinyLlama (Section 5). The additive slack of Theorem 3.3 does not guarantee this, but the gap admits a per-library certificate, since models reduced by the shared projector satisfy the (r+1)-core identity exactly; the supremum over tasks and the chart of the discrepancy between the full and reduced models is computable by enumerating the simplex faces; and any query whose reduced score clears the threshold by more than that supremum provably receives the verdict of the full models (Proposition D.5). At m=8 the certificate settles 23% of the 211 subset queries, all subsets of sizes two to five plus the full eight-set, at the operating threshold (retention deficit zero, the feasibility boundary of the protocol in Section 5) on the Qwen chart, with a median per-task bound of 16.3 on scores that span 78. On the task-diverse TinyLlama chart, whose rank reduction is lossy, it settles none. The measured gaps never exceed 9.9 on Qwen and 0.37 on TinyLlama, and no certified decision disagreed at any threshold in a 201-point sweep.

A verdict is valid for the base and adapter versions recorded with it. The minimax score is 1-Lipschitz in the task models under the sup norm, so measured model drift can be compared directly with the certificate margin when deciding whether to refit. Adding an adapter or a protected task extends the fitted models without a rebuild (median drift 0.16 nats per added coordinate; nats are natural-log units of loss), and consolidating an accepted merge into the base moves the remaining models by up to 0.41 nats, comparable to certificate margins; Appendix B.5 gives the update rules and measurements. This yields a margin-triggered refit rule that is correct by construction: a stored verdict with certificate margin γ stays valid while the accumulated drift bound $\max_t d_t$ stays below γ, so maintenance re-solves exactly the queries whose margins

[1] Both figures are at zero floor, the production decide default, and are reproduced by the released query-timing scripts; the d=49 measurement uses the fifty-adapter Mistral index of Section 5.1, the released fifty-adapter chart whose fitted models are on disk. At the default floor 0.2 a five-way query exhausts the coefficient mass and reduces to the single equal-floor point (Section 5.3), so the zero-floor timing is the upper bound; lower floors restore the full optimization.

Algorithm 1 Atlas pipeline (used in all experiments): 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 $\ell \in \mathbb{R}^m_{\geq 0}$ supported on U with $\|\ell\|_1 \leq 1$, threshold τ, margin η; linear map R with adapter atoms $a_i{=}Re_i$ (simplex coordinates throughout; a general atomic chart is handled by the remark after Theorem 2.5), per-task calibration prompts $\{\mathcal{D}_t\}_{t\in T}$.
Output: Verdict $\in \{\text{accept}, \text{reject}, \text{inconclusive}\}$ with primal coefficient $c^\star$, adapter core $V^\star{=}\{i \in U : c^\star_i > \ell_i\}$, critical task core $A^\star \subseteq T$, dual $\lambda^\star$, gap $U(c^\star){-}L(\lambda^\star)$, and a robust primal-dual envelope.
Fit or reuse task models. Regime 1 (atlas index, fixed chart): fit one model per protected task once on the shared chart $C(U,\ell)$ and reuse it across every later query, at amortized cost nN_d scalar probes over the library for $N_d{=}1 + d + d(d{+}1)/2$ poised design points (Appendix D.9). Regime 2 (restricted atlas, shared ambient chart): fit the models once on the full ambient simplex and answer the query by restriction to $C(U,\ell)$ at no additional model cost. Regime 3 (local quadratic screen): refit the models for the current chart at cost $|T|N_d$, as in the subset-specific PEFT runs. In every regime the anchored model is $\widehat{q}_t(c) = \alpha_t + b_t^\top(c-a_t) + \frac{1}{2}(c-a_t)^\top H_t(c-a_t)$ (Eq. (2)) for $t \in T$, with $H_t \succeq 0$. *Exact mode* solves on these full models, and its cores reproduce their score. *Reduced mode* first restricts them to the top-r eigenspace Π_r of $\bar{H}{=}|T|^{-1}\sum_{t\in T} H_t$ (Appendix C.1), when its strict residual check passes; its cores, of order at most $r{+}1$, are exact for the rank-reduced score $\rho_r = \min_{c\in C(U,\ell)} \max_{t\in T} \widehat{q}_t^{(r)}(c)$, and $|\rho - \rho_r| \leq \delta$ for the uniform reduction error $\delta = \sup_{t,c} |\widehat{q}_t(c) - \widehat{q}_t^{(r)}(c)|$ (Proposition D.5).
Solve the minimax. Write q_t° for the models in force, $\widehat{q}_t$ in exact mode and $\widehat{q}_t^{(r)}$ in reduced mode, and ρ° for their minimax score, ρ or ρ_r. Compute $\rho^\circ = \min_{c\in C(U,\ell)} \max_{t\in T} q_t^\circ(c)$ (Eq. (4)), for a primal minimizer $c^\star$ and an optimal dual $\lambda^\star$ (Theorem 2.1).
Latent optimum. Write the primal in latent coordinates, $x^\star{=}Rc^\star$, and split off the floor along the free mass $s_\ell{=}1 - \|\ell\|_1$: the residual $w^\star{=}(x^\star - R\ell)/s_\ell$ lies in $\operatorname{conv}\{a_i : i \in U\}$ with adapter atoms $a_i{=}Re_i$ (when $s_\ell{=}0$ the chart is the single point ℓ, so the adapter core $V^\star$ is empty, the mandatory support $\{i : \ell_i > 0\}$ is returned separately, and the next step is skipped).
Carathéodory elimination. Constructively reduce the representation of $w^\star$ to at most $k_U{+}1$ atoms: while more than $k_U{+}1$ atoms carry positive weight, solve for an affine dependence among the active a_i and shift the weights along it until one vanishes, holding $w^\star$ fixed. This returns $w^\star = \sum_{i\in V^\star} z_i a_i$ with $z \in \Delta_{V^\star}$, $|V^\star| \leq k_U{+}1$, and $k_U = \dim \operatorname{aff}\{a_i : i \in U\}$. Set $c^\star{=}\ell + s_\ell z$, reading the weights $z \in \Delta_{V^\star}$ as a vector in $\mathbb{R}^m$ supported on $V^\star$; the residual core $V^\star{=}\{i : c^\star_i > \ell_i\}$ reproduces the score on $C_\ell(V^\star)$ (Theorem 2.5) and is an acceptance witness when its robust upper bound clears τ.
Critical task core. Extract a Helly task core by greedily deleting tasks from the active set $\{t : q_t^\circ(c^\star) \geq \rho^\circ - \varepsilon_{\text{act}}\}$, with ε_{act} a numerical tolerance in place of exact equality, while the score of the remaining tasks on the models in force stays within $\varepsilon_{\text{core}}$ of ρ°, then re-solving the dual on the remainder (an arbitrary optimal dual need not be sparse). This returns $A^\star \subseteq T$ with $|A^\star| \leq k_U{+}1$ whose restricted score equals ρ° (Theorem 2.5); Section 5.2 uses this construction. This is the critical task core, and an obstruction core when $\rho^\circ > \tau$.
Verify the cores. Recompute the score on the residual-support chart with the models in force and confirm it equals ρ°, so the task and residual adapter cores jointly reconstruct the solved score (the complete adapter support is $\operatorname{supp}\ell \cup V^\star$); in reduced mode $|\rho - \rho_r| \leq \delta$ carries the check to the full models.
Robust intervals and verdict. Form the target-certificate bounds $U_{\text{true}}(c^\star){=}\max_t(\widehat{q}_t(c^\star){+}\epsilon_t^+)$ and $L_{\text{true}}(\lambda^\star){=}\min_c \sum_t \lambda_t^\star \widehat{q}_t(c){-}\sum_t \lambda_t^\star \epsilon_t^-$ (Theorem 2.3) from the per-task envelopes $\epsilon_t^\pm$, which collect the sampling, interpolation, nonlinear-remainder, and projection terms of Theorem 2.3; the solver gap stays in the primal-dual bracket $U(c^\star){-}L(\lambda^\star)$. In reduced mode the final verdict is checked against the full fitted models rather than the reduced ones, since U_{true} evaluates $\widehat{q}_t$ at the reduced-mode coefficient $c^\star$ and L_{true} recomputes $\min_c \sum_t \lambda_t^\star \widehat{q}_t(c)$ on the full models, so the dual stays a valid lower bound on the full fitted problem; the rank-reduction slack δ enters only the separately reported reduced score ρ_r. With empty envelopes $\epsilon_t^\pm{=}0$ 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.

the drift has crossed. Because the measured drifts sit at the scale of typical margins rather than far below it, the rule currently triggers often; its present value is the soundness of every served verdict rather than refit avoidance. The schedule that keeps a whole index current at library scale remains open (Section 6).

5 Experiments

The six public panels evaluate the local quadratic screen, with task models refitted for each queried subset. The Helly probe over 146 triples and the m=8 core-table certification hold the chart fixed. The eight-task panel scored on the answer loss and the m=16, m=30, and m=50 builds of Section 5.1 fit once on the ambient simplex and answer later queries on its faces (Section 4).

All evaluations share one configuration that fixes the feasibility rule, calibration/held-out splits, the four merge operators (linear; TIES (Yadav et al., 2023); DARE-TIES (Yu et al., 2024); and *magnitude pruning*), the retention thresholds $\tau \in \{0.55, 0.65, 0.70, 0.75, 0.85\}$ used in the stability sweep, and the per-adapter coefficient floor ℓ_i=0.2. Magnitude pruning is our own baseline, defined operationally as task-vector averaging after zeroing per-task-vector entries with magnitude below the median; it is TIES with the trim step but without sign-conflict resolution. The Qwen queries are drawn on four predeclared slices that balance adapter participation, with pair queries drawn cyclically over the adapter ordering. 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).

5.1 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. We test this on five classification tasks from the released Mistral-7B panel, each with recoverable answers (the correct answer occupies identifiable tokens in the target text). The panel has 45 equal-weight pair and triple records over 20 distinct merges, and every adapter improves on its base by +0.20 to +0.97 accuracy. The answer loss tracks measured accuracy (Pearson −0.72 pooled, −0.70 after standardizing within task, and −0.89 to −0.96 on the three tasks whose merge accuracy varies most), while the prompt-scored loss used in our panels does not (+0.37 pooled). The correlation weakens when merges barely change accuracy or when the adapter's gain over the base is small; in both cases there is little signal to track (Appendix A).

The same losses also rank merges. With retention defined as keeping at least 90% of the adapter's accuracy gain (21 of 45 records retained), the answer loss recovers 48% of the retained records at 71% precision in the top 30% of the ranking, with pooled AUROC 0.73, within-task 0.96 to 1.00 on tasks 1391, 190, and 280, 0.64 on task290, and below-chance performance on task391, the task with the smallest gain. The prompt-scored loss at the same budget returns 24% recall at 36% precision, under the 47% base rate (AUROC 0.38).

The fitted task models retain this association. Quadratic models fitted once on the answer loss over one shared chart of the five adapters predict the measured answer loss of the panel's merges (fidelity, Pearson 0.81; 0.79 on the 25 off-design triples) and rank accuracy retention at AUROC 0.78, versus 0.73 for the measured loss itself on the same retention labels. An ablation in Appendix A re-scores projected events with the unprojected model and recovers part of the ranking cost of projection (+0.04 pooled AUROC, largest for DARE-TIES).

Table 2 and Figure 2 summarize six panels scored on the answer loss, across two bases, three operators, and library sizes from four to sixteen adapters; the sixteen-adapter library consists of classification tasks from the Lots-of-LoRAs collection (Brüel-Gabrielsson et al., 2024) on Mistral-7B. Every panel has positive adapter gains, and its measured answer loss moves with task accuracy (capability Pearson between −0.54 and −0.80); the m=8 Mistral gains run 2.1 to 11.0 nats (median 5.7). Fidelity depends on how much off-design variation the chart contains. The exactly poised m=8 linear panel gives an overall Pearson correlation of 0.52 and a correlation on triples alone of 0.47; both are higher for TIES and DARE-TIES, whose merge losses span a wider range (spread in answer loss 0.35, 1.21, 1.87 nats for linear, DARE-TIES, TIES). The denser m=16 chart reaches 0.96 on triples alone, and the same criterion carries to the generative panel, whose four free-form tasks' answer loss correlates with exact match and token-F1 at −0.80 pooled.

The Qwen-1.5B panel marks the limit of the local approximation. Several merges leave the calibration region and reach answer losses near 25 nats, and triple fidelity then falls to 0.04. Threshold screening

Table 2: 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 over the four negative triples 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	$0.98^{\dagger}$	–	−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

nonetheless remains informative on this panel (AUROC 0.76), because the fitted score still separates near-threshold feasible from infeasible merges even where it cannot predict the magnitude of the extreme failures.

Exhaustive validation is expensive at m=16, since the $\binom{16}{3}$=560 triples cost hours to validate while one 136-point build scores them all and every higher order; 120 independently evaluated off-design triples give fidelity 0.96 (pairs on the design at 0.98; off-design means unseen coefficient settings, with evaluation items drawn from the task's probe items, so fidelity measures generalization across coefficients rather than across items, Appendix B), and the predicted losses screen these merges at AUROC 0.99.

Theorem 2.9 constrains gauge-invariant merge maps, so we also ran the sixteen-adapter panel under the exact weighted sum of updates, which is in that class. Both operators were fit and validated on the same panel with the same items and protocol, differing only in the combination type. 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 adapter-block bootstrap interval $[-0.04, +0.16]$ contains zero, while screening AUROC moves by less than 0.01 in either direction across the retention thresholds. The direction is the one Eq. (1) predicts, since that derivation is exact when $\Delta(c)$ is linear in the coefficients, but sixteen adapters do not resolve an effect this small. That rerun also puts the factor-space arm at 0.912 against the 0.957 of the original run on different hardware and library versions, 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 base, adding fourteen and then twenty further classification tasks to the library. Each build fits the singles-and-pairs design once (465 and 1,275 design merges on one ambient simplex) and answers every subset query on its faces; three tasks at m=30 and six at m=50 drop out because their prompts exceed the context limit, leaving 27 and 44 scored tasks. Validation covers every pair plus 150 and 200 sampled triples, of which 443 and 1,096 subsets have every member scored. At the 70% retention threshold the screen gives AUROC 0.94 at m=30 and 0.95 at m=50 (adapter-block bootstrap 95% intervals $[0.87, 0.99]$ and $[0.89, 0.98]$, leave-one-adapter-out ranges $[0.93, 0.97]$ and $[0.94, 0.96]$; triples alone 0.93 and 0.94), and pooled fidelity on off-design triples is 0.86 and 0.85. Two of the forty-four scored tasks have negative single-adapter gains, so their labels measure only near-base retention rather than preservation of a positive adapter gain; 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 the triple fidelity falls from 0.69 to 0.54, and at the strictest threshold (τ=0.85) 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.

