# RESCUERS MATTER: RETHINKING CONFLICT IN LORA MERGING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When merging LoRA adapters, a pair that violates a task-loss cap can become feasible after another adapter is added. We call this a rescue. We ask when rescuers can be abundant and how often they occur in trained adapter banks. An exact softmax construction realizes rescue with rank-one LoRA: opposing useful-token preferences can cancel in a pair and be restored by a third adapter. In a robust two-type regime, every opposite-type pair is rescued by every remaining adapter; under balanced iid types, the certified fraction of distinct rescued triples tends to three quarters. A rank-one quadratic construction also gives a sharp $2/n$ cardinality ratio for sound downward-closed policies. Exhaustive audits of six new CLIP adapter banks find rescuers for most bad pairs overall, with rescue in every bank even under stricter margins. Earlier complete-cube audits show that permanent exclusion can reduce merge accuracy. The theory concerns explicit constructions and sampling laws; the empirical results do not provide population-risk certificates.

## 1 INTRODUCTION

Low-rank adaptation (LoRA) represents task-specific weight updates as products $BA$ (Hu et al., 2022). Task arithmetic combines these updates into a single model without a router or multiple inference paths (Ilharco et al., 2023). The difficulty is deciding which adapters to combine: a merge that helps one task may harm another, and the effect of an adapter depends on which others are present. We study this dependence under fixed task-loss thresholds.

Suppose a selection rule marks a harmful pair as an edge in a conflict graph and forbids every coalition containing that edge. Its accepted family is then downward closed: every subset of an accepted coalition must also be accepted. But adding another adapter can restore feasibility. We call this a rescue. A permanent edge loses that possibility. This restriction comes from the subset-selection rule; parameter-level methods for resolving interference need not impose it.

Consider three released adapters. On EuroSAT, at merge scale $0.5$, a MNIST+SVHN LoRA pair has mean cross-entropy $2.1017$; adding the EuroSAT adapter reduces it to $1.6772$. A signed quadratic fitted only to empty, singleton, and pair losses predicts $1.7020$ for the unseen triple. The fit anticipates the improvement, while a permanent edge excludes the triple. This example concerns relative empirical loss, not population-cap satisfaction. It illustrates contextual improvement; the cap-based rescue definition below requires the pair and triple to be compared against the same task caps.

Rescue need not require irreducible higher-order interactions: signed quadratic terms can already make an added adapter beneficial in context. Our main questions are when rescuers can be abundant and how often they occur in trained adapter banks. We first define rescue under fixed task caps, then give an exact softmax mechanism and sufficient conditions for abundance, and finally test its prevalence through exhaustive coalition audits. The cost of permanent exclusion is a consequence of this context dependence.

**Rescue and its abundance.** We define a rescuer as an adapter that turns an infeasible pair into a feasible triple. An exact rank-one softmax construction gives a robust regime in which every opposite-type pair has every remaining adapter as a rescuer (Theorem 1). A general theorem extends abundance to explicit attainable-score laws under positive pair-cancellation and rescuer probabilities (Theorem 3).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

**Rescue in trained adapter banks.** Exhaustive audits find rescue in all six new CLIP banks across two backbones, with and without rank truncation. At the primary cap, 76.0% of bad pairs are rescuable without truncation and 70.5% at rank eight. We report direction frequencies, rescuer counts, and two-sided buffers separately. The experiments establish recurrence in these banks, not that learned adapters follow the theory's sampling law.

**The cost of permanent exclusion.** In a single-layer rank-one LoRA construction, the optimal coalition contains all $n$ adapters, while every sound downward-closed policy retains at most two. The bound is attained and holds even with complete loss information. Keeping the signed pairwise differences instead recovers the full optimum within the quadratic class (Theorem 2). The earlier complete-cube experiments measure the corresponding empirical accuracy cost; auxiliary selection studies are retained in the appendix.

## 2 Related work

**Parameter composition.** Model soups average fine-tuned weights without increasing inference cost (Wortsman et al., 2022); task arithmetic combines updates relative to a shared base (Ilharco et al., 2023). Fisher merging weights parameters by estimated importance (Matena and Raffel, 2022), while RegMean uses input second moments to match individual models' linear-layer outputs (Jin et al., 2023). We fix the merge operator and study which subsets meet the task-loss caps.

**Interference and LoRA alignment.** TIES resolves parameter-sign conflicts (Yadav et al., 2023), and DARE drops and rescales delta parameters (Yu et al., 2024). KnOTS aligns LoRA updates through a joint SVD (Stoica et al., 2025); TSV and OSRM address singular-direction interference and orthogonal subspaces (Gargiulo et al., 2025; Zhang and Zhou, 2025). These operations can change coalition losses without imposing downward closure. Our lower bound applies when a selection rule permanently forbids subsets.

**Adaptive merging and selection.** AdaMerging learns task-wise or layer-wise coefficients through entropy minimization on unlabeled samples (Yang et al., 2024). SimMerge learns from offline two-way merge utilities and similarity features to select operators, subsets, and multi-way merge plans (Bolton et al., 2026). Our auxiliary selection study fits low-order task losses from the current adapter bank while holding the operator, scales, and final rank fixed. Our query counts exclude offline supervision of the kind used by SimMerge, so the budgets are not directly comparable.

**Functional structure and interaction theory.** Tangent-space task arithmetic links composition to weight disentanglement (Ortiz-Jimenez et al., 2023); Li et al. (2025) derive generalization guarantees for task vectors and low-rank approximations in a nonlinear Transformer model. Our main question is when rescuers are abundant. We use exact softmax competition, building on the attention formulation and its token-selection behavior (Vaswani et al., 2017; Tarzanagh et al., 2023); the auxiliary analysis concerns exclusion costs and low-order observations. Hodge-based diagnostics and TwistedMerge study higher-order alignment or certified obstructions (Zheng and Allen-Blanchette, 2026; Gong and Xu, 2026); we study task loss in fixed base-model coordinates. For this analysis we use mixed conflict hypergraphs, interaction expansions, and local-moment relaxations (Shapoval and Lee, 2021; Sundararajan et al., 2020; Sherali and Adams, 1990).

## 3 Rescue and its abundance

Fix a base model $\theta_0$ and $n$ adapters. At layer $\ell$, adapter $i$ contributes

$$D_{i\ell} = \alpha_i B_{i\ell} A_{i\ell}, \qquad W_\ell(z) = W_{0\ell} + \sum_{i=1}^{n} z_i D_{i\ell}, \qquad z \in \{0,1\}^n. \tag{1}$$

Write $S = \{i : z_i = 1\}$ and let $F_t(S)$ be task $t$'s population risk. Freeze the dictionary, scales $\alpha_i$, task caps $\tau_t$, utility $U(S)$, and selection budget independently of the final audit. A vertex may also denote a frozen spectral block; each point of a finite scale grid defines a separate candidate family. No monotonicity or submodularity is assumed.

**Definition 1** (Rescue and rescuer count)**.** For the fixed merge operator, let $\mathcal{F} = \{S : F_t(S) \le \tau_t \ \forall t\}$. A pair $P$ is rescued by $k \notin P$ if $P \notin \mathcal{F}$ and $P \cup \{k\} \in \mathcal{F}$. Its rescuer count is $d_R(P) = |\{k \notin P : P \cup \{k\} \in \mathcal{F}\}|$.

This ordinary rescue does not require singleton failure or failure of every pair in the triple. A strictly minimal rescue, where every nonempty proper subset is infeasible, is a separate event. We distinguish many pairs having a rescuer from each pair having many rescuers. A directed event $(P, k)$ and a distinct rescued triple also have different denominators.

### 3.1 Abundant rescue from exact softmax competition

Ordinary rescue need not require any singleton to fail. Consider an exact one-head attention classifier with two useful tokens of value $+1$ and one distracting token of value $-1$. Query-only rank-one LoRA updates can realize the attention scores $(-\log 2 + Z_S, -\log 2 - Z_S, -b + H_S)$, where $Z_S = \sum_{i \in S} a_i$ and $H_S = \sum_{i \in S} h_i$. The distracting-token mass and correct-class margin are exactly

$$p_B(S) = \sigma(\Lambda(S)), \qquad \Lambda(S) = -b + H_S - \log\cosh Z_S, \qquad m(S) = 1 - 2p_B(S). \tag{2}$$

Here $\sigma$ is the logistic sigmoid; the task classifies the fixed input against a zero competing logit, so it is correct exactly when $\Lambda(S) < 0$. Equation (22) in Appendix A.5 gives the query-only rank-one realization. Opposing useful-token preferences can cancel in a pair, exposing the accumulated distracting bias. A third adapter can restore a useful preference and correct the prediction. No softmax linearization is used.

**Theorem 1** (A robust regime with abundant ordinary rescue)**.** *In this classifier, let $n \ge 3$, $b = 1.5$, $h_i \in [0.95, 1.05]$, and $a_i = 3s_i + \epsilon_i$, where $s_i \in \{-1, +1\}$ and $|\epsilon_i| \le 0.1$. The base and every singleton are correct. Every opposite-type pair is incorrect, and every remaining adapter rescues it, with pair log-odds at least $0.25$ and triple log-odds at most $-0.25$. Thus, at any classification-error cap in $[0, 1)$, the number $N$ of distinct rescued triples satisfies*

$$N \ge \frac{n_+ n_- (n-2)}{2},$$

*where $n_+, n_-$ count the two types. For iid types with $\Pr(s_i = +1) = p \in (0, 1)$, the certified mixed-triple fraction converges in probability to $3p(1-p)$, and $\Pr(N > 0) \ge 1 - p^n - (1-p)^n$ for $n \ge 3$.*

*Proof sketch.* Singleton log-odds are at most $-0.45$. An opposite pair has $|Z_S| \le 0.2$, so $\Lambda(S) \ge 0.4 - 0.02 > 0.25$, using $\log\cosh z \le z^2/2$. A containing triple is within $0.3$ of $+3$ or $-3$; since $\log\cosh$ is 1-Lipschitz, $\Lambda(S) \le 1.65 - [\log\cosh 3 - 0.3] < -0.25$. Count triples containing both types and apply the law of large numbers.

At $p = 1/2$, the certified fraction tends to $3/4$, and the eight-adapter existence bound is $127/128$. The corresponding certified direction fraction tends to $1/2$. These are conclusions for the explicit sampling model, not measured frequencies in trained banks. Its base-linearized margin decreases with $H_S$ and cannot produce rescue when all $h_i > 0$. The general softmax argument replaces $\log\cosh$ by a centered log-partition function: positive pair-cancellation probability and a uniform positive rescuer probability yield quadratically many conflicting pairs, each with linearly many rescuers, with probability tending to one (Theorem 3). Applying this mechanism to a trained multilayer model requires attainable score coordinates and control of downstream, multi-task, and compression margins; the empirical results below do not establish those conditions.

Rescue makes permanent subset exclusion potentially costly. The following construction isolates this consequence from the abundance mechanism above.

### 3.2 A consequence: the price of permanent exclusion

For a loss cap $\tau$, let $\mathcal{F} = \{S : L(S) \le \tau\}$. A policy $\mathcal{A}$ is *sound* if $\mathcal{A} \subseteq \mathcal{F}$, and *downward closed* if $S \in \mathcal{A}$ implies $T \in \mathcal{A}$ for every $T \subseteq S$. Put $V(\mathcal{A}) = \max(\{0\} \cup \{|S| : S \in \mathcal{A}\})$, $\mathrm{OPT} = V(\mathcal{F})$, and let $\mathrm{OPT}_{\mathrm{downward}}$ maximize $V$ over sound downward-closed families. Let $\mathrm{OPT}_{\mathrm{signed}}$ denote the maximum cardinality allowed by the quadratic interpolant at the same cap.

**Theorem 2** (Sharp price of downward closure for rank-one LoRA merging). *For every $n \geq 3$ and $\gamma > 0$, there is an exactly quadratic Boolean coalition loss $L$, realized by rank-one LoRA updates to a single $2 \times 2$ linear layer with one input example and squared loss, and a cap $\tau$ such that*

$$\boxed{\mathrm{OPT}_{\mathrm{downward}} = \frac{2}{n}\,\mathrm{OPT}, \qquad \mathrm{OPT}_{\mathrm{signed}} = \mathrm{OPT} = n.} \tag{3}$$

*Both representations can use the same $1 + n + \binom{n}{2}$ empty, singleton, and pair observations; these uniquely determine the signed interpolant within the degree-at-most-two class. The downward-closed optimum remains two even with all $2^n$ losses. The full coalition has feasibility margin $\gamma$, and the ratio $2/n$ is sharp on this family: a sound downward-closed policy attains it.*

*Proof sketch.* Write $n = m + 1$, with rescuer $r = n$, and take the excess risk

$$G_\gamma(z) = \sum_{1 \leq i < j \leq m} z_i z_j - \frac{m-1}{2} z_r \sum_{i=1}^{m} z_i - \gamma z_r. \tag{4}$$

Without $r$, a size-$k$ coalition has score $\binom{k}{2}$; with $r$, its score is $-k(m-k)/2 - \gamma$. Thus every ordinary pair is infeasible, every rescued coalition is feasible, and the largest sound downward-closed family is $\{S : |S \cap [m]| \leq 1\}$, of utility two. Quadratic interpolation recovers the full-set score $-\gamma$, hence utility $n$. Appendix H.2 constructs rank-one $D_i = v_i e_1^\top$ with $x = e_1$, $W_0 = I_2$ and target $y$ satisfying

$$L(z) = \left\| \left( W_0 + \sum_i z_i D_i \right) x - y \right\|_2^2 = \tau + G_\gamma(z), \qquad \tau = \gamma + \frac{m^2}{8}. \tag{5}$$

Appendix H gives the construction and complete proof.

## 4 Empirical rescue: prevalence and consequences

We evaluate every pair and triple in six new banks to measure rescue frequency. The original six complete-cube banks then quantify a consequence of rescue: the feasible utility lost to permanent exclusion. The two cohorts are reported separately.

### 4.1 Rescue frequency in six new adapter banks

We train three new eight-adapter banks on each of CLIP ViT-B/32 and ViT-B/16 (Radford et al., 2021), paired across three new data repartitions. The tasks are MNIST, SVHN, GTSRB, EuroSAT, Stanford Cars, DTD, RESISC45, and SUN397. Each rank-8 adapter trains for 1,200 steps. At scale 0.5, we evaluate both the dense sum and its per-layer rank-8 truncation in FP32. Each split has 15,329 training images and, per task, 256 discovery and 512 audit images. RGB hashes separate training from holdout and discovery from audit within each split. The public image pools are repartitioned; splits overlap one another and the earlier cohort, so these are repeated bank observations rather than independent datasets or official benchmark scores. Appendix B gives the frozen protocol and full diagnostics.

All $1 + 8 + 28 + 56 = 93$ coalitions through size three are evaluated. For audit error $\widehat{e}_t(S)$ and cap $c = 0.02$, define

$$v(S) = \max_t \{\widehat{e}_t(S) - \widehat{e}_t(\varnothing) - c\}, \qquad I_\gamma(P, k) = \mathbf{1}\{v(P) > \gamma,\ v(P \cup \{k\}) \leq -\gamma\}. \tag{6}$$

Ordinary rescue uses $\gamma = 0$; a positive $\gamma$ is a two-sided empirical buffer, not a confidence bound. Each bank has 168 pair–addition directions and 56 distinct triples. Bad-pair coverage asks whether any of a bad pair's six candidate additions rescues it. Shared adapters and images make these events dependent; 1,008 directions are not 1,008 independent trials.

Figure 1 shows rescue in all six banks, with and without truncation. Both representations have 126/1,008 rescue directions (12.5%) and 99/336 distinct rescued triples (29.5%). Rescuers exist for 73/96 bad pairs (76.0%) dense and 67/95 (70.5%) at rank eight. Of the directions, 104 are shared, 22 are dense-only, and 22 are rank-8-only. Thus the phenomenon does not depend on truncation, although individual events change.

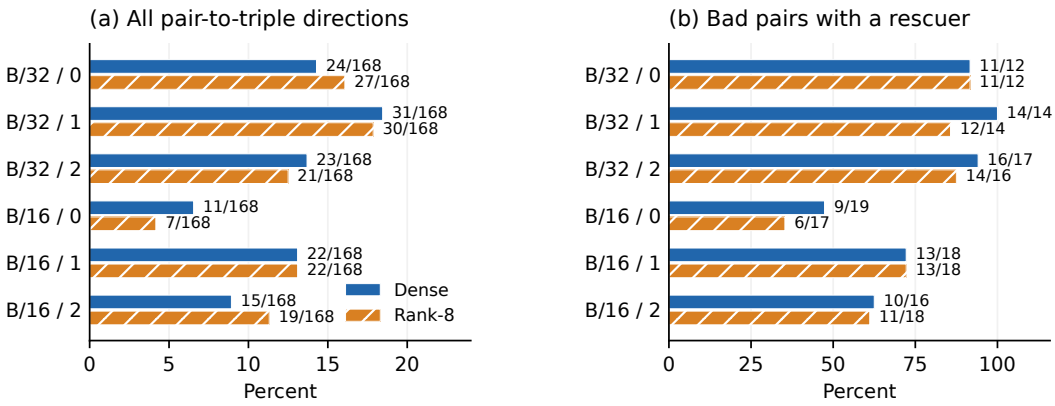


Figure 1: Rescue across all six new banks at the 2 pp empirical cap. Left: rescues among all pair–addition directions. Right: bad pairs with at least one rescuer. Labels give exact counts. The two representations are evaluated on the same images; the banks share public image pools.

Table 1: Number of rescuers per bad pair in the six new banks. Each bad pair has six candidate additions. Counts are pairs, not independent trials. Mean includes the pairs with no rescuer.

| | Rescuer count $d_R(P)$ | | | | | |
|---|---|---|---|---|---|---|
| Representation | 0 | 1 | 2 | $\geq 3$ | Bad pairs | Mean |
| Dense | 23 | 35 | 28 | 10 | 96 | 1.31 |
| Rank-8 | 28 | 25 | 30 | 12 | 95 | 1.33 |

Table 1 distinguishes prevalence across bad pairs from the number of available rescuers. Including zero-rescuer pairs, the mean count is 1.31 dense and 1.33 at rank eight. Uniformly sampling a bad-pair–addition choice succeeds in $126/(96 \cdot 6) = 21.9\%$ and $126/(95 \cdot 6) = 22.1\%$ of cases. The empirical conclusion is that many bad pairs have a rescuer, not that most random additions succeed or that almost every pair has many rescuers.

At 0.5 pp and 1 pp buffers, 62 and 36 directions remain in each representation, respectively, with rescue still present in all six banks (Table 2). All six also contain ordinary rescue at each tested cap of 1, 2, 3, and 5 pp. These sensitivities support recurrence while showing that many zero-buffer events lie near a cap. Strict-minimal rescue is absent in the primary new-bank setting: 0/336 triples in both representations.

The rescuer is itself singleton-infeasible in 26/126 dense and 24/126 rank-8 directions, showing that feasibility in isolation does not determine contextual value. Specific witnesses are less stable than bank-level recurrence: each representation has 98 discovery positives, of which 47 replicate on audit; audit adds 79 other positives. Local tangent/GGN and quadratic diagnostics also leave substantial errors (Appendix B.3). These results do not identify the exact token mechanism or establish the sampling conditions of Theorem 1 in trained CLIP models.

### 4.2 Original six banks and complete-cube protocol

In the original cohort, we train three eight-adapter banks on each of CLIP ViT-B/32 (E7) and ViT-B/16 (E8) (Radford et al., 2021). The tasks are MNIST, SVHN, GTSRB, EuroSAT, Stanford Cars, DTD, RESISC45, and SUN397. Each rank-8 adapter trains for 1,200 steps. Every primary coalition uses the FP32 per-layer merge $W_0 + \Pi_8(0.5 \sum_{i \in S} D_i)$, with checkpoint, scale, rank, tasks, and caps fixed before audit. The common split has 15,329 training images and, per task, 256 discovery and 512 audit images; global RGB hashes separate training from holdout. Public test pools are repartitioned, so these are not official benchmark scores. E8 reuses the E7 split after its results were seen; seeds and backbones are not independent dataset replications. Appendix M gives dataset sources, the protocol, and preceding negative experiments.

Table 2: Two-sided empirical buffers in the new banks. A pair exceeds the 2 pp cap by more than $\gamma$, and its triple passes every cap by at least $\gamma$. Denominators retain all original directions and original bad pairs. The last column holds separately for each representation.

| | Rescue directions | | Rescuable bad pairs | | Banks |
|---|---|---|---|---|---|
| $\gamma$ (pp) | Dense | Rank-8 | Dense | Rank-8 | with rescue |
| 0 | 126/1008 | 126/1008 | 73/96 | 67/95 | 6/6 |
| 0.5 | 62/1008 | 62/1008 | 40/96 | 40/95 | 6/6 |
| 1 | 36/1008 | 36/1008 | 24/96 | 26/95 | 6/6 |

Table 3: Complete-cube empirical optima, 512 audit images per task. Accuracy (%) is maximized subject to all task caps; cardinality is optimized separately. Down denotes the largest sound downward-closed family. Gap is the accuracy cost in pp, not a cardinality ratio.

| Backbone | Bank | Full acc. | Down acc. | Gap | Full $\lvert S\rvert$ | Down $\lvert S\rvert$ |
|---|---|---|---|---|---|---|
| B/32 | 0 | 67.85 | 64.65 | 3.20 | 7 | 4 |
| B/32 | 1 | 67.43 | 67.43 | 0.00 | 7 | 4 |
| B/32 | 2 | 69.75 | 67.46 | 2.29 | 7 | 4 |
| B/16 | 0 | 73.41 | 70.31 | 3.10 | 7 | 3 |
| B/16 | 1 | 74.24 | 70.80 | 3.44 | 7 | 3 |
| B/16 | 2 | 73.66 | 69.14 | 4.52 | 6 | 2 |

Let $\widehat{e}_t(S)$ be audit error and define empirical feasibility by $\widehat{e}_t(S) - \widehat{e}_t(\varnothing) \le 0.02$ on every task. The auxiliary selection study (Appendix G) minimizes average error $J(S) = \frac{1}{8}\sum_t e_t(S)$ under the discovery version of these caps. Methods access cached predictions through a budgeted discovery-query interface. Only after choices are frozen does a separate evaluator read all $2^8 = 256$ audit coalitions. For a feasible output, constrained regret is $\widehat{J}(S) - \min_{T\in\widehat{\mathcal{F}}} \widehat{J}(T)$; it is undefined for an infeasible output. We distinguish base returns from feasible nonempty merges. Audit point estimates are not population guarantees.

### 4.3 A CONSEQUENCE: THE COST OF EXCLUSION IN THE ORIGINAL BANKS

Every original bank contains pairs that violate a task cap while a containing triple passes all eight caps, including rescue directions shared by the truncated and untruncated models (Table 19). To quantify the resulting representation cost, define the largest sound downward-closed family from the complete audit table:

$$\widehat{\mathcal{D}} = \{S : T \in \widehat{\mathcal{F}} \text{ for every } T \subseteq S\}. \tag{7}$$

Every empirically sound downward-closed policy lies in $\widehat{\mathcal{D}}$. Its optimum measures the cost of representation with complete information, separately from the 37-query exclusion baseline. In Table 3, the accuracy cost is positive in five banks, including all three B/16 banks (3.10–4.52 percentage points, pp). Cardinality falls from seven to four on B/32 and from seven/seven/six to three/three/two on B/16. Accuracy and cardinality are optimized separately. These multilayer rank-8 models have nonquadratic classification error; the table does not verify the theoretical $2/n$ ratio.

**Removing truncation.** Evaluating all 256 untruncated coalitions at the same scale on the original audit images leaves a positive accuracy cost in all six banks (0.81–5.22 pp). Among 1,008 pair-to-triple directions, 85 are rescues with and without truncation, 12 only without it, and 20 only at rank 8. Thus rescue recurs in every bank and persists beyond truncation, although its effect on optimal accuracy varies. Unequal ranks and reused images make this an exploratory mechanism comparison (Appendix M.8).

**A frozen rescuing triple.** E7 freezes a previously discovered MNIST–GTSRB–EuroSAT public-adapter triple and all eight subsets before evaluating 512 new images per task. At scale 0.5 and final rank 16, the triple has 58.15% average accuracy, versus 47.27% for the base; its worst task regression is 1.17 pp. All six nonempty proper subsets violate a 2 pp cap, some by only one image. Under paired image resampling, the triple passes in 72.1% of draws, but all six subset failures persist in

only 16.5% (Appendix M.10). This empirical family is not downward closed; it establishes neither third-order interaction nor population feasibility.

## 5 LIMITATIONS AND CONCLUSION

An infeasible subset need not make a larger merge infeasible. We formalize this rescue event and give an exact softmax regime in which a positive fraction of pairs each have linearly many rescuers. Exhaustive audits in six new trained CLIP banks find a rescuer for most bad pairs, with rescue in every bank even under two-sided empirical buffers. These observations support prevalent ordinary rescue in the studied banks; they do not establish the score-distribution assumptions of the theorem. As a consequence, permanent exclusion can lose useful merges, as shown by the original sharp $2/n$ construction and complete-cube experiments. A harmful pair is therefore evidence about that pair. Treating it as a ban on every larger coalition requires further justification.

The empirical study covers eight vision tasks and two related CLIP backbones, with shared public image pools and overlapping repartitions. Transfer to independent datasets or language-model LoRAs remains untested. Earlier public-adapter experiments have unresolved source-split overlap, and scale $0.5$ was chosen after exploratory diagnostics. All selection comparisons fix the merge operator; they do not compare TIES, KnOTS, TSV, or OSRM as alternative algorithms. The abundance theory concerns explicit attainable-score laws, not the distribution produced by adapter training. Classification error need not satisfy the auxiliary degree or uniform-tail assumptions; the sharp ratio measures cardinality rather than accuracy, and the information theorem excludes parameter inspection. General signed optimization is hard. Finally, the empirical cap results supply no population safety certificate, and specific rescue labels can change between discovery and audit. Testing independent data, additional architectures, and other merge operators would determine how broadly these observations transfer.

## AI USE STATEMENT

We used GPT-5.6 Sol and GPT-6 to assist with manuscript writing and polishing, literature search, and the development and checking of mathematical proofs. The authors take full responsibility for the content of this paper.

## REPRODUCIBILITY STATEMENT

The supplementary archive contains editable LaTeX, the CPU prototype and results, frozen CLIP protocols, per-example error caches, selection records, and analysis code. Its README gives offline checks for the reported tables and discovery-only selection replay. Appendix B documents the new frequency protocol and diagnostics; Appendix A gives the complete softmax abundance argument. Appendix M gives original training, query accounting, full seed tables, and negative results. Appendix P separates the earlier evaluation protocols; Appendices M.8–M.10 document untruncated inference and paired resampling. Large trained checkpoints and the complete CLIP logit archives are not bundled. Regenerating model outputs requires the recorded public models, datasets, training code, and GPU environment.

## ETHICS STATEMENT

The experiments use public vision benchmarks and constructed networks. Rescue under the stated loss does not certify fairness, privacy, general safety, or deployment suitability. A bounded-risk guarantee applies only to its declared loss, distribution, candidates, and thresholds.

## REFERENCES


Theodore W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955. https://www.jstor.org/stable/2032333.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. https://arxiv.org/abs/2110.01052.

Oliver Bolton, Aakanksha, Arash Ahmadian, Sara Hooker, Marzieh Fadaee, and Beyza Ermis. SimMerge: Learning to select merge operators from similarity signals. *arXiv preprint arXiv:2601.09473*, 2026. https://arxiv.org/abs/2601.09473.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. https://arxiv.org/abs/1703.00121.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. https://www.robots.ox.ac.uk/~vgg/data/dtd/.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18695–18705, 2025. https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Stefan Glock, Felix Joos, Jaehoon Kim, Marcus Kühn, and Lyuben Lichev. Conflict-free hypergraph matchings. In *Proceedings of the 2023 ACM–SIAM Symposium on Discrete Algorithms*, pp. 2991–3005, 2023. https://doi.org/10.1137/1.9781611977554.ch115. Full preprint: https://arxiv.org/abs/2205.05564.

Ting Gong and Shitan Xu. TwistedMerge: Certified higher-order diagnostics and abstention for model merging. *arXiv preprint arXiv:2607.20887*, 2026. https://arxiv.org/abs/2607.20887.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. https://arxiv.org/abs/1709.00029.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. https://doi.org/10.1080/01621459.1963.10500830.

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. https://arxiv.org/abs/2106.09685.

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. https://arxiv.org/abs/2212.04089.

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. https://arxiv.org/abs/2212.09849.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. https://doi.org/10.1109/ICCVW.2013.77.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. https://doi.org/10.1109/5.726791.

Hongkang Li, Yihua Zhang, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. When is task vector provably effective for model editing? A generalization analysis of nonlinear Transformers. In *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2504.10957.

Michael S. Matena and Colin A. Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, pp. 148–188. Cambridge University Press, 1989. https://doi.org/10.1017/CBO9781107359949.008.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. http://ufldl.stanford.edu/housenumbers/.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/d28077e5ff52034cd35b4aa15320caea-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. https://proceedings.mlr.press/v139/radford21a.html.

scikit-learn developers. `sklearn.datasets.load_digits`: Dataset documentation. Official documentation, accessed September 11, 2026. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html. The executed experiments use scikit-learn 1.8.0.

Andriy Shapoval and Eva K. Lee. Generalizing 0-1 conflict hypergraphs and mixed conflict graphs: Mixed conflict hypergraphs in discrete optimization. *Journal of Global Optimization*, 80(4):805–840, 2021. https://doi.org/10.1007/s10898-021-01012-3.

Hanif D. Sherali and Warren P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990. https://doi.org/10.1137/0403036.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *The 2011 International Joint Conference on Neural Networks*, pp. 1453–1460, 2011. https://doi.org/10.1109/IJCNN.2011.6033395.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the Knots. In *International Conference on Learning Representations*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d4f8a5109c5083b5307fcd0bddae7a7-Abstract-Conference.html.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268, 2020. https://proceedings.mlr.press/v119/sundararajan20a.html.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-Margin Token Selection in Attention Mechanism. In *Advances in Neural Information Processing Systems*, 2023. https://arxiv.org/abs/2306.13596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. https://arxiv.org/abs/1706.03762.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya 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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, 2022. https://proceedings.mlr.press/v162/wortsman22a.html.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. https://doi.org/10.1109/CVPR.2010.5539970.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.02575.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775, 2024. https://proceedings.mlr.press/v235/yu24p.html.

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 (Volume 1: Long Papers)*, pp. 26459–26472, 2025. https://aclanthology.org/2025.acl-long.1284/.

Dongzhe Zheng and Christine Allen-Blanchette. Beyond pairwise: Diagnosing higher-order merge failures via Hodge decomposition. *ICML 2026 Workshop on Weight-Space Symmetries*, 2026. https://openreview.net/forum?id=5IsDyRzKks. Author code: https://github.com/ContinuumCoder/Higher-Order-Model-Merge.

# A SOFTMAX COMPETITION CREATES ABUNDANT RESCUE

The main text gives a robust two-type specialization. Here we derive the mechanism and general abundance conditions directly from softmax competition. A pair can lose complementary preferences among useful tokens, leaving an accumulated distracting-token bias exposed; a third adapter can restore a useful preference. No output return to the base is required. We study ordinary rescue: an infeasible pair contained in a feasible triple. Singletons may all be feasible. All probability statements concern independent adapter-bank draws conditional on a fixed network, task, and merge rule. They do not assert that trained banks satisfy the sampling or margin conditions.

Within this appendix, $z_i$ denotes attention-score coordinates, rather than the Boolean membership indicator in (1).

## A.1 AN EXACT LOG-ODDS IDENTITY

Fix an attention query, useful tokens $G$, and a distracting token $B$. Let their base scores be $\beta_j$ and set

$$\pi_j = \frac{e^{\beta_j}}{\sum_{l\in G} e^{\beta_l}}, \qquad b = \log \sum_{l\in G} e^{\beta_l} - \beta_B > 0.$$

The scores include the standard attention scale. For adapter $i$, let $\delta_{ij}$ be its additive score increment, and define

$$\bar{\delta}_i = \sum_{j\in G} \pi_j \delta_{ij}, \qquad h_i = \delta_{iB} - \bar{\delta}_i, \qquad r_{ij} = \delta_{ij} - \bar{\delta}_i. \tag{8}$$

Assume that the centered score effects have a fixed coordinate representation

$$r_{ij} = w_j^\top z_i, \quad z_i \in \mathbb{R}^d, \quad \sum_{j\in G} \pi_j w_j = 0, \quad \operatorname{span}\{w_j : j \in G\} = \mathbb{R}^d. \tag{9}$$

Write $Z_S = \sum_{i\in S} z_i$ and $H_S = \sum_{i\in S} h_i$.

**Proposition 1** (Exact softmax competition)**.** *The distracting-token attention mass satisfies*

$$p_B(S) = \sigma(\Lambda(S)), \qquad \Lambda(S) = -b + H_S - K(Z_S), \qquad K(z) = \log \sum_{j\in G} \pi_j e^{w_j^\top z}. \tag{10}$$

*This identity is exact for additive scores; the softmax is not linearized. Moreover $K(0) = 0$, $\nabla K(0) = 0$, and $K(z) \geq 0$.*

*Proof.* Take the logarithm of the ratio of the distracting mass to the total useful mass. Substitute $\delta_{ij} = \bar{\delta}_i + w_j^\top z_i$. The common increment $\sum_i \bar{\delta}_i$ cancels, leaving (10). The three properties of $K$ follow from normalization, centering, and Jensen's inequality, respectively. □

For a query-only LoRA update with fixed inputs and keys, $\delta_{ij} = k_j^\top (D_i x)/\sqrt{d_h}$ is exactly linear in the realized product $D_i = B_i A_i$. This is not a tangent approximation of the attention output. The construction uses standard softmax attention Vaswani et al. (2017). Token selection and its dependence on value embeddings have been studied in attention optimization Tarzanagh et al. (2023); those results do not assert the rescue law proved here.

For an exact classification model, give useful tokens value $+1$ and the other token value $-1$, and take the attention readout as the correct-class margin against a zero competing logit. Then

$$m(S) = 1 - 2p_B(S) = -\tanh(\Lambda(S)/2). \tag{11}$$

Thus $\Lambda \leq -\gamma$ and $\Lambda \geq \gamma$, with $\gamma > 0$, certify correct and incorrect predictions with strict margins. The base is correct. At any error cap in $[0, 1)$, a bad pair followed by a correct triple is an actual classification-error rescue on this conceptual task.

**Proposition 2** (The base tangent misses the mechanism)**.** *If all $h_i > 0$, the base-linearized margin is*

$$m_{\tan}(S) = \tanh(b/2) - 2\sigma(-b)(1 - \sigma(-b))H_S.$$

*It is strictly decreasing under adding an adapter. Its classification-cap family is downward closed, while the exact margin can exhibit the rescues below.*

*Proof.* Differentiate (10) and use $\nabla K(0) = 0$. Every added adapter increases $H_S$. The tangent model therefore cannot turn a negative margin into a positive margin by adding an adapter. □

### A.2 THE RESCUE REGION FOLLOWS FROM KEY GEOMETRY

Define

$$W = \max_j \|w_j\|, \quad \pi_{\min} = \min_j \pi_j > 0, \quad \kappa = \min_{\|u\|=1} \max_j w_j^\top u.$$

**Lemma 1** (Local cancellation and escape at large preference)**.** *Under (9), $\kappa > 0$ and*

$$0 \le K(z) \le \tfrac{1}{2} W^2 \|z\|^2, \qquad K(z) \ge \kappa \|z\| - \log(1/\pi_{\min}), \qquad |K(z) - K(z')| \le W\|z - z'\|. \tag{12}$$

*Consequently every positive-level sublevel of $K$ contains a neighborhood of zero and is bounded.*

*Proof.* For any nonzero $u$, if all $w_j^\top u \le 0$, their strictly positive weighted mean being zero would force every projection to vanish. This contradicts the span assumption. Compactness of the unit sphere gives $\kappa > 0$. For the upper bound, let $X$ equal $w_j^\top z$ with probability $\pi_j$. Then $\mathbb{E}X = 0$ and $X \in [-W\|z\|, W\|z\|]$. Hoeffding's elementary moment-generating-function bound Hoeffding (1963) gives $\log \mathbb{E}e^X \le W^2\|z\|^2/2$. Retaining a maximizer in the log-sum proves the lower bound. Finally, $\nabla K$ is a convex combination of the $w_j$, and hence has norm at most $W$. □

Allow a uniform log-odds discrepancy

$$\Lambda_{\text{act}}(S) = -b + H_S - K(Z_S) + e_S, \qquad |e_S| \le \eta \quad (1 \le |S| \le 3). \tag{13}$$

Assume $0 < h_- \le h_i \le h_+$ and fix $\gamma > 0$. Set

$$T_2 = 2h_- - b - \gamma - \eta > 0, \qquad T_3 = 3h_+ - b + \gamma + \eta, \quad C_2 = \{K \le T_2\}, \quad C_3 = \{K < T_3\}. \tag{14}$$

Then

$$z_i + z_j \in C_2 \implies \Lambda_{\text{act}}(ij) \ge \gamma, \tag{15}$$

$$z_i + z_j + z_k \notin C_3 \implies \Lambda_{\text{act}}(ijk) \le -\gamma. \tag{16}$$

If additionally $h_+ \le b - \gamma - \eta$, every singleton is correct. These conditions are compatible whenever $b > 3(\gamma + \eta)$, by choosing $h_-, h_+$ in an interior subinterval of $((b + \gamma + \eta)/2, b - \gamma - \eta)$. No equality such as a prescribed nonlinear return point is imposed.

### A.3 MANY CONFLICTING PAIRS, EACH WITH MANY RESCUERS

The next theorem only uses two score-distribution quantities. It allows independent nonidentically distributed adapters and dependence between $h_i$ and $z_i$ within each adapter.

**Theorem 3** (Quantitative abundance of ordinary rescues)**.** *Suppose (13)–(14) hold and the pairs $(h_i, z_i)$ are independent. Let $p_0, q_0 > 0$ satisfy*

$$\Pr(z_i + z_j \in C_2) \ge p_0 \quad (i \ne j), \qquad \sup_{s \in C_2} \Pr(z_i + s \in C_3) \le 1 - q_0 \quad (i \le n). \tag{17}$$

*Let $N$ count distinct triples containing at least one pair certified by (15) and satisfying (16). For $n \ge 3$,*

$$\Pr\left\{ N \ge \frac{p_0 q_0}{4} \binom{n}{3} \right\} \ge 1 - e^{-np_0^2/8} - \binom{n}{2} e^{-(n-2)q_0/8}. \tag{18}$$

*On the same event, at least $(p_0/2)\binom{n}{2}$ pairs are certified conflicts, and each has at least $q_0(n-2)/2$ distinct rescuers. With fixed $p_0, q_0$, there are therefore $\Theta(n^2)$ certified bad pairs and $\Theta(n^3)$ rescued triples with probability tending to one. The lower bound is understood as vacuous if its right side is negative.*

*Proof.* Let $M = \sum_{i<j} \mathbf{1}\{z_i + z_j \in C_2\}$. Then $\mathbb{E}M \ge p_0\binom{n}{2}$. Changing one adapter changes $M/\binom{n}{2}$ by at most $(n-1)/\binom{n}{2} = 2/n$. The bounded-differences inequality McDiarmid (1989) gives

$$\Pr\{M < (p_0/2)\binom{n}{2}\} \le e^{-np_0^2/8}.$$

For a fixed pair, condition on its two adapters. If it is certified, its sum lies in $C_2$. The indicators $\mathbf{1}\{z_i + z_j + z_k \notin C_3\}$ for $k \notin \{i, j\}$ are independent and each has conditional mean at least $q_0$. The

multiplicative Chernoff bound (also obtained directly from the product Bernoulli generating function) gives a lower-tail probability at most $e^{-(n-2)q_0/8}$ for fewer than $q_0(n-2)/2$ successes. A union bound over the events that a certified pair has too few rescuers gives the second error term. On the intersection of the events there are at least

$$\frac{p_0}{2}\binom{n}{2}\frac{q_0}{2}(n-2)$$

certified pair-to-triple directions. A triple has only three pairs. Division by three and $\binom{n}{2}(n-2) = 3\binom{n}{3}$ give (18). □

No independence between different pair or triple events is asserted. The conditional argument accounts for their shared adapters. If the score-error bound holds on an external event of probability at least $1-\delta_n$, subtract $\delta_n$ from the right side, without conditioning the sampling argument on that event. For a genuine downstream task, counted events are actual rescues only after the task-margin conditions in Section A.8 are met.

For a fixed iid bank law, the indicator that a triple satisfies the stated criterion is a symmetric bounded kernel. Changing one adapter changes its normalized total by at most $3/n$, so the normalized count converges in probability to its single-triple probability. Thus abundance is stronger than an expected-count statement. Directed rescue probabilities must not be confused with probabilities for distinct rescued triples.

### A.4 Broad Gaussian laws and full-support laws

Assume first that $z_i \sim N(0,\Sigma)$ iid with $\Sigma \succ 0$, while $h_i$ lies in the stated interval (the joint score law must be realizable). For $C_2, C_3$ from (14), define

$$\begin{aligned} p_0 &= \Pr\{N(0,2\Sigma) \in C_2\} > 0, \qquad R = \frac{T_3 + \log(1/\pi_{\min})}{\kappa}, \\ q_0 &= \Pr\{\|N(0,\Sigma)\| > R\} > 0. \end{aligned} \tag{19}$$

The first assertion follows because $C_2$ contains a ball. The set $C_3$ is contained in the radius-$R$ ball. Anderson's Gaussian translation inequality Anderson (1955) says that a translate of a centered symmetric convex ball has no larger centered Gaussian mass, giving the required uniform escape bound. No isotropic covariance is assumed. Replacing $w_j$ by $\Sigma^{1/2}w_j$ yields whitened constants $W_\Sigma, \kappa_\Sigma$ and

$$p_0 \ge \Pr\{\chi_d \le \sqrt{T_2}/W_\Sigma\}, \qquad q_0 \ge \Pr\{\chi_d > (T_3 + \log(1/\pi_{\min}))/\kappa_\Sigma\}.$$

These lower bounds can be conservative and can be very small. Low dimension alone does not guarantee a substantial finite-bank probability.

**Corollary 1** (No prescribed drift or tuned Gaussian center is necessary)**.** *Suppose the exact model (10), the classification readout (11), and $b > 0$ hold. If the attainable coordinates $(h_i, z_i) \in \mathbb{R}^{d+1}$ are iid with any nondegenerate Gaussian law, then there is a constant $c_* > 0$ such that*

$$\Pr\{N^{\text{actual}} \ge c_*\binom{n}{3}\} \longrightarrow 1.$$

*All witnessing pairs and triples may be restricted to a subbank whose singletons are correct. The constant depends on the score law, key geometry, and base margin; it is not uniform over Gaussian means or covariances.*

*Proof.* Choose $0 < \gamma < b/3$ and a compact interval $I = [h_-, h_+]$ in $((b+\gamma)/2, b-\gamma)$. Its Gaussian marginal probability is $\alpha > 0$. There are $\alpha n(1 + o_{\Pr}(1))$ accepted adapters. Conditional on acceptance, they are iid; the Gaussian conditional law of $z$ given $h \in I$ is a mixture of Gaussians with a common positive-definite Schur-complement covariance and means in a compact set. Pair sums therefore put a uniform positive probability on a fixed ball contained in $C_2$. Every such conditional Gaussian, whatever its mean, has a uniform probability of escaping every translate of the radius-$R$ ball, by Anderson's inequality. This remains true after mixing over $h \in I$. Apply Theorem 3 to the accepted subbank. Its size is linear in $n$ with high probability. The singleton condition follows from the choice of $I$. □

The same asymptotic conclusion holds if the joint coordinate law has a density strictly positive everywhere, without a Gaussian assumption. Indeed the accepted conditional $z$ law has full support, so its pair sum gives positive mass to $C_2$. If $C_3 \subset B(0,R)$ and $Z$ has full support and a density,

$$\sup_s \Pr\{Z \in B(s,R)\} < 1.$$

To see this, the probability as a function of $s$ is continuous and vanishes at infinity. Its maximum is attained, and is strictly less than one because every such ball has a nonempty open complement with positive probability. This proves a uniform $q_0 > 0$. The Gaussian version is useful because its constants are more quantitative. A density with full support in unrealizable coordinates is not an admissible assumption; rank of the actual score map must be checked. Nor does full support alone imply a useful value at $n = 8$.

**Connection to rank-one LoRA sampling.** For a fixed nonzero query input $x$, let $D_i = u_i x^\top / \|x\|^2$. Then $D_i x = u_i$ and each nonzero update has rank one. If the linear map from $u_i$ to $(h_i, z_i)$ is surjective, nondegenerate Gaussian $u_i$ induces the score law in Corollary 1. The needed condition is that the distracting-versus-mean-useful score direction adds an independent mode to the centered useful-score directions. This is a condition on keys and accessible query updates, not a consequence of LoRA rank alone. It provides a sampling class without hand-designed return points, but not a derivation of the distribution of optimizer-trained adapters.

### A.5 An explicit one-head example with non-negligible constants

Take two useful tokens with centered score vectors $+1, -1$ and equal weights. Then $K(z) = \log\cosh z$. Choose

$$b = 1.5, \quad h_i = 1, \quad \gamma = 0.25, \quad \eta = 0, \quad z_i \overset{\text{iid}}{\sim} N(0,4). \tag{20}$$

Every singleton has $\Lambda(i) \le -0.5$. A certified pair satisfies $|z_i + z_j| \le r_2$, and a certified full triple satisfies $|z_i + z_j + z_k| \ge r_3$, where

$$r_2 = \operatorname{arcosh}(e^{0.25}) = 0.73690459\ldots, \quad r_3 = \operatorname{arcosh}(e^{1.75}) = 2.43551088\ldots.$$

Writing $\Phi$ for the standard normal CDF,

$$p_2 = 2\Phi(r_2/(2\sqrt{2})) - 1 = 0.20554897\ldots,$$
$$q_0 = 2\Phi(-r_3/2) = 0.22331692\ldots.$$

The second bound holds uniformly over every pair sum, not only over certified bad pairs. The specified pair-to-triple direction has certified probability

$$p_\to = \int_{-r_2}^{r_2} \phi_{2\sqrt{2}}(s)\left[\overline{\Phi}((r_3 - s)/2) + \overline{\Phi}((r_3 + s)/2)\right] ds = 0.04801421\ldots \ge p_2 q_0. \tag{21}$$

The numerical values are evaluations in this conceptual model, not empirical estimates for trained LoRAs. Conditional on a *fixed certified pair* in an eight-adapter bank, its six remaining candidates include a rescuer with probability at least $1-(1-q_0)^6 = 0.78048558\ldots$. This is not an unconditional bound for the whole eight-adapter bank. The mean number of certified directed rescues is $3\binom{n}{3}p_\to$; an actual rescued triple may correspond to several directions.

For an exact standard attention realization, use query input $e_1 \in \mathbb{R}^3$, base query projection $W_{Q,0} = \operatorname{diag}(1,0,0)$, and keys

$$k_+ = \sqrt{3}(-\log 2, 1, 0)^\top, \quad k_- = \sqrt{3}(-\log 2, -1, 0)^\top, \quad k_B = \sqrt{3}(-b, 0, 1)^\top.$$

Give these tokens values $+1, +1, -1$ and use

$$D_i = (0, z_i, h_i)^\top e_1^\top. \tag{22}$$

The merged query is $(1, Z_S, H_S)^\top$ and the scaled attention scores are $(-\log 2 + Z_S, -\log 2 - Z_S, -b + H_S)$. Thus the log-odds is exactly $-b + H_S - \log\cosh Z_S$. Keys can be obtained from three fixed independent source tokens by a fixed key projection, as in cross-attention; a masked self-attention embedding is also possible. No claim that this is an entire pretrained CLIP architecture is intended. Each update and each merged update has rank at most one, so rank truncation retaining one direction is exact.

### A.6 A ROBUST TASK-MIXTURE REGIME: THREE QUARTERS OF TRIPLES

The Gaussian escape argument need not rely on a remote tail. A separated mixture supplies an alternative regime with large, explicit probabilities.

**Theorem 4** (Two preference groups produce abundant rescue)**.** *Assume the preceding one-distractor model, the drift bounds, and $h_+ \le b - \gamma - \eta$. Let*

$$z_i = S_i a + \epsilon_i, \quad S_i \in \{-1, +1\}, \quad \|\epsilon_i\| \le \delta,$$

*where no exact antipodality is required of individual updates. Suppose*

$$2W^2\delta^2 \le T_2, \qquad \min\{K(a), K(-a)\} - 3W\delta \ge T_3. \tag{23}$$

*Every opposite-type pair is a certified conflict, and* every *third adapter in the bank rescues it. If $n_+$ and $n_-$ are the type counts, then*

$$N^{\text{actual}} \ge \binom{n}{3} - \binom{n_+}{3} - \binom{n_-}{3} = \frac{n_+ n_- (n-2)}{2}. \tag{24}$$

*For iid types with $\Pr(S_i = +1) = p \in (0,1)$, the certified mixed-triple fraction tends to $3p(1-p)$. For $n \ge 3$,*

$$\Pr(N^{\text{actual}} > 0) \ge 1 - p^n - (1-p)^n.$$

*In particular the asymptotic lower fraction is $3/4$ at $p = 1/2$, and the occurrence bound at $n = 8$ is $127/128$.*

*Proof.* An opposite pair has sum norm at most $2\delta$, so Lemma 1 gives $K \le 2W^2\delta^2 \le T_2$. After any third adapter is added, the sum is within $3\delta$ of either $a$ or $-a$, and the Lipschitz bound gives $K \ge \min\{K(a), K(-a)\} - 3W\delta \ge T_3$. Count triples that contain both types. The law of large numbers gives the limiting fraction, and absence of both types has probability $p^n + (1-p)^n$. □

For $K(z) = \log\cosh z$, take $b = 1.5$, $\gamma = 0.25$, $\eta = 0$, $a = 3$, $\delta = 0.1$, and $h_i \in [0.95, 1.05]$. Then $T_2 = 0.15$, $T_3 = 1.9$, $2\delta^2 = 0.02 < T_2$, and $K(3) - 0.3 = 2.00932850\ldots > T_3$. All inequalities have slack. More general centers $m_+, m_-$ are allowed: replace $2\delta$ in the pair norm by $\|m_+ + m_-\| + 2\delta$, and $3\delta$ in the triple perturbation by $\|m_+ + m_-\| + 3\delta$, with centers $m_+$ and $m_-$. Thus the result persists under perturbations of cluster locations and weights. These explicit mixture probabilities are model conclusions, not measured frequencies in the paper's vision banks.