Two further experiments examine the m=8 atlas directly. Refitting the eight task models on progressively denser designs and predicting 40 off-design four- and five-way merges raises the Pearson fidelity from 0.70 at the poised design of singles and pairs to 0.76 at 2.6× redundancy (all 56 triples added). The coefficient minimizes the worst member's answer loss, and evaluated against equal weights that loss drops on 45 of 56 triples (mean 0.23 nats); the accuracy effect concentrates on the triples where equal weights leave a member weak. On the four triples with a member below 0.6 accuracy under equal weights the coefficient improves that member (three of four, +0.07 accuracy), while on the triples equal weights already handle, reweighting

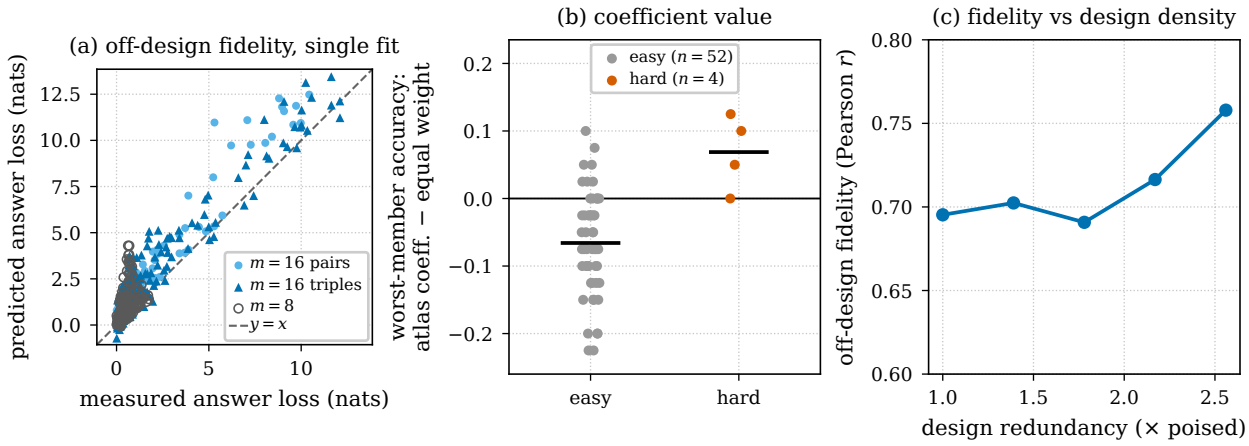

Figure 2: The index fitted once on the answer loss. (a) Predicted versus measured answer loss on off-design merges: a single 136-point build at m=16 predicts all 120 off-design triples (triangles) at Pearson 0.96, with the 120 pairs on the design (circles) at 0.98, while the exactly poised m=8 chart sits nearer the fit noise floor (0.52). (b) Accuracy of the worst member task at the atlas minimax coefficient minus equal weights, by triple: the coefficient improves the four hard triples, those where equal weights leave a member below 0.6 accuracy (mean bar shown), and changes little elsewhere. (c) Off-design fidelity rises overall with design redundancy; the largest gain arrives above 2×.

slightly lowers accuracy, so the coefficient is worth applying on the former and unnecessary on the rest.

On three-adapter charts sampled densely (91 points each) the quadratic model, fit on only the six poised design points, reproduces the full answer-loss surface at median R^2=0.91 (minimum 0.75 across nine task surfaces), and six of the nine surfaces are convex as fitted, so the PSD projection behind accept-soundness is active on the other three. The same grids yield empirical finite-design envelopes for Theorem 2.3. Over all 91 grid points of each chart, the design points included since the PSD projection means the fitted model need not interpolate them, 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, on nine surfaces spanning 0.6 to 8.3 nats. Padding the per-task retention thresholds by these envelopes certifies 41 of the 42 task-subset verdicts for the deployed score over the evaluated 91-point coefficient design at thresholds 0.70 and 0.85, and every bracket contains the deployed score measured on the grid; the remaining query, one three-task set at 0.85, returns inconclusive rather than a wrong verdict. Three further dense charts over nine tasks new at fifty adapters repeat the measurement on the scale substrate, with wider envelopes (ϵ_t^+ of 0.3 to 2.6 and ϵ_t^- of 0.8 to 5.4 nats on surfaces spanning 2.5 to 10.7): padding certifies 32 of the 42 verdicts over the evaluated grid, all 32 correct, the other ten inconclusive, and all 42 brackets contain the deployed score.

5.2 Coefficient and adapter-core intervention

We construct the cores of Theorem 2.5 directly on three libraries of fitted task models, the homogeneous Qwen chart (m=8) and the Mistral charts at m=8 and m=16, fitted on the answer loss. On all three the exact shared dimension $r_{\text{exact}}=\text{rank}(M_C)$ (Eq. (8)) equals the chart dimension d $(7, 7, 15)$ and the Hessian-only rank, so the linear terms add no direction and no exact rank reduction is available; the cores are small regardless. On the Qwen m=8, Mistral m=8, and Mistral m=16 charts 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, as do the two together ($\beta(A^\star, V^\star)=\rho$), and every core is inside the $r_{\text{exact}}+1$ bound. This critical 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 (mean deficit 0.5 to 47 nats across the three charts). Projecting the models to a shared rank-two subspace gives cores of size at most $r+1$=3 (adapter cores exactly 3) whose score sits within the certified band of the full models; $|\rho - \rho_r| \le \delta$ holds in every case, with gaps of 0.8 to 1.9 far inside δ bounds of 5 to 59.

Table 3: Deployed adapter-core intervention on one eight-way Mistral-7B m=8 query (eight Lots-of-LoRAs tasks, linear merge, zero floor, 40 held-out items per task disjoint from the fit split). *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 over the 56 deployed worst losses, with interquartile range in brackets, and the median over the 56 support-level mean accuracies.

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

On one eight-way Mistral-7B query (all eight adapters, linear merge, zero floor, 40 held-out items per task disjoint from the fit split) we test whether the fitted score transfers from the surrogate to the model (Table 3). The score-optimal merge is naturally sparse, using 5 of the 8 adapters because its optimum lies on a face (r_{exact}=d=7 here, so this is not a forced rank reduction); 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. Its worst-task loss is the lowest of all evaluated supports, though the best three-adapter support reaches slightly higher mean accuracy (0.74).

A fifteen-query replication on the sixteen-adapter chart removes the single-query scope. Five queries each at orders four, five, and six, sampled with a fixed seed, solved at zero floor on the released m=16 fitted task models, and deployed on held-out instances disjoint from the fit split, give the atlas coefficient a lower deployed worst-member answer loss than equal weights on thirteen of the fifteen (median improvement 2.25 nats, range -0.21 to $+8.25$), with a median change in mean member accuracy of $+0.01$; the fitted models predict an improvement on all fifteen, so two predicted improvements fail to deploy and none reverses by more than 0.21 nats.

Restricting to three adapters, we reoptimize and deploy all $\binom{8}{3}$=56 subsets. The fitted worst score ranks them against the deployed worst at Spearman 0.88 (Pearson 0.85), and the best-fitted subset is the deployed optimum (worst loss 2.03, rank 1 of 56). The two atlas-derived reductions rank 5th (the top-3 weight truncation) and 18th (the rank-2 Carathéodory core) of 56, both above the median reoptimized three-adapter subset (5.84, interquartile $[3.30, 8.28]$) and the equal-weight baseline over all eight adapters (3.65). The fitted score therefore orders supports without a new model evaluation, while dropping to three adapters costs worst-task loss, since the best three-adapter support (2.03) does not reach the five-adapter minimizer (1.59). This is one query. Two coefficients with the same fitted latent image receive identical scores, and whether they also behave identically on the deployed model is untested.

5.3 Screening across public libraries

Pair queries run on $C_{\{i,j\}} \subset \mathbb{R}^2$ (a Δ^1 chart with intrinsic affine dimension 1) and triples on $C_{\{i,j,k\}} \subset \mathbb{R}^3$ (a Δ^2 chart with intrinsic affine dimension 2). For every query we evaluate a coefficient design for held-out labels, fit retention models on calibration prompts, project indefinite Hessians before solving the convex certificate, and record the projection size; manifests, configurations, per-query records, and dependency ranges are included in the released archive (Appendix B). 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. Balanced slices draw triples so that each adapter appears equally often; circulant slices take the pairs $(i, i{+}k)$ at fixed offsets k, indices modulo m.

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 of the chart, and runs 146 unique triples from the m=50 linear merge panel (Appendix B.2). The probe records a single Helly witness at d=2, with all three face-pairs feasible and the triple infeasible at τ=0; one witness in 146 triples (0.68%) is a lower bound on the rate. Most mistakes of the pairwise screen on the subset-specific charts occur because the pair and triple charts differ (Table 6). The synthetic instance below, a numerical case of Theorem 2.6, isolates the $d{+}1$ obstruction at d=2, 3, 4.

We verify the d+1 obstruction order of Theorem 2.4 on synthetic libraries at d=2, 3, 4. With normalized scores $q_t(c) = \|c - v_t\|_2^2 / R_0^2$ at threshold 1 and the d+1 centers v_t at the vertices of a regular d-simplex of unit circumradius, the full-simplex retention score is $\rho_S = 1/R_0^2$ and each d-vertex face gives $\rho_T = (1 - 1/d^2)/R_0^2$, both analytic. Table 4 reports a Helly witness per d, with $\rho_S = 1.156, 1.108, 1.063$ at $R_0 = 0.93, 0.95, 0.97$ (and $\rho_T = 0.867, 0.985, 0.996$) and the obstructing core size exactly d+1 in every case. This is the numerical instance of Theorem 2.6, whose adapter-core bound the same construction also makes tight, and a numerical grid matches the analytic values to within 1.5×10^{-2} for $d \leq 3$, the resolution of its 160-point mesh.

Table 4: Synthetic Helly verification across d=2, 3, 4. For each d we place d+1 unit-norm centered-quadratic sublevel sets $F_t = \{c : \|c - v_t\|_2^2 \leq R^2\}$ at the vertices of a regular d-simplex of unit circumradius. By symmetry the Chebyshev center of the full simplex is the origin, so $\rho_S = 1/R^2$; the Chebyshev center of a d-vertex face is the face circumcenter, so $\rho_T = (1 - 1/d^2)/R^2$. Every d-subset is feasible and the full (d+1)-tuple is infeasible at the listed R, with obstructing core size exactly d+1.

d	R	n=d+1	full ρ_S	all d-subsets feasible	core size
2	0.93	3	1.156	✓	3
3	0.95	4	1.108	✓	4
4	0.97	5	1.063	✓	5

On the Qwen m=50 run, whose probes are generic prompts, the atlas AUROC is linear 0.89, magnitude pruning 0.85, TIES 0.79, and DARE-TIES 0.76, with a DARE-Linear (Yu et al., 2024) sensitivity run at 0.87; on the four main operators the pooled AUPRC is 0.74 (per-operator 0.68 to 0.79). Every per-operator bootstrap and clustered (slice, adapter-block) 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). Per-operator Spearman correlations are 0.82, 0.77, 0.66, 0.57 and regression slopes 0.77, 0.73, 0.61, 0.39; DARE-TIES has the lowest slope, consistent with additional variation from its drop-and-rescale step.

Linear and magnitude pruning are the two strongest operators on three of the five smaller panels (Table 11), while the two Pythia panels deviate and AUPRC is the more informative metric at their low base rates (Pythia-410M AUROC 0.57 to 0.87, AUPRC 0.18 to 0.52, at a 14% base rate; Pythia-1B AUROC 0.62 to 0.75, AUPRC 0.23 to 0.69; per-panel certificate diagnostics in Table 7). The Phi-2 cells at AUROC 1.00 rest on 68 feasible events across four operators and should be read as point estimates rather than stable rankings.

Because the certificates use PSD-projected models, we also split screening quality by projection status; pooled AUROC is 0.886 on naturally PSD events and 0.783 on projected events (Appendix A). Figure 3 summarizes the per-operator AUROC, the recall and precision available at each validation budget, and the low-cost baselines.

On the TinyLlama-12 library (one adapter per task, spanning a GSM8K math adapter, text-to-SQL, physics, Docker help, and eight others), selecting the lowest-score 20% of candidates by the atlas and merging them clears the retention threshold on every constituent task in 30% of pairs and 12% of triples, versus 7% and 4% under random selection, a 3 to 4.3× rate over a 4.6% base. The atlas-picked merges meet the GSM8K adapter's threshold in every case, versus 73% at random, and the fragile SQL task's in 11%, versus 7%.

On the Qwen m=50 panel (4000 events, full validation ≈ 33 MI300X-hours at 30 s per event), at a 20% validation budget, screening with the atlas and validating the top-k recovers 45.5% of the subsets that pass held-out validation at 77.6% precision and saves 73% of the GPU-hours (adapter-resampling bootstrap: recall [34.2, 59.6]%, precision [60.8, 91.3]%); Table 5 reports the full curve from the 5% to the 50% budget. Screening cost is ∼2 s per query (calibration evaluations for the model fit plus the certificate optimization) versus ∼30 s per full evaluation; the overhead is below 7% of one full evaluation per query, and the reported 73% saving at the 20% budget includes it. 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. At the default floor of 0.2, five selected adapters exhaust the coefficient mass, so a five-way query at that floor scores the single equal-floor merge; lower floors restore free weight for coefficient selection. Exhaustive screening of large subset spaces therefore needs the amortized atlas index (Algorithm 1) rather than the local quadratic screen used in these runs.

Atlas screening outperforms the label-free low-cost screens on Qwen (Table 9), 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, with pooled Spearman correlation 0.700 with the held-out deficit. Clique screening

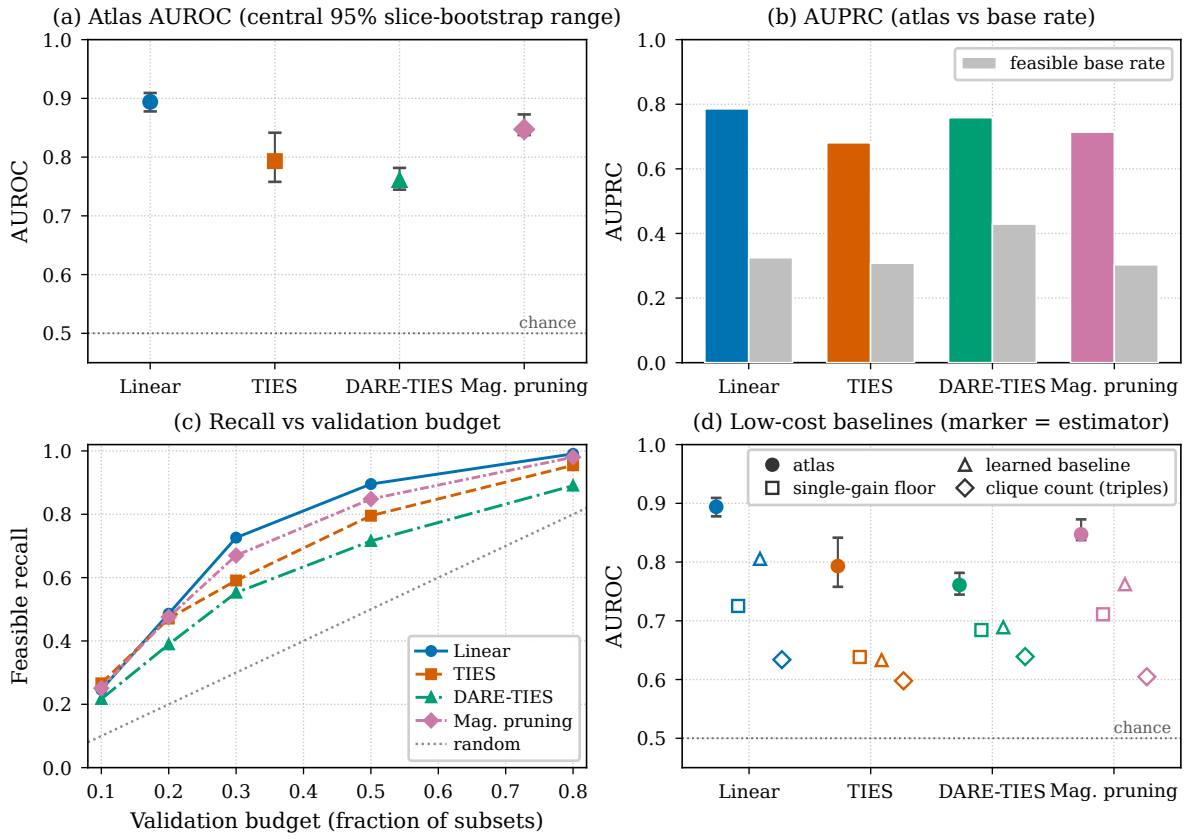

Figure 3: The Qwen m=50 panel (4000 pair and triple events over four predeclared slices). (a) Atlas AUROC per merge operator, with central 95% ranges from a cluster bootstrap over the four slices (1000 draws; with four clusters these are stability ranges rather than calibrated intervals, Appendix B.4). (b) AUPRC per operator, with the feasible base rate for reference. (c) Recall of top-k screening across validation budgets, per operator, with the random-selection line. (d) The atlas versus the low-cost baselines on triples. AUROC is highest for linear merging and lowest for DARE-TIES.

over a fully validated pair table ranks triples at 0.77 to 0.87, near the atlas per operator, but presumes the $\binom{50}{2}$ held-out pair evaluations the other screens do not receive; Section 6 and the hybrid of Appendix B.1 treat that budget explicitly. At a 20% validation budget it recovers 39 to 49% of the pair and triple events that pass held-out validation at 72 to 84% precision per operator (Table 18). Cosine baselines on raw (B, A) and on $\Delta W{=}BA$ both trail the atlas (Table 16; Section 6).