### A.7 SEVERAL DISTRACTING TOKENS AND EXACT NONLINEAR SCORE WIDTHS

For two token groups $G, B$, normalize base scores separately to $\pi^G, \pi^B$. Center each adapter's increments within each group. If their centered effects are $w_j^\top z_i$ and $v_l^\top z_i$, and $h_i$ is the difference between the two group means, then the exact identity is

$$\begin{aligned} \Lambda(S) &= -b + H_S - \mathcal{K}(Z_S), \qquad \mathcal{K} = K_G - K_B, \\ K_G(z) &= \log \sum_j \pi_j^G e^{w_j^\top z}, \qquad K_B(z) = \log \sum_l \pi_l^B e^{v_l^\top z}. \end{aligned} \tag{25}$$

A difference of log-partitions is not generally convex or nonnegative. The following geometric condition suffices for abundance, without either property:

$$\kappa_{GB} := \min_{\|u\|=1} \left( \max_{j \in G} w_j^\top u - \max_{l \in B} v_l^\top u \right) > 0. \tag{26}$$

It says that the convex hull of distracting score effects lies strictly inside the useful-effect hull, with a uniform directional gap. In this case

$$\mathcal{K}(z) \le \tfrac{1}{2} W_G^2 \|z\|^2, \qquad \mathcal{K}(z) \ge \kappa_{GB}\|z\| - \log(1/\pi_{\min}^G).$$

The first inequality uses $K_B \ge 0$; the second retains a maximizing useful term and bounds $K_B$ by its support function. Thus $C_2$ contains a ball and $C_3$ is bounded as before. Theorem 3 and the Gaussian sampling argument apply with $K$ replaced by $\mathcal{K}$. Automatic singleton feasibility requires an additional lower bound on $\mathcal{K}$; for example, $K_G \ge K_B$ suffices. One sufficient route to this last inequality is that the distracting-effect distribution is a mean-preserving contraction of the useful-effect distribution, by Jensen's inequality. Arbitrary token geometry need not satisfy (26); softmax alone does not imply rescue.

**Functional dimension before the softmax.** Let $r$ have the generating law $\mu$ of centered useful-token score increments on a fixed query or declared probe. In its weighted-mean-zero space define

$$d_{\mathrm{att}}(\epsilon) = \min\{\dim E : \operatorname*{ess\,sup}_{r\sim\mu} \|r - P_E r\|_\infty \le \epsilon\}. \tag{27}$$

Here $P_E$ is a fixed weighted orthogonal projection; $E$ is a population subspace or is selected on an independent pilot sample. Because log-sum-exp is 1-Lipschitz in the sup norm,

$$\left|\log \sum_j \pi_j e^{\sum_i r_{ij}} - \log \sum_j \pi_j e^{\sum_i (P_E r_i)_j}\right| \le |S|\epsilon.$$

For one distracting token this contributes at most $3\epsilon$ to $\eta$; for two projected groups the two errors add. The nonlinearity is kept exact within the retained score space. This is not an approximation of the output by its base Jacobian. A spectral participation ratio alone does not establish (27). A nonzero mean preference $r_i = a + Wz_i$ produces $|S|a$ inside the log-sum; it must be retained, or explicitly bounded as an additional score error, rather than silently mean-centered away.

If both query and key projections change, the exact score includes

$$\frac{(q_0 + \sum_i \delta q_i)^\top (k_{0j} + \sum_i \delta k_{ij})}{\sqrt{d_h}} = \beta_j + \sum_i \ell_{ij} + \frac{(\sum_i \delta q_i)^\top (\sum_i \delta k_{ij})}{\sqrt{d_h}}.$$

Under $\|\delta q_i\| \le u$ and $\|\delta k_{ij}\| \le v$, the last term is at most $|S|^2 uv/\sqrt{d_h}$ in magnitude. Log-odds contains two group log-sums, so a common per-token score error $\epsilon_s$ contributes at most $2\epsilon_s$. This bound preserves arbitrary softmax nonlinearity, but does not by itself bound upstream hidden-state changes, value/output updates, other heads, or downstream loss. Those effects need actual margin controls.

### A.8 What is required for actual multi-task losses

The classification construction is exact but simplified. In a general network, "useful" and "distracting" must be justified by the actual downstream map, not assigned solely from attention weights.

**Proposition 3** (Nonlinear downstream margin transfer)**.** *Fix an input and all other contextual state. Let $\Psi$ be the true downstream correct-class margin as a function of this head's value mixture. Suppose $\Psi \ge m > 0$ on the convex hull of useful values and $\Psi \le -m$ on the convex hull of distracting values. If $\Psi$ is $L$-Lipschitz on the relevant mixture segments, their cross-group diameter is at most $D$, and other perturbations change the final margin by at most $\zeta$, then $p_B \le \epsilon$ certifies a positive final margin and $p_B \ge 1 - \epsilon$ certifies a negative final margin whenever*

$$m > LD\epsilon + \zeta.$$

*In particular use $\epsilon = \sigma(-\gamma)$ with the log-odds certificates.*

*Proof.* A mixture with distracting mass at most $\epsilon$ is within $D\epsilon$ of some point in the useful convex hull. Apply the Lipschitz bound and the other-perturbation bound. The reverse case is identical. □

For an affine downstream readout, endpoint value margins give the convex-hull margin condition. For a nonlinear downstream network, correct signs only at individual token values do not suffice: mixtures must also be controlled. Neither condition requires the complete network to be close to its base tangent.

For a population classification task, suppose the routing certificates hold on an input region $E$ of probability $r$, the base is correct there, the pair is wrong there, and the triple is correct there. If the pair's improvement over the base outside $E$ is at most $v$ in total probability and the triple's excess error outside $E$ is at most $u$, then

$$F(ij) - F(\varnothing) \ge r - v, \qquad F(ijk) - F(\varnothing) \le u. \tag{28}$$

An excess-error cap $s$ is crossed in the rescue direction whenever $u \le s < r - v$. Other tasks must each satisfy their own cap after the third adapter is added. Routing success on one token or one task cannot certify an eight-task feasible triple. Score and downstream conditions can instead be imposed jointly on a declared multi-task input family; their joint sampling mass, not a product of unjustified taskwise probabilities, enters the abundance bound. Compression must preserve the relevant actual margins or be included in the deployed map. It is exact only in the shared-right-factor construction above, not in a general trained bank.

**A uniform multi-task escape bound without task independence.** As an optional sufficient condition, suppose each task's bad region in a projection $P_t Z$ is contained in a ball of radius $R_t$, and a candidate's common preference vector is Gaussian with covariance $\Sigma$. Choose a unit vector $u$ for which $a_t = P_t \Sigma^{1/2} u \neq 0$ for every task; such a vector exists for finitely many nonzero maps. Write a standard Gaussian as $ru + v$, where $r \sim N(0,1)$ is independent of the orthogonal component $v$. Conditional on $v$, entry into task $t$'s ball confines $r$ to an interval of length at most $2R_t/\|a_t\|$, whatever the common mean or prior pair sum. The union of all bad intervals has length at most $2A$, with $A = \sum_t R_t/\|a_t\|$. A standard normal places at most $2\Phi(A) - 1$ mass on any measurable set of this length, because its density is symmetric decreasing. Hence simultaneous escape has probability at least $2\overline{\Phi}(A) > 0$. This gives a possible $q_0$ for joint routing certificates. It can be extremely small and does not replace the downstream/population margin conditions.

### A.9 Interpretation and limits

The mechanism gives a direct route from token geometry to a rescue region, and from score-distribution mass to abundance. In the fixed-margin regime, the important statement is not only that a bank has a witness: a positive fraction of its pairs can each have a linear number of rescuers. The exact two-head construction in Appendix C gives a separate illustration. The abundance theorem does not require minimal rescue: it studies ordinary pair-to-triple rescue and permits all singletons to be safe. That is the event needed to contradict downward closure.

The broad Gaussian and full-support corollaries are asymptotic results for an explicit class of *attainable attention-score laws*. They are not proofs that trained LoRA coordinates are Gaussian or have full support. Uniformly small sampling constants can make an asymptotic claim irrelevant for a small bank. Useful-token redundancy, distracting drift, and diversity of adapter preferences are substantive properties to test. One-token monotone attention, identical value embeddings, or a score law confined to a harmless set supplies immediate exceptions to any architecture-only universality claim.

The experiments in Section 4 adapt query, key, value, and output matrices across many blocks, use rank truncation, and enforce several classification-error caps. Connecting those experiments to this theorem requires the exact score charts, their approximation errors, and the declared downstream margins. A low-dimensional base-output Jacobian is not enough. A promising experimental diagnostic is to compare full-softmax and frozen/tangent-softmax coalition predictions while measuring $h_i$, centered preferences, and task-cap margins on independently selected probes. Such ablations identify a candidate mechanism; they do not by themselves establish population safety or an optimizer's output law.

The probability tools are standard, with attribution above. Existing nonlinear Transformer task-vector theory includes conceptual learning guarantees Li et al. (2025); no claim to be the first nonlinear theory of task arithmetic is made. The result concerns a specific exact softmax mechanism, its sampling conditions, and quantitative abundance. The accompanying program checks the explicit network identities and numerical examples; it neither replaces the proofs nor evaluates pretrained banks.

## B Protocol and complete diagnostics for the new frequency study

This study measures ordinary pair-to-triple rescue in six newly trained banks. It is separate from the earlier E7/E8 complete-cube and selection studies in Section 4.2 and Appendix G. All frequency denominators in Section 4 refer to the new cohort unless explicitly marked otherwise. The study enumerates a fixed local domain; no discovery procedure selects its reported audit triples.

### B.1 Training, evaluation, and data dependence

The task order is MNIST, SVHN, GTSRB, EuroSAT, Stanford Cars, DTD, RESISC45, and SUN397. Dataset sources and the CLIP classification-head construction are as in Appendix M. Each backbone has three new banks, indexed $b = 0, 1, 2$, with zero-based task index $t$. Holdout and training-image sampling seeds are $2026092600 + 100b + t$ and $2026094600 + 100b + t$, respectively, paired across

backbones. Adapter training seeds are $2026092700 + 100b + t$ for B/32 and $2026093700 + 100b + t$ for B/16. Each split contains 2,048 training examples for each task except EuroSAT (1,932) and DTD (1,109), totaling 15,329. Discovery and audit contain 256 and 512 images per task, respectively. Each phase contains unique RGB images, global training–holdout RGB-hash overlap is zero within a split, and discovery–audit overlap is zero within a split.

The public image pools are repartitioned rather than used as official benchmark train/test splits. Repartitioning does not produce independent datasets: the three audit-set pairwise overlaps are 395, 347, and 369 images, and the corresponding training overlaps are 4,338, 4,313, and 4,302. The new holdouts share 2,715, 2,755, and 2,674 images with the earlier manifest. Both backbones use the same split for a given bank index. We therefore report finite-bank counts, not binomial confidence intervals treating banks or directions as independent replications.

All 48 adapters use LoRA rank 8, LoRA alpha 8, and dropout 0.1 on the query, key, value, and output matrices in all 12 vision blocks (48 matrices). Training uses AdamW, learning rate $10^{-4}$, weight decay 0.01, batch size 32, a warmup of at most 50 steps followed by cosine decay, gradient clipping at 1, and 1,200 steps. The last checkpoint is used. Training uses BF16 autocast with FP32 parameters and logits; merging, SVD, and evaluation use FP32. The frozen primary scale is 0.5, inherited from prior exploratory work. The dense update is $0.5 \sum_{i \in S} D_i$ and the rank-8 update is its per-layer rank-8 SVD truncation; here $D_i$ denotes the trained update before this common merge scale.

Every bank evaluates 93 coalitions through size three. Audit errors determine the eight relative-base cap constraints in (6). Caps of 1, 2, 3, and 5 pp and buffers of 0, 0.5, and 1 pp are reported as predefined sensitivities. Comparisons use a numerical tolerance of $10^{-12}$. The primary event requires some pair task to exceed its cap and the triple to pass all eight caps; it does not require singleton feasibility, pairwise improvement on every task, or failure of all three contained pairs. The strict-minimal event additionally requires all six nonempty proper subsets to be infeasible, with the same buffer convention when applicable.

Table 4: New banks at scale 0.5, empirical cap 2 pp and zero buffer. Directions use all 168 pair–addition choices per bank. Coverage uses the number of infeasible pairs in that bank as its denominator.

| | | Rescue directions | | Rescuable / bad pairs | |
|---|---|---|---|---|---|
| Backbone | Bank | Dense | Rank-8 | Dense | Rank-8 |
| B/32 | 0 | 24/168 | 27/168 | 11/12 | 11/12 |
| B/32 | 1 | 31/168 | 30/168 | 14/14 | 12/14 |
| B/32 | 2 | 23/168 | 21/168 | 16/17 | 14/16 |
| B/16 | 0 | 11/168 | 7/168 | 9/19 | 6/17 |
| B/16 | 1 | 22/168 | 22/168 | 13/18 | 13/18 |
| B/16 | 2 | 15/168 | 19/168 | 10/16 | 11/18 |

### B.2 Margins, scale, and bank composition

Table 4 gives the exact bank-level counts underlying Figure 1. Tables 1 and 2 use the original 96 dense or 95 rank-8 bad pairs even when a positive buffer is imposed. At 0.5 pp, buffered rescuer coverage is 40/96 dense and 40/95 rank-8; at 1 pp it is 24/96 and 26/95. Strict-minimal rescue is absent at the primary 2 pp cap. At a 1 pp cap and zero buffer there is one such triple dense and two at rank eight, but these events disappear at a 0.5 pp buffer. They do not support a claim of prevalent strict minimality.

Secondary scales 0.25 and 0.75 were evaluated only for the prespecified bank 0 on each backbone, conditional on the experiment's time budget. Table 5 shows that scale effects differ across backbones. There is no monotone trend that can be extrapolated to other banks.

We also enumerate subbanks inside the six fixed new banks and ask whether each contains any primary rescue direction. Table 6 describes this coverage structure. Larger subbanks expose more candidate directions; the subbanks overlap heavily. Their coverage is neither an independent-bank experiment nor a test of the asymptotic iid sampling theorem.

Table 5: Scale sensitivity in the two prespecified bank-0 models, at the 2 pp cap and zero buffer. Each cell counts rescues among 168 audit directions.

| Backbone | Representation | Scale 0.25 | Scale 0.5 | Scale 0.75 |
|---|---|---|---|---|
| B/32 | Dense | 5 | 24 | 15 |
| B/32 | Rank-8 | 4 | 27 | 11 |
| B/16 | Dense | 22 | 11 | 6 |
| B/16 | Rank-8 | 20 | 7 | 6 |

Table 6: Subbanks containing at least one rescue, within the observed new banks. The denominator is $6\binom{8}{m}$ for subbank size $m$.

| Subbank size $m$ | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| Dense | 99/336 | 280/420 | 299/336 | 164/168 | 48/48 |
| Rank-8 | 99/336 | 274/420 | 287/336 | 160/168 | 48/48 |

### B.3 Contextual gain and stability diagnostics

For each rescue direction, take the task with the largest pair cap violation as the bottleneck task (with deterministic task-order tie breaking). Its contextual gain is the pair error minus the triple error. Its singleton gain is the base error minus the rescuer's singleton error. The contextual gain is larger in 84/126 dense and 75/126 rank-8 events. This comparison conditions on observed rescue and uses the same audit images; it is descriptive rather than causal. In 65/126 events in each representation, the rescuer's training task is absent from the set of tasks on which the pair violates a cap. The rescuer's singleton is infeasible in 26/126 dense and 24/126 rank-8 events.

Applying the primary definition to the disjoint discovery images yields 98 positives per representation. Audit retains 47, loses 51, and adds 79. The identical aggregate numbers for dense and rank-8 do not imply identical events. A paired image bootstrap uses 1,000 resamples per bank, sharing resampled image indices across coalitions within each task. The saved 2.5%–97.5% quantiles describe fixed-bank image-resampling sensitivity; for new B/32 bank 0 at rank eight, the rescue-direction range is 13–41.025 around the observed count of 27. These quantiles are not confidence limits for a distribution of newly trained banks, and the repeated directions cannot be treated as independent data.

**Quadratic set-function diagnostic.** For each task, fit the unique quadratic to the 37 empty, singleton, and pair error measurements on the same audit images, then predict triple errors. Using actual pair labels and predicted triple feasibility gives TP/FP/FN/TN counts of 81/12/45/870 dense and 89/16/37/866 rank-8, over all 1,008 directions. This is a held-out-coalition diagnostic with reused images, not independent prediction or a candidate-selection benchmark. The exact gain decomposition separates singleton gain, two pairwise extra gains, and a third-order set-function remainder as in (40). That remainder is not a third derivative with respect to network parameters.

### B.4 Local functional geometry: completed negative diagnostics

The two prespecified bank-0 probes use the first 128 audit images per task. We compute base-model logit Jacobian–vector products (JVPs), a cross-entropy generalized Gauss–Newton (GGN) approximation including its first-order term, and actual coalition outputs. Classification predictions are obtained from linearized logits; the GGN approximation predicts cross-entropy (CE). Their thresholds and outputs must not be conflated. The CE diagnostic's 0.1 nat normalization is not the classification-error cap.

Table 7 shows substantial errors in finite-amplitude predictions. Maximum CE errors range from about 3.06 to 5.10 nats. Finite-difference validation covers 64 task–adapter pairs per backbone; the maximum relative JVP discrepancies, taking the better of two step sizes, are approximately 0.20% and 0.63%. Accurate local derivatives therefore do not imply accurate coalition predictions at the merge scale.

Table 7: Local approximation diagnostics on 128-image-per-task probes. TP/FP/FN/TN compare rescue labels from linearized logits with actual labels on the same probes (168 directions per row). CE MAE compares GGN with actual cross-entropy. These probe labels differ from the 512-image primary audit.

| Backbone | Actual representation | TP | FP | FN | TN | CE MAE (nat) |
|---|---|---|---|---|---|---|
| B/32 | Dense | 4 | 20 | 4 | 140 | 0.357 |
| B/32 | Rank-8 | 8 | 16 | 5 | 139 | 0.346 |
| B/16 | Dense | 1 | 25 | 8 | 134 | 0.462 |
| B/16 | Rank-8 | 1 | 25 | 8 | 134 | 0.451 |

The first four functional principal components retain 90.85% of the trace for B/32 and 90.33% for B/16, but their maximum residuals relative to the largest feature norm are 30.27% and 37.20%. These are PCA residuals under the declared empirical normalization, not optimal worst-case subspace widths. Eight observed adapters spanning at most eight dimensions is not evidence that the adapter-generating distribution has low intrinsic dimension. Neither this geometry nor the tangent predictions measure the exact useful-token preferences and distracting drift in Theorem 1.

### B.5 Provenance and verification

The saved protocol fixes the task order, masks, scales, seeds, training settings, and source hashes. Its analysis includes frequencies, individual directions, decompositions, discovery–audit replication, bootstrap summaries, and subbank counts. The completed run covers 32 source conditions and 58 condition–phase analyses; primary counts use only the six new banks at scale 0.5 on audit.

The original execution checked logits, labels, errors, and CE; the largest CE recomputation discrepancy was $1.81 \times 10^{-6}$. Frozen source, manifest, and classification-head hashes remained unchanged. A separate export of per-task error and CE means from the original prediction caches supports an offline recheck of all 480 new-bank frequency rows, their denominators, direction labels, strict-minimal counts, and exact decompositions. The largest decomposition discrepancy was $4.44 \times 10^{-16}$. GGN positive semidefiniteness, decomposition identities, and functional residual calculations were also checked. The final offline check did not rerun GPU inference or decode the original images; image-hash and logits consistency rely on the saved execution records.

The companion materials include the frozen frequency-analysis tables, exported per-task means, and the offline checker `verify_final.py`. The numerical attention checker is `check_softmax_competition.py` in the softmax theory bundle. The accompanying README identifies the exact source locations. These numerical checks support identities and table transcription; they do not replace the proofs, an independent reproduction, or population-risk analysis.

## C A contrasting mechanism with positively parallel updates

The ordinary-rescue mechanism in Section 3.1 uses cancellation of useful-token score preferences. The following separate example shows that rescue can also occur with positively parallel parameter updates. It concerns strictly minimal rescues, not the frequency study's primary event. We retain only its exact construction and margin stability; no trained-bank distribution is inferred from it.

### C.1 Exact softmax rescue without adapter cancellation

Use two tokens $x_1 = e_1, x_2 = e_2 \in \mathbb{R}^2$, two scalar attention heads, and evaluate their output at query token $x_1$. Stacking head projections as rows, set

$$W_{Q,0} = \begin{pmatrix} -1 & 0 \\ -2 & 0 \end{pmatrix}, \qquad W_K = W_V = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad w_O = (1, -1). \tag{29}$$

These are ordinary scaled dot-product softmax heads with head dimension one; concatenation is followed by a signed linear readout Vaswani et al. (2017). Only the query projection is adapted.

Adapter $i$ has

$$D_i = a_i \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = a_i \begin{pmatrix} 1 \\ 1 \end{pmatrix} e_1^\top, \qquad a_i > 0. \tag{30}$$

Thus every adapter is rank one, all realized updates are positively parallel, and all merge coefficients are fixed at one. For a coalition $S$, write $t_S = \sum_{i \in S} a_i$. The two heads have scores $(t_S - 1, 0)$ and $(t_S - 2, 0)$ and values $(1, 0)$. Their exact readout is

$$h(t_S) = \sigma(t_S - 1) - \sigma(t_S - 2), \qquad \sigma(u) = (1 + e^{-u})^{-1}. \tag{31}$$

Choose regression target $h(0)$ and define

$$g(t) = h(t) - h(0), \qquad F(S) = g(t_S)^2. \tag{32}$$

**Proposition 4** (Nonlinearity creates a rescue absent from the tangent model)**.** *For $a_1 = a_2 = a_3 = 1$, the base and the full triple have zero loss, while all six nonempty proper subsets have loss $\delta_*^2 > 0$, where*

$$\delta_* := g(1) = g(2) = \frac{1}{2} - 2\sigma(-1) + \sigma(-2) = 0.081320079282\ldots. \tag{33}$$

*Consequently the triple is a minimal rescue at every squared-loss cap $0 \le \tau < \delta_*^2$. In contrast, the base-linearized output gives*

$$F_{\tan}(S) = c^2 t_S^2, \qquad c := g'(0) = \sigma'(1) - \sigma'(2) = 0.091618347838\ldots > 0, \tag{34}$$

*which is strictly increasing under addition of any positive adapter. Its cap families are all downward closed.*

*Proof.* The identity $\sigma(-u) = 1 - \sigma(u)$ implies $h(3-t) = h(t)$, so $g(0) = g(3) = 0$ and $g(1) = g(2)$. Moreover, $\sigma'(u) = 1/(4\cosh^2(u/2))$ is even and strictly decreases with $|u|$. It follows that

$$h'(t) > 0 \quad (t < 3/2), \qquad h'(t) < 0 \quad (t > 3/2).$$

Thus $g(t) > 0$ for $0 < t < 3$, proving the claimed gap and the exact subset losses. Differentiating at zero gives (34). Since all $a_i > 0$, every strict superset has larger $t_S$ and hence larger tangent loss. □

The full update equals $3\binom{1}{1}e_1^\top \neq 0$; this is not cancellation of LoRA products or their base-Jacobian images. The two softmax heads enter their transition regimes at different cumulative update values. Their difference first increases and then decreases, returning to its base value at a nonzero update. The signed readout is an ordinary network operation, but the nonmonotone dependence on the adapter sum is created by the softmax forward map.

*Remark* 1 (Nonlinear forward geometry is not Boolean interaction order). At the eight Boolean points with $a_1 = a_2 = a_3 = 1$, the same exact nonlinear network happens to satisfy

$$F(z) = \delta_*^2\big(z_1 + z_2 + z_3 - z_1 z_2 - z_1 z_3 - z_2 z_3\big).$$

Thus nonlinearity-induced failure of the base tangent model does not imply an irreducible third-order Boolean loss term. The origin of a rescue and the interaction order of its finite coalition table are different questions.