On triples, ranking by the number of pair sub-queries the atlas *predicts* feasible ($\widehat{\rho} \leq 0$), with ties broken by the triple score, improves the 20%-budget operating point from 0.54/0.67 to 0.56/0.69 recall/precision at no extra validation cost (Appendix B.1).

The atlas build cost is amortized across the whole query space. At m=50 a pair table costs $\binom{50}{2}$ full held-out evaluations, about 10 GPU-hours at 30 s each, and then answers only pair queries, whereas the atlas build scores 1,275 design merges on every protected task's probes (nN_d scalar evaluations of task losses with $n{=}m$, at probe cost instead of full-validation cost) and then answers every subset query at every order without further evaluation. Higher-order query volume widens that gap, and the predicted-pair hybrid inherits the same build. The measured wall-clock break-even still depends on probe size and batching, so we do not reduce the two to a single figure.

We train gradient-boosted trees and a small MLP on Qwen m=50 under leave-one-slice-out cross-validation, in three feature regimes that model a fresh library (cold-start), a library whose pairs have been screened (pair-aware), and a fully labeled copy of the target library (full-history); Table 8 in Appendix B.1 specifies the features and the sign-flip stability diagnostics.

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 ∼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%

The atlas outperforms the supervised baselines on new or pair-screened libraries; a GBM overtakes it only in the transductive setting, after receiving adapter identities and feasibility labels from the target library itself. In cold-start, the atlas outperforms the GBM by +0.263 (TIES) to +0.392 (magnitude pruning), and adding the realized pair-feasibility count narrows the gap to +0.223 to +0.330 without flipping it. With the 50-dimensional adapter one-hot the GBM exceeds the atlas by 0.025 (linear) to 0.107 (DARE-TIES). Training-data size plays no role in this break-even, since all three regimes use the same 3000 events per fold. The cold-start and pair-aware comparisons are asymmetric in the inputs the methods receive, since the atlas consumes calibration evaluations of the actual merged model while the GBM and MLP receive only metadata features, so those regimes measure the value of the calibration probes. Granting equal access removes the asymmetry. A GBM and an MLP given the atlas's probe-derived outputs, the predicted deficit and the certificate fields, shift per-operator AUROC by -0.02 to $+0.02$ relative to the atlas rule under the same leave-one-slice-out folds, so the fixed convex rule extracts essentially all the screening signal its probes carry. Replicating the cold-start comparison on the TinyLlama stability sweep of 5,720 queries, under leave-one-cell-out folds over the twenty threshold-and-seed cells, gives atlas 0.93 versus GBM 0.70 on linear (+0.24 AUROC).

Performance is stable over the tested range of both free parameters. A five-point floor sweep over $\{0.10, \ldots, 0.30\}$ on the Qwen linear cell keeps AUROC between 0.916 and 0.97 (Table 19), the panel with pairs evaluated on chart faces (α=0) keeps it at 0.80 to 0.90 across operators, and sweeping the retention threshold so the feasible base rate runs from 9% to 67% leaves pooled AUROC in $[0.79, 0.87]$.

A TinyLlama-v0.3 sweep over 20 full-support cells (five τ thresholds × four seeds, 5720 queries) gives mean AUROC 0.938 (range 0.843 to 0.976) and top-quartile enrichment 3.858 over a 2.41% base rate. In the four τ=0.55 cells, 17 accepted triples are non-cliques in the pair graph, false negatives of the pairwise screening step on subset-specific charts rather than fixed-chart Helly violations.

A DARE-Linear (Yu et al., 2024) re-run on the same Qwen m=50 balanced/circulant slices (four slices, 1000 events total, 14.9% feasibility) reaches atlas AUROC 0.870, between linear and magnitude pruning; the merged results and per-slice manifests are included in the released archive.

5.4 Measured effective dimension

We bound $\sigma\text{-}\operatorname{adim}_\epsilon$ from the top eigenspace of the mean Hessian $\bar{H}$ on two real sub-libraries with m=5 adapters on a shared 4-simplex (d=4), fitting a PSD-projected quadratic Hessian $H_t \in \mathbb{R}^{4\times4}$ per adapter over 70 evaluations on a poised design (N_4=15 required). For each r we compute the *strict max-task* residual of Theorem 3.3,

$$\frac{\max_t \|H_t - \Pi_r H_t \Pi_r\|_F}{\max_t \|H_t\|_F},$$

where Π_r is the rank-r projector onto the top eigenspace of $\bar{H}=n^{-1}\sum_t H_t$, and certify $\sigma\text{-}\operatorname{adim}_\epsilon \le r$ when it falls below ϵ (Figure 4). TinyLlama-v0.3 concentrates 82.8% of $\operatorname{tr}\bar{H}$ in the top direction and 97.7% in the top two, certifying $\sigma\text{-}\operatorname{adim}_{0.20} \le 2$, while at ϵ=0.10 no rank below the chart dimension passes the strict residual; the Qwen five-adapter sample concentrates 70.4% and 91.3%, certifying ≤ 3 and ≤ 2, with affine residual 8.00 at r=3, versus the tolerance $\epsilon L_F D^2 \approx 112$ for $L_F = \max_t \|H_t\|_F$ and $D = \operatorname{diam}(C)$. Across the full library the same diagnostic on 20 random five-adapter cores gives median $\sigma\text{-}\operatorname{adim}_{0.20}$=2 ($r$=2 in 11 cores, r=1 in 3,

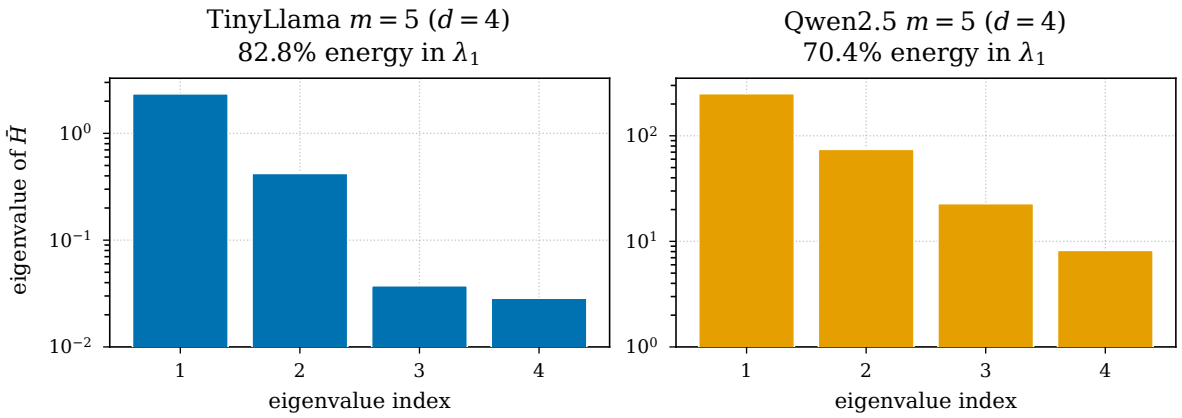

Figure 4: Eigenvalue spectrum of the mean Hessian $\bar{H}$ on the two m=5, d=4 sub-libraries (log scale). TinyLlama (chat and creative adapters) concentrates 82.8% of $\operatorname{tr} \bar{H}$ in the top eigenvalue and certifies $\sigma\text{-}\operatorname{adim}_{0.20} \leq 2$; the Qwen five-adapter sample concentrates 70.4% and 91.3% in the top one and two eigenvalues, certifying $\sigma\text{-}\operatorname{adim}_{0.10} \leq 3$ and $\sigma\text{-}\operatorname{adim}_{0.20} \leq 2$.

r=3 in 6), so the rank-2 proxy gives three-task approximate cores on the typical five-adapter chart, with the stated residual slack; each core is fitted on its own locally chosen chart and projector.

On a single shared chart the picture holds at m=8, where eight adapters on a single 7-simplex, with one projector from the mean Hessian, certify $\sigma\text{-}\operatorname{adim}_{0.20} \leq 2$ and $\sigma\text{-}\operatorname{adim}_{0.10} \leq 3$ across the sub-library (strict-max residuals 0.14 and 0.09), and the (r+1)-core tables from the same models reproduce all 126 four- and five-way scores exactly at r=3; the certified reduction stays per-core (Section 7). The more heterogeneous panels have larger certified ranks. Eight TinyLlama adapters with diverse tasks give $\sigma\text{-}\operatorname{adim}_{0.20}$=5 (residual 0.40 at r=3) with core-table errors below 10^{-3} at r=2, and an m=10 TinyLlama sub-library first falls below 0.20 at r=4 and 0.10 at r=6, still below the worst-case d+1=10. Threshold decisions are preserved on both shared charts, but vacuously, since every four- and five-way query there is feasible for the fitted surrogate (Section 7); sweeping the threshold across the score range makes the test discriminative, with full and reduced-core decisions agreeing at every threshold the interval certificate clears and about half the time inside its uncertified band.

At m=50 the certification is negative. Fitting all fifty adapters on one shared 49-simplex (1,275 singles-and-pairs design evaluations) with one projector from the mean Hessian leaves the strict max-task residual at 0.48 for r=3 and 0.29 at r=10, first crossing 0.20 at r=23 and 0.10 at r=42, so no small shared rank is certified at library scale. The certified rank rises with heterogeneity across these shared charts, from 2 on the homogeneous m=8 chart, to 4 and 5 on the more heterogeneous m=10 and task-diverse m=8 charts, to 6 on the sixteen-task Mistral chart, whose strict residual first falls below 0.20 at r=6, to 23 at m=50. Library size alone does not drive this growth: on the task-homogeneous Mistral classification substrate the same diagnostic certifies rank 6 at m=30 and rank 5 at m=50 (first crossings of the 0.20 residual on the ambient charts of Section 5.1), flat from sixteen adapters through fifty, so the rank-23 requirement on the Qwen chart tracks task heterogeneity rather than adapter count.

Clustering the adapters by the affinity of their curvature subspaces does not lower the shared rank. The adapters' full-dimensional curvatures share no common low-rank subspace, and partitioning the fifty adapters into equal-size charts certifies no materially lower sub-block rank than a random partition (half a rank lower at size twenty-five, none elsewhere), with median sub-block rank 4 at size five rising to about 12 at size twenty-five. These sub-blocks reuse the shared library Hessians rather than locally refit models, which is why their medians exceed the locally refit median of 2. A partition also covers only 6 to 37% of the library's subset queries inside one chart.

Table 6: 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

Table 7: Public PEFT panel summary. Public model and adapter cards reported Apache-2.0 or MIT-compatible licenses when checked; repository identifiers are recorded in the shipped configs where available, and weights/adapters are not redistributed. Qwen uses four predeclared m=50 balanced/circulant slices; the other base panels enumerate all pairs and triples. Per-operator rows and certificate fields are released in the supplementary CSVs.

Base panel	Events	Feasible	Atlas AUROC range	Atlas AUPRC range
Qwen2.5-0.5B-Instruct (m=50)	4000	1365	0.76–0.89	0.68–0.79
TinyLlama-v0.3 (m=12)	1144	53	0.73–0.95	0.30–0.65
Mistral-7B-Instruct-v0.2 (m=10)	660	439	0.64–0.81	0.66–0.91
Phi-2 (m=10)	660	68	0.78–1.00	0.25–1.00
Pythia-410M (m=10)	660	90	0.57–0.87	0.18–0.52
Pythia-1B (m=10)	660	156	0.62–0.75	0.23–0.69

6 Discussion and related work

Subspace and routing-based mergers. 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 Null-Space Compression (NSC, arXiv preprint; Lee et al., 2026) go further and compress the updates into low-rank or null-space coordinates. Pico (arXiv preprint; Tang and Yang, 2026) and OSRM (Zhang and Zhou, 2025) calibrate the LoRA factor spaces directly, LoRA Soups (Prabhakar et al., 2025) composes adapter checkpoints, and LoraHub (Huang et al., 2024a) searches for composition coefficients on a fixed subset; Fisher merging (Matena and Raffel, 2022) and RegMean (Jin et al., 2023) instead compute the merged weights themselves, from per-parameter Fisher information and from closed-form least squares on layer statistics respectively, again for a supplied subset. The atlas screens the subsets themselves, and its coefficient comes from one convex optimization over the fitted models rather than from black-box search; a practitioner can screen with the atlas first, then merge the subsets that pass with any of these operators.

KnOTS as a screener. An alignment-based screener from KnOTS's SVD step, scoring each event by the energy of its stacked updates outside the top singular component, is at chance on 999 linear Qwen pair and triple events (AUROC 0.500, versus 0.894 for the atlas score on the same events), as the tested alignment score contains no task-loss information. The strongest signal in the concurrent line of work on mergeability prediction, per-adapter gradient alignment, is also at chance (0.50) on the eight-task panel scored on the answer loss, where the atlas scores 0.95 to 0.98 across central retention thresholds on the same 84 subsets; the panel's retention base rate is high, and a different alignment estimator could behave otherwise. Gauge invariance (Theorem 2.9) separately rules out any universally exact rule on raw factor geometry, without by itself predicting chance AUROC on a particular panel.

Concurrent full-model fusers and pairwise graph heuristics. Full-model fusers such as FuseChat (Wan et al., 2025), Model Stock (Jang et al., 2024), EMR-Merging (Huang et al., 2024b), and Twin-Merging (Lu et al., 2024) target the merge itself and do not produce a feasibility score for a subset. At the screening end, heuristics built on the pairwise mergeability graph or on task-vector similarity between pairs (Ilharco et al., 2023; Yadav et al., 2023) have low recall as accept rules, with precision that varies widely across panels, and cannot rank pairs (Table 12). As a budgeted ranker on triples the clique count is strong, and Section 5 folds it into a hybrid with the atlas. Cosine baselines on raw (B, A) or on ΔW trail the atlas by ≥ 25 AUROC points (Table 16). Task grouping for multi-task learning (Fifty et al., 2021) predicts which tasks should train together using gradient affinities, a different problem from screening already-trained adapters for merging. Our supervised baselines (§5) characterize the break-even, in which a 50-dimensional adapter one-hot lets a GBM exceed the atlas by 0.03 to 0.11 AUROC on the same library it was trained on, while the atlas is ahead in the cold-start and pair-aware regimes that match screening on new libraries.