**Proposition 5** (An exact classification-error witness)**.** *Give the correct class logit $\gamma - g(t)$ and the other class logit zero, where $0 < \gamma < \delta_*$. At $t = 0, 1, 2, 3$, respectively, the classification errors are $0, 1, 1, 0$. Therefore the same triple is a minimal rescue for classification error at every cap $0 \le \tau < 1$ relative to its zero-error base. This remains true in a sufficiently small open neighborhood of the network weights, the three coefficients, and the input-token pair. Furthermore, for any sufficiently small fixed $\eta > 0$, there is a $\gamma > 0$ such that* every *triple in* every *bank with $a_i \in [1 - \eta, 1 + \eta]$ is a minimal classification-error rescue. A sufficient condition, with $0 < \eta < 1/6$, is*

$$g(3 - 3\eta) < \gamma < \min\{g(1 - \eta), g(2 + 2\eta)\}. \tag{35}$$

*For example, $\eta = 0.1$ and $\gamma = 0.04$ satisfy this condition.*

*Proof.* The four correct-class margins are $\gamma$, $\gamma - \delta_*$, $\gamma - \delta_*$, and $\gamma$. All are separated from zero. There are only eight coalitions, and all logits depend continuously on the stated variables; intersect their finitely many margin-preserving neighborhoods. For any input distribution supported inside a sufficiently small such input neighborhood, the same error pattern holds pointwise and hence in population. For the interval statement, singletons and pairs lie in $[1 - \eta, 2 + 2\eta]$, where the hump shape gives the right-hand residual lower bound in (35). Triple sums lie in $[3 - 3\eta, 3 + 3\eta]$, where $g$ decreases, and hence their residual is at most $g(3 - 3\eta)$. Thus all subset logits have the required signs. The interval condition is nonempty for small $\eta$ because its left end tends to zero and its right end tends to $\delta_* > 0$. This is a constructed classification distribution, not a guarantee for a pretrained model or a benchmark distribution. □

The equality $g(3) = 0$ is also stable as a *nearby root*: since $g'(3) = -c \neq 0$, the implicit function theorem gives a return point near 3 under sufficiently small smooth perturbations of this scalar-output network. The positive proper-subset gaps survive at the correspondingly shifted coefficient centers. For fixed positive caps, continuity alone suffices for label stability. Neither statement asserts persistence of a many-output exact return without its corresponding rank conditions.

# D Additional consequences and information boundaries

The following statements retain the original representation, information, and selection analysis. They describe consequences and possible uses of rescue; the abundance claim is established separately in Section 3.1.

## D.1 Consequences and boundary cases

**Corollary 2** (Representation barrier)**.** *For Theorem 2, the unique largest sound downward-closed family is* $\mathcal{A}^{\star}_{\mathrm{dc}} = \{S : |S \cap [m]| \leq 1\}$. *Every sound policy formed by prohibiting fixed adapter subsets, including ordinary conflict hypergraphs, has utility at most two, regardless of its information or computation.*

The corollary applies to fixed-subset exclusion; graph-based algorithms with other selection rules can fall outside it.

**Corollary 3** (Pairwise identifiability)**.** *The empty, singleton, and pair losses of Theorem 2 uniquely determine $L$ within the quadratic class. Retaining their signed mixed differences reconstructs the entire feasible family and recovers* $\mathrm{OPT} = n$.

The model-class premise is necessary: $G_\gamma + \lambda z_1 z_2 z_r$ has the same pair observations but makes the full coalition infeasible for $\lambda > \gamma$.

**Corollary 4** (Rank-one realization and minimality)**.** *The exact single-layer realization (5) uses output dimension two, which is minimal for squared loss with an affine residual on one example. In that class, $\tau = \gamma + m^2/8$ is the smallest cap for the exact unscaled identity in any output dimension. Linearly independent rank-one effects realize the same excess risk in $n + 1$ outputs at a larger cap.*

The two-dimensional construction uses identical ordinary updates. For fixed $\gamma$, the rescuer norm grows as $O(n)$ and the cap as $O(n^2)$.

**Corollary 5** (Robustness with positive margins)**.** *Let $\mathcal{F}_\gamma = \{S : G_\gamma(\mathbf{1}_S) \leq 0\}$. If $\|\widehat{G} - G_\gamma\|_\infty \leq \varepsilon < 1$ and $\varepsilon \leq \gamma$, then*

$$\{S : r \in S\} \subseteq \{S : \widehat{G}(S) \leq 0\} \subseteq \mathcal{F}_\gamma.$$

*Signed selection still attains $n$, while every true-sound downward-closed policy has utility at most two. Under the quadratic assumption, observation errors at most $\delta$ yield the sharp uniform interpolation-error bound* $(2n^2 - 4n + 1)\delta$.

Appendix H.3 proves this statement and distinguishes conservative certification, which accepts the full set when $2\varepsilon \leq \gamma$, from boundary-label stability.

### D.2 THE INFORMATION BOUNDARY OF SIGNED SELECTION

Let $F$ denote excess loss, with feasibility $F(S) \le 0$, and fix $1 \le q < n$. An order-$q$ algorithm knows the model class and obtains instance-specific information only from $\mathcal{O}_q(F) = (F(S) : |S| \le q)$. Adaptive queries to these sets are allowed; parameter inspection is not. Let $\mathcal{P}_p$ be the Boolean multilinear polynomials of degree at most $p$, and put

$$\mathcal{H}_q = \{F : F(S) \le 0 \text{ for all } |S| \le q\}, \qquad \mathcal{C}_{p,q} = \mathcal{P}_p \cap \mathcal{H}_q.$$

Write $\mathrm{OPT}(F) = \max_{F(S)\le 0} |S|$. Define

$$V_q^\star(\mathcal{C}) = \sup_{\mathcal{A}} \inf_{F\in\mathcal{C}} \frac{\mathbb{E}|\mathcal{A}(\mathcal{O}_q(F))|}{\mathrm{OPT}(F)}, \qquad \star \in \{\det, \mathrm{as}\}, \tag{36}$$

where algorithms must be safe for every $F \in \mathcal{C}$, deterministically (det) or almost surely under randomization (as).

**Theorem 5** (Sharp interaction-order information dichotomy)**.** *For $0 \le p \le n$,*

$$\boxed{V_q^{\det}(\mathcal{C}_{p,q}) = V_q^{\mathrm{as}}(\mathcal{C}_{p,q}) = \begin{cases} 1, & p \le q, \\ q/n, & p \ge q+1. \end{cases}} \tag{37}$$

*Every $F \in \mathcal{P}_q$ is reconstructed exactly by its $N_q(n) = \sum_{j=0}^{q} \binom{n}{j}$ low-order observations. This count is worst-case optimal for deterministic exact reconstruction on $\mathcal{P}_q$, even with arbitrary-size queries. The minimax value $q/n$ is also attained on a class of depth-$(q+1)$, width-two squared-loss LoRA systems with common input, target and base weights, nonzero unit-norm rank-one updates at every layer, and zero loss on all observed coalitions.*

*Proof sketch.* Möbius inversion gives exact reconstruction when $p \le q$; unrestricted computation permits exhaustive optimal selection. For $p \ge q+1$, the losses 0 and $E_{q+1}(z) = \sum_{|e|=q+1} \prod_{i\in e} z_i$ agree on every observed set, but have optima $n$ and $q$. Safety on the latter forces the same output distribution to select at most $q$ adapters on the former. Returning a fixed $q$-set attains the bound because $\mathcal{C}_{p,q} \subseteq \mathcal{H}_q$. Appendix I proves all claims and realizes the indistinguishable pair 0 and

$$F_{\mathrm{bad}}(S) = \left[\prod_{\ell=0}^{q} (|S| - \ell)\right]^2$$

in the stated LoRA architecture. This squared construction has Boolean degree $\min(n, 2q+2)$; $E_{q+1}$ supplies the abstract degree-$(q+1)$ boundary. Depth $q+1$ is minimal for a nonzero deep-linear squared loss with an all-zero order-$q$ transcript, regardless of width or update rank. A raw-loss cap $1/2$ preserves the hard pair with strict margins.

The hard pair gives the upper bound; a common feasible $q$-set gives the matching lower bound. At $q = 2$, both theorems give $2/n$, but for different reasons: Theorem 2 restricts the accepted family, whereas Theorem 5 restricts what the algorithm observes. Both allow unrestricted computation.

### D.3 RESCUE, INTERACTION ORDER, AND PERMANENT EXCLUSION

For a fixed task and scale, write $F(ij) = F(\{i,j\})$ and define

$$C_{ij} = F(ij) - \min\{F(i), F(j)\}, \qquad R_{ij;k} = F(ij) - F(ijk). \tag{38}$$

The scan substitutes empirical means for $F$, calls a pair conflicting when $C_{ij} > 0.02$, and calls a triple a strong rescue when also $R_{ij;k} > 0.02$, in cross-entropy nats. These relative contrasts do not assert population-cap satisfaction. A permanent edge $ij$ would forbid every containing coalition.

To isolate interaction order, let $D_0 = F(ij) - F(i) - F(j) + F(\varnothing)$, $D_k = F(ijk) - F(ik) - F(jk) + F(k)$, and $Q = D_0 - D_k$. Then $Q$ is the negative third-order mixed difference, symmetric in $i, j, k$. In contrast, for a signed quadratic $P_2 = c_\varnothing + \sum_i c_i z_i + \sum_{i<j} c_{ij} z_i z_j$,

$$R^{(2)}_{ij;k} = -c_k - c_{ik} - c_{jk}, \qquad Q^{(2)} = 0. \tag{39}$$

Thus rescue can arise from pairwise interactions with the rescuer. Even improvement beyond its singleton gain does not establish $Q > 0$: with $G_0 = F(\varnothing) - F(k)$, $G_i = F(i) - F(ik)$, and $G_j = F(j) - F(jk)$, inclusion–exclusion gives

$$R_{ij;k} - G_0 = (G_i - G_0) + (G_j - G_0) + Q. \tag{40}$$

A smaller $D_k$ attenuates a positive interaction only when $D_0 > 0$; positive $Q$ alone need not imply positive $R$.

### D.4 Literal conflicts and low-rank realizability

To express a conflict that disappears in a larger coalition, allow required absences. For example,

$$I(z) = \kappa z_1 z_2 (1 - z_3), \qquad \kappa > 0, \tag{41}$$

penalizes the pair only when adapter 3 is absent. This example is cubic; the rescue in Theorem 2 is quadratic.

**Definition 2** (Literal conflict)**.** A pattern $(P, N)$ consists of disjoint sets of required presences and absences. Its face is $\mathcal{F}(P, N) = \{S : P \subseteq S,\ S \cap N = \varnothing\}$. It is a task-$t$ conflict when $\inf_{S \in \mathcal{F}(P,N)} F_t(S) > \tau_t$. Equivalently, it is a hyperedge on literal vertices $\{i^+ : i \in P\} \cup \{j^- : j \in N\}$, with exactly one literal chosen for each adapter.

A forbidden pattern gives the valid no-good inequality

$$\sum_{i \in P} z_i + \sum_{j \in N} (1 - z_j) \leq |P| + |N| - 1. \tag{42}$$

This standard representation is related to mixed conflict hypergraphs (Shapoval and Lee, 2021). Certifying it requires a bound on every completion of the face.

For literal penalties of maximum scope $d$ and vertex incidence $\Delta$, Theorem 7 gives a depth-$(d+1)$ realization with pairwise Frobenius-orthogonal updates of per-layer rank at most $\max\{1, \Delta\}$. Disjoint scopes need only rank one. Corollary 6 shows that sub-$k$ loss queries cannot identify pure $k$-way penalties. Low rank and parameter orthogonality therefore do not imply low interaction order.

**Proposition 6** (Why unsigned conflicts are insufficient)**.** *A nonempty family of feasible subsets is the independent-set family of an ordinary hypergraph on the original adapter vertices if and only if it is downward closed. For $0 \leq \tau < \kappa$, the feasible family of (41) is not downward closed, so no such ordinary hypergraph represents it exactly.*

Literal patterns represent arbitrary feasible families by forbidding their infeasible full assignments. Quadratics already represent some non-downward-closed families, including Theorem 2.

## E Signed interactions and context-valid certificates

For a single task, define the anchored mixed difference and order-$q$ polynomial

$$c(e) = \sum_{V \subseteq e} (-1)^{|e| - |V|} F(V), \qquad P_q(S) = \sum_{\substack{e \subseteq S \\ |e| \leq q}} c(e). \tag{43}$$

The identity $F(S) = \sum_{e \subseteq S} c(e)$ is classical Möbius inversion, related to cooperative-game interaction decompositions (Sundararajan et al., 2020). Keep coefficients signed and task-specific: taking a maximum over tasks before expansion changes the function and can introduce interaction terms.

To certify an exclusion, we need a uniform radius $|F_t(S) - \widehat{P}_{t,q}(S)| \leq \epsilon_t$ on the candidate domain. Theorem 8 bounds the tail through next-order differences across contexts; small differences at the empty coalition do not suffice. Appendix K adds simultaneous bounded-loss probe errors and the mass of discarded coefficients. The resulting radius can be too large to certify a decision.

The local-moment relaxation (63) contains every completion of a literal face $\mathcal{F}(P, N)$ and preserves signed cancellations (Sherali and Adams, 1990). Subtracting $\epsilon_t$ from a verified lower bound gives $\underline{F}_t(P, N)$ in (64). If this exceeds $\tau_t$, every completion violates the cap, so Theorem 9 certifies the exclusion (42). Cuts chosen adaptively on the same uniform event remain sound. The appendix gives the relaxation and rational dual verification. Without a valid radius, the fit can only propose candidates.

Table 8: The two certificates require different evidence; empirical fit alone supplies neither.

| | Structural exclusion | Final-model risk audit |
|---|---|---|
| Certifies | A whole presence–absence face | Only the frozen deployed candidates |
| Needs | Uniform approximation/tail radius | Independent bounded-loss audit data |
| Compression | Must be accounted for explicitly | Evaluates the compressed model itself |
| May conclude | No valid cut is established | No candidate can be certified |

## F Applying the analysis to selection

**Selection.** Given certified radii $\epsilon_t$, maximize $U(S)$ subject to $\widehat{P}_{t,q}(S) + \epsilon_t \le \tau_t$, the budget, and certified literal cuts. For additive utility, standard product constraints $y_e \le z_i$ $(i \in e)$ and $y_e \ge \sum_{i \in e} z_i - |e| + 1$ give a pseudo-Boolean mixed-integer program. Without a valid radius, its solutions form a heuristic shortlist.

**Proposition 7** (Conservative utility guarantee)**.** *On a uniform error event, an exact optimizer of the sufficient constraints above is task-feasible and has utility at least that of every budget-feasible $S$ satisfying $F_t(S) \le \tau_t - 2\epsilon_t$ for all $t$. Certified literal cuts do not weaken this comparison.*

The comparison requires a margin. If the sufficient constraints are infeasible, no certified fallback is implied. Appendix K also gives the $2\epsilon$ plug-in regret bound for unconstrained risk minimization.

**Compression.** Fusion in (1) is invariant under $B_i \mapsto B_i Q_i$, $A_i \mapsto Q_i^{-1} A_i$ for invertible $Q_i$ (Stoica et al., 2025; Gong and Xu, 2026). Let $\Pi_{R_\ell}$ truncate the fused layer update to rank $R_\ell$. Its aggregate Frobenius error is

$$e_R(S)^2 = \sum_{\ell} \sum_{j > R_\ell} \sigma_j^2 \left( \sum_{i \in S} D_{i\ell} \right). \tag{44}$$

If task risk is $K_t$-Lipschitz along the dense-to-compressed segment, it increases by at most $K_t e_R(S)$. This bound requires a valid $K_t$; a pre-compression certificate alone is insufficient. Appendix L gives an orthogonal rank-one example whose truncation residual first appears at order $R + 1$, although its total loss is nonincreasing.

**Independent final audit.** Freeze $K$ fully specified compressed models using discovery data only. For $T$ tasks, $0 < \delta_{\mathrm{aud}} < 1$, and $m_t^{\mathrm{aud}}$ independent fresh task-$t$ examples with loss in $[0, 1]$, compute

$$u_{kt} = \widehat{F}_{kt}^{\mathrm{aud}} + \sqrt{\frac{\log(KT/\delta_{\mathrm{aud}})}{2m_t^{\mathrm{aud}}}}. \tag{45}$$

Clip $u_{kt}$ at 1 if desired. Return the highest-utility candidate satisfying $u_{kt} \le \tau_t$ for every task; otherwise abstain.

**Theorem 6** (Post-compression risk audit)**.** *Conditional on any discovery procedure producing the frozen list, with probability at least $1 - \delta_{\mathrm{aud}}$, every accepted candidate satisfies every population cap. This holds without a low-order tail assumption and allows selection among the candidates after observing the audit. It does not certify the learned hypergraph or the optimality of the returned merge.*

This is a one-sided Hoeffding union bound (Hoeffding, 1963), an instance of established risk calibration by multiple testing (Angelopoulos et al., 2021). Dependence across tasks is allowed. Adding candidates after inspecting the audit requires fresh data or a sequential correction.

**The CH-Merge procedure.** Query all subsets through order $q$ of the frozen dictionary, retain signed coefficients, and extract literal cuts only when their bounds are valid. Solve the certified constraints or form a heuristic shortlist, fuse and compress, then freeze and audit the candidates. Discovery costs $\sum_{j \le q} \binom{n}{j}$ coalition evaluations per scale on all calibration tasks. Table 8 distinguishes

Table 9: Mean audit accuracy (%) across all outputs, followed by the number of empirically feasible nonempty outputs. Means include violations and base fallbacks, so accuracy alone does not rank constrained success. Each primary query uses 128 images per task. Verification adds at most eight calls on a disjoint 128-image block (41 total for B/16 exclusion, 45 otherwise). Random counts include all 20 search seeds per bank; none are independent new datasets.

| | ViT-B/32 | | ViT-B/16 | |
|---|---|---|---|---|
| Method | 37 queries | + verification | 37 queries | + verification |
| F1-LS37 | 60.18 (0/3) | 52.74 (0/3) | 70.04 (0/3) | 65.78 (1/3) |
| Signed F2 | 64.12 (2/3) | 64.22 (3/3) | 70.94 (3/3) | 68.61 (3/3) |
| Permanent exclusion | 62.70 (3/3) | 62.70 (3/3) | 60.60 (2/3) | 60.60 (2/3) |
| Best observed37 | 62.70 (3/3) | 62.70 (3/3) | 67.52 (3/3) | 68.08 (3/3) |
| Greedy37 | 67.00 (2/3) | 65.69 (2/3) | 67.76 (2/3) | 70.38 (3/3) |
| Random37 | 66.31 (36/60) | 60.83 (39/60) | 68.06 (31/60) | 68.45 (48/60) |

the two certificates. The experiments below evaluate proposals and empirical caps; they claim neither structural nor population-risk certification.

# G Original auxiliary selection comparisons

These comparisons use the earlier complete-cube E7/E8 cohort described in Section 4.2. All outcomes, including cap violations and negative calibration results, are retained. They do not determine which triples enter the new exhaustive frequency study.

## G.1 Matched query protocol

The primary setting uses 128 proposal images per task. F1-LS37, signed F2, permanent exclusion, and Best observed37 share $1 + 8 + \binom{8}{2} = 37$ empty/singleton/pair queries. F1 fits an additive model by least squares with the base fixed; F2 interpolates quadratically. Permanent exclusion uses F2 but forbids supersets of every observed cap-violating singleton or pair. Best observed chooses a measured coalition. Greedy37 and Random37 may query larger sets within the same budget; random search uses 20 fixed seeds per bank. F2 versus exclusion isolates representation; the search comparisons match evaluation cost.

## G.2 Application: recovering excluded candidates

Signed F2 recovers an excluded feasible coalition in each B/16 bank. All three primary choices are nonempty and empirically feasible; each lies outside $\widehat{\mathcal{D}}$ and contains a discovery-forbidden subset. Their mean accuracy is $70.94\%$, versus $67.52\%$ for Best observed and $60.60\%$ for exclusion. The three F2 choices have regrets of $0.32$, $2.00$, and $6.18$ pp. Table 9 reports every output, including violations and base fallbacks. F2 is nevertheless $0.20$ pp below Best observed in B/16 bank 2 and feasible in only two B/32 banks. F1 violates a cap in all six primary outputs.

## G.3 Calibration and verification of proposals

A separately frozen setting evaluates the base and at most seven proposed candidates on the other 128 discovery images per task. Repeated evaluations count toward the budget: 41 total calls for B/16 exclusion, 45 otherwise. The 512-image audit remains excluded from selection. Verification makes B/32 signed F2 feasible in all three banks, with $64.22\%$ mean accuracy, $1.52$ pp above Best observed.

On B/16, verification reduces signed F2 accuracy from $70.94\%$ to $68.61\%$; verified greedy reaches $70.38\%$ and is feasible and more accurate in every bank. The original signed candidates incur three, four, and three extra DTD errors on 128 verification images and are rejected, although all pass the later empirical audit. A few errors near the cap can thus change the selected coalition. Random search also gains feasible outputs while losing mean accuracy on B/32 (Table 9).

Using all 256 discovery images with 37 queries doubles the primary image cost. B/16 signed F2 then has 73.13% mean accuracy, three feasible outputs, and 0.64 pp mean regret; B/32 has only one feasible signed output. Appendix M.5 reports four predefined 64-image blocks and prediction errors. More calibration and smaller overall error do not guarantee correct threshold decisions. No frozen nonempty output passes the multiplicity-corrected nominal iid paired audit diagnostic. Thus these experiments do not establish the deployment guarantee of Theorem 6.

**Relation to the mechanism controls.** In the earlier four-public-adapter study, a graph excludes 40 unseen larger coalitions; 30 improve on a contained conflicting pair, and a quadratic predicts 27. Frozen predictions recover 26 of 29 positive margins on disjoint examples. Cubic fitting lowers held-out cross-entropy MAE but raises strong-rescue false positives from 6 to 18 at unchanged recall. Appendix Q gives these relative-loss results; Appendix R gives planted and learned-digit controls.

# H Proof of the representation theorem and its corollaries

## H.1 Sharp price and pairwise recovery

We prove the utility and information claims of Theorem 2, together with Corollaries 2 and 3. The next subsection supplies the single-layer LoRA realization. For $T \subseteq [m]$, put $k = |T|$. Directly from (4),

$$G_\gamma(T) = \binom{k}{2}, \qquad G_\gamma(T \cup \{r\}) = -\frac{k(m-k)}{2} - \gamma.$$

Hence $\mathcal{F}_\gamma = \{S : r \in S\} \cup \{S \subseteq [m] : |S| \le 1\}$. In particular $[n]$ is feasible with excess $-\gamma$ and is the unique cardinality maximizer, with utility $n$.

If $\mathcal{A} \subseteq \mathcal{F}_\gamma$ is downward closed and $S \in \mathcal{A}$ contains two ordinary adapters, their pair belongs to $\mathcal{A}$ but has score $1 > 0$, contradicting soundness. Every such policy is therefore contained in $\mathcal{A}^\star_{\mathrm{dc}} = \{S : |S \cap [m]| \le 1\}$. This family is sound and downward closed, and it contains $\{i, r\}$, so its utility is two. It is the unique largest sound downward-closed subfamily; other policies may tie its maximum utility. For the empty family our definition gives $V = 0$. No part of this argument restricts the observations, computation, or representation size. Complete access to the $2^n$ losses does not change the conclusion.

For a degree-at-most-two Boolean polynomial $Q$, the observation table determines

$$c_\varnothing = Q(\varnothing), \qquad c_i = Q(i) - Q(\varnothing), \qquad c_{ij} = Q(ij) - Q(i) - Q(j) + Q(\varnothing). \tag{46}$$

For $G_\gamma$ the table is

$$G_\gamma(\varnothing) = G_\gamma(i) = 0, \quad G_\gamma(r) = -\gamma, \quad G_\gamma(ij) = 1, \quad G_\gamma(ir) = -\frac{m-1}{2} - \gamma \quad (i < j \le m).$$

Thus (46) gives the exact coefficients of (4). The $1 + n + \binom{n}{2}$ evaluations recover the entire risk within the quadratic class. Evaluating its prediction at $[n]$ gives $-\gamma < 0$, and returning that set attains the maximum possible utility. This requires no general optimization-runtime claim. With the known cap $\tau$, the same conclusion follows from observing $L = \tau + G_\gamma$ instead of excess risk.