Concurrent mergeability predictors. A concurrent line of work from 2026 asks whether merge success can be predicted before merging. Zhou et al. (2026) regress merge success on interpretable properties and find gradient alignment the strongest signal. Wang et al. (2026) score a model pool from limited unlabeled data, and Rahamim et al. (2026) trace mergeability to base-model knowledge. Cao et al. (2026) bound task-level merging collapse with rate–distortion arguments, and Tang et al. (2026) predict LoRA-update mergeability from early training signals, both at the set level. These methods estimate whether a supplied set will merge, whereas the atlas fits one quadratic model per protected task and then returns, from the fitted models alone, a coefficient certified for the fitted models, together with its supporting adapters and the tasks that block it (Theorems 2.1, 2.4, 2.5; Section 4). Theorem 2.8 limits rules determined only by pair information and does not bound a set-level learned predictor, and we make no matched-information performance comparison with these predictors, whose calibration access and outputs differ from ours.

7 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 5.1 evaluates finite-design versions on six dense Mistral charts, the original three and three over tasks new at fifty adapters, certifying deployed scores over the evaluated 91-point designs, while the larger panels' certificates remain statements about the fitted models. The deployed tests of coefficient selection (Section 5.2) cover one eight-way query and fifteen further queries at orders four to six on the sixteen-adapter chart.

Weak probe targets on Qwen. The generic prompts in the m=50 Qwen run give weak single-adapter gains (Appendix B), so those results measure retention relative to the base on generic prompts; the capability claims rest on the Mistral panels scored on the answer loss, which now reach fifty adapters (Section 5.1). Screening the public adapter library for capability would need a library built from tasks with recoverable answers, which the public library lacks; because the atlas records where each channel's loss is scored, that change is a refit rather than a redesign.

Dependent events. Pair and triple records share adapters and tasks. The adapter-resampling and leave-one-adapter-out analyses (Appendix B.4) measure stability within the observed libraries and make no claim of population-level coverage, since the events form a dependent hypergraph and the leave-one-out folds overlap.

Index scale. The fit-once index is validated to fifty adapters on one base and one merge operator. The Mistral ambient builds hold AUROC 0.94 to 0.95 at thirty and fifty adapters (Section 5.1), but the four-operator evidence at fifty adapters still comes from the local quadratic screen on Qwen, the 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 to meet the 0.20 residual criterion, so direct convex queries remain the default for heterogeneous libraries and $(r{+}1)$-core reductions are certified per-core.

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 along with 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. The strongest current evidence comes from the Mistral atlas, whose one 136-point build at sixteen tasks predicts 120 off-design triples at Pearson 0.96 (interval $[0.75, 0.98]$) and whose one 1,275-point linear-merge build at fifty screens 1,096 pair and triple subsets at AUROC 0.95, and from the fifty-adapter Qwen run across four operators, which returns 45.5% of the retention-passing merges at 78% precision while spending 27% of the validation compute. The dual certificates returned by rejected queries are constructive minimal blocking sets, so the same machinery supports synthesizing the largest jointly retainable sub-library and covering a library with few merged models; we develop this, with further theory, in a companion paper. The main open problem is dynamic maintenance after accepted merges change the base model.

Ethical considerations

The method is intended to reduce validation waste by screening candidate merges before a practitioner spends GPU time 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, with per-adapter identifiers recorded in the per-panel manifests (Appendix B). No human subjects, scraped user data, 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

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.

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

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.

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.

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.

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*, 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*, 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*, 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 *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*, 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, December 2023. `https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/`; model card `https://huggingface.co/microsoft/phi-2`; accessed 2026-05-25.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 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*, 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. GitHub repository `https://github.com/huggingface/peft`; experiments use `peft` version 0.19.1; accessed 2026-05-25.

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

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, Simone Calderara, and Joost van de Weijer. Accurate and efficient low-rank model merging in core space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Proceedings PDF: `https://proceedings.neurips.cc/paper_files/paper/2025/file/5913278289e8de3e8f9ac22b47f3ae06-Paper-Conference.pdf`. arXiv:2509.17786.

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*, 2023.

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: Industry Track*, pages 644–655, 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*, 2025.

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.

Yixuan Tang and Yi Yang. Crowded in B-Space: Calibrating shared directions for LoRA merging. *arXiv preprint arXiv:2604.16826*, 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. DOI: 10.1080/01621459.1927.10502953.

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.

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

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 (Table 15), 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), in line with the drop-and-rescale variation noted in Section 5, 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 (balanced and circulant designs, 637 projected events, bf16), re-scoring each projected event by the unprojected model $\max_t \widetilde{q}_t(\widehat{c})$ at the same certificate solution, in place of the projected score, raises pooled AUROC from 0.839 to 0.879 (+0.040). The gain grows with the size of the projection correction and holds in both slices: DARE-TIES 0.756→0.842 (+0.086, per-slice +0.066 and +0.111) and magnitude pruning +0.012, against a flat linear (−0.008) and TIES (−0.010). The certificate retains the soundness guarantee (Proposition 2.2), 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 $|A| \le d{+}1$, and let $\widehat{c}$ be a matched primal. Not every optimal dual is sparse, but a sparse one always exists: at the saddle point, dual stationarity expresses 0 as a convex combination of active-task gradient terms in $\mathbb{R}^d$ (plus a normal-cone element of C), and Carathéodory's theorem selects an optimal dual supported on at most $d{+}1$ active tasks. 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 $\widehat{\rho}_S(C) = \widetilde{\rho}_S(C)$ at the same primal-dual pair.*

Proof. For the convex projected problem $(\widehat{c}, \widehat{\lambda})$ is an exact saddle of Theorem 2.1 whose dual support A is contained in the active set at $\widehat{c}$ (complementary slackness). The dual objective puts no weight outside A, so $\sum_t \widehat{\lambda}_t \widehat{q}_t = \sum_{t\in A} \widehat{\lambda}_t \widetilde{q}_t$ once $\widetilde{H}_t \succeq 0$ on A, leaving $L(\widehat{\lambda})$ unchanged. The core tasks bind, $\widehat{q}_t(\widehat{c}) = U(\widehat{c})$ for $t \in A$, where $\widetilde{q}_t(\widehat{c}) = \widehat{q}_t(\widehat{c})$; for $t \notin A$, $\widetilde{q}_t(\widehat{c}) \le \widehat{q}_t(\widehat{c}) \le U(\widehat{c})$ by Proposition 2.2. Thus $\max_t \widetilde{q}_t(\widehat{c}) = U(\widehat{c})$. Since $L(\widehat{\lambda}) \le \widetilde{\rho}_S(C) \le \max_t \widetilde{q}_t(\widehat{c})$ by weak duality alone (no convexity of the unprojected models required), the true bracket closes at $(\widehat{c}, \widehat{\lambda})$ and the two scores agree. 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. □

Certificate-margin reliability. Binning the 4,000 events by certificate margin quintile gives feasibility rates $\{8.6, 13.9, 22.5, 48.0, 77.6\}$% from least-feasible to most-feasible quintile, with Wilson intervals (Wilson, 1927); Figure 5 plots the same relation per panel, binning each panel's linear-merge events into six margin buckets. Under leave-one-slice-out Platt scaling (Platt, 2000) the score calibrates to a Brier score (Brier, 1950) of 0.164 on the Qwen m=50 panel, 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. The retained score is a language-model loss on the task probes, which the certificates fit; we test whether it tracks task 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 scored on 40 held-out items, the input-prompt loss is flat and does not track accuracy (Pearson +0.43), while the same loss restricted to the answer token tracks it (Pearson −0.96), separating higher-accuracy pairs (≈ 0.8) from lower-accuracy triples (≈ 0.55). The probe's data and fitting script are in the released archive.

The five-task panel replicates this pattern. Five classification tasks from the same released Mistral panel (tasks 190, 280, 290, 391, 1391; 40 held-out items each containing a mix of labels; adapter gains over the base of +0.20 to +0.97 accuracy) give 45 deduplicated pair and triple merges with base and single-adapter reference rows per task. Pooled over merges, $r(L_{\text{answer}}, \text{acc}) = -0.72$ (−0.70 after within-task standardization) versus $r(L_{\text{input}}, \text{acc}) = +0.37$; per task, −0.96 (task280), −0.95 (task1391), −0.89 (task190), −0.55 (task290, whose merge accuracies span only 0.075), and −0.15 (task391, the smallest gain at +0.20).

B Protocols and additional data

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, so the chart coordinate parameterizes this operator rather than the convex path $\sum_i c_i\, \Delta W_i$ through the individual updates (the exact weighted sum is `combination_type=cat`).

Every fit and every validation merge in a panel use the same operator, so all fidelity and screening 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 probe items, so those numbers measure generalization across coefficient settings, while the deployed intervention runs (Section 5.2) evaluate on held-out instances disjoint from the fit split. The Qwen panels' held-out scores come from validation prompts separate from the calibration probes throughout.

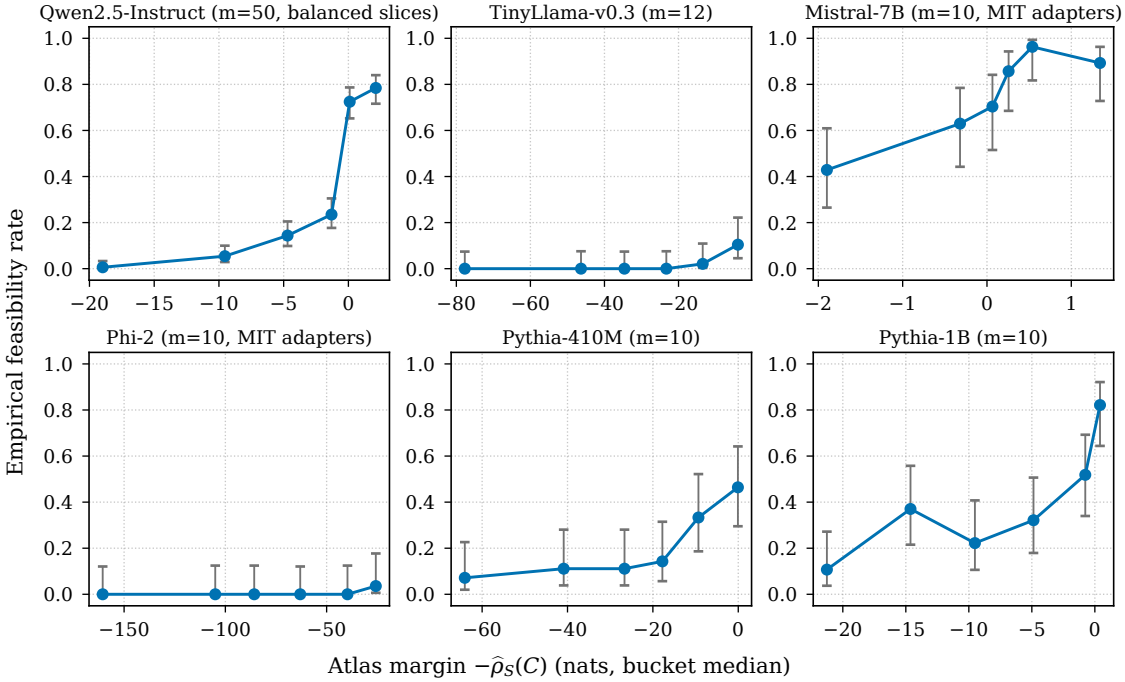

Figure 5: Observed feasibility rate within six certificate-margin buckets per panel, for linear merges, with 95% Wilson intervals and bucket medians on the horizontal axis. The broadly increasing curves show that a larger certificate margin predicts a higher held-out feasibility rate; buckets hold 166 to 167 events on Qwen and 27 to 48 on the smaller panels.

B.1 Qwen baselines and clustered uncertainty

We compare the atlas score with low-cost geometry-free screens (Table 9), with gradient-boosted trees and a small MLP under three feature regimes (Table 8), and with clustered uncertainty diagnostics (Table 10); Tables 11 to 13 carry the same diagnostics across the public panels.

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. Whether this comparison charges both methods the same validation budget depends on where the pair-feasibility signal comes from. Using the atlas's own *predictions* ($\widehat{\rho} \le 0$ on each pair) costs no extra validation, and the resulting hybrid improves on the atlas score alone at every budget on the 4,400 pooled public triples (1,080 feasible): 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 the *held-out* pair labels instead gives 0.34/0.84, 0.61/0.75, 0.78/0.64, 0.92/0.45 (with the hybrid's deterministic atlas tie-break; the pair-graph row of Table 14 breaks ties in expectation, which shifts its entries slightly), but that count presumes a fully validated pair table (as does Table 14, whose pair-graph row uses held-out labels and whose budget therefore excludes pair-table construction); we report it only to bound the value of a pre-existing pair table. Both hybrids keep the atlas's coefficients and certificates for every candidate they forward.

Table 8: Supervised baselines on the balanced Qwen m=50 run. GBM = gradient-boosted trees (500 estimators, depth 3); MLP = 64→32 ReLU. Three feature regimes: *cold-start* (arity + operator + slice; 10 features), *pair-aware* (adds the count of held-out-feasible pair sub-queries per triple; presumes a validated pair table), *full-history* (adds a 50-dim adapter one-hot). Leave-one-slice-out cross-validation. The last column drops every event involving each adapter in turn and reports the leave-one-adapter-out median and central 95% range of the paired atlas−GBM AUROC difference (a sensitivity range, not a confidence interval). The sign-flip p_{MC} (10,000 Monte Carlo permutations) is below 10^{-4} in every cell for both the GBM and MLP comparisons; the fifty leave-one-adapter-out folds overlap, so p_{MC} is a consistency diagnostic for the sign of the gap, not a calibrated test. The atlas−MLP ranges and per-cell aggregates are in the released archive.

Feature regime	Operator	Atlas AUROC	GBM AUROC	MLP AUROC	Atlas−GBM (LOO med, 95% range)
cold-start	linear	0.894	0.513	0.485	+0.381 [+0.370, +0.396]
	TIES	0.793	0.530	0.542	+0.264 [+0.252, +0.279]
	DARE-TIES	0.761	0.492	0.530	+0.270 [+0.260, +0.278]
	magnitude prune	0.847	0.455	0.496	+0.392 [+0.384, +0.404]
pair-aware	linear	0.894	0.570	0.514	+0.325 [+0.312, +0.337]
	TIES	0.793	0.570	0.558	+0.226 [+0.212, +0.234]
	DARE-TIES	0.761	0.536	0.518	+0.228 [+0.215, +0.237]
	magnitude prune	0.847	0.517	0.478	+0.331 [+0.322, +0.341]
full-history	linear	0.894	0.919	0.875	−0.025 [−0.029, −0.019]
	TIES	0.793	0.848	0.824	−0.054 [−0.061, −0.048]
	DARE-TIES	0.761	0.868	0.829	−0.107 [−0.113, −0.100]
	magnitude prune	0.847	0.900	0.869	−0.053 [−0.058, −0.046]

Table 9: Cheap screening baselines on the balanced Qwen m=50 campaign. Single-gain floor uses only the weakest selected adapter's calibration gain. The learned-cheap baseline is leave-slice-out logistic regression over arity and single-adapter calibration-gain features, excluding atlas scores. The clique count ranks each triple by its number of feasible pair faces under the face-closure pair run; it uses held-out pair labels and so presumes a validated pair table, a budget the other columns do not receive. All rows regenerate from the released events via `scripts/qwen_cheap_baselines.py`.

operator	atlas AUROC	single-gain floor	learned baseline	clique count (triples)	random
Linear	0.894	0.725	0.805	0.872	0.500
TIES	0.793	0.638	0.625	0.772	0.500
DARE-TIES	0.761	0.684	0.687	0.774	0.500
Mag. pruning	0.847	0.711	0.759	0.834	0.500

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 3-adapter subsets (intrinsic affine dimension 2, ambient $\mathbb{R}^3$) with no coefficient floor, so each pair sub-query $\{i, j\}$ is the c_k=0 face of the same chart. This makes the probe one-sided: pair feasibility certified on a face carries over to the full chart, so each recorded event is a valid fixed-chart Helly witness, but pair feasibility attained only in the interior of Δ^2 is not tested, and the recorded count is a lower bound on the number of fixed-chart Helly events among these triples. The probe runs on 146 unique triples from the Qwen m=50 linear merge panel; the run produced 292 raw records across paired slices, and the released aggregate deduplicates by adapter key. For each candidate, all three pair sub-queries and the full triple are evaluated on the same Δ^2 chart and its faces. The probe records the single fixed-chart Helly event reported in Section 5; Table 17 gives the deduplicated breakdown (48 triples with all four sub-queries feasible, 66 with all four infeasible, 31 mixed, 1 Helly event).