Without the quadratic model class, $G_\gamma + \lambda z_1 z_2 z_r$ has the identical observation table and an infeasible full coalition when $\lambda > \gamma$. Pairwise recovery therefore requires the quadratic assumption. The downward-closure bound holds regardless of how much the policy observes. □

## H.2 Single-layer realization and minimality

We complete the realization claim of Theorem 2 and prove Corollary 4. With standard basis $e_1, e_2 \in \mathbb{R}^2$, choose $\beta > 0$ and set

$$\beta^2 = \gamma + \frac{m^2 - 1}{8}, \qquad v_i = \frac{e_1}{\sqrt{2}} \quad (i \le m), \qquad v_r = -\frac{m-1}{2\sqrt{2}} e_1 + \beta e_2, \tag{47}$$
$$r_0 = -\frac{e_1}{2\sqrt{2}} - \beta e_2, \qquad x = e_1, \quad y = e_1 - r_0, \qquad D_i = v_i e_1^\top.$$

These vectors satisfy

$$\|v_i\|^2 + 2\langle r_0, v_i\rangle = \tfrac{1}{2} - \tfrac{1}{2} = 0, \qquad 2\langle v_i, v_j\rangle = 1 \quad (i \neq j \leq m),$$
$$2\langle v_i, v_r\rangle = -\tfrac{m-1}{2}, \qquad \|v_r\|^2 + 2\langle r_0, v_r\rangle = \tfrac{m^2-1}{8} - \beta^2 = -\gamma.$$

Also $\|r_0\|^2 = 1/8 + \beta^2 = \tau$. Because $z_i^2 = z_i$ on the Boolean cube,

$$\left\| r_0 + \sum_i z_i v_i \right\|^2 - \tau = \sum_i (\|v_i\|^2 + 2\langle r_0, v_i\rangle) z_i + 2\sum_{i<j} \langle v_i, v_j\rangle z_i z_j = G_\gamma(z).$$

With $x = e_1$, $W_0 = I_2$, $y = e_1 - r_0$, and $D_i = v_i e_1^\top$, the model residual equals $r_0 + \sum_i z_i v_i$. Each $v_i$ is nonzero, $D_i e_1 = v_i$, and $D_i$ has image $\operatorname{span}\{v_i\}$, proving rank exactly one and the LoRA factorization. The equality is on Boolean coalitions, not on the continuous gate cube.

**Positive-semidefinite Gram matrix.** With $\mathbf{1} = \mathbf{1}_m$, the effect Gram matrix is

$$K = \begin{pmatrix} \frac{1}{2}\mathbf{1}\mathbf{1}^\top & -\frac{m-1}{4}\mathbf{1} \\ -\frac{m-1}{4}\mathbf{1}^\top & \frac{m(m-1)}{4} + \gamma \end{pmatrix}.$$

For any $(a, t) \in \mathbb{R}^m \times \mathbb{R}$,

$$(a,t)^\top K (a,t) = \frac{1}{2}\left(\sum_i a_i - \frac{m-1}{2} t\right)^2 + \beta^2 t^2 \geq 0.$$

The two independent linear forms and $\beta > 0$ give rank two. The zero-diagonal matrix of multilinear cross-coefficients need not itself be positive semidefinite: Gram diagonal terms become linear coefficients on the Boolean cube.

**Direct losses and parameter dependence.** Writing $k = \sum_{i=1}^m z_i$ and $h = z_r$, the residual and loss are

$$r(z) = \frac{k - \frac{1}{2} - \frac{m-1}{2} h}{\sqrt{2}} e_1 + \beta(h-1) e_2, \qquad L(z) = \begin{cases} \tau + k(k-1)/2, & h = 0, \\ \frac{1}{2}(k - m/2)^2, & h = 1. \end{cases}$$

Thus $\tau - L(T \cup \{r\}) = \gamma + k(m-k)/2 \geq \gamma$, and $L(ij) - \tau = 1$ for ordinary pairs. The empty set and ordinary singletons lie on the cap. Moreover,

$$\|D_i\|_{\mathrm{F}}^2 = \tfrac{1}{2} \ (i \leq m), \qquad \|D_r\|_{\mathrm{F}}^2 = \gamma + \tfrac{m(m-1)}{4}, \qquad \|r_0\|^2 = \tau = \gamma + \tfrac{m^2}{8}.$$

The model dimension is fixed, but the rescuer and residual norms grow linearly in $n$ for fixed $\gamma$. The construction thus requires growing update norms and an unscaled cap that depends on $n$.

**Minimal output dimension.** Suppose a scalar residual $t + \sum_i a_i z_i$ gave the exact same excess risk. Every ordinary effect is nonzero because $2a_i a_j = 1$ for distinct ordinary indices. Its zero singleton coefficient forces $a_i^2 + 2ta_i = 0$, so every ordinary effect equals $a = -2t$ and $a^2 = 1/2$. Writing the rescuer effect as $b$, the cross-coefficient forces $b = -(m-1)a/2$. Its singleton coefficient would be

$$b^2 + 2tb = b^2 - ab = \frac{m^2 - 1}{8} > 0,$$

contradicting $-\gamma < 0$. This lower bound concerns one-example squared loss of an affine residual; multiple scalar observations form a higher-dimensional stacked residual. Together with the explicit construction it proves minimal output dimension two.

**Minimal threshold for the exact unscaled identity.** For any output dimension and residual representation $\|r_0 + \sum_i z_i v_i\|^2 - \tau = G_\gamma(z)$, let $d = \sum_{i=1}^m \|v_i\|^2$. The ordinary-pair coefficients imply

$$0 \leq \sum_{i<j\leq m} \|v_i - v_j\|^2 = (m-1)d - \frac{m(m-1)}{2}, \qquad d \geq \frac{m}{2}.$$

The constant term gives $\tau = \|r_0\|^2$. Expanding a squared norm with the required coefficients then gives

$$0 \le \left\| r_0 + v_r + \tfrac{1}{2}\sum_{i=1}^{m} v_i \right\|^2 = \tau - \gamma - \frac{d}{4} - \frac{m(m-1)}{8}.$$

Hence $\tau \ge \gamma + d/4 + m(m-1)/8 \ge \gamma + m^2/8$, with equality in our construction. A cap is constant over coalitions; it cannot be constant over $n$ for the exact unscaled polynomial. □

**Linearly independent adapter effects.** Identical ordinary effects are not necessary for realizability. For any $\rho > 0$, append $m$ output coordinates and set

$$\bar{v}_i = (v_i, \rho e_i)\ (i \le m), \quad \bar{v}_r = (v_r, 0), \quad \bar{r}_0 = (r_0, -\tfrac{\rho}{2}\mathbf{1}_m).$$

The extra Boolean squared loss is $\rho^2 \sum_i (z_i - 1/2)^2 = m\rho^2/4$. Thus $\bar{\tau} = \tau + m\rho^2/4$ preserves the exact excess risk. The appended coordinates force each ordinary coefficient in a linear dependence to vanish; then the nonzero rescuer forces its coefficient to vanish. All $n$ effects are independent. Taking $\bar{x} = e_1 \in \mathbb{R}^2$, $\bar{W}_0 = (I_2, 0)^\top$, $\bar{y} = \bar{W}_0\bar{x} - \bar{r}_0$, and $\bar{D}_i = \bar{v}_i e_1^\top$ yields rank-one matrices with output dimension $m + 2 = n + 1$. This variant does not claim output dimension two or pairwise orthogonality.

### H.3 Uniform robustness and noisy pairwise observations

We prove Corollary 5 and clarify the role of the margin. Suppose $\|\widehat{G} - G_\gamma\|_\infty \le \varepsilon < 1$ and $\varepsilon \le \gamma$, uniformly over all Boolean coalitions. Every truly infeasible set is an unrescued set of size at least two and has score at least one; its estimate is therefore positive. Every rescued set has estimated score at most $-\gamma + \varepsilon \le 0$. Consequently

$$\{S : r \in S\} \subseteq \widehat{\mathcal{F}} := \{S : \widehat{G}(S) \le 0\} \subseteq \mathcal{F}_\gamma.$$

Maximizing cardinality over $\widehat{\mathcal{F}}$ returns $[n]$ with true utility $n$, while the best downward-closed policy sound for the *true* family still has utility two. The margin argument does not require a quadratic estimate. A generic conservative rule $\widehat{G}(S) + \varepsilon \le 0$ is also sound and accepts $[n]$ whenever $2\varepsilon \le \gamma$.

The original empty set and ordinary singletons have zero margin. If soundness is instead measured against the perturbed family, the best downward-closed utility may drop below two. For complete label stability, let $H_\gamma = G_\gamma - 1/2$: its feasible family is unchanged, every feasible score is at most $-1/2$, and every infeasible score is at least $1/2$. Uniform perturbations strictly below $1/2$ preserve all labels and the exact separation. The same LoRA realization implements this version with cap $\tau + 1/2$.

Under the quadratic model assumption, let observed low-order values $q_T$ obey $|q_T - G_\gamma(T)| \le \delta$ for $|T| \le 2$. For a coalition of size $s \ge 2$, its quadratic interpolant is

$$\widehat{Q}(S) = \sum_{\{i,j\}\subseteq S} q_{ij} - (s-2)\sum_{i\in S} q_i + \binom{s-1}{2} q_\varnothing.$$

The same identity holds for $G_\gamma$. Summing absolute observation weights gives

$$|\widehat{Q}(S) - G_\gamma(S)| \le (2s^2 - 4s + 1)\delta, \quad \|\widehat{Q} - G_\gamma\|_\infty \le (2n^2 - 4n + 1)\delta. \tag{48}$$

At sizes zero and one the error is at most $\delta$. The constant is sharp: at $[n]$, take all pair and empty-set errors $+\delta$ and all singleton errors $-\delta$. Thus $\delta < \min\{1, \gamma\}/(2n^2 - 4n + 1)$ suffices for strict rescued feasibility and soundness. This distinguishes exact identifiability from stable extrapolation; accurate pair observations alone do not control arbitrary higher-order tails.

### H.4 Deriving the diagnostic identities

For a quadratic, adding $k$ to $\{i, j\}$ changes risk by $c_k + c_{ik} + c_{jk}$, proving (39). For any set function, expand the right side of (40):

$$\begin{aligned}
&(F(i) - F(ik)) + (F(j) - F(jk)) - 2(F(\varnothing) - F(k)) \\
&\quad + F(ij) - F(i) - F(j) + F(\varnothing) - F(ijk) + F(ik) + F(jk) - F(k) \\
&= F(ij) - F(ijk) - F(\varnothing) + F(k) = R_{ij;k} - G_0.
\end{aligned}$$

The third finite difference on $\{i, j, k\}$ is $-Q$; changing the pair/rescuer assignment does not create a new interaction coefficient. These equalities apply to empirical losses as algebraic identities, without a statistical interpretation unless sampling assumptions are added.

### H.5 Independent noise and the boundary of exact recovery

The stored controlled extension uses $G_0$, the $\gamma = 0$ boundary case of (4), and adds an independent $\mathcal{N}(0, \sigma^2)$ perturbation to each distinct coalition value, reusing it if that coalition is queried again. The graph and interpolant read the same noisy low-order observations. For a set of size $r$, the interpolant expands as

$$\widehat{P}_2(S) = \binom{r-1}{2}\widehat{F}(\varnothing) - (r-2)\sum_{i\in S}\widehat{F}(i) + \sum_{\{i,j\}\subseteq S}\widehat{F}(ij), \qquad r \geq 2.$$

Consequently its full-set noise standard deviation is

$$\sigma\sqrt{\binom{n-1}{2}^2 + n(n-2)^2 + \binom{n}{2}}.$$

This grows quadratically in $n$. In this $\gamma = 0$ experiment the true full coalition has zero noiseless margin, and independent evaluation noise can change its feasibility. These failure rates do not evaluate the new $\gamma > 0$ robustness statement in Appendix H.3, whose uniform-error premises must be checked separately. Gaussian noise is unbounded, so its standard deviation alone does not give that deterministic event.

For each positive noise level and $n$, the stored experiment uses 100 trials with matched seeds across noise levels, plus six noiseless instances, totaling 1,806 trials. Exact cardinality branch-and-bound checks the signed selection objective; rejected cardinalities and the branch without the rescuer receive lower-bound certificates. The observed graph is verified to remain a clique plus isolated rescuer. Independent noisy higher-coalition values are generated only after selection. Evaluating the full set and, when needed, finding a feasible set of size $n - 1$ certifies the true optimum in every stored trial. The solver was checked against 360 exhaustive small cardinality problems. The score assigning zero to an infeasible selected set is explicitly failure-penalized, not an algorithm that repairs feasibility.

### H.6 Local scale bounds versus fitted exponents

For $F_\lambda(S) = L(\theta_0 + \lambda\sum_{i\in S} D_i)$ and $e = \{i_1, \ldots, i_k\}$, repeated integration gives

$$c_\lambda(e) = \lambda^k \int_{[0,1]^k} \nabla^k L\left(\theta_0 + \lambda\sum_{j=1}^{k} t_j D_{i_j}\right)[D_{i_1}, \ldots, D_{i_k}]\, dt.$$

A uniform derivative bound on these segments gives $|c_\lambda(e)| \leq M_k\lambda^k \prod_{i\in e}\|D_i\|$. This is an upper bound near zero, not an assertion that every coefficient has a nonzero order-$k$ leading term or that a finite-interval fitted slope equals $k$. Contextual tail control requires the corresponding bound across all relevant contexts, as in Theorem 8.

## I Proof of the interaction-order information dichotomy

Throughout this section, $1 \leq q < n$, $z_e = \prod_{i\in e} z_i$, and $z_\varnothing = 1$. An algorithm knows the declared class and common experimental setup. Its only information that varies with the unknown instance is the exact table $\mathcal{O}_q(F)$, or answers to adaptive queries on $\mathcal{B}_q = \{S : |S| \leq q\}$. The full table contains every possible answer to those queries. Adapter parameters, gradients, logits, and additional data are not observations in this oracle model. We prove Theorem 5, including its stated reconstruction and LoRA claims.

### I.1 Exact reconstruction, optimal query count, and the missing information

For $F(z) = \sum_{|e|\le q} c_e z_e$, Boolean-lattice Möbius inversion gives

$$c_e = \sum_{T\subseteq e} (-1)^{|e|-|T|} F(T), \qquad F(S) = \sum_{\substack{e\subseteq S \\ |e|\le q}} c_e. \tag{49}$$

Indeed, substituting $F(T) = \sum_{a\subseteq T} c_a$ into the first expression leaves the coefficient of $c_a$ equal to $\sum_{a\subseteq T\subseteq e}(-1)^{|e|-|T|}$. This is zero unless $a = e$, when it is one. Every value needed for $|e| \le q$ is in $\mathcal{O}_q(F)$, proving uniqueness. The count is $N_q(n) = |\mathcal{B}_q| = \sum_{j=0}^{q} \binom{n}{j}$. These identities are classical; the oracle theorem uses them to separate reconstruction from safe selection.

The monomials of degree at most $q$ are linearly independent, so $\dim \mathcal{P}_q = N_q(n)$. Suppose a deterministic algorithm exactly reconstructed every member with fewer evaluations, even at arbitrary coalition sizes. Run it on the zero function and record its $t < N_q(n)$ queries $S_1, \dots, S_t$. The homogeneous constraints $H(S_j) = 0$ leave a nonzero $H \in \mathcal{P}_q$. On $H$, the same zero answers force the same subsequent queries, stopping time, and reconstruction, which cannot be correct for both $0$ and $H$. This proves the worst-case necessity claim, including adaptive deterministic algorithms. It is a reconstruction bound for the real-valued class $\mathcal{P}_q$, not a lower bound on every selection problem or a runtime guarantee.

For arbitrary set functions the same inversion also yields

$$\ker \mathcal{O}_q = \operatorname{span}\{z_e : |e| > q\}, \qquad \dim \ker \mathcal{O}_q = 2^n - N_q(n). \tag{50}$$

A function vanishing on every observed coalition has zero coefficients through order $q$; conversely, each higher-order monomial vanishes there. Thus identical low-order loss tables are equivalent to agreement of all low-order Möbius coefficients, leaving the higher-order coefficients unrestricted.

### I.2 The hard pair and exact minimax value

Let $H$ be any function that is zero on $\mathcal{B}_q$ and positive on every larger coalition. Then $0$ and $H$ share the all-zero transcript and have optimal cardinalities $n$ and $q$. Every algorithm safe on both must output at most $q$ adapters on $H$. Since its output distribution is identical on the two instances, it also outputs at most $q$ on the zero instance. This holds deterministically or almost surely for randomized algorithms, including adaptive querying.

For any class satisfying $\{0, H\} \subseteq \mathcal{C} \subseteq \mathcal{H}_q$, this proves the upper bound in

$$V_q^{\det}(\mathcal{C}) = V_q^{\mathrm{as}}(\mathcal{C}) = q/n. \tag{51}$$

For the lower bound, fix a $q$-element set $Q$ and always return it. The class premise makes it feasible for every instance, while $\mathrm{OPT}(F) \le n$ implies $|Q|/\mathrm{OPT}(F) \ge q/n$. The deterministic rule is also admissible in the randomized class. Merely containing the hard pair gives the upper bound; it does not establish this matching lower bound for an arbitrary larger class. The feasibility premise guarantees positive optima and a common safe $q$-set.

To obtain the exact degree boundary, take

$$H(z) = E_{q+1}(z) := \sum_{|e|=q+1} z_e, \qquad E_{q+1}(\mathbf{1}_S) = \binom{|S|}{q+1}. \tag{52}$$

This has Boolean degree exactly $q+1$ and the required zero/positive pattern. For $p \ge q+1$, the class $\mathcal{C}_{p,q} = \mathcal{P}_p \cap \mathcal{H}_q$ contains this pair and is contained in $\mathcal{H}_q$, proving value $q/n$. For $p \le q$, inversion reconstructs every class member; exhaustive selection of a maximum-cardinality feasible set attains one. Safety bounds every accepted cardinality by the optimum, so one is also an upper bound. This proves (37). It concerns these full degree classes, not every class that happens to contain a high-degree function.

### I.3 A common-base two-dimensional rank-one LoRA pair

Set $d = q+1$, $k(z) = \sum_i z_i$, and $p_d(k) = \prod_{\ell=0}^{q}(k-\ell)$. The squared-loss pair

$$F_{\mathrm{safe}}(S) = 0, \qquad F_{\mathrm{bad}}(S) = p_d(|S|)^2 \tag{53}$$

again agrees on $\mathcal{B}_q$. For $|S| \geq d$, $p_d(|S|) \geq d!$, so its optima are $n$ and $q$. A scalar depth-$d$ realization of $F_{\text{bad}}$ uses $x = 1$, target zero, base weights $W_\ell^0 = -\ell$ and updates $D_{\ell,i} = 1$ for $\ell = 0, \ldots, q$. Its output is the product $p_d(k)$. Each scalar update is nonzero rank one, but is full rank in dimension one.

For strictly rank-deficient updates and a common architecture for both instances, let $e_1, e_2$ be the standard basis of $\mathbb{R}^2$ and put

$$P = e_1 e_1^\top, \qquad N = e_2 e_1^\top, \qquad P^2 = P, \quad PN = N^2 = 0.$$

Both matrices have rank one and Frobenius norm one. Use the same input $x = e_1$, target $y = 0$, depth $d$, two-dimensional states, and base weights $W_\ell^0 = -\ell P$. Every adapter is a bundle of updates activated by the same $z_i$ in every layer. Choose

$$D_{\ell,i}^{\text{bad}} = P, \quad W_\ell^{\text{bad}}(z) = (k(z) - \ell)P; \qquad D_{\ell,i}^{\text{safe}} = N, \quad W_\ell^{\text{safe}}(z) = -\ell P + k(z)N. \tag{54}$$

The bad network has output

$$W_q^{\text{bad}}(z) \cdots W_0^{\text{bad}}(z) e_1 = p_d(k(z)) e_1,$$

so its squared loss is (53). Since $q \geq 1$, the safe network has at least two layers and

$$W_1^{\text{safe}}(z) W_0^{\text{safe}}(z) = (-P + kN)(kN) = -kPN + k^2 N^2 = 0.$$

Its final output and loss vanish for every coalition. The first hidden state $ke_2$ is nonzero for every nonempty coalition; the second layer annihilates it. Thus no zero adapter update is used. Explicit LoRA factors are $(B, A) = (e_1, e_1^\top)$ in the bad instance and $(e_2, e_1^\top)$ in the safe instance. The common bases satisfy $\|W_\ell^0\|_{\text{F}} = \ell \leq q$; all update norms are one, independent of $n$ and $q$. The largest loss is $F_{\text{bad}}([n]) = [n!/(n-q-1)!]^2$.

Define $\mathcal{C}_{n,q}^{\text{LoRA}}$ to be the squared-loss functions generated by this fixed depth, dimension, input, target and common bases, with every per-layer update nonzero rank one of norm one, restricted to zero loss on $\mathcal{B}_q$. Both constructed instances lie in this class, which is contained in $\mathcal{H}_q$ at raw-loss cap zero. Therefore (51) proves

$$V_q^{\text{det}}(\mathcal{C}_{n,q}^{\text{LoRA}}) = V_q^{\text{as}}(\mathcal{C}_{n,q}^{\text{LoRA}}) = q/n. \tag{55}$$

The class, bases, and ranks may be known to the algorithm. The differing update parameters remain unobserved: inspecting them would distinguish this pair immediately and falls outside the theorem's oracle model.

### I.4 Boolean degree and minimal depth

On the Boolean cube, $p_d(k(z)) = d! E_d(z)$. Squaring and reducing repeated variables gives

$$F_{\text{bad}}(z) = (d!)^2 \sum_{|A|=|B|=d} z_{A \cup B}, \qquad \deg_{\text{B}}(F_{\text{bad}}) = \min(n, 2q+2). \tag{56}$$

Every union has size at most $\min(n, 2d)$, unions attaining that size exist, and all coefficients are positive. Hence there is no cancellation at the highest degree. The width-two squared realization and the abstract $E_{q+1}$ construction have the same feasible families but generally different interaction orders.

**Proposition 8** (Minimal depth for a nonzero zero-transcript squared loss)**.** *Consider a depth-$D$ deep-linear network with layer matrices affine in the coalition indicators, a fixed input and target, arbitrary finite widths, and arbitrary update ranks. If $D \leq q$ and its squared loss vanishes for every $|S| \leq q$, then it vanishes for every coalition. Consequently, the depth $q+1$ in (54) is minimal for a nonzero squared loss with the specified zero transcript.*

*Proof.* Each layer increases the total polynomial degree of an output coordinate by at most one, so every residual coordinate has degree at most $D$. Multilinear reduction on the Boolean cube cannot increase degree. A zero squared loss forces every residual coordinate to be zero on $\mathcal{B}_q$. Since its degree is at most $q$, (49) forces that coordinate to vanish everywhere. The explicit depth-$(q+1)$ bad instance is nonzero, proving sharpness. The conclusion concerns the all-zero squared-loss transcript; it does not assert minimal depth for indistinguishable instances at arbitrary nonzero loss levels. □

### I.5 Strict margins and probabilistic safety

The obstruction also holds away from a feasibility boundary. Keep the same raw squared losses but use cap $1/2$, equivalently centered scores $\bar{F}_\star = F_\star - 1/2$. Both instances give the observed score $-1/2$. Every safe-instance score is $-1/2$; the bad instance has score $-1/2$ for $|S| \le q$ and at least $(d!)^2 - 1/2 \ge 1/2$ otherwise. Feasible families are unchanged. Uniform score perturbations strictly below $1/2$ preserve all feasibility labels. The shifted class $\{F - 1/2 : F \in \mathcal{C}^{\mathrm{LoRA}}_{n,q}\}$ still has minimax value $q/n$: its hard pair gives the upper bound and every fixed $q$-set is feasible. Label stability alone does not assert that arbitrary perturbed observations remain identical.

If the algorithm is instead required to be safe with probability at least $1 - \eta$ on every instance, the common probability of selecting more than $q$ elements on the hard pair is at most $\eta$. Thus

$$\frac{\mathbb{E}|\mathcal{A}(\mathcal{O}_q(F_{\mathrm{safe}}))|}{n} \le \frac{q}{n} + \eta\left(1 - \frac{q}{n}\right). \tag{57}$$

On the two-instance class this safe-instance bound is attained by returning a fixed $q$-set with probability $1 - \eta$ and $[n]$ with probability $\eta$. This is not an asserted minimax equality for probabilistic safety on every larger class. Almost-sure safety is the case $\eta = 0$ used in Theorem 5.

**Relation to the representation theorem and numerical checks.** Theorem 2 uses an exactly quadratic instance where pairwise observations suffice and downward closure discards feasible rescuing coalitions. Theorem 5 removes the representation restriction but limits observations and demands safety across a declared class. Its hard pair has no rescue and its feasible families are downward closed; it isolates missing information. The shared ratio $2/n$ at $q = 2$ does not make these the same impossibility. The standard-library check in `code/test_narrative.py` enumerates $n = 2, \ldots, 8$ and $q = 1, \ldots, n - 1$, verifies the actual two-dimensional layer products and their squared losses, low-order transcript equality, exact Boolean degrees, reconstruction, kernel counts, and strict-margin labels. These finite checks supplement the general proofs.