Table 10: Clustered uncertainty diagnostics for the balanced Qwen m=50 campaign. Slice intervals bootstrap the four predeclared slices; adapter-block intervals repeatedly keep 80% of adapters and recompute metrics on events fully contained in the kept block.

operator	slice AUROC 95%	slice AUPRC 95%	adapter-block AUROC 95%	adapter-block AUPRC 95%
Linear	[0.878, 0.909]	[0.758, 0.810]	[0.866, 0.921]	[0.650, 0.837]
TIES	[0.758, 0.842]	[0.633, 0.729]	[0.746, 0.835]	[0.538, 0.745]
DARE-TIES	[0.745, 0.782]	[0.729, 0.791]	[0.731, 0.806]	[0.681, 0.814]
Mag. pruning	[0.838, 0.873]	[0.702, 0.741]	[0.808, 0.886]	[0.539, 0.779]

Table 11: Multi-base-panel baselines. Random has AUROC 0.50 and AUPRC equal to the feasible rate. The single-gain screen ranks by the weakest calibration single-adapter gain in the subset. The learned feature baseline is blocked by slice when available and otherwise by deterministic hash folds; it uses only subset size and calibration single-gain summaries (deg. = degenerate slice: one class only).

Base panel	Operator	Atlas ROC	Atlas PR	Single ROC	Single PR	Learned ROC	Learned PR
Qwen2.5-0.5B-Instruct	Linear	0.89	0.79	0.73	0.58	0.80	0.71
	TIES	0.79	0.68	0.64	0.42	0.65	0.42
	DARE-TIES	0.76	0.76	0.68	0.59	0.70	0.62
	Mag. pruning	0.85	0.71	0.71	0.53	0.76	0.62
TinyLlama-v0.3	Linear	0.92	0.31	0.88	0.18	0.90	0.19
	TIES	0.81	0.33	0.75	0.20	0.68	0.14
	DARE-TIES	0.73	0.30	0.84	0.29	0.83	0.30
	Mag. pruning	0.95	0.65	0.94	0.22	0.94	0.27
Mistral-7B-Instruct-v0.2	Linear	0.78	0.90	0.72	0.88	0.75	0.91
	TIES	0.72	0.66	0.60	0.53	0.65	0.53
	DARE-TIES	0.64	0.83	0.55	0.78	0.57	0.77
	Mag. pruning	0.81	0.91	0.74	0.91	0.80	0.94
Phi-2	Linear	1.00	1.00	0.84	0.03	deg.	deg.
	TIES	0.78	0.51	0.65	0.47	0.66	0.45
	DARE-TIES	0.81	0.25	0.73	0.16	0.85	0.38
	Mag. pruning	1.00	0.91	0.65	0.09	0.87	0.21
Pythia-410M	Linear	0.74	0.41	0.74	0.42	0.75	0.53
	TIES	0.76	0.18	0.51	0.13	0.51	0.08
	DARE-TIES	0.57	0.31	0.59	0.30	0.64	0.30
	Mag. pruning	0.87	0.52	0.56	0.21	0.58	0.17
Pythia-1B	Linear	0.75	0.69	0.79	0.70	0.78	0.68
	TIES	0.62	0.23	0.38	0.10	0.75	0.26
	DARE-TIES	0.73	0.46	0.71	0.34	0.78	0.47
	Mag. pruning	0.64	0.45	0.59	0.41	0.69	0.46

B.3 Operating points and practitioner decision table

Table 18 reports the screening trade-off of Figure 3(c) as a validation-budget percentile, the recall of held-out feasible subsets, and an absolute false-negative count, in units a practitioner can use to plan a validation budget. 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 the per-query screen (quadratic models fitted at the N_d poised design points plus solving the certificate problem), about 2 s versus 30 s of full validation per query, or roughly 7%; the savings in Table 5 include it.

B.4 Leave-one-adapter-out stability diagnostics

As an adapter-block diagnostic (Table 20), recomputing AUROC and AUPRC 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 for all four operators and the per-operator AUROC range spans 0.021 to 0.043, so no single adapter accounts for the screening signal.

Table 12: Pair-graph baseline on triple queries. The score is the fraction of evaluated pair faces that are feasible; the clique decision accepts only triples whose three pair faces are feasible (deg. = degenerate slice: one class only).

Base panel	Operator	triples	graph AUROC	precision	recall	missing faces
Qwen2.5-0.5B-Instruct	Linear	400	0.90	0.97	0.27	0
	TIES	400	0.81	0.64	0.24	0
	DARE-TIES	400	0.77	0.81	0.33	0
	Mag. pruning	400	0.85	0.90	0.22	0
TinyLlama-v0.3	Linear	220	0.96	0.00	0.00	0
	TIES	220	0.84	1.00	0.06	0
	DARE-TIES	220	0.92	0.67	0.17	0
	Mag. pruning	220	0.96	0.00	0.00	0
Mistral-7B-Instruct-v0.2	Linear	120	0.88	0.97	0.71	0
	TIES	120	0.85	0.74	0.61	0
	DARE-TIES	120	0.64	0.79	0.58	0
	Mag. pruning	120	0.92	0.99	0.72	0
Phi-2	Linear	120	deg.	0.00	0.00	0
	TIES	120	0.58	1.00	0.02	0
	DARE-TIES	120	0.86	0.00	0.00	0
	Mag. pruning	120	1.00	0.00	0.00	0
Pythia-410M	Linear	120	0.77	0.50	0.04	0
	TIES	120	0.73	0.00	0.00	0
	DARE-TIES	120	0.57	0.00	0.00	0
	Mag. pruning	120	0.78	1.00	0.10	0
Pythia-1B	Linear	120	0.82	0.89	0.36	0
	TIES	120	0.86	0.25	0.12	0
	DARE-TIES	120	0.72	0.18	0.22	0
	Mag. pruning	120	0.74	0.62	0.30	0

Adapter-resampling bootstrap. Following the pigeonhole idea of Owen (2007), we resample the adapters instead of the events, drawing adapter counts from a multinomial over the fifty-adapter pool, weighting every pair or triple event by the product of its adapters' counts, and recomputing each statistic. With B=2,000 replicates and a fixed seed, the 95% percentile AUROC intervals on the Qwen m=50 panel are linear $[0.821, 0.946]$, magnitude pruning $[0.758, 0.918]$, TIES $[0.691, 0.883]$, DARE-TIES $[0.673, 0.835]$, and pooled $[0.748, 0.884]$, and the pooled 20%-budget operating point spans recall $[34.2, 59.6]$% and precision $[60.8, 91.3]$%; every interval excludes 0.5. Owen's coverage guarantees are for crossed random-effects designs, so these are a stability analysis over the adapter population, not calibrated coverage for the pair/triple hypergraph.

Public PEFT protocol. The released archive includes the runner, manifests, and analysis scripts, and every number in the appendix tables is reproduced by running the archive's entry point, `reproduce.py`, with the corresponding panel's configuration. A small smoke configuration, `configs/smoke.json`, pairs the public TinyLlama base with three of the panel's adapters, for conversation, mathematical reasoning, and SQL generation.

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 manifests of the non-Qwen panels enumerate every pair and triple per merge operator with minimum per-adapter coefficient 0.2.

All runs use PyTorch on ROCm/CUDA-class accelerators with 4-bit or 8-bit loading where needed; the released runs used a single AMD MI300X with ROCm 7.0, and Table 21 lists each panel's base model, license, and event count.

All runs use a single fixed random seed for the calibration/held-out split, declared in each manifest, and we take the split before computing any feasibility labels. A query is labeled held-out feasible when some

Table 13: Leave-one-adapter-out clustered diagnostic for the atlas score. Each entry is the 2.5/50/97.5 percentile of AUROC after dropping all events that contain one adapter.

Base panel	Operator	LOO AUROC percentiles
Qwen2.5-0.5B-Instruct	Linear	0.89/0.89/0.90
	TIES	0.78/0.79/0.80
	DARE-TIES	0.75/0.76/0.77
	Mag. pruning	0.84/0.85/0.86
TinyLlama-v0.3	Linear	0.90/0.93/0.99
	TIES	0.78/0.81/0.84
	DARE-TIES	0.67/0.73/0.81
	Mag. pruning	0.94/0.95/1.00
Mistral-7B-Instruct-v0.2	Linear	0.69/0.78/0.84
	TIES	0.68/0.71/0.81
	DARE-TIES	0.61/0.64/0.68
	Mag. pruning	0.66/0.81/0.85
Phi-2	Linear	1.00/1.00/1.00
	TIES	0.72/0.76/0.84
	DARE-TIES	0.65/0.77/0.87
	Mag. pruning	0.99/0.99/1.00
Pythia-410M	Linear	0.68/0.73/0.79
	TIES	0.71/0.73/0.89
	DARE-TIES	0.50/0.57/0.65
	Mag. pruning	0.81/0.87/0.92
Pythia-1B	Linear	0.70/0.74/0.80
	TIES	0.58/0.64/0.71
	DARE-TIES	0.70/0.74/0.85
	Mag. pruning	0.61/0.64/0.74

Table 14: Validation-budget yield on public triple queries. At each budget, entries are recall/precision among evaluated triples after sorting by the indicated screen; random is its expectation under uniform sampling. The pair-graph row ranks by the count of *held-out-feasible* pair sub-queries and so presumes a validated pair table; its budget excludes that pair-table construction. Section B.1 gives a hybrid that ranks by *atlas-predicted* pair feasibility instead and so needs no pair validation.

Screen	n (n_+)	10%	20%	30%	50%
Atlas	4400 (1080)	0.29/0.71	0.54/0.67	0.66/0.54	0.82/0.40
Learned baseline	4280 (1080)	0.21/0.54	0.37/0.46	0.51/0.43	0.75/0.38
Single-gain	4400 (1080)	0.22/0.54	0.41/0.51	0.53/0.43	0.80/0.39
Pair graph	4400 (1080)	0.33/0.80	0.57/0.70	0.78/0.64	0.91/0.45
Random	4400 (1080)	0.10/0.25	0.20/0.25	0.30/0.25	0.50/0.25

coefficient in its evaluated design satisfies the retention rule below on the held-out prompts; a coefficient is scored by its worst selected-task deficit. The supplementary README gives the commands.

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

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

The denominator is the single-adapter gain $L_0(t) - L_1(t)$, which must be positive for the normalization to be meaningful. We clamp it below at 0.005 nats; a task whose adapter does not measurably improve on the base then contributes the criterion $L_c(t) \lesssim L_0(t)$, the merge must not degrade the base loss, rather than a division by a small or negative number. This clamp is often active on the Qwen panel, whose calibration probes are generic instruction prompts, so the single-adapter gains there are small (median -0.02 nats, 43 of 50 below the floor; base and single-adapter losses are in the released summaries) and for most Qwen tasks the label reduces to holding the merge near the 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 over the fitted chart, and label feasibility on held-out prompts from the evaluated

Table 15: PSD-projection stress test. Full reports atlas AUROC on all rows; low-75% drops the largest quartile of PSD corrections within each base panel; high-25% reports the largest-correction quartile alone.

Base panel	rows	proj. rate	full ROC	low-75% ROC	high-25% ROC
Qwen2.5-0.5B-Instruct	4000	0.65	0.82	0.87	0.73
TinyLlama-v0.3	1144	0.63	0.83	0.91	0.71
Mistral-7B-Instruct-v0.2	660	0.81	0.72	0.72	0.64
Phi-2	660	0.75	0.90	0.94	0.85
Pythia-410M	660	0.91	0.72	0.75	0.70
Pythia-1B	660	0.93	0.67	0.73	0.51

Table 16: Cosine baselines: per-operator AUROC for the atlas, mean-pairwise cosine, min-pairwise cosine, and random on the cached comparison slices. For each cosine baseline we report $\max(\mathrm{AUROC}(\cos), \mathrm{AUROC}(-\cos))$. The atlas exceeds both cosine baselines by at least 25 AUROC points on the cached Qwen comparison slice and TinyLlama-v0.3.

Base	Operator	n (n_+)	Atlas AUROC	Cosine AUROC (mean / min)	Random
TinyLlama-v0.3	linear	286 (8)	**0.87**	0.60 / 0.52	0.50
Qwen2.5-0.5B-Instruct	linear	250 (106)	**0.88**	0.54 / 0.57	0.50
m=50 cached comparison	TIES	250 (143)	**0.88**	0.52 / 0.56	0.50
	DARE-TIES	250 (172)	**0.81**	0.56 / 0.54	0.50
	Mag. pruning	250 (118)	**0.84**	0.55 / 0.51	0.50

coefficient design. Every fresh query returns a coefficient, primal upper value, dual lower value, dual weights, certificate gap, per-task margins, and, when the held-out worst deficit is positive, an obstructing task core. Table 22 reports the certificate and PSD-projection diagnostics per panel and operator.

Stability summary. Across the threshold-and-seed sweep (Section 5), the atlas score accepts 79 pair cells and 59 triple cells under a frozen manifest, with mean AUPRC 0.471; the per-event records are included with the code.

B.5 Maintenance of the index

Maintenance: adding records. Each verdict records the base and adapter versions it was computed against. If maintenance changes each task model by at most d_t on the chart, $\sup_c |q'_t(c) - q_t(c)| \le d_t$, then $|\rho'(T;U,\ell) - \rho(T;U,\ell)| \le \max_{t\in T} d_t$ by the 1-Lipschitz dependence of the minimax on the models, which is the formal content of the retention rule in Section 4. Insertion has three cases. (a) A new adapter adds a coordinate: on the full simplex each existing model acquires $m{+}1$ new coefficients (one linear, one square, $m{-}1$ cross terms), all fixed by the same $m{+}1$ insertion evaluations scored on the existing protected tasks' probes, so the models are extended and only their PSD projections recomputed; replaying the build one adapter at a time moves the shared models by a median 0.16 nats per added coordinate, measured on the released index-update records, and on a chart of fixed dimension the coordinate count does not grow. (b) A new protected task adds no coordinate: it fits one new model by scoring that task on the N_d existing design points, with no new merge and no other model touched. (c) A new adapter and a new task together add the coordinate and then fit the new task's model on the enlarged design.

Maintenance: changing the base. 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 (released with the consolidation records), small relative to the multi-nat gains but of the same size as typical certificate margins; this comparison supplies the index's refit criterion.