## J Proofs of the representation and information results

**Theorem 7** (Low-rank realization of literal conflicts)**.** *Let $\mathcal{H} = \{(P_e, N_e) : e \in E\}$ have $P_e \cap N_e = \varnothing$, maximum scope size $d \ge 1$, and maximum vertex incidence degree $\Delta$. Given $a_i \ge 0$ and $w_e \ge 0$, there is a depth-$(d+1)$ diagonal linear network of width $n + |E|$, a fixed input, and a weighted squared loss whose binary LoRA coalition risk is exactly*

$$F_{\mathcal{H}}(z) = \sum_i a_i(1 - z_i) + \sum_{e \in E} w_e \prod_{i \in P_e} z_i \prod_{j \in N_e} (1 - z_j). \tag{58}$$

*Every adapter has rank at most $\max\{1, \Delta\}$ at each layer. Distinct adapters' updates are Frobenius-orthogonal at every layer. If $a_i > 0$ and $|P_e| \ge 2$ for every edge, every singleton strictly improves on the base.*

*Construction.* The first layer creates $n$ utility channels $z_i$. Each remaining path channel corresponds to one conflict and multiplies its literals across layers: a presence gate has base entry $0$ and update $+1$; an absence gate has base entry $1$ and update $-1$. Unused stages are identities. Utility targets are $1$ and conflict targets $0$. Binary indicators equal their squares. Each adapter touches at most one entry per incident edge in a layer, and no two adapters touch the same entry. Appendix J gives the factorization and proof.

The theorem uses a constructed base. It shows that low rank and parameter orthogonality alone cannot guarantee low interaction order; the frequency of such interactions in trained transformers is an empirical question. Frobenius orthogonality is also weaker than the data-conditioned properties sought by orthogonal-subspace methods (Zhang and Zhou, 2025).

**Corollary 6** (Observation and graph-reduction barriers)**.** *For any $k \ge 3$, pure $k$-way conflicts are realizable with all coalition losses on fewer than $k$ adapters identical to an additive loss. Consequently, an oracle restricted to those coalition losses cannot decide whether an unseen $k$-adapter merge has an additional penalty $\kappa$; its worst-case prediction error is at least $\kappa/2$. Moreover, for disjoint $k$-edges with equal utilities, replacing each known hyperedge by its clique retains only $1/(k-1)$ of the optimal conflict-free utility. These examples need only rank-one adapters per layer and depth $k + 1$.*

The information bound applies to the stated loss queries; inspecting weights or evaluating larger coalitions can distinguish the instances. The utility bound applies to clique reduction. Applying conflict-free hypergraph matching results to a trained-adapter bank would additionally require their degree and regularity assumptions (Glock et al., 2023).

### J.1 Proof of Theorem 7

Let $m = |E|$, and index the width-$n + m$ coordinates by the $n$ utility channels and one channel for every literal edge. All layers are square and diagonal. Feed the vector of all ones into the network. There are $d + 1$ layers, indexed by $0, 1, \ldots, d$.

In layer 0, put base diagonal value 0 on utility coordinate $i$, and let adapter $i$ add 1 there. No other adapter changes that coordinate. On every conflict coordinate, the layer-0 base value is 1 and no adapter changes it. Thus utility outputs after this layer are $z_i$ and conflict outputs are all 1.

For edge $e$, list the members of $P_e \cup N_e$ in any order as $i_{e,1}, \ldots, i_{e,k_e}$, where $k_e \le d$. On conflict coordinate $e$ at layer $\ell \le k_e$, put

$$(W_{0\ell})_{ee} = \begin{cases} 0, & i_{e,\ell} \in P_e, \\ 1, & i_{e,\ell} \in N_e, \end{cases} \qquad (D_{i_{e,\ell},\ell})_{ee} = \begin{cases} +1, & i_{e,\ell} \in P_e, \\ -1, & i_{e,\ell} \in N_e. \end{cases}$$

All other adapters have zero at this diagonal position. At layers $\ell > k_e$ the base entry is 1 and every update entry is zero. Utility coordinates have base entry 1 and zero updates at all layers after layer 0. Multiplying the diagonal layers therefore gives outputs

$$o_i(z) = z_i, \qquad o_e(z) = \prod_{i \in P_e} z_i \prod_{j \in N_e} (1 - z_j).$$

Take target 1 on utility coordinates and target 0 on conflict coordinates, with weights $a_i$ and $w_e$. For binary $z$, $(z_i - 1)^2 = 1 - z_i$ and $o_e(z)^2 = o_e(z)$, proving (58). The equality is asserted on the binary cube, not on its continuous relaxation.

For any layer after layer 0, adapter $i$ touches at most one diagonal entry per incident edge, hence at most $\Delta$ entries. In layer 0 it touches exactly its utility entry. A diagonal matrix supported on positions $j_1, \ldots, j_r$, with entries $s_1, \ldots, s_r \in \{-1, +1\}$, has the explicit LoRA factorization

$$B = [s_1 e_{j_1} \ \cdots \ s_r e_{j_r}], \qquad A = [e_{j_1} \ \cdots \ e_{j_r}]^\top.$$

Its rank is $r$. The zero matrix needs rank zero. Absorb any desired fixed nonzero scale into $B$. Thus every layer update has rank at most $\max\{1, \Delta\}$. Two adapters never occupy the same diagonal entry at a given layer, so their Frobenius inner product is zero. Orthogonality also holds after concatenating all layer parameters.

Finally, if every edge has at least two required presences, then all conflict indicators vanish at the base and at each singleton. Selecting adapter $i$ reduces loss by exactly $a_i > 0$. This proves the last assertion. For disjoint scopes, $\Delta = 1$, giving the rank-one special case. A large output width and several adapted layers are part of the construction; the theorem does not claim that the architecture is an efficient way to train a useful model. □

### J.2 Proof of Corollary 6

Take an ordinary $k$-uniform hypergraph, so $N_e = \varnothing$ and $|P_e| = k$. Every coalition of size below $k$ has loss $\sum_i a_i(1 - z_i)$, regardless of the hyperedges and their weights. Equivalently, the coefficients of orders $2, \ldots, k - 1$ in its Möbius expansion all vanish. Compare two systems, one with no conflict penalty and another with a single $k$-edge of weight $\kappa$. Any deterministic or randomized procedure receiving only sub-$k$ coalition-loss observations sees the same transcript in the two systems. At a coalition containing that edge, the two losses differ by $\kappa$. If its prediction is $v$, the triangle inequality gives

$$\max\{|v - F_0(S)|, |v - F_1(S)|\} \ge \kappa/2.$$

For randomized predictions, the same inequality holds for the maximum of the two expected absolute errors. A threshold strictly between the two values makes the safety decision unidentifiable from that transcript. Both losses can be put in a bounded range by a common positive rescaling; the gap

rescales with them. Nothing in this argument hides the model parameters from a method that is allowed to inspect them: the result explicitly concerns the stated loss-oracle observation class.

For the clique statement, take $m$ disjoint edges of size $k$, with equal utility per vertex. True hypergraph independence allows $k-1$ vertices per edge; replacing each edge by a clique allows only one. Their utility ratio is exactly $1/(k-1)$. If conflict weights exceed the total available utility, any coalition containing a conflict has greater loss than the base, and minimizing (58) is equivalent to maximum-utility conflict-free selection. Disjointness implies $\Delta = 1$, and the realization has depth $k+1$. □

### J.3 Proof of Proposition 6

If $S$ contains no hyperedge, neither does any subset of $S$, so an independent-set family is downward closed. Conversely, let $\mathcal{A}$ be a nonempty downward-closed family. Form a hypergraph from its inclusion-minimal nonmembers. A set belongs to $\mathcal{A}$ if and only if it contains none of these nonmembers: otherwise any nonmember contains a minimal nonmember by finiteness. This represents $\mathcal{A}$ as a hypergraph independent-set family.

For (41), $\{1,2,3\}$ has interference zero, while its subset $\{1,2\}$ has interference $\kappa$. At any cap $0 \le \tau < \kappa$, downward closure fails. In literal form, the one forbidden pattern is $(\{1,2\},\{3\})$, and its inequality is

$$z_1 + z_2 + (1 - z_3) \le 2, \quad \text{or equivalently} \quad z_1 + z_2 - z_3 \le 1.$$

The all-present assignment satisfies this inequality; the harmful pair does not. □

## K Full approximation and exclusion arguments

### K.1 Formal bounds and exclusion certificate

For disjoint $e, U$, let $D_e F(U) = \sum_{V \subseteq e} (-1)^{|e|-|V|} F(U \cup V)$. Give $[n]$ any fixed order and set $S_{<i} = \{j \in S : j < i\}$.

**Theorem 8** (Contextual next-order identity)**.** *For every set function and $0 \le q < n$,*

$$F(S) - P_q(S) = \sum_{\substack{e \subseteq S \\ |e| = q+1}} D_e F(S_{<\min e}). \tag{59}$$

*If $|D_e F(U)| \le \beta_e$ uniformly over the relevant disjoint contexts, then*

$$|F(S) - P_q(S)| \le \sum_{\substack{e \subseteq S \\ |e| = q+1}} \beta_e. \tag{60}$$

*In particular, if $F(S) = L(\theta_0 + \sum_{i \in S} D_i)$ and $\|\nabla^{q+1} L(\theta)\|_{\mathrm{op}} \le M_{q+1}$ on the whole cumulative-update cube, then the right-hand side is at most $M_{q+1}(\sum_{i \in S} \|D_i\|)^{q+1}/(q+1)!$.*

The identity follows by assigning each omitted Möbius support to its largest $q+1$ elements. At order $q = 2$, it states exactly what pair-only fitting leaves out: third-order mixed differences *in contexts*. Small cubic coefficients at the empty coalition alone do not bound the fourth-order tail, as Corollary 6 shows. The identity uses discrete calculus and requires no smoothness; applying its derivative corollary to clipped losses would require a separate smoothness check.

Suppose losses lie in $[0,1]$ and all adapters are frozen before calibration. With $m_t$ independent calibration examples on task $t$, let $M_{\mathrm{probe}}$ count all distinct frozen probes, including any scale grid. Hoeffding's inequality and a union bound yield simultaneous probe error

$$\eta_t = \sqrt{\frac{\log(2T M_{\mathrm{probe}}/\delta)}{2m_t}}, \qquad |\widehat{c}_t(e) - c_t(e)| \le 2^{|e|}\eta_t. \tag{61}$$

For coalitions of size at most $b$, a valid contextual tail bound $B_t$ therefore gives

$$\sup_{|S| \le b} |F_t(S) - \widehat{P}_{t,q}(S)| \le \epsilon_t := B_t + \eta_t \sum_{j=0}^{\min(q,b)} 2^j \binom{b}{j}. \tag{62}$$

The bound can be too large to certify a decision. A sparse approximation must add the absolute mass of discarded coefficients. Without a justified $B_t$, the fit can propose candidates but cannot certify exclusions.

### K.2 Certifying a forbidden face without deleting rescuers

Bounding each unfixed monomial separately by the minimum of zero and its coefficient discards cancellations. For five disjoint copies of $z_1z_2 - z_1z_2z_3$, this gives $-3$ on a face where the true minimum is 1. A local-moment relaxation preserves the cancellations (Sherali and Adams, 1990).

Let $y_e$ denote a relaxed monomial for every $|e| \le q$, with $y_\varnothing = 1$. Require nonnegative local assignment masses

$$\mu_{A,B}(y) := \sum_{J \subseteq B} (-1)^{|J|} y_{A \cup J} \ge 0, \quad A \cap B = \varnothing, \quad |A| + |B| \le q. \tag{63}$$

Together with $0 \le y_e \le 1$, fixing all moments forced by $(P, N)$ defines $\mathcal{L}_q(P, N)$. Every actual completion induces a feasible moment vector. In particular, $\mu_{\{1,2\},\{3\}} = y_{12} - y_{123} \ge 0$ preserves the rescue cancellation. Define

$$\underline{F}_t(P, N) = \min_{y \in \mathcal{L}_q(P,N)} \sum_{|e| \le q} \widehat{c}_t(e) y_e - \epsilon_t. \tag{64}$$

If a selection budget is used, one may add valid moment constraints for it; the unconstrained relaxation remains a valid, possibly weaker, bound on budget-feasible completions. The error radius in (64) must cover those completions.

**Theorem 9** (Context-safe exclusion)**.** *Assume $|F_t(S) - \widehat{P}_{t,q}(S)| \le \epsilon_t$ on the candidate domain. If the lower bound in (64) exceeds $\tau_t$, inequality (42) excludes no task-feasible candidate in that domain. Any number of such cuts, chosen adaptively from the same uniformly valid coefficient event, remain simultaneously sound. No monotonicity assumption is required.*

An interaction coefficient alone does not certify an exclusion: the lower bound must cover every completion. A cut can be shortened by dropping a presence or absence and recomputing the bound; the shorter cut is valid only if that bound still exceeds the cap. The LP has $\sum_{j \le q} \binom{n}{j}$ moments and at most $\sum_{j \le q} 2^j \binom{n}{j}$ local inequalities. It need not be exact on cyclic interaction structures. Our implementation verifies a dual lower bound in rational arithmetic for the stored coefficients.

### K.3 Möbius inversion and proof of Theorem 8

For any finite set function, inclusion–exclusion gives

$$\sum_{e \subseteq S} c(e) = \sum_{V \subseteq S} F(V) \sum_{V \subseteq e \subseteq S} (-1)^{|e| - |V|} = F(S),$$

because the inner sum is $(1-1)^{|S \setminus V|}$. For $e \cap U = \varnothing$, another application of the same cancellation shows

$$D_e F(U) = \sum_{V \subseteq U} c(e \cup V). \tag{65}$$

Expand the right-hand side of (59) with (65). Every resulting support has size at least $q + 1$. Conversely, for a support $A \subseteq S$ with $|A| \ge q + 1$, take $e$ to be its largest $q + 1$ elements and $V = A \setminus e$. Then $V \subseteq S_{<\min e}$, so $c(A)$ occurs. This representation is unique, since any additional elements of $V$ must lie below every member of $e$. Thus each omitted coefficient appears exactly once, proving the identity for every $q$, including $q = 0$. If $|S| \le q$, both sides are zero.

The triangle inequality proves (60). For the differentiable corollary, write $e = \{i_1, \dots, i_{q+1}\}$. Repeated applications of the fundamental theorem of calculus give

$$D_e F(U) = \int_{[0,1]^{q+1}} \nabla^{q+1} L \left( \theta_0 + \sum_{i \in U} D_i + \sum_{j=1}^{q+1} t_j D_{i_j} \right) [D_{i_1}, \dots, D_{i_{q+1}}] \, dt.$$

The norm assumption on the *whole cumulative-update cube* bounds its magnitude by $M_{q+1}\prod_{i\in e}\|D_i\|$. In the expansion of $(\sum_{i\in S}\|D_i\|)^{q+1}$, the distinct-index products each appear $(q+1)!$ times; all repeated-index products are nonnegative. Summing the bound over $e$ proves the final inequality. The norm here may be the Euclidean norm of all vectorized layer updates; using a different norm requires the matching multilinear operator norm. □

**Why base-only derivatives do not suffice.** A pure $(q+1)$-way indicator has zero anchored coefficients through order $q$ but is nonzero on a larger coalition. More generally, a bound on $\nabla^{q+1}L(\theta_0)$ alone does not control the integrals above, whose evaluation points span contexts and interpolation coordinates. Neither observing a few low-order terms nor fitting them accurately supplies a uniform derivative bound.

### K.4 Statistical error event

Condition on the frozen model dictionary and scale families, constructed independently of calibration. For each probe and task, Hoeffding's two-sided bound gives $\Pr(|\widehat{F}_t(S)-F_t(S)|>\eta_t)\le 2e^{-2m_t\eta_t^2}$. Using (61) and summing over $TM_{\mathrm{probe}}$ probe–task pairs gives a simultaneous event of probability at least $1-\delta$. No independence between probes is required. On this event,

$$|\widehat{c}_t(e)-c_t(e)|\le\sum_{V\subseteq e}|\widehat{F}_t(V)-F_t(V)|\le 2^{|e|}\eta_t.$$

For $|S|\le b$, summing over its supports through order $q$ proves the stochastic term in (62). Adding the uniform tail bound proves the claim. If a coefficient is omitted from the fitted representation, its absolute value contributes to a deterministic bound on the additional prediction error. When its true value is unknown, the corresponding estimated magnitude plus its estimation radius gives a conservative replacement. A scale grid must be included in $M_{\mathrm{probe}}$; optimizing over a continuous, uncountable scale set would require an additional uniformity argument.

The bound grows rapidly in $b$ and $q$ because it sums separate worst-case coefficient errors. This calculation does not compare the sample requirements of all possible estimators. Variance-aware or joint coefficient estimates may give tighter bounds.

### K.5 Local-moment validity and proof of Theorem 9

For an actual binary assignment $z$, set $y_e=\prod_{i\in e}z_i$. Then

$$\mu_{A,B}(y)=\prod_{i\in A}z_i\prod_{j\in B}(1-z_j)\ge 0.$$

Moreover, $y_e=0$ if $e$ intersects an absent literal, and $y_e=1$ if every member of $e$ is required present. Thus the moment vector of each face completion lies in $\mathcal{L}_q(P,N)$. For any completion in the candidate domain,

$$\min_{y\in\mathcal{L}_q(P,N)}\sum_e\widehat{c}_t(e)y_e-\epsilon_t\le\widehat{P}_{t,q}(S)-\epsilon_t\le F_t(S).$$

If the left side exceeds $\tau_t$, every such completion violates the task cap, proving validity of (42). Soundness holds deterministically on the same uniform error event for *every* face; selecting faces adaptively does not require an extra union bound over the faces. A valid but loose lower bound may fail to produce a cut; no converse is asserted.

For the five-block rescue example, each block contributes a nonnegative local assignment mass $y_{ab}-y_{abc}$. Fixing $a,b$ present and $c$ absent forces that block's contribution to 1. The four other contributions remain nonnegative. A compatible completion of the other blocks makes them all zero, proving that the LP minimum is exactly 1. By contrast, independent term bounds count the fixed block as 1, and may independently set the negative cubic from each other block to $-1$ while setting its positive pair to zero. This yields $1-4=-3$. The local-moment improvement is therefore structural, not a tuning effect. □

### K.6 Verified dual lower bounds

A floating-point primal minimizer is generally an upper bound on a minimization problem and is not an exclusion certificate. Our LP implementation obtains inequality multipliers from HiGHS and verifies a lower bound separately. Write the local LP as

$$\min c^\top y \quad \text{subject to } Ay \le b, \quad l \le y \le u.$$

For any $\lambda \le 0$, put $r = c - A^\top \lambda$. Every feasible $y$ satisfies

$$c^\top y \ge \lambda^\top b + \sum_j \min\{r_j l_j, r_j u_j\}. \tag{66}$$

Indeed, $\lambda^\top Ay \ge \lambda^\top b$, and each residual coordinate is minimized on its box interval. Dual optimality or even exact dual feasibility in a separate equality formulation is not needed. In our moment LP, $b = 0$, $A$ has integer entries and the box endpoints are 0 or 1, including fixed moments. We truncate returned positive multipliers to zero, interpret all stored coefficient and multiplier floats as exact binary rationals, and evaluate (66) with Python's `Fraction`. The final floating return is rounded downward. A user-supplied radius must independently be a valid upper bound; exact arithmetic does not establish a statistical premise.

### K.7 Proof of Proposition 7

If $\widehat{P}_t(S) + \epsilon_t \le \tau_t$, then $F_t(S) \le \tau_t$ by the uniform error bound. If $F_t(S) \le \tau_t - 2\epsilon_t$, then $\widehat{P}_t(S) + \epsilon_t \le F_t(S) + 2\epsilon_t \le \tau_t$. Thus every candidate with this margin lies in the sufficient-constraint feasible set. Certified cuts do not remove it by Theorem 9. Exact utility maximization over that set therefore achieves at least its utility. If the sufficient set is empty, no existence is claimed. For unconstrained risk minimization, with $\widehat{S} \in \arg\min_S \widehat{P}(S)$ and $S^* \in \arg\min_S F(S)$,

$$F(\widehat{S}) \le \widehat{P}(\widehat{S}) + \epsilon \le \widehat{P}(S^*) + \epsilon \le F(S^*) + 2\epsilon.$$

An approximate optimizer adds its optimization gap. □

## L Compression and audit proofs

### L.1 Gauge invariance and the truncation bound

An invertible gauge change preserves each realized update: $(B_i Q_i)(Q_i^{-1} A_i) = B_i A_i$. Consequently, dense fusion and the SVD of the fused matrix depend only on the realized updates. At a singular-value tie, an implementation may choose among several optimal rank truncations; a deterministic choice based only on the fused matrix remains gauge-invariant. Naively averaging factors separately need not preserve products and is outside our formulation.

By the SVD best-approximation identity, the squared Frobenius error at layer $\ell$ is the sum of squared singular values beyond $R_\ell$. Euclidean aggregation across vectorized layers gives (44). A $K_t$-Lipschitz risk then changes by at most $K_t e_R(S)$. This assumes that the same task loss and the actual compressed parameter vector are used. It does not automatically extend to quantization, routing, retokenization or other deployment changes; those must also be audited or bounded.

**Proposition 9** (An exact compression-induced capacity interaction)**.** *In one linear layer, let $D_i = e_i e_i^\top$ for $i \in [n]$, use base zero and target identity on the standard-basis input batch, and sum squared errors over that batch. Before compression, $F(S) = n - |S|$. After an SVD rank-$R$ truncation of the selected update,*

$$F_R(S) = n - \min\{|S|, R\}, \qquad F_R(S) - F(S) = (|S| - R)_+.$$

*Thus all probes of size at most $R$ miss the compression residual, while the first nonzero anchored interaction in that residual occurs at order $R + 1$.*

*Proof.* The selected matrix is the orthogonal projector onto the coordinates in $S$. It has $|S|$ unit singular values. Its rank-$R$ truncation is a rank-$\min\{|S|, R\}$ orthogonal projector within that coordinate subspace, regardless of the choice among tied singular directions. Its squared Frobenius distance from $I_n$ is $n - \min\{|S|, R\}$. The residual set function vanishes on every set of size at most $R$, and equals 1 on each set of size $R + 1$, so its first nonzero Möbius coefficients have that order. □

This proposition describes saturation of attainable utility, not a claim that adding an adapter necessarily raises aggregate loss. The dense sums and fixed-rank compressed models are different comparison classes.

### L.2 Proof of Theorem 6

Condition on all discovery data, model training and the resulting frozen candidate list. For candidate $k$ and task $t$, a one-sided Hoeffding bound gives

$$\Pr\Big(F_{kt} > \widehat{F}^{\mathrm{aud}}_{kt} + r_t \mid \text{discovery}\Big) \le e^{-2m_t^{\mathrm{aud}} r_t^2}.$$

Take $r_t = \sqrt{\log(KT/\delta_{\mathrm{aud}})/(2m_t^{\mathrm{aud}})}$ and union-bound over $KT$ pairs. With the advertised conditional probability, every upper bound is valid simultaneously. Any candidate selected after observing the audit is therefore safe with respect to the declared caps if all its upper bounds pass. If none passes, returning no certified candidate is necessary; the base is not an automatic certified fallback. Finally integrate over discovery randomness to obtain the unconditional statement. Dependence across candidates or tasks is harmless to the union bound, but the within-task concentration premise must hold. □

## M Controlled eight-adapter CLIP selection: protocol and full results

### M.1 Training, data, and final-rank evaluation

E7 trains CLIP ViT-B/32 and E8 trains CLIP ViT-B/16 (Radford et al., 2021). Each backbone has three independently initialized training banks, each with eight single-task adapters. The mask-bit order is MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), GTSRB (Stallkamp et al., 2011), EuroSAT (Helber et al., 2019), Stanford Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), RESISC45 (Cheng et al., 2017), SUN397 (Xiao et al., 2010); bit $i$ has value $2^i$, starting at $i = 0$. All 48 adapters use LoRA rank 8, $\alpha = 8$, and dropout $0.1$, targeting the query, key, value, and output projections in each of 12 visual transformer blocks (48 matrices). Other weights, the visual projection, and the text-derived classification heads are frozen. E8 rebuilds its heads with the B/16 text encoder using the same fixed prompt texts; it does not reuse B/32 head vectors.

Training uses AdamW, learning rate $10^{-4}$, weight decay $0.01$, batch size 32, a warmup of at most 50 steps followed by cosine decay, and gradient clipping at 1. Every adapter trains for exactly 1,200 steps and uses the last checkpoint, without early stopping. Training uses BF16 autocast with FP32 parameters, optimizer state, logits, and cross-entropy. Bank $b$ and task $i$ use seed $2026091400+100b+i$ in E7 and $2026091600+100b+i$ in E8. No task, failed output, or training seed is removed. Each primary coalition evaluation uses the FP32 per-layer model $W_0 + \Pi_8(0.5 \sum_i D_i)$; inference, SVD, logits, and loss computation use FP32. All primary selection comparisons within a backbone therefore use the same final rank and merge scale, including the oracle.