Table 17: Fixed-chart Qwen probe on 146 unique triples from the Qwen m=50 linear-merge panel. The script probed 292 raw records across paired slices and the released aggregate deduplicates by adapter key. Each candidate's three pair sub-queries and the full triple are evaluated on the same shared 3-adapter Δ^2 chart (intrinsic affine dimension 2, ambient $\mathbb{R}^3$) with no coefficient floor, so each pair sub-query sits as a face $\{c_k{=}0\}$ of the same Δ^2 chart. "Mixed" rows are triple-infeasible with one or two feasible pair sub-queries. The probe finds one fixed-chart d=2 Helly event: all three pair sub-queries are feasible at τ=0 while the full triple is infeasible. An earlier 50-triple run found no such event.

sample	unique triples	raw records	all 4 compatible	all 4 incompatible	mixed	fixed-chart Helly events
Scaled Qwen probe	146	292	48	66	31	1

Table 18: Operating-point decision table for atlas-driven screening on the balanced Qwen m=50 campaign. Each cell reports feasible-subset recall, precision among validated subsets, and absolute false negatives at the stated validation budget.

operator	10% budget rec / prec / FN	20% budget rec / prec / FN	30% budget rec / prec / FN	50% budget rec / prec / FN	80% budget rec / prec / FN
Linear (n_+=325)	25% / 80% / 245	49% / 79% / 167	73% / 79% / 89	90% / 58% / 34	99% / 40% / 3
TIES (n_+=308)	27% / 82% / 226	47% / 72% / 163	59% / 61% / 126	80% / 49% / 63	95% / 37% / 14
DARE-TIES (n_+=429)	22% / 93% / 336	39% / 84% / 262	55% / 79% / 192	72% / 61% / 122	89% / 48% / 47
Mag. pruning (n_+=303)	25% / 76% / 227	48% / 72% / 159	67% / 68% / 100	85% / 51% / 46	98% / 37% / 6

C Additional theory

C.1 Effective dimension and estimation

Let $\mathcal{L} = \{H_1, \dots, H_n\} \subset \mathbb{R}^{d\times d}$ be the per-task PSD Hessians, one per protected task. Define the *spectral effective dimension at residual* ϵ (the smallest number of directions needed to hold every task's Hessian to within an ϵ fraction of its size) as

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

and let $D = \operatorname{diam}(C)$, $L_F = \max_t \|H_t\|_F$.

Remark C.1 (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 $\max_t \|H_t - \Pi_r H_t \Pi_r\|_F \le \epsilon L_F$ gives a computable upper bound on $\sigma\text{-}\operatorname{adim}_\epsilon(\mathcal{L})$ and, by Theorem 3.3, an obstruction order at most $r+1$ with additive certificate slack $\epsilon L_F D^2$ (Frobenius-residual definition; a non-zero linear term b_t adds $2\max_t \|(I-\Pi_r)b_t\|_2 D$, which vanishes when every linear term is zero). This certifies the residual of one specific projector; it is not a solver for the minimum-rank projector, and Section 5 measures the residual directly.

Estimation. Estimated models $\widehat{q}_t$ with uniform error $\sup_{t,c}|\widehat{q}_t(c) - q_t(c)| \le \eta$ propagate linearly, so $|\widehat{\rho}_S - \rho_S| \le \eta$ for every S and the two atlas filtrations are η-interleaved (Chazal et al., 2016). Fitting one quadratic model uses $N_d = 1 + d + d(d+1)/2$ coefficient-design points; reaching sup-norm precision η under poised quadratic interpolation needs $N = O\big(\kappa^2 B^2 \eta^{-2} \log(nN_d/\delta_0)\big)$ example evaluations per task at each design point, at failure probability δ_0, hence nN_dN scalar task-example evaluations across the n protected tasks, with κ the conditioning constant of the poised design (Appendix D.9). Any accept/reject decision, and any persistent feature with margin larger than η, is then certified (Appendix D.9).

C.2 Gauge invariance and the impossibility of raw-coordinate rules

The ellipsoidal sublevel sets depend on the task metric, which already rules out raw task-vector cosine as a universal rule; Theorem 2.9 (stated in Section 2) shows that the sublevel-set family and threshold complex of Eq. (5) are exactly the gauge-invariant content of the fitted local models.

Table 19: Sensitivity sweep of the coefficient floor α on Qwen2.5-0.5B-Instruct, linear merge, leave-slice-out evaluation. The minimum per-adapter coefficient floor α defines the subset-specific simplex $\Delta^{|S|-1} \cap \{c_i \geq \alpha\}$. AUROC is computed against the held-out retention feasibility label.

α	n	pairs	triples	feasible rate	atlas AUROC
0.10	200	120	80	0.255	0.918
0.15	200	120	80	0.200	0.922
0.20	200	120	80	0.185	0.916
0.25	200	120	80	0.130	0.966
0.30	200	120	80	0.095	0.943

Table 20: Leave-one-adapter-out clustered diagnostics on the balanced Qwen m=50 campaign. For each adapter, AUROC/AUPRC are recomputed after dropping every event involving that adapter.

operator	full AUROC	LOO AUROC [min, median, max]	full AUPRC	LOO AUPRC [min, median, max]
Linear	0.894	$[0.881, 0.894, 0.902]$	0.786	$[0.690, 0.787, 0.803]$
TIES	0.793	$[0.759, 0.795, 0.803]$	0.681	$[0.562, 0.682, 0.693]$
DARE-TIES	0.761	$[0.742, 0.762, 0.772]$	0.758	$[0.709, 0.761, 0.771]$
Mag. pruning	0.847	$[0.821, 0.849, 0.857]$	0.714	$[0.580, 0.715, 0.735]$

D Proofs and additional details

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 $(c, \lambda) \mapsto F(c, \lambda)$ is continuous, convex in c, and affine (hence concave) 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, $U(c) \leq \tau$ 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 2.4

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

Lower bound ($\geq$). Fix any $r < \rho_S$. Then $\bigcap_{t \in S} \{c \in C : q_t(c) \leq r\} = \varnothing$ by definition of ρ_S. Each sublevel set is a convex subset of $C \subseteq \mathbb{R}^d$. By Helly's theorem there is some $T \subseteq S$ with $|T| \leq d + 1$ already having empty intersection at level r, so $\rho_T > r$. Letting $r \uparrow \rho_S$ gives $\max_{|T| \leq d+1} \rho_T \geq \rho_S$. Combined with monotonicity,

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

Tightness. We construct the witness inside a compact chart, matching the class of sublevel sets used by the fitted task models. Place d+1 vertices $v_0, \ldots, v_d$ at the vertices of a regular d-simplex on the unit sphere in $\mathbb{R}^d$ (so the Chebyshev center of the full simplex is the origin and the Chebyshev center of any d-vertex face is its circumcenter at distance $1/d$). Choose any compact convex chart $C \subset \mathbb{R}^d$ containing the unit ball, e.g. $C = [-2, 2]^d$; this contains every vertex and every face circumcenter. Set $F_t = C \cap \{c : \|c - v_t\|_2^2 \leq R_0^2\}$ with $\sqrt{1 - 1/d^2} < R_0 < 1$. Each F_t is the intersection of C with a centered-quadratic sublevel set, hence compact convex. Every d-subset of $\{F_0, \ldots, F_d\}$ has non-empty intersection inside C (the corresponding face circumcenter is in C and within distance R_0 of each face vertex), but $\bigcap_{t=0}^{d} F_t = \varnothing$ since no point in $\mathbb{R}^d$ is within distance $R_0 < 1$ of all d+1 vertices simultaneously. Table 4 reports an analytic and numerical realization at d=2, 3, 4. Pairwise insufficiency is the d=2 instance, embedded into higher dimensions by product with a compact ball.

Table 21: Reproducibility manifest summary: the base model, license, library size, and event count for each panel. Per-adapter Hugging Face repository identifiers and configurations are in the released archive.

Panel	Base model (citation)	License family	m	Events
Qwen m=50	Qwen2.5-0.5B-Instruct (Qwen Team, 2024)	Apache 2.0 (adapter mix Apache/MIT)	50	4000
TinyLlama m=12	TinyLlama-v0.3 (Zhang et al., 2024)	Apache 2.0	12	full enumeration
Mistral m=10	Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	Apache 2.0	10	full enumeration
Phi-2 m=10	Phi-2 (Javaheripi and Bubeck, 2023)	MIT	10	full enumeration
Pythia-410M m=10	Pythia-410M (Biderman et al., 2023)	Apache 2.0	10	full enumeration
Pythia-1B m=10	Pythia-1B (Biderman et al., 2023)	Apache 2.0	10	full enumeration
TinyLlama stability sweep	TinyLlama-v0.3 (Zhang et al., 2024)	Apache 2.0	12	5,720 queries ($5\tau \times 4$ seeds)
Qwen Δ^2 Helly probe	Qwen2.5-0.5B-Instruct (Qwen Team, 2024)	Apache 2.0	50	146 unique triples (§B.2)
Effective rank diagnostic (TinyLlama)	TinyLlama-v0.3 (Zhang et al., 2024)	Apache 2.0	5	70 design points
Effective rank diagnostic (Qwen)	Qwen2.5-0.5B-Instruct (Qwen Team, 2024)	Apache 2.0	5	70 design points

D.3 Proof of Theorem 2.5

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 $w^\star = \sum_{i\in U} z_i^\star a_i \in \operatorname{conv}\{a_i : i \in U\}$.

Task core. Write $k_C = \dim \operatorname{aff} R\,C(U,\ell)$ for the dimension of the chart's latent image, so $k_C = k_U$ when $s_\ell > 0$ and $k_C = 0$ when $s_\ell = 0$. Each sublevel set $\{c \in C(U,\ell) : q_t(c) \le r\}$ is the preimage under the affine map $c \mapsto Rc$ of a convex set inside the k_C-dimensional affine image $R\,C(U,\ell)$. Applying Helly in that image (Theorem 2.4 with ambient dimension k_C) gives $A^\star \subseteq T$, $1 \le |A^\star| \le k_C{+}1$, 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): if $A \subseteq T$ attains $\rho(A;U,\ell) = \rho$ while every $A\setminus\{t\}$ attains strictly less, pick r strictly between $\max_t \rho(A\setminus\{t\};U,\ell)$ and ρ; then A is minimally infeasible at level r, so Helly in the k_C-dimensional image forces $|A| \le k_C{+}1$.

Adapter core. The hull $\operatorname{conv}\{a_i : i \in U\}$ lies in an affine space of dimension k_U, so Carathéodory's theorem writes $w^\star = \sum_{i\in V^\star} z'_i a_i$ with $z' \in \Delta_{V^\star}$ and $|V^\star| \le k_U{+}1$. Put $c' = \ell + s_\ell \sum_{i\in V^\star} z'_i v_i$ (so $c' \in C_\ell(V^\star)$). Then $Rc' = R\ell + s_\ell w^\star = Rc^\star$, so $q_t(c') = \tilde q_t(Rc') = \tilde q_t(Rc^\star) = q_t(c^\star)$ for every t; hence c' is also a minimizer and $\{i : c'_i > \ell_i\} \subseteq V^\star$ has at most $k_U{+}1$ elements. Since $C_\ell(V^\star) \subseteq C_\ell(U) = C(U,\ell)$ (the free mass is spread over a subset), $\rho(T;V^\star,\ell) \ge \rho(T;U,\ell)$; but c' attains $\max_{t\in T} q_t(c') = \rho(T;U,\ell)$ inside $C_\ell(V^\star)$, giving equality. The complete support is $\operatorname{supp}\ell \cup V^\star$.

Discrete saddle. Task monotonicity ($A \subseteq T \Rightarrow \rho(A;V,\ell) \le \rho(T;V,\ell)$) and adapter antitonicity ($V \subseteq U \Rightarrow \rho(A;V,\ell) \ge \rho(A;U,\ell)$, since $C_\ell(V) \subseteq C_\ell(U)$), together with $\rho(A^\star;U,\ell) = \rho(T;U,\ell) = \rho(T;V^\star,\ell) = \rho$, give

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

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

D.4 Proof of Theorem 2.6

The stated properties of the regular r-simplex are standard: unit vectors $v_0,\dots,v_r$ summing to zero satisfy $0 = \|\sum_i v_i\|^2 = (r{+}1) + (r{+}1)r\,\gamma$, where $\gamma = \langle v_i, v_{i'}\rangle$ is the common inner product of distinct vertices, so $\gamma = -1/r$. Throughout, c ranges over the relevant chart, $x := Rc$ over its latent image (so x ranges over $\operatorname{conv}\{v_0,\dots,v_r\}$ as c ranges over $\Delta_{[m]}$), and $q_t(c) = \|x - v_t\|^2$.

Full score. For any c, write $x = Rc$; averaging the $r{+}1$ centers gives

$$\max_t \|x - v_t\|^2 \ge \tfrac{1}{r+1}\textstyle\sum_t \|x - v_t\|^2 = \|x\|^2 - 2\big\langle x, \tfrac{1}{r+1}\sum_t v_t\big\rangle + \tfrac{1}{r+1}\sum_t \|v_t\|^2 = \|x\|^2 + 1,$$

using $\sum_t v_t = 0$ and $\|v_t\| = 1$. Thus $\max_t \|x - v_t\|^2 \ge \|x\|^2 + 1 \ge 1$, and equality in the last step forces $x = 0$; the coefficient barycenter $c = \frac{1}{r+1}\sum_i e_i \in C$ has $x = Rc = 0$ and gives every $\|x - v_t\|^2 = 1$, so it attains the value. Hence $\rho([n];[m],0) = 1$, with $x = 0$ (the barycenter) the unique minimizer.

Task core tight. Fix j and restrict the tasks to $T = [n] \setminus \{j\}$. The r centers $\{v_i : i \ne j\}$ are the vertices of a regular $(r{-}1)$-simplex whose circumcenter is their barycenter $\frac{1}{r}\sum_{i\ne j} v_i = -v_j/r$, the latent image of the coefficient point with $c_i = 1/r$ for $i \ne j$ and $c_j = 0$. At $x = -v_j/r$, for each $i \ne j$,

$$\|x - v_i\|^2 = \|v_j/r + v_i\|^2 = \tfrac{1}{r^2} + \tfrac{2}{r}\langle v_i, v_j\rangle + 1 = \tfrac{1}{r^2} - \tfrac{2}{r^2} + 1 = 1 - \tfrac{1}{r^2},$$

Table 22: Certificate and convexity diagnostics for the public base-panel evidence. PSD proj. counts rows whose fitted Hessian was projected onto the PSD cone before solving the continuous certificate. Isolated rows with a large max gap (e.g. Pythia-410M linear at 1.2) are treated as inconclusive and are not used as hard accept/reject guarantees; the AUROC analysis uses the surrogate score as a ranker.