The training counts in task order are 2,048, 2,048, 2,048, 1,932, 2,048, 1,109, 2,048, and 2,048, totaling 15,329. Every task contributes 256 discovery and 512 audit images. E7 excludes previously evaluated images from its holdout pool, and global RGB hashes show no training/holdout overlap. The native public test pools were explicitly repartitioned into training/discovery/audit; these results must not be presented as canonical benchmark test scores. The three banks share one split. E8 reuses that split after the E7 results were inspected, making it a second-backbone replication within the CLIP family, not a new untouched final test. Hash separation does not establish absence from CLIP pretraining. The public-adapter witness additionally inherits the source-split limitations in Appendix P.

**Resource versions and usage terms.** Table 10 identifies the frozen model and dataset snapshots. MNIST uses the canonical test IDX files, SVHN uses `test_32x32.mat`, and GTSRB uses the official final-test image and label archives. The remaining five test pools use the listed `tanganke` mirrors; the original dataset papers are cited above. Model revisions and file hashes are recorded in each run's `model_sources.json`; the initial data checksums and EuroSAT snapshot are in `results/experiment_metadata.json`, and the four additional dataset snapshots are in

E7's `new_asset_sources.json`. These snapshots identify the source pools, while `sample_manifest.csv` identifies the actual repartition used here.

The upstream CLIP software is MIT-licensed[1]; this software license does not grant rights to the benchmark images. SVHN's source page specifies non-commercial use[2], and DTD's source page makes the data available for research purposes[3]. Third-party models and datasets remain subject to their respective upstream terms; use of a public mirror does not replace those terms or establish unrestricted redistribution rights.

Table 10: Frozen upstream snapshots used for the controlled CLIP experiments. Model identifiers refer to Hugging Face model repositories; the five `tanganke` identifiers refer to dataset repositories. All dataset snapshots use their public test pools before our declared repartition.

| Resource | Revision |
|---|---|
| `openai/clip-vit-base-patch32` | `3d74acf9a28c67741b2f4f2ea7635f0aaf6f0268` |
| `openai/clip-vit-base-patch16` | `57c216476eefef5ab752ec549e440a49ae4ae5f3` |
| `tanganke/eurosat` | `43750fc422c7fbd4289e2df3f7473617a5937acc` |
| `tanganke/stanford_cars` | `9abf6cf7d6dfa7b95152a0d6e791ea9435b47a40` |
| `tanganke/dtd` | `d2afa97d9f335b1a6b3b09c637aef667f98f966e` |
| `tanganke/resisc45` | `5c484982a196dcb501b8e79c6cf4680dc6f2ccdf` |
| `tanganke/sun397` | `9c4afd6174afa3974f209a0ef099dbbe244dfd27` |

### M.2 Frozen selection and evaluation accounting

The primary objective minimizes the unweighted average of eight task error rates under a per-task excess-error cap of $0.02$ relative to the same backbone's base. Selection breaks ties by the smaller integer mask, not by cardinality. The audit oracle uses the same objective and cap on its own images. A cap violation gives undefined constrained regret (NA); a base fallback has a defined regret but is not a nonempty success. Average-accuracy summaries include all outputs. Pass counts in the detailed tables count tasks, not independent trials.

For each task, F1-LS37 solves least squares on the common 37 low-order observations while fixing its intercept to the observed base error. Signed F2 uses exact Boolean quadratic interpolation of those observations without clipping extrapolated error predictions. Both search all 256 predicted coalitions. Permanent exclusion deletes every superset of an observed nonempty singleton or pair whose error increase exceeds $0.02$ on any task, then uses the same F2 objective and predicted caps in the remaining domain. This baseline differs from the earlier cross-entropy conflict rule, and from the complete-information downward core in Eq. (7). Best observed37 chooses the empirically best feasible measured coalition among the common 37.

Random37 queries the base and 36 distinct nonempty masks, sampled uniformly without replacement, for each of 20 fixed seeds $2026091200, \ldots, 2026091219$ per bank. All seeds are retained. Forward greedy37 queries the base, then all extensions by one adapter at each of eight successive steps; it extends the current mask by minimum average observed error, without imposing the cap on intermediate steps. This costs $1 + 8 + 7 + \cdots + 1 = 37$ distinct queries. Its final output is the best cap-feasible coalition anywhere in the queried set, rather than necessarily the last mask. Random and greedy can observe high-order coalitions; only the information shared by F1, F2, exclusion, and Best observed is identical.

The primary protocol uses the first 128 discovery images per task. For verification, each method freezes the base plus its seven highest-ranked available nonempty candidates, or fewer if its domain is smaller. Predicted-feasible candidates precede predicted-infeasible ones; the latter are ordered by maximum cap violation, then all ties by mean error and mask. Surrogate methods rank their allowed domains; search and Best observed rank only queried masks. The list is frozen before the second 128-image block is used to select its directly measured best feasible member. The B/16 exclusion domain gives four verification masks and 41 total calls; all other method/bank cases use eight and

[1] https://github.com/openai/CLIP/blob/main/LICENSE.
[2] http://ufldl.stanford.edu/housenumbers/.
[3] https://www.robots.ox.ac.uk/~vgg/data/dtd/.

45 calls. A repeated mask on different images incurs a new call. The primary, verification, and 256-image settings respectively cost 4,736, at most 5,760, and 9,472 image evaluations per task. A query evaluates one coalition on all eight tasks; full-cube audit cost is additional offline evaluation, never a free selection budget.

The predefined secondary settings are all 256 discovery images and each of four nonoverlapping 64-image blocks. The latter overlap the larger settings and are sensitivity diagnostics, not four extra datasets. Each setting has five deterministic methods and 20 random outputs: $25 \times 7 = 175$ frozen decisions per bank, or 525 per backbone. Full-cube inference is collected offline into discovery and audit caches before selection. The reported query budgets are enforced during replay by a discovery-query interface and charged per method, rather than measured by separately rerunning each method's GPU calls. For each bank, candidate lists, selections, coefficients, source, and protocol hashes are frozen before the separate evaluator opens its audit cache. Discovery-only replay reproduces all 525 decisions per backbone. Complete-cube audit data are used only after this freeze for oracle comparisons and diagnostics, without changing any choice.

### M.3 Complete-information representation comparison and frozen witness

The feasible/core family sizes for B/32 banks 0–2 are $104/33$, $100/41$, and $115/36$; for B/16 they are $78/16$, $95/26$, and $81/16$, including the base. Table 3 optimizes average accuracy separately from cardinality. The complete-information family is defined by all subsets, including high-order ones, so the table measures the best possible sound downward restriction on the empirical table, not the performance of a learned graph. The full accuracy-oracle masks are 143, 77, 79 for B/32 and 111, 79, 94 for B/16; the downward-core accuracy-oracle masks are 74, 77, 77 and 74, 74, 10, respectively.

The additional E7a witness uses the previously released KnOTS adapters for MNIST, GTSRB, and EuroSAT, scale $0.5$, and final rank 16. Its eight subset masks were fixed after the exploratory E6 result but before the E7 fresh audit. Table 11 evaluates all eight tasks, not only the three specialist tasks. The triple gains $10.89$ pp over the base while every nonempty proper subset fails a task cap. Several subset failures are $11/512 = 2.1484$ pp, only one excess error beyond the largest passing integer count. These are point estimates, not proof that the corresponding population family has the same boundary. Only these eight public-adapter masks were re-evaluated; this is not a fresh 256-mask public-adapter oracle.

Table 11: Frozen public-adapter witness (E7a): scale 0.5, final rank 16, all eight subsets, 512 new audit images per task. Accuracy is averaged over all eight tasks. Worst regression is excess error over the base in pp; Feasible denotes the empirical 2 pp cap on every task.

| Coalition | Mask | Accuracy (%) | Worst regression | Feasible |
|---|---|---|---|---|
| Base | 0 | 47.27 | 0.00 | Yes |
| MNIST | 1 | 53.17 | 2.15 | No |
| GTSRB | 4 | 49.37 | 16.80 | No |
| EuroSAT | 8 | 53.93 | 2.15 | No |
| MNIST + GTSRB | 5 | 54.25 | 16.99 | No |
| MNIST + EuroSAT | 9 | 58.13 | 2.15 | No |
| GTSRB + EuroSAT | 12 | 54.86 | 2.73 | No |
| MNIST + GTSRB + EuroSAT | 13 | 58.15 | 1.17 | Yes |

### M.4 All deterministic selections and random-search outcomes

Tables 12, 13, and 14 report all deterministic choices in the primary, verification, and full-discovery settings. "Worst" is the minimum task accuracy, "Pass" counts tasks meeting the relative cap, and regret is in pp. A row with fewer than eight passing tasks is infeasible even if its mean accuracy is high. Mask 0 denotes a base fallback, which is excluded from nonempty-success counts. The common base accuracies are $47.27\%$ (B/32) and $55.35\%$ (B/16). All 20 random outputs per bank are summarized without selecting a favorable seed in Table 15; their individual masks and task metrics remain in the result records.

The B/32 verified signed masks 202, 204, and 43 all contain discovery-banned subsets and lie outside the audit downward core. Their maximum task regressions are 1.56, 1.95, and 1.37 pp. The B/16

primary signed masks 78, 206, and 83 have the same two exclusion properties, with maximum regressions 1.17, 1.76, and $-0.20$ pp. A negative maximum means every task improves, but does not imply near-optimal average accuracy. Verified B/16 greedy is both empirically feasible and more accurate than verified signed F2 in all three banks. Recovering excluded candidates therefore does not guarantee an advantage over search.

Table 12: Primary setting: 37 queries, 128 images per task per query. All deterministic outputs. Acc. and Worst are average and minimum task accuracy (%); Pass counts empirical task caps; Regret is constrained average-error regret (pp). NA denotes a violation. Mask 0 is a base fallback and is not a nonempty success.

| Backbone, bank | Method | Mask | $\|S\|$ | Acc. | Worst | Pass | Regret |
|---|---|---|---|---|---|---|---|
| B/32, 0 | F1-LS37 | 255 | 8 | 61.08 | 46.09 | 6/8 | NA |
| | Signed F2 | 106 | 4 | 64.06 | 38.28 | 7/8 | NA |
| | Permanent exclusion | 10 | 2 | 63.04 | 43.36 | 8/8 | 4.81 |
| | Best observed37 | 10 | 2 | 63.04 | 43.36 | 8/8 | 4.81 |
| | Greedy37 | 111 | 6 | 67.53 | 51.76 | 7/8 | NA |
| B/32, 1 | F1-LS37 | 255 | 8 | 59.42 | 38.48 | 5/8 | NA |
| | Signed F2 | 106 | 4 | 64.45 | 36.13 | 8/8 | 2.98 |
| | Permanent exclusion | 10 | 2 | 62.33 | 39.45 | 8/8 | 5.10 |
| | Best observed37 | 10 | 2 | 62.33 | 39.45 | 8/8 | 5.10 |
| | Greedy37 | 110 | 5 | 65.87 | 54.49 | 8/8 | 1.56 |
| B/32, 2 | F1-LS37 | 255 | 8 | 60.03 | 41.80 | 6/8 | NA |
| | Signed F2 | 251 | 7 | 63.84 | 34.57 | 8/8 | 5.91 |
| | Permanent exclusion | 10 | 2 | 62.72 | 42.58 | 8/8 | 7.03 |
| | Best observed37 | 10 | 2 | 62.72 | 42.58 | 8/8 | 7.03 |
| | Greedy37 | 78 | 4 | 67.60 | 42.38 | 8/8 | 2.15 |
| B/16, 0 | F1-LS37 | 159 | 6 | 70.78 | 41.02 | 7/8 | NA |
| | Signed F2 | 78 | 4 | 73.10 | 41.80 | 8/8 | 0.32 |
| | Permanent exclusion | 192 | 2 | 59.20 | 42.38 | 7/8 | NA |
| | Best observed37 | 12 | 2 | 67.36 | 44.14 | 8/8 | 6.05 |
| | Greedy37 | 64 | 1 | 59.47 | 44.14 | 8/8 | 13.94 |
| B/16, 1 | F1-LS37 | 191 | 7 | 69.38 | 48.83 | 7/8 | NA |
| | Signed F2 | 206 | 5 | 72.24 | 41.21 | 8/8 | 2.00 |
| | Permanent exclusion | 72 | 2 | 64.06 | 42.77 | 8/8 | 10.18 |
| | Best observed37 | 12 | 2 | 67.53 | 43.95 | 8/8 | 6.71 |
| | Greedy37 | 222 | 6 | 71.73 | 41.21 | 8/8 | 2.51 |
| B/16, 2 | F1-LS37 | 254 | 7 | 69.95 | 49.41 | 7/8 | NA |
| | Signed F2 | 83 | 4 | 67.48 | 43.36 | 8/8 | 6.18 |
| | Permanent exclusion | 192 | 2 | 58.52 | 42.38 | 8/8 | 15.14 |
| | Best observed37 | 12 | 2 | 67.68 | 44.53 | 8/8 | 5.98 |
| | Greedy37 | 207 | 6 | 72.07 | 41.21 | 7/8 | NA |

Table 13: Verification setting: 128-image proposals plus up to eight direct calls on a separate 128-image block. All deterministic outputs. Acc. and Worst are average and minimum task accuracy (%); Pass counts empirical task caps; Regret is constrained average-error regret (pp). NA denotes a violation. Mask 0 is a base fallback and is not a nonempty success.

| Backbone, bank | Method | Mask | $\|S\|$ | Acc. | Worst | Pass | Regret |
|---|---|---|---|---|---|---|---|
| B/32, 0 | F1-LS37 | 239 | 7 | 63.70 | 50.00 | 7/8 | NA |
| | Signed F2 | 202 | 4 | 63.57 | 38.48 | 8/8 | 4.27 |
| | Permanent exclusion | 10 | 2 | 63.04 | 43.36 | 8/8 | 4.81 |
| | Best observed37 | 10 | 2 | 63.04 | 43.36 | 8/8 | 4.81 |
| | Greedy37 | 239 | 7 | 63.70 | 50.00 | 7/8 | NA |
| B/32, 1 | F1-LS37 | 0 | 0 | 47.27 | 29.30 | 8/8 | 20.17 |
| | Signed F2 | 204 | 4 | 62.67 | 42.77 | 8/8 | 4.76 |
| | Permanent exclusion | 10 | 2 | 62.33 | 39.45 | 8/8 | 5.10 |
| | Best observed37 | 10 | 2 | 62.33 | 39.45 | 8/8 | 5.10 |
| | Greedy37 | 11 | 3 | 65.77 | 39.65 | 8/8 | 1.66 |
| B/32, 2 | F1-LS37 | 0 | 0 | 47.27 | 29.30 | 8/8 | 22.49 |
| | Signed F2 | 43 | 4 | 66.41 | 41.60 | 8/8 | 3.34 |
| | Permanent exclusion | 10 | 2 | 62.72 | 42.58 | 8/8 | 7.03 |
| | Best observed37 | 10 | 2 | 62.72 | 42.58 | 8/8 | 7.03 |
| | Greedy37 | 78 | 4 | 67.60 | 42.38 | 8/8 | 2.15 |
| B/16, 0 | F1-LS37 | 189 | 6 | 69.26 | 50.78 | 7/8 | NA |
| | Signed F2 | 12 | 2 | 67.36 | 44.14 | 8/8 | 6.05 |
| | Permanent exclusion | 192 | 2 | 59.20 | 42.38 | 7/8 | NA |
| | Best observed37 | 10 | 2 | 69.04 | 44.53 | 8/8 | 4.37 |
| | Greedy37 | 26 | 3 | 69.36 | 43.16 | 8/8 | 4.05 |
| B/16, 1 | F1-LS37 | 207 | 6 | 72.73 | 42.19 | 8/8 | 1.51 |
| | Signed F2 | 74 | 3 | 70.80 | 42.19 | 8/8 | 3.44 |
| | Permanent exclusion | 72 | 2 | 64.06 | 42.77 | 8/8 | 10.18 |
| | Best observed37 | 12 | 2 | 67.53 | 43.95 | 8/8 | 6.71 |
| | Greedy37 | 30 | 4 | 72.24 | 41.60 | 8/8 | 2.00 |
| B/16, 2 | F1-LS37 | 0 | 0 | 55.35 | 42.97 | 8/8 | 18.31 |
| | Signed F2 | 12 | 2 | 67.68 | 44.53 | 8/8 | 5.98 |
| | Permanent exclusion | 192 | 2 | 58.52 | 42.38 | 8/8 | 15.14 |
| | Best observed37 | 12 | 2 | 67.68 | 44.53 | 8/8 | 5.98 |
| | Greedy37 | 26 | 3 | 69.53 | 42.58 | 8/8 | 4.13 |

Table 14: Secondary setting: 37 queries, 256 images per task per query; twice the primary image cost. All deterministic outputs. Acc. and Worst are average and minimum task accuracy (%); Pass counts empirical task caps; Regret is constrained average-error regret (pp). NA denotes a violation. Mask 0 is a base fallback and is not a nonempty success.

| Backbone, bank | Method | Mask | $\|S\|$ | Acc. | Worst | Pass | Regret |
|---|---|---|---|---|---|---|---|
| B/32, 0 | F1-LS37 | 255 | 8 | 61.08 | 46.09 | 6/8 | NA |
| | Signed F2 | 27 | 4 | 65.53 | 40.82 | 7/8 | NA |
| | Permanent exclusion | 11 | 3 | 65.75 | 41.60 | 7/8 | NA |
| | Best observed37 | 10 | 2 | 63.04 | 43.36 | 8/8 | 4.81 |
| | Greedy37 | 15 | 4 | 68.38 | 40.43 | 7/8 | NA |
| B/32, 1 | F1-LS37 | 255 | 8 | 59.42 | 38.48 | 5/8 | NA |
| | Signed F2 | 30 | 4 | 64.82 | 41.02 | 7/8 | NA |
| | Permanent exclusion | 74 | 3 | 64.28 | 37.70 | 8/8 | 3.15 |
| | Best observed37 | 10 | 2 | 62.33 | 39.45 | 8/8 | 5.10 |
| | Greedy37 | 78 | 4 | 67.94 | 41.80 | 7/8 | NA |
| B/32, 2 | F1-LS37 | 255 | 8 | 60.03 | 41.80 | 6/8 | NA |
| | Signed F2 | 14 | 3 | 65.62 | 42.77 | 8/8 | 4.13 |
| | Permanent exclusion | 27 | 4 | 65.65 | 39.65 | 7/8 | NA |
| | Best observed37 | 10 | 2 | 62.72 | 42.58 | 8/8 | 7.03 |
| | Greedy37 | 95 | 6 | 68.63 | 41.02 | 7/8 | NA |
| B/16, 0 | F1-LS37 | 255 | 8 | 68.92 | 48.05 | 6/8 | NA |
| | Signed F2 | 110 | 5 | 72.63 | 52.54 | 8/8 | 0.78 |
| | Permanent exclusion | 74 | 3 | 70.31 | 42.58 | 8/8 | 3.10 |
| | Best observed37 | 10 | 2 | 69.04 | 44.53 | 8/8 | 4.37 |
| | Greedy37 | 207 | 6 | 71.97 | 41.60 | 7/8 | NA |
| B/16, 1 | F1-LS37 | 255 | 8 | 69.17 | 47.66 | 7/8 | NA |
| | Signed F2 | 94 | 5 | 73.10 | 41.02 | 8/8 | 1.15 |
| | Permanent exclusion | 202 | 4 | 69.21 | 41.80 | 8/8 | 5.03 |
| | Best observed37 | 10 | 2 | 68.80 | 43.75 | 8/8 | 5.44 |
| | Greedy37 | 30 | 4 | 72.24 | 41.60 | 8/8 | 2.00 |
| B/16, 2 | F1-LS37 | 255 | 8 | 67.85 | 48.24 | 6/8 | NA |
| | Signed F2 | 94 | 5 | 73.66 | 41.60 | 8/8 | 0.00 |
| | Permanent exclusion | 73 | 3 | 68.68 | 42.97 | 8/8 | 4.98 |
| | Best observed37 | 10 | 2 | 69.14 | 42.97 | 8/8 | 4.52 |
| | Greedy37 | 30 | 4 | 72.85 | 42.58 | 8/8 | 0.81 |

Table 15: All Random37 search seeds, 20 per row. The 37 and +verify settings use 128 images per query; 256 uses 256. Accuracy, adapter count, and worst-task accuracy are means over all outputs, including violations/fallbacks. Valid means feasible and nonempty; Base and Invalid complete the denominator 20.

| Backbone | Setting | Bank | Acc. | $\lvert S\rvert$ | Worst | Valid | Base | Invalid |
|---|---|---|---|---|---|---|---|---|
| B/32 | 37 | 0 | 66.47 | 5.10 | 45.48 | 15 | 0 | 5 |
| B/32 | 37 | 1 | 65.64 | 4.80 | 45.61 | 11 | 0 | 9 |
| B/32 | 37 | 2 | 66.82 | 4.75 | 42.64 | 10 | 0 | 10 |
| B/32 | +verify | 0 | 59.28 | 2.95 | 40.90 | 7 | 3 | 10 |
| B/32 | +verify | 1 | 62.03 | 2.75 | 40.07 | 18 | 0 | 2 |
| B/32 | +verify | 2 | 61.19 | 2.60 | 38.90 | 14 | 4 | 2 |
| B/32 | 256 | 0 | 67.31 | 4.35 | 43.37 | 14 | 0 | 6 |
| B/32 | 256 | 1 | 66.81 | 4.10 | 42.14 | 8 | 0 | 12 |
| B/32 | 256 | 2 | 67.44 | 5.10 | 41.80 | 8 | 0 | 12 |
| B/16 | 37 | 0 | 63.63 | 2.45 | 43.41 | 8 | 5 | 7 |
| B/16 | 37 | 1 | 70.95 | 5.35 | 42.43 | 19 | 0 | 1 |
| B/16 | 37 | 2 | 69.60 | 4.80 | 43.14 | 4 | 0 | 16 |
| B/16 | +verify | 0 | 68.73 | 2.95 | 43.08 | 18 | 0 | 2 |
| B/16 | +verify | 1 | 69.42 | 4.50 | 43.93 | 16 | 1 | 3 |
| B/16 | +verify | 2 | 67.21 | 2.85 | 44.71 | 14 | 2 | 4 |
| B/16 | 256 | 0 | 70.49 | 4.75 | 43.98 | 10 | 0 | 10 |
| B/16 | 256 | 1 | 71.37 | 5.35 | 43.57 | 15 | 0 | 5 |
| B/16 | 256 | 2 | 70.50 | 4.20 | 44.38 | 14 | 0 | 6 |

### M.5 Calibration size, surrogate error, and statistical limits

Table 16 retains all four 64-image blocks. Table 17 pools the absolute error of taskwise predictions over every coalition of size at least three and all eight tasks; size-specific means are weighted by the number of coalitions, rather than averaged equally across sizes. These errors compare discovery predictions with audit error rates and include both sampling error and surrogate misspecification. They cannot be attributed solely to high-order interactions. On B/32, 256-image F2 has smaller MAE than F1 in every bank, yet only one final F2 choice is feasible. On B/16 the 256-image F2 setting performs well, but costs twice as many image evaluations as the primary setting. No block or sample size replaces the predefined primary result after audit inspection.

Table 16: Feasible nonempty outputs for every predefined 64-image discovery block. Each method uses 37 queries. Counts pool three training banks; random search additionally includes 20 search seeds per bank. Shared images and banks make these sensitivity diagnostics dependent.

| Backbone | Method | Block 0 | Block 1 | Block 2 | Block 3 |
|---|---|---|---|---|---|
| B/32 | F1-LS37 | 0/3 | 1/3 | 1/3 | 0/3 |
| | Signed F2 | 1/3 | 2/3 | 1/3 | 1/3 |
| | Permanent exclusion | 3/3 | 2/3 | 2/3 | 3/3 |
| | Best observed37 | 2/3 | 2/3 | 3/3 | 3/3 |
| | Greedy37 | 1/3 | 1/3 | 2/3 | 1/3 |
| | Random37 | 37/60 | 24/60 | 25/60 | 25/60 |
| B/16 | F1-LS37 | 1/3 | 0/3 | 0/3 | 0/3 |
| | Signed F2 | 0/3 | 3/3 | 2/3 | 1/3 |
| | Permanent exclusion | 3/3 | 3/3 | 1/3 | 1/3 |
| | Best observed37 | 3/3 | 3/3 | 3/3 | 2/3 |
| | Greedy37 | 3/3 | 2/3 | 3/3 | 1/3 |
| | Random37 | 40/60 | 19/60 | 36/60 | 24/60 |

For the nominal paired audit diagnostic, let $h$ count examples where a candidate errs and the base is correct, and $g$ count the converse. The excess error is the difference of these two event probabilities. For each bank, the union of its $K$ distinct frozen nonbase outputs is evaluated using a Clopper–

Table 17: Mean absolute error of predicted task error rates versus audit rates (pp), pooling all 219 coalitions of size at least three and eight tasks. This includes sampling error and model misspecification. The 256-image setting doubles the 128-image query cost.

| Backbone | Bank | 128: F1 | 128: F2 | 256: F1 | 256: F2 |
|---|---|---|---|---|---|
| B/32 | 0 | 7.41 | 8.48 | 7.91 | 5.09 |
| B/32 | 1 | 7.63 | 7.81 | 8.40 | 5.05 |
| B/32 | 2 | 7.42 | 6.66 | 8.00 | 5.07 |
| B/16 | 0 | 8.30 | 6.79 | 7.95 | 5.32 |
| B/16 | 1 | 7.79 | 6.32 | 7.70 | 4.97 |
| B/16 | 2 | 7.76 | 9.65 | 7.94 | 6.96 |

Pearson upper bound on $p_h$ minus a lower bound on $p_g$, allocating $\alpha = 0.05/(2K \cdot 8)$ to each tail. No nonbase output passes all task bounds in any bank. This is a nominal iid diagnostic on a repartitioned pool, not an asserted finite-population or deployment certificate. It is distinct from the 128-image empirical verification rule and from the absolute-risk Hoeffding procedure of Theorem 6. Audit images are not used to repair a failed selection.