Base panel	Operator	cont.	fallback	PSD proj.	min eig.	max PSD adj.	med. gap	max gap
Qwen2.5-0.5B-Instruct	Linear	1000	0	554	-4.0×10^2	4.0×10^2	0	5.3×10^{-6}
	TIES	1000	0	737	-3.8×10^2	3.8×10^2	0	5.0×10^{-7}
	DARE-TIES	1000	0	682	-1.2×10^3	1.2×10^3	0	8.5×10^{-4}
	Mag. pruning	1000	0	637	-4.2×10^2	4.2×10^2	0	1.3×10^{-5}
TinyLlama-v0.3	Linear	286	0	109	-5.5×10^2	5.5×10^2	0	6.9×10^{-9}
	TIES	286	0	241	-4.1×10^2	4.1×10^2	0	1.8×10^{-10}
	DARE-TIES	286	0	236	-2.5×10^3	2.5×10^3	0	7.5×10^{-3}
	Mag. pruning	286	0	136	-3.7×10^2	3.7×10^2	0	5.4×10^{-8}
Mistral-7B-Instruct-v0.2	Linear	165	0	121	-2.7×10^2	2.7×10^2	0	5.5×10^{-11}
	TIES	165	0	143	-2.7×10^2	2.7×10^2	0	9.4×10^{-10}
	DARE-TIES	165	0	137	-1.2×10^3	1.2×10^3	0	1.4×10^{-8}
	Mag. pruning	165	0	132	-2.9×10^2	2.9×10^2	0	3.9×10^{-11}
Phi-2	Linear	165	0	108	-1.5×10^3	1.5×10^3	0	1.5×10^{-8}
	TIES	165	0	130	-9.2×10^2	9.3×10^2	0	3.2×10^{-8}
	DARE-TIES	165	0	134	-2.7×10^3	2.7×10^3	0	8.8×10^{-10}
	Mag. pruning	165	0	120	-7.9×10^2	7.9×10^2	0	5.6×10^{-7}
Pythia-410M	Linear	165	0	149	-2.0×10^3	2.0×10^3	0	1.2
	TIES	165	0	149	-1.5×10^3	1.5×10^3	0	2.2×10^{-2}
	DARE-TIES	165	0	153	-2.5×10^3	2.5×10^3	0	1.2×10^{-6}
	Mag. pruning	165	0	150	-1.9×10^3	1.9×10^3	0	1.2×10^{-4}
Pythia-1B	Linear	165	0	153	-1.0×10^3	1.0×10^3	0	3.5×10^{-8}
	TIES	165	0	157	-6.4×10^2	6.4×10^2	0	6.3×10^{-8}
	DARE-TIES	165	0	155	-7.3×10^2	7.3×10^2	0	3.4×10^{-8}
	Mag. pruning	165	0	149	-5.3×10^2	5.3×10^2	0	4.1×10^{-6}

so $\max_{i \neq j} \|x - v_i\|^2 = 1 - 1/r^2$ and $\rho(T; [m], 0) \leq 1 - 1/r^2$. For the reverse bound, average over $\{v_i : i \neq j\}$, using $\sum_{i \neq j} v_i = -v_j$:

$$\max_{i \neq j} \|x - v_i\|^2 \geq \tfrac{1}{r} \textstyle\sum_{i \neq j} \|x - v_i\|^2 = \|x\|^2 + \tfrac{2}{r}\langle x, v_j \rangle + 1 \geq 1 - \tfrac{1}{r^2},$$

the final step minimizing the quadratic in x at $x = -v_j/r$, which lies in the latent image of the chart. Hence $\rho(T; [m], 0) = 1 - 1/r^2 < 1$: no r-task subquery reaches the full score, and the task core needs all $r{+}1$ tasks.

Adapter core tight. Fix j and restrict the adapters to $U' = [m] \setminus \{j\}$, so the chart is the opposite coefficient face $C(U', 0)$, whose latent image is $\mathrm{conv}\{v_i : i \neq j\}$. Every $c \in C(U', 0)$ has latent image $x = Rc = \sum_{i \neq j} \theta_i v_i$ with $\theta \in \Delta$, whence $\langle x, v_j \rangle = \sum_{i \neq j} \theta_i \langle v_i, v_j \rangle = -1/r$. Therefore

$$\|x - v_j\|^2 = \|x\|^2 - 2\langle x, v_j \rangle + 1 = \|x\|^2 + \tfrac{2}{r} + 1 \geq 1 + \tfrac{2}{r},$$

and since $j \in [n]$, $\max_{t \in [n]} \|x - v_t\|^2 \geq \|x - v_j\|^2 \geq 1 + 2/r$ for every c in the face. Hence $\rho([n]; U', 0) \geq 1 + 2/r > 1$. No convex hull of only r latent atoms $\{v_i : i \neq j\}$ contains the circumcenter 0, whose barycentric representation uses all $r{+}1$ vertices, so every r-adapter restriction is strictly above the full score and the adapter core needs all $r{+}1$ adapters.

Both cores of Theorem 2.5 therefore have order exactly $k_{[m]}{+}1 = r{+}1$ on this instance, and the same $r{+}1$ points witness both bounds simultaneously. □

D.5 Proof of Theorem 3.1

Work in the intrinsic coordinate z of $\mathrm{aff}(C) = c_0 + L$, where Q has orthonormal columns spanning L and $c = c_0 + Qz$. Substituting $c - a_t = (c_0 - a_t) + Qz$ into the anchored model (2) and expanding,

$$\widehat{q}_t(c_0 + Qz) = \kappa_t + \big(Q^\top (b_t + H_t(c_0 - a_t))\big)^\top z + \tfrac{1}{2} z^\top (Q^\top H_t Q) z = \kappa_t + g_t^\top z + \tfrac{1}{2} z^\top \widehat{H}_t z,$$

where $\kappa_t = \alpha_t + b_t^\top(c_0 - a_t) + \frac{1}{2}(c_0 - a_t)^\top H_t(c_0 - a_t)$ is a constant, $g_t = Q^\top(b_t + H_t(c_0 - a_t))$, and $\widehat{H}_t = Q^\top H_t Q \succeq 0$ inherits positive semidefiniteness from H_t. Each restricted model is thus a convex quadratic in z on the compact convex set $K := \{z : c_0 + Qz \in C\} \subset \mathbb{R}^{d_C}$, $d_C := \dim L$, and K has non-empty interior by the relative interior hypothesis.

Constant directions. For a unit $u \in \mathbb{R}^{d_C}$ say $\widehat{q}_t$ is constant along u if $s \mapsto \widehat{q}_t(c_0 + Q(z + su))$ is constant wherever $z, z + su \in K$. Fix an interior $z_0 \in K$; for small s the derivative in s is $g_t^\top u + z_0^\top \widehat{H}_t u + s\, u^\top \widehat{H}_t u$. Constancy forces the s-coefficient $u^\top \widehat{H}_t u = 0$, hence $\widehat{H}_t u = 0$ since $\widehat{H}_t \succeq 0$, and the remaining term then gives $g_t^\top u = 0$; conversely $\widehat{H}_t u = 0$ and $g_t^\top u = 0$ make $\widehat{q}_t$ invariant under every translation by u. The constant directions of task t are therefore exactly $\ker \widehat{H}_t \cap \ker g_t^\top$, and because K has non-empty interior this local characterization is global on $\operatorname{aff}(C)$.

Kernel of M_C. Each summand of $M_C = \sum_t (\widehat{H}_t^2 + g_t g_t^\top)$ is PSD, so $u^\top M_C u = \sum_t (\|\widehat{H}_t u\|^2 + (g_t^\top u)^2)$, using $\ker \widehat{H}_t^2 = \ker \widehat{H}_t$ for the PSD matrix $\widehat{H}_t$. Hence $u \in \ker M_C$ iff $\widehat{H}_t u = 0$ and $g_t^\top u = 0$ for every t, that is iff u is a common constant direction of the whole family on C. Consequently $\dim \ker M_C = d_C - r_{\text{exact}}(C)$, and $\widehat{q}_t$ depends on z only through the orthogonal projection Π_{exact} of z onto $(\ker M_C)^\perp$: setting $\tilde{q}_t(\Pi_{\text{exact}} z) := \widehat{q}_t(c_0 + Qz)$ is well defined because the right-hand side is unchanged when z is shifted within $\ker M_C$, and $\tilde{q}_t$ is convex since $\widehat{q}_t$ is.

Minimality. Suppose $\widehat{q}_t(c_0 + Qz) = \psi_t(\Pi z)$ for every t and $z \in K$, with Π a linear map of rank r'. Every $u \in \ker \Pi$ is then a constant direction of each task, so $\ker \Pi \subseteq \ker M_C$ and $r' = d_C - \dim \ker \Pi \geq d_C - \dim \ker M_C = r_{\text{exact}}(C)$. No factorization of rank below $r_{\text{exact}}(C)$ exists.

Finite core and query dimension. Write $\Pi_{\text{exact}} = WW^\top$ with $W \in \mathbb{R}^{d_C \times r_{\text{exact}}(C)}$ having orthonormal columns; then $\tilde{q}_t$ depends on z only through $W^\top z$ and is convex, so Theorem 3.2 applies to the compact convex chart K with the map $\Pi : z \mapsto W^\top z$ and zero slack, giving $\rho_S(C) = \max_{T \subseteq S,\, |T| \leq r_{\text{exact}}(C)+1} \rho_T(C)$ for every S. In coefficient coordinates $c = c_0 + Qz$ gives $z = Q^\top(c - c_0)$, so the models factor through the composite $R_C := W^\top Q^\top$ via $c \mapsto R_C(c - c_0)$, which realizes the shared linear map R of Theorem 2.5. Its adapter atoms $a_i = R_C v_i = R_C e_i$ ($v_i = e_i$, with the common anchor term $R_C c_0$ absorbed) lie in the $r_{\text{exact}}(C)$-dimensional range of R_C, so $k_U = \dim \operatorname{aff}\{a_i : i \in U\} \leq r_{\text{exact}}(C)$ and both cores of Theorem 2.5 have order at most $r_{\text{exact}}(C) + 1$.

Why the ambient matrix fails. The naive $M = \sum_t (H_t^2 + b_t b_t^\top)$ is formed in ambient coordinates, and its kernel is the set of *ambient* constant directions, which need not meet L. On $C = \{(x, 0) : |x| \leq 1\} \subset \mathbb{R}^2$ with $\operatorname{aff}(C) = c_0 + L$, $c_0 = 0$, $L = \operatorname{span}\{e_1\}$, $Q = e_1$, take the centered quadratic $q(x, y) = y^2$, that is $\alpha = 0$, $b = 0$, $H = \operatorname{diag}(0, 2)$, $a = 0$. Then $\widehat{H} = Q^\top H Q = 0$ and $g = Q^\top(b + H(c_0 - a)) = 0$, so $M_C = 0$ and $r_{\text{exact}}(C) = 0$, matching the fact that $q \equiv 0$ on C. The ambient matrix $M = H^2 + bb^\top = \operatorname{diag}(0, 4)$ has rank 1 and counts the y-direction the chart never enters. □

D.6 Proof of Theorem D.2 (core sufficiency)

Exact fixed-chart threshold decisions are the thresholded values of the low-order score table, which the next corollary reads as birth times.

Corollary D.1 (Scores as birth times)**.** *The score ρ_S is the birth time of simplex S in the thresholded atlas:*

$$K_r = \{S : \rho_S \leq r\}, \qquad \rho_S = \inf\{r : S \in K_r\}.$$

The entire fixed-chart score function is the upper envelope of its low-order table:

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

Thus the threshold decision $\rho_S(C) \leq \tau$ for any subset S is answered by a max-core lookup over its $\leq (d+1)$-cores. Returning an accepting coefficient for S requires one convex minimax optimization over all tasks in S, since the primal witness of the worst core need not satisfy the remaining tasks; cached primal witnesses can be verified against S as a faster path when they exist.

Theorem D.2 (Sufficient statistic for fixed-chart threshold decisions)**.** *Fix $C \subset \mathbb{R}^d$. For every screening rule $\Phi(S, \tau) \in \{0, 1\}$ on the fixed chart, measurable in the fitted models, whose output exactly decides $\rho_S(C) \leq \tau$,*

the output is extensionally determined by the $(d{+}1)$-core score table $\{\rho_T(C)\}_{|T|\le d+1}$ (the implementation may inspect the models further but cannot produce a different exact bit). Equivalently, there exists an exact rule with the same input-output behavior that factors through the $(d{+}1)$-core table $\mathfrak{A}(\mathcal{L},C)=(\{\rho_T(C), c_T^*, \lambda_T^*\}_{|T|\le d+1})$. The numeric pair-score table $\{\rho_{\{i,j\}}(C)\}$ alone does not suffice (Proposition D.3).

Proof of Theorem D.2. By Theorem 2.4,

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

Hence any exact threshold rule satisfies

$$\begin{aligned}\Phi(S,\tau) &= \mathbf{1}[\rho_S(C)\le\tau]\\ &= \mathbf{1}\left[\max_{T\subseteq S,\,|T|\le d+1}\rho_T(C)\le\tau\right],\end{aligned}$$

so the rule depends on the library only through the $(d{+}1)$-core score table. The primal-dual witnesses stored in $\mathfrak{A}(\mathcal{L},C)$ are not needed to decide the bit, but Theorem 2.1 shows they certify the returned value. The thresholded pairwise graph does not determine the triple threshold when $d\ge 2$ by Proposition D.4 and Theorem 2.8: two libraries can have the same complete pairwise graph at threshold τ while disagreeing on the triple threshold decision. □

Proposition D.3 (Pair scores do not determine triple scores)**.** *Let $C=[0,1]^2$. Library $\mathcal{A}$ has task models $q_t(c)=\|c-v_t\|^2$ with $v_1=(0,0)$, $v_2=(1,0)$, $v_3=(\frac{1}{2},\frac{\sqrt{3}}{2})$; library $\mathcal{B}$ keeps q_1,q_2 and replaces the third model by $q_3'(c)=4\,\|c-v_3'\|^2$ with $v_3'=(\frac{1}{2},\frac{\sqrt{5}}{4})$. Every singleton score is 0 and every pair score equals $\frac{1}{4}$ in both libraries, while*

$$\rho_{\{1,2,3\}}(\mathcal{A})=\tfrac{1}{3},\qquad \rho_{\{1,2,3\}}(\mathcal{B})=\tfrac{37-8\sqrt{10}}{36},$$

which differ by $\frac{8\sqrt{10}-25}{36}\approx 0.0083$.

Proof. For two models $w_i\|c-v_i\|^2$ and $w_j\|c-v_j\|^2$ with $\|v_i-v_j\|=L$, the pair minimax is attained on the segment between the centers (projection onto it decreases both distances) at the equalization point, with value $w_iw_jL^2/(\sqrt{w_i}+\sqrt{w_j})^2$. With unit weights and $L=1$ every pair of $\mathcal{A}$ scores $\frac{1}{4}$; in $\mathcal{B}$ the pairs with weights 1 and 4 at distance $\frac{3}{4}$ score $4\cdot\frac{9}{16}/9=\frac{1}{4}$ as well. For the triples, the maximum of the three models is convex and invariant under the reflection $c_1\mapsto 1-c_1$, so a minimizer lies on the axis $c_1=\frac{1}{2}$. On the axis the $\mathcal{A}$ maximum is minimized at the circumcenter, giving the squared circumradius $\frac{1}{3}$. For $\mathcal{B}$ at $c=(\frac{1}{2},y)$ the models read $\frac{1}{4}+y^2$ (twice, increasing in $y\ge 0$) and $4(y-\frac{\sqrt{5}}{4})^2$ (decreasing up to $\frac{\sqrt{5}}{4}$), with larger values outside $[0,\frac{\sqrt{5}}{4}]$, so the minimax sits at their crossing $y=\frac{\sqrt{5}-\sqrt{2}}{3}$, the relevant root of $3y^2-2\sqrt{5}\,y+1=0$, giving $\frac{1}{4}+y^2=\frac{37-8\sqrt{10}}{36}$. Every optimizer above lies in C, so the chart constraint is inactive. □

Proposition D.4 (Compact ellipsoidal triple obstruction)**.** *In $d=2$ there is a compact convex chart C and three centered quadratic sublevel sets $F_i=C\cap\{c:\|c-v_i\|_2^2\le R_0^2\}$ such that every pair intersects but $F_1\cap F_2\cap F_3=\emptyset$.*

D.7 Proof of Proposition D.4

Place v_1,v_2,v_3 at the vertices of an equilateral triangle of side length 2 with circumcenter at the origin, and take the compact square $C=[-2,2]^2$. Pick R_0 with $1<R_0<2/\sqrt{3}$ and set $F_i=C\cap\{x:\|x-v_i\|_2^2\le R_0^2\}$. Each pairwise center distance is $2<2R_0$, and the pairwise lens contains a point on the corresponding edge of the triangle, hence lies inside C. The smallest enclosing disk of the three centers has radius $2/\sqrt{3}$, so no point in C is within distance R_0 of all three centers and the triple intersection is empty. Each F_i is the intersection of C with a centered-quadratic sublevel set, and the obstruction therefore sits inside the class of compact convex sublevel sets used in the main theorems. Taking $R_0=1.1$ (so the sublevel radius is $r=1.1^2$) gives the numeric instance: every pair still overlaps, the smallest enclosing disk of the three centers has radius $2/\sqrt{3}\approx 1.155>1.1$ so the triple is empty, and the triple score is $\rho_{\{1,2,3\}}(C)=4/3$, with the primal witness at the circumcenter and symmetric dual weighting $\lambda^*=(1/3,1/3,1/3)$; pairwise screening admits but the atlas rejects.

D.8 Proof of Theorem 2.8

The theorem is a graph-level claim and we prove it at that level. Proposition D.4 gives one library $\mathcal{L}_1$ in d=2 with three quadratic sublevel sets at the vertices of an equilateral triangle of side 2, radius R_0 chosen so that all three pairs intersect (complete pair graph) but the triple is empty. Take $\mathcal{L}_2$ to be three identical balls at the same vertices with a slightly larger radius $R_0' > R_0$ chosen so that the triple does intersect (the smallest enclosing-circle radius is $2/\sqrt{3}$, so $R_0' > 2/\sqrt{3}$ suffices); the pair graph is still complete. The two libraries are therefore pairwise-graph-indistinguishable (both pair graphs are the same complete graph K_3) yet disagree on the triple-feasibility decision, so any rule that depends on the library only through the thresholded pairwise-mergeability graph is non-exact on this pair.

D.9 Proof of the estimation claims

Uniform error $\sup_{t,c} |\widehat{q}_t(c) - q_t(c)| \le \eta$ gives the sandwich $\max_t q_t(c) - \eta \le \max_t \widehat{q}_t(c) \le \max_t q_t(c) + \eta$ for every c. Two facts follow: $|\widehat{\rho}_S - \rho_S| \le \eta$ for every S, and $F_t(r-\eta) \subseteq \widehat{F}_t(r) \subseteq F_t(r+\eta)$ for every task. Intersecting the sets and taking nerves gives the η-interleaving $K_{r-\eta} \subseteq \widehat{K}_r \subseteq K_{r+\eta}$, with persistent-homology stability inherited from the standard interleaving theorem (Chazal et al., 2016). Quadratic functions on a k-dimensional chart form a $1 + k + k(k+1)/2$-dimensional vector space, so N_d scalar evaluations are enough for identification. For a poised design with matrix X of quadratic monomial features $\phi(c)$ at the design points, set $\kappa = \sup_{c \in C} \|\phi(c)^\top X^{-1}\|_1$. Hoeffding (Hoeffding, 1963) (for bounded per-prompt deficits in $[-B, B]$) with a union bound over the nN_d task/design values gives $\sup_{t,c} |\widehat{q}_t(c) - q_t(c)| \le O(\kappa B \sqrt{\log(nN_d/\delta_0)/N})$ from N examples per task at each design point, with failure probability δ_0; the same bound holds for sub-Gaussian per-prompt noise with parameter σ via the standard sub-Gaussian Hoeffding inequality (replacing B by σ). A nonlinear remainder ϱ between the true loss f_t and the quadratic model q_t pads every margin by ϱ uniformly. Under quadratic growth μ, an ϵ-optimal estimated coefficient $\widehat{c}$ satisfies $\|\widehat{c} - c^*\| \le \sqrt{2(\epsilon + 2(\eta + \varrho))/\mu}$.

D.10 Proof of Theorem 2.9

By definition, the models determine every sublevel set, score, optimizer, nerve, core, and dual certificate. Conversely, the sublevel-set family $\{F_t(r) : r \in \mathbb{R}\}$ recovers $q_t(c)$ through $q_t(c) = \inf\{r : c \in F_t(r)\}$. The abstract complexes K_r alone need not determine the functions, because nested convex families with different geometries can share the same intersection pattern. Let the merge map be gauge invariant, so that $\Delta(c)$ is a function of the represented updates $\{\Delta W_i\}$ alone. A LoRA gauge transformation leaves each represented update fixed, hence leaves $\Delta(c)$ fixed, so the tangent action $J_t \Delta(c)$ is unchanged. Any non-gauge-invariant rule can move while the underlying function, and hence its answers, stays fixed. The hypothesis is needed: for a factor-space operator such as PEFT's `linear` combination the merged update carries cross terms $B_i A_j$, and the gauge $(B_i, A_i) \mapsto (B_i R_i, R_i^{-1} A_i)$ moves $B_i R_i R_j^{-1} A_j$ while fixing every ΔW_i, so $\Delta(c)$ is not gauge invariant and 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. That rules out any universally exact rule built only on raw parameter geometry.

D.11 Proof of Theorem 3.2

Define $\tilde{F}_t(\tau) = \{\tilde{c} \in \mathbb{R}^r : \tilde{q}_t(\tilde{c}) \le \tau\}$, with convexity inherited from $\tilde{q}_t$. Let $K = \Pi(C)$, which is compact convex in $\mathbb{R}^r$ as the linear image of a compact convex set. We claim that for every $\tau \in \mathbb{R}$ and $S \subseteq [n]$,

$$\bigcap_{t \in S} F_t(\tau) \cap C = \emptyset \iff K \cap \bigcap_{t \in S} \tilde{F}_t(\tau) = \emptyset. \tag{$*$}$$

($\Leftarrow$) Suppose some $\tilde{c} \in K \cap \bigcap_t \tilde{F}_t(\tau)$. Then $\tilde{c} = \Pi c$ for some $c \in C$, and $q_t(c) = \tilde{q}_t(\Pi c) = \tilde{q}_t(\tilde{c}) \le \tau$ for every $t \in S$, so $c \in \bigcap_t F_t(\tau) \cap C$. ($\Rightarrow$) Suppose some $c \in \bigcap_t F_t(\tau) \cap C$. Then $\Pi c \in K$ and $\tilde{q}_t(\Pi c) = q_t(c) \le \tau$ for every $t \in S$, so $\Pi c \in K \cap \bigcap_t \tilde{F}_t(\tau)$. Consider now the family $\{\tilde{F}_t(\tau) \cap K\}_{t \in S}$ of compact convex sets in $\mathbb{R}^r$. By Helly's theorem in $\mathbb{R}^r$ (Helly, 1923), if the full intersection is empty then some $T \subseteq S$ with $|T| \le r + 1$ has empty intersection. By ($*$) we have $\bigcap_{t \in T} F_t(\tau) \cap C = \emptyset$, that is, $\rho_T(C) > \tau$. Applying this at every $\tau < \rho_S(C)$ yields the theorem's equality.

For tightness, we place r-simplex vertices $v_0, \dots, v_r \in \mathbb{R}^r$ inside a closed ball $\tilde{C}$, set $\tilde{q}_t(\tilde{c}) = \|\tilde{c} - v_t\|^2$, and lift to $\mathbb{R}^d$ via $\Pi(c_1, \dots, c_d) = (c_1, \dots, c_r)$ and $C = \tilde{C} \times [0,1]^{d-r}$. Pick any $\tau \in (\rho^{(r)}, \rho^{(r+1)})$, where $\rho^{(k)}$ is the smallest threshold at which every k-subset is compatible. Every r-subset is compatible at τ, but the full $(r+1)$-subset is not.

D.12 Proof of Theorem 3.3

Define $\bar{q}_t(c) := \tilde{q}_t(\Pi c)$. By hypothesis $\sup_{c \in C} |q_t(c) - \bar{q}_t(c)| \le \delta$, which carries over to $\sup_c |\max_t q_t(c) - \max_t \bar{q}_t(c)| \le \delta$. Minimizing in c gives $|\rho^q_S(C) - \rho^{\bar{q}}_S(C)| \le \delta$ for every $S \subseteq [n]$. The proxy $\bar{q}$ has cylindrical form, so Theorem 3.2 applies and yields $\rho^{\bar{q}}_S(C) = \max_{|T| \le r+1} \rho^{\bar{q}}_T(C)$. Combining,

$$\begin{aligned} \rho^q_S(C) &\le \rho^{\bar{q}}_S(C) + \delta = \max_{|T| \le r+1} \rho^{\bar{q}}_T(C) + \delta \\ &\le \max_{|T| \le r+1} \rho^q_T(C) + 2\delta, \end{aligned}$$

where the last step uses $\rho^{\bar{q}}_T \le \rho^q_T + \delta$ taken inside the max. The reverse inequality $\max_{|T| \le r+1} \rho^q_T(C) \le \rho^q_S(C)$ is immediate from monotonicity of ρ^q in S. Hence $0 \le \rho^q_S(C) - \max_{|T| \le r+1} \rho^q_T(C) \le 2\delta$.

D.13 Proof of Theorem 3.4

Write $\bar{q}_t(c) := \tilde{q}_t(Rc)$ for the cylindrical proxy, let $\beta^{\bar{q}}(A, V) := \rho^{\bar{q}}(A; V, 0)$ be its score at zero floor, and keep $\beta = \beta^q$ for the true score. By hypothesis $\sup_{t,c} |q_t(c) - \bar{q}_t(c)| \le \delta$; minimizing $\max_{t \in A}(\cdot)$ over $c \in C(V, 0)$ as in the proof of Theorem 3.3 gives

$$|\beta^q(A, V) - \beta^{\bar{q}}(A, V)| \le \delta \qquad \text{for every nonempty } A \subseteq T,\ V \subseteq U. \tag{$\dagger$}$$

Exact saddle for the proxy. The proxy models factor through R, so $k_V \le \operatorname{rank} R = r$ for every $V \subseteq U$, and in particular $k_U \le r$. Fix $A \subseteq T$. The inner problem $\min_{V \subseteq U} \beta^{\bar{q}}(A, V)$ equals $\rho^{\bar{q}}(A; U, 0)$ and, by the adapter core of Theorem 2.5, is attained by some $V \subseteq U$ with $|V| \le k_U + 1 \le r+1$, so restricting the inner minimization to $|V| \le r+1$ leaves it unchanged. The outer problem $\max_{A \subseteq T} \rho^{\bar{q}}(A; U, 0)$ equals $\rho^{\bar{q}}(T; U, 0)$ and, by the task core, is attained by some A with $|A| \le k_U + 1 \le r+1$, so restricting to $|A| \le r+1$ leaves it unchanged. Hence $\underline{\rho}^{\bar{q}}_r(T, U) = \rho^{\bar{q}}(T; U, 0)$, where $\underline{\rho}^{\bar{q}}_r$ is the restricted maximin formed from $\beta^{\bar{q}}$. The symmetric argument on the minimax, using the task core on each chart $C(V, 0)$ (latent dimension $k_V \le r$) and then the adapter core of the T-problem, gives $\overline{\rho}^{\bar{q}}_r(T, U) = \rho^{\bar{q}}(T; U, 0)$, with $\overline{\rho}^{\bar{q}}_r$ the restricted minimax formed from $\beta^{\bar{q}}$; the discrete saddle of Theorem 2.5 equates the two.

Transfer. Both $f \mapsto \max_{|A| \le r+1} \min_{|V| \le r+1} f(A, V)$ and $f \mapsto \min_{|V| \le r+1} \max_{|A| \le r+1} f(A, V)$ are 1-Lipschitz in f under the sup norm, since a uniform δ change in the payoff moves each nested extremum by at most δ. With ($\dagger$) this yields $|\underline{\rho}_r(T, U) - \underline{\rho}^{\bar{q}}_r(T, U)| \le \delta$ and $|\overline{\rho}_r(T, U) - \overline{\rho}^{\bar{q}}_r(T, U)| \le \delta$. Using $\underline{\rho}^{\bar{q}}_r = \overline{\rho}^{\bar{q}}_r = \rho^{\bar{q}}(T; U, 0)$ together with $|\rho - \rho^{\bar{q}}(T; U, 0)| \le \delta$ (which is ($\dagger$) at $(A, V) = (T, U)$),

$$|\rho - \underline{\rho}_r(T, U)| \le 2\delta, \qquad |\rho - \overline{\rho}_r(T, U)| \le 2\delta.$$

Bracket. Weak duality of the finite game gives $\overline{\rho}_r(T, U) \ge \underline{\rho}_r(T, U)$. Both scores lie in the interval $[\rho^{\bar{q}}(T; U, 0) - \delta,\ \rho^{\bar{q}}(T; U, 0) + \delta]$ by the transfer bounds, so their difference is at most 2δ; hence $0 \le \overline{\rho}_r(T, U) - \underline{\rho}_r(T, U) \le 2\delta$. □

D.14 An interval certificate for rank-reduced models

Proposition D.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, let $\widehat{q}_t(c) = \alpha_t + b_t^\top c + \frac{1}{2} c^\top H_t c$ with $H_t \succeq 0$ be the fitted models, with (α_t, b_t) the coefficients in the uncentered parametrization (Eq. (2) expanded about the origin), and for an orthogonal projector Π of rank r let $\widehat{q}^{(r)}_t$ be the reduced model with coefficients $(\alpha_t, \Pi b_t, \Pi H_t \Pi)$. Write $\epsilon_t := \sup_{c \in C} |\widehat{q}_t(c) - \widehat{q}^{(r)}_t(c)|$ and $\epsilon_S := \max_{t \in S} \epsilon_t$, and let $\rho^{(r)}_S(C)$ denote the score computed from the reduced models. Then $\epsilon_t \le \max_i |((I - \Pi) b_t)_i| + \frac{1}{2} \max_{i,j} |(H_t - \Pi H_t \Pi)_{ij}|$, computable exactly by face enumeration when C is a polytope; the scores transfer as $|\rho_S(C) - \rho^{(r)}_S(C)| \le \epsilon_S$ for every $S \subseteq [n]$; each $\widehat{q}^{(r)}_t$ is convex and depends on c only*

through Πc*, so Theorem 3.2 gives* $\rho_S^{(r)}(C) = \max_{T\subseteq S,\,|T|\le r+1} \rho_T^r(C)$ *exactly for the reduced models; and whenever the reduced core-table score differs from* τ *by more than* ϵ_S*, the full-model decision* $\rho_S(C) \le \tau$ *coincides with the reduced core-table decision.*

Proof sketch. The difference $e_t(c) = ((I-\Pi)b_t)^\top c + \frac{1}{2}c^\top(H_t - \Pi H_t \Pi)c$ is bounded on $\|c\|_1 \le 1$ by $\max_i |((I-\Pi)b_t)_i| + \frac{1}{2}\max_{i,j} |(H_t - \Pi H_t\Pi)_{ij}|$, giving (i); its exact value follows from enumerating the polytope faces (a quadratic is constant on each stationary set). Claim (ii) is the uniform-transfer argument of Theorem 3.3 with $\delta = \epsilon_S$; (iii) writes $\Pi = WW^\top$ so each $\widehat{q}_t^{(r)}$ factors through $c \mapsto W^\top c$ and Theorem 3.2 applies; (iv) combines (ii) and (iii). Reporting the reduced scores with the primal-dual bracket of Theorem 2.1 folds the solver gap into the margin. □

Section 4 reports the measured values on both m=8 shared charts; the face enumeration costs $O(2^d)$ linear solves, so at larger d the closed form in (i) replaces it.