Training-seed, search-seed, and block repetitions all share observations, and the two backbones share the data split. We give complete seed results and empirical counts without treating them as independent Bernoulli trials or reporting a significance test from six independent datasets. In particular, empirical feasibility of a 512-image output is not a known population label against which the 128-image verification decision can be declared statistically wrong.

### M.6 Preceding negative experiments and choice of scale

We chose the fixed $0.5$ scale after earlier diagnostics. E3/E4 evaluated eight public rank-16 adapters at scale 1 and final rank 16, with 256 audit images per task. Only the base satisfied all eight task caps, so both feasible-cardinality optima were zero and this criterion gave no non-downward-closure witness. Across three discovery blocks, high-order F2 error-rate MAE was approximately $0.438$–$0.561$, worse than F1. The later controlled-bank experiments changed the scale in response to these diagnostics.

E5 then fixed seven diagnostic masks $\{0, 1, 8, 9, 13, 128, 255\}$ after seeing E3/E4 and compared scale 1 versus $0.5$, each with dense and rank-16 merges on the same images. Dense scale-1 merges also failed badly, so compression alone did not explain the failure. Native PEFT and explicitly merged adapter evaluation agreed. Scale $0.5$ improved the triple, but it still regressed by 3.125 pp on DTD in those samples. These were post-hoc diagnostics, not an independent confirmation or a complete new oracle.

E6 froze scale $0.5$ and final rank 16 and enumerated all 256 public-adapter coalitions on new images: three 64-image discovery blocks and 256 audit images per task. The empirical feasible masks were $\{0, 8, 13, 64, 65\}$; the oracle triple (mask 13) reached $59.33\%$ average accuracy versus $47.46\%$ for the base, with cardinality optima three versus one. Signed F2 selected masks 0, 13, and 77 across the three 64-image blocks; mask 77 violated the cap. The predefined first block yielded no signed-feasible, graph-excluded mechanism witness. Thus E7a freezes a triple found in exploratory E6 audit analysis and confirms it on new examples, while E7/E8 train separate controlled adapter banks. None of these stages justifies omitting the unsuccessful scale, block, or method results.

### M.7 Evidence provenance and reproduction

The primary E7/E8 tables are generated from completed, validated records. The additional untruncated inference and post-hoc diagnostics are described in Appendices M.8–M.10; no adapters were retrained. The supplementary ZIP preserves the relative paths of the analysis code, protocols, tabular results and per-example error caches. It supports offline table checks and selection replay; large trained checkpoints and the complete coalition-logit archives are omitted to meet the upload size limit. The inventory below describes the full experiment workspace, whose omitted files are distinguished in the supplementary README. Paths are relative to the workspace root:

- `real_lora_rescue_scan/E7_20260914_complete/` contains E7's validation record and `E7_controlled_clip_3seed_20260914/`, including the frozen witness.
- `real_lora_rescue_scan/E8_complete/` contains E8's validation record and `E8_clip_b16_3seed/`.
- Each run retains `protocol.json`, `sample_manifest.csv`, training checkpoints, source snapshots, classification heads/prompts, and all 525 choices in `all_selection_results.csv`. Each bank retains `oracle.csv`, frozen selections/shortlists/fits, per-example discovery/audit caches, raw coalition predictions, and paired diagnostics.
- Under `real_lora_rescue_scan/`, the folders `E1_E4_20260912_complete/`, `E5_20260912_diagnostics/`, and `E6_20260913_complete/` retain the preceding negative and diagnostic stages.
- `workflow/review_e7.py` independently verifies both result formats. `workflow/paper_e7_e8_tables.py` rebuilds the marked manuscript tables, or checks them with `--check`, using the stored CSV/JSON records without model inference.

Validation passed for 24 final checkpoints and 776 raw coalitions in E7 (including eight witness subsets), and 24 checkpoints and 768 coalitions in E8. All 525 choices per backbone replay without audit access; source/protocol hashes and training/holdout separation match. Maximum cross-entropy recomputation discrepancies from saved logits are below $1.9 \times 10^{-6}$, and numerical selection replay differences are below $9 \times 10^{-16}$. These are implementation checks, not independent external reproduction. Recorded training/evaluation times are approximately 1 h 41 min and 3 h 14 min, respectively, and are not normalized hardware comparisons. Model/data acquisition and the frozen GPU environment are needed to rerun; rebuilding the paper uses only the stored records.

### M.8 Untruncated compression control

To test whether the observed representation cost requires low-rank truncation, we evaluate $W_0 + 0.5 \sum_{i \in S} D_i$ for all 256 coalitions in each of the three E7 B/32 and three E8 B/16 banks, adding 1,536 coalition evaluations. We retain the original adapters, heads, preprocessing, FP32 evaluation and 512 audit images per task. Eight rank-8 updates have total rank at most 64; the original collector with rank 64 retains all singular directions of the concatenated update. No adapter is retrained and no selection is retuned. Both protocols were frozen before inspecting the new control outcomes, but the original audit results had already been seen. B/32 execution was conditional only on runtime and resources, not on B/16 performance. Both controls completed.

A synthetic full-factor reconstruction check validates the untruncated implementation. Before collection, original rank-8 masks 0, 1, 3 and 7 reproduce their saved labels and predictions; final base and singleton logits agree exactly across representations. Each backbone's numerical validation checks 768 new coalition files and 6,144 paired task slices, reproduces the original rank-8 oracles, and confirms unchanged frozen-source hashes. Cross-entropy recomputation errors are below $1.9 \times 10^{-6}$. Complete raw archives and source/input snapshots are hash-verified locally.

Without truncation, all six banks have positive accuracy costs of downward closure, ranging from 0.81 to 5.22 pp (Table 18). Thus rank truncation is not necessary for this empirical restriction in these banks. Compression changes its magnitude in both directions: B/16 bank 1 and B/32 bank 2 have larger gaps after compression, whereas B/32 bank 1 has zero rank-8 gap. Unequal final ranks make this a mechanism comparison, not an equal-capacity method comparison; it also does not identify the interaction order of the underlying losses.

We retain all $3\binom{8}{3} = 168$ pair-to-triple directions per bank. A direction is a rescue when the pair violates at least one task cap and the triple passes all eight, each using the same base and empirical 2 pp thresholds. Of 1,008 directions, 85 are rescues in both representations, 12 only without truncation, 20 only at rank 8, and 891 in neither (Table 19). The totals are 97 untruncated and 105 rank-8 rescues. These are cap-crossing events, distinct from the taskwise cross-entropy contrasts in Eq. (38); both records are retained. Shared combinations and observations preclude treating the directions as independent trials.

The source snapshots, protocols, per-example predictions and validation records are stored in `real_lora_rescue_scan/submission_compression_20260924/` (B/16) and

Table 18: Complete-cube compression control on the original audit images. Full and Down are the best eight-task mean accuracies (%) in the empirical feasible family and its downward core; Gap is their difference in pp. Maximum cardinalities are optimized separately. Rank-8 rows reproduce the primary results; untruncated updates have coalition-dependent rank at most 64.

| Backbone | Bank | Representation | Full acc. | Down acc. | Gap | Full $\|S\|$ | Down $\|S\|$ |
|---|---|---|---|---|---|---|---|
| B/32 | 0 | Untruncated | 69.63 | 64.40 | 5.22 | 7 | 4 |
| B/32 | 0 | Rank-8 | 67.85 | 64.65 | 3.20 | 7 | 4 |
| B/32 | 1 | Untruncated | 69.24 | 68.43 | 0.81 | 7 | 4 |
| B/32 | 1 | Rank-8 | 67.43 | 67.43 | 0.00 | 7 | 4 |
| B/32 | 2 | Untruncated | 70.90 | 68.68 | 2.22 | 7 | 4 |
| B/32 | 2 | Rank-8 | 69.75 | 67.46 | 2.29 | 7 | 4 |
| B/16 | 0 | Untruncated | 75.27 | 70.56 | 4.71 | 7 | 3 |
| B/16 | 0 | Rank-8 | 73.41 | 70.31 | 3.10 | 7 | 3 |
| B/16 | 1 | Untruncated | 75.15 | 72.34 | 2.81 | 6 | 4 |
| B/16 | 1 | Rank-8 | 74.24 | 70.80 | 3.44 | 7 | 3 |
| B/16 | 2 | Untruncated | 75.10 | 69.87 | 5.22 | 6 | 3 |
| B/16 | 2 | Rank-8 | 73.66 | 69.14 | 4.52 | 6 | 2 |

Table 19: Every pair-to-triple direction under the empirical eight-task 2 pp caps. Both denotes rescue in both representations; U only and R8 only denote rescue only before or after rank-8 truncation. Neither retains all non-rescue directions. The 1,008 directions share coalitions and images and are not independent trials.

| Backbone | Bank | Both | U only | R8 only | Neither |
|---|---|---|---|---|---|
| B/32 | 0 | 11 | 1 | 6 | 150 |
| B/32 | 1 | 8 | 2 | 2 | 156 |
| B/32 | 2 | 16 | 3 | 1 | 148 |
| B/16 | 0 | 18 | 5 | 0 | 145 |
| B/16 | 1 | 11 | 0 | 5 | 152 |
| B/16 | 2 | 21 | 1 | 6 | 140 |
| Total | – | 85 | 12 | 20 | 891 |

`real_lora_rescue_scan/submission_compression_b32_20260924/` (B/32). The B/32 validation file inherits an E8 scope-label typo from the shared analyzer; its recorded input directory, source hashes and rank-8 replay identify E7 correctly. The original record is preserved with a separate provenance note. These controls reuse inspected observations and supply neither independent test evidence nor a population-risk certificate.

### M.9 Post-hoc sensitivity and selection diagnosis

These diagnostics reuse the already inspected E7/E8 audit observations. They do not change the predefined 2 pp cap, any frozen selection, or the primary tables. Table 20 recomputes the complete feasible family and its downward core at five caps. The empirical accuracy cost is positive in all six banks at 1, 3 and 5 pp, and in five at 2 pp. Thus the observed restriction is not confined to the original empirical cutoff, although its magnitude changes sharply with the cap. These correlated sensitivity settings provide neither independent replications nor population guarantees.

Table 20: Post-hoc cap sensitivity of the complete-cube empirical accuracy cost of downward closure (pp). Each column recomputes both feasible families and both accuracy optima on the same 512 audit images per task. The predefined main cap remains 2 pp; no column is an independent replication.

| Backbone | Bank | 0 pp cap | 1 pp | 2 pp | 3 pp | 5 pp |
|---|---|---|---|---|---|---|
| B/32 | 0 | 16.36 | 5.18 | 3.20 | 1.83 | 1.83 |
| B/32 | 1 | 13.99 | 6.45 | 0.00 | 2.15 | 2.15 |
| B/32 | 2 | 16.16 | 5.40 | 2.29 | 2.29 | 2.15 |
| B/16 | 0 | 14.01 | 13.35 | 3.10 | 3.69 | 3.69 |
| B/16 | 1 | 13.62 | 5.98 | 3.44 | 2.49 | 2.49 |
| B/16 | 2 | 13.33 | 5.15 | 4.52 | 3.05 | 2.20 |

Table 21 separates the quality of the already frozen candidate list from its verification output. In B/16 banks 0 and 1, the signed list contains an empirically feasible candidate respectively 5.74 and 1.44 pp more accurate than the output selected by verification. In bank 2, the best feasible member of the signed list reaches only 67.68%, below the greedy list's 69.53%. The first two cases locate an observed loss within the re-selection stage; the third also exposes a proposal limitation. These are post-hoc audit comparisons, not permission to replace the actual selections. An audit point estimate is not a known population label, so the comparison does not establish that verification made a statistically incorrect rejection.

Table 21: Frozen-shortlist diagnosis. Best is the highest audit accuracy (%) among empirically feasible members of the already frozen shortlist, including the base. Selected is the actual verification output; its feasibility is assessed on the audit. Gap (pp) is defined only when that output is feasible. Best is a post-hoc diagnostic upper bound within this list, not the method's achieved accuracy.

| Backbone | Bank | Method | Best | Selected | Feasible | Gap |
|---|---|---|---|---|---|---|
| B/32 | 0 | F2 | 63.67 | 63.57 | Yes | 0.10 |
| B/32 | 0 | Greedy | 67.72 | 63.70 | No | NA |
| B/32 | 1 | F2 | 64.45 | 62.67 | Yes | 1.78 |
| B/32 | 1 | Greedy | 65.87 | 65.77 | Yes | 0.10 |
| B/32 | 2 | F2 | 66.43 | 66.41 | Yes | 0.02 |
| B/32 | 2 | Greedy | 67.60 | 67.60 | Yes | 0.00 |
| B/16 | 0 | F2 | 73.10 | 67.36 | Yes | 5.74 |
| B/16 | 0 | Greedy | 70.31 | 69.36 | Yes | 0.95 |
| B/16 | 1 | F2 | 72.24 | 70.80 | Yes | 1.44 |
| B/16 | 1 | Greedy | 72.24 | 72.24 | Yes | 0.00 |
| B/16 | 2 | F2 | 67.68 | 67.68 | Yes | 0.00 |
| B/16 | 2 | Greedy | 69.53 | 69.53 | Yes | 0.00 |

### M.10 Paired image-bootstrap stability

We generate 1,000 image-bootstrap replicates with NumPy random seed 20260924, resampling 512 images with replacement within each task. All coalitions and methods share the same resampling

counts, as do the three banks and both backbones, whose audit identity manifests are verified to match. Tasks are resampled separately and retain equal weight. We recompute each complete feasible family and downward core for Table 22; the candidates in Tables 23 and 24 remain frozen. No retraining or discovery re-selection is performed.

The bootstrap uses images as the resampling unit; it does not model unrecorded source-image clustering, training uncertainty or a new dataset. The empirical optimum difference is nonnegative by construction, and small label changes can exclude many supersets. Consequently a positive resampling distribution is not a significance test for a population representation gap. For example, the zero observed B/32 bank-1 gap has a positive bootstrap median. We report the quantiles as descriptive sensitivity, not calibrated confidence limits for this nonsmooth constrained optimum. All three B/16 banks retain a positive gap in all 1,000 resamples; the zero-gap frequencies for B/32 banks 0–2 are 0.3%, 2.9%, and 3.0%.

Table 22: Image-bootstrap stability of the empirical accuracy cost (pp), at the original 2 pp cap. Every resample reconstructs the feasible family, downward core and their optima. The 2.5th and 97.5th percentiles describe the resampling distribution; they are not asserted confidence limits for a population optimum difference or a significance test.

| Backbone | Bank | Observed | 2.5th pct. | Median | 97.5th pct. |
|---|---|---|---|---|---|
| B/32 | 0 | 3.20 | 0.95 | 4.25 | 10.35 |
| B/32 | 1 | 0.00 | 0.00 | 3.56 | 9.82 |
| B/32 | 2 | 2.29 | 0.00 | 3.17 | 8.55 |
| B/16 | 0 | 3.10 | 2.71 | 7.23 | 14.70 |
| B/16 | 1 | 3.44 | 2.15 | 5.57 | 12.01 |
| B/16 | 2 | 4.52 | 2.61 | 5.41 | 13.70 |

Table 23 shows that empirical passage and resampling stability differ. B/16's primary signed candidates pass all task caps in 68.8%, 38.7%, and 85.7% of the resamples; the verified candidates pass in 86.4%, 87.7%, and 94.1%. The accompanying accuracy losses remain those of the original tables. These frequencies do not establish population safety or that verification universally improves selection. All five deterministic methods in the primary and verification settings are retained in the saved 68-row selection-stability file (60 controlled outputs and eight witness subsets). The accompanying 544 taskwise rows give harmful/helpful paired counts, cap slack, and excess-error resampling quantiles.

Table 23: Frozen signed F2 outputs: mask, observed worst task regression (pp), and percentage of resamples passing all eight empirical 2 pp caps. Choices remain fixed during resampling. Pass frequencies are descriptive stability measures, not probabilities of population safety.

| | | Primary | | | Verified | | |
|---|---|---|---|---|---|---|---|
| Backbone | Bank | Mask | Worst | Pass (%) | Mask | Worst | Pass (%) |
| B/32 | 0 | 106 | 2.15 | 46.7 | 202 | 1.56 | 54.8 |
| B/32 | 1 | 106 | 1.76 | 41.4 | 204 | 1.95 | 31.8 |
| B/32 | 2 | 251 | 1.17 | 48.6 | 43 | 1.37 | 64.9 |
| B/16 | 0 | 78 | 1.17 | 68.8 | 12 | 0.59 | 86.4 |
| B/16 | 1 | 206 | 1.76 | 38.7 | 74 | 0.78 | 87.7 |
| B/16 | 2 | 83 | -0.20 | 85.7 | 12 | 0.39 | 94.1 |

For the frozen public witness, the triple remains empirically feasible in 72.1% of resamples, but the joint pattern in which it passes and all six nonempty proper subsets fail appears in only 16.5%. The single-image subset violations therefore make the complete boundary pattern fragile, even though the original measured inequalities and average-accuracy gain are unchanged. We retain the witness as an empirical example and do not treat it as a stable population characterization. The 16.5% joint pattern is stronger than non-downward-closure, which requires only one failing subset of a passing triple. Masks 4 and 5 fail in every resample, so the triple supplies such a witness in the same 72.1% of resamples in which it passes. Neither frequency is a probability of population feasibility.

Table 24: Image-bootstrap stability of every frozen public-adapter subset (E7a). Worst is observed excess error over the base (pp); slack is 2 pp minus Worst. Pass is the percentage of resamples meeting all eight caps. These are repeated draws from the original audit, not new examples.

| Mask | Worst | Slack | Pass (%) |
|---|---|---|---|
| 0 | 0.00 | 2.00 | 100.0 |
| 1 | 2.15 | -0.15 | 42.4 |
| 4 | 16.80 | -14.80 | 0.0 |
| 5 | 16.99 | -14.99 | 0.0 |
| 8 | 2.15 | -0.15 | 33.8 |
| 9 | 2.15 | -0.15 | 35.4 |
| 12 | 2.73 | -0.73 | 16.6 |
| 13 | 1.17 | 0.83 | 72.1 |

The supplemental protocol, source hashes, reproducible analysis and all resampling summaries are stored under `outputs/submission_night_20260924/`. The cap and shortlist tables come from `outputs/experiment_self_check_20260915/` and were recomputed against the original caches before inclusion. These additions do not supersede the unsuccessful multiplicity-corrected audit diagnostics above.

# N EXPERIMENTAL PROTOCOL AND ADDITIONAL RESULTS

## N.1 SCOPE AND COMPUTATIONAL RECORD

The original controlled experiments described in this section were executed on a CPU with one numerical thread and no downloaded pretrained model. The released-CLIP experiments have a separate GPU protocol in Appendix P. Reported versions are Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0 and PyTorch 2.10.0+cpu. The planted experiment took approximately 3.36 seconds in this environment; the five digit seeds each took approximately 2.45–2.75 seconds. These are measured local runtimes, not transferable hardware benchmarks. Exact software versions and raw observations are included. The source package contains no font files.

## N.2 PLANTED EXPERIMENT

Seeds are the integers $0, \ldots, 19$. Generate a random permutation of all $\binom{16}{3}$ triples and scan it, accepting a triple only if none of its members already has degree 6, until 16 triples are accepted. Generate utilities independently from $\mathrm{Uniform}[0.7, 1.3]$. Every penalty is $1 + \sum_i a_i$, which makes any conflict-containing assignment worse than the base. Evaluate the exact set loss on all masks $0, \ldots, 2^{16} - 1$; these observations form the diagnostic oracle only.

Anchored order-$q$ fitting receives all masks of cardinality at most $q$. Its prediction is minimized over all masks, but those additional predictions are not network queries. Global quadratic least squares uses 697 distinct masks, including zero, and all monomials of degree at most two. Random search chooses the lowest true queried loss among the same 697 observations. The clique baseline receives all true triples, inserts all pairs in them, and chooses its maximum-utility independent set exactly. It has privileged structural information and is not listed as a zero-data practical method. Results are in `constructed_raw.csv` and `constructed_summary.json`.

## N.3 FIVE-BLOCK RESCUE EXPERIMENT

The adapter scopes are $\{0, 1, 2\}, \{3, 4, 5\}, \ldots, \{12, 13, 14\}$, with the first two members required present and the third required absent. The risk is the sum of their five unit indicators. The utility is the number of retained adapters and the risk cap is zero. The actual network realization uses width 20 and four diagonal layers. All adapter updates have rank one or zero in each layer. A mixed-integer solver maximizes utility with the exact signed coefficients and literal cuts. Its solution retains 15 vertices; the unsigned positive-pair policy retains 10. The local-moment lower-bound checks use cap $0.5$ to distinguish the harmful face's minimum 1 from a compatible completion's risk 0. No statistical uncertainty is present in this exact polynomial experiment. Results are in `rescue_results.json`.

## N.4 LEARNED DIGIT ADAPTERS

The dataset is loaded with `sklearn.datasets.load_digits`, whose original images have shape $8 \times 8$ (scikit-learn developers, 2026). Pixel values are divided by 16. For each seed $s \in \{0, 1, 2, 3, 4\}$, use stratified splits with random states $s$, $s + 100$ and $s + 200$: first allocate 60% of identities to training, then 37.5% of the remainder to discovery, then 40% of the new remainder to audit. The exact counts are 1,078/269/180/270. All transformed versions of one identity stay in the same split. The six domains are right/left/down/up shifts by one pixel with zero padding, and rotations by $+25$ and $-25$ degrees using bilinear interpolation, fixed image size, zero padding and no prefiltering. No untransformed domain is included among the six merge tasks.

The base is a three-linear-layer MLP of widths 64, 64, 32 and 10, with GELU after the first two layers. Train all base weights and biases on original training images with full-batch Adam, learning rate $0.005$, weight decay $10^{-4}$ and 150 epochs. Freeze the base completely. Each domain gets an independent adapter with rank 4 at all three layers: $A$ is initialized from $0.05\mathcal{N}(0, 1)$ and $B = 0$. Train with full-batch Adam, learning rate $0.01$, weight decay $10^{-4}$ and 160 epochs on that domain's training identities. The update is $BA$ without an additional training-time scale. In common LoRA notation this corresponds to setting the factor $\alpha/r$ to one. Merging subsequently uses the declared four-point scale grid.

All candidate losses on discovery are computed as the mean over six task-specific means of $\min(\mathrm{CE}, 5)/5$. Anchored order 2 and order 3 fit that *average* risk for proposal generation; they do not produce taskwise structural certificates in this experiment. The final audit separately checks every task. Each surrogate yields one predicted minimizer per scale; these four proposals are then evaluated directly on discovery data and the best is selected. Exhaustive discovery minimizes true discovery loss over all $4 \times 64$ scale–mask combinations. It is not a test oracle. Global quadratic fitting uses 42 distinct random masks per scale, with a shared mask design across scales. Random search selects the best of those observations. Best-single and all-adapter baselines also choose their scales using discovery only. Zero is included as a separate base candidate.

The same selected scale and mask are evaluated before and after rank-4 SVD compression at each layer. Therefore the rank-4 comparison isolates a deployment change; it is not a compression-aware oracle search. We do not claim that its exhaustive-discovery row is optimal among compressed models. Eight labeled rank-4 outputs (including possible duplicate models) form the frozen audit list. The task cap $0.30$ is set in the script before inspecting results, and no failed run is retuned to make the audit accept.

Table 25: Mean absolute prediction error on the 22 coalitions of size at least four. All quantities are in normalized clipped-loss units; five-seed mean $\pm$ standard deviation.

| Scale | Anchored order 2 | Signed order 3 |
|---|---|---|
| 0.25 | $0.00775 \pm 0.00210$ | $0.00426 \pm 0.00049$ |
| 0.50 | $0.03733 \pm 0.01968$ | $0.02695 \pm 0.01053$ |
| 0.75 | $0.07952 \pm 0.03835$ | $0.05774 \pm 0.02343$ |
| 1.00 | $0.10636 \pm 0.04749$ | $0.07418 \pm 0.02142$ |

At large scales both surrogates become less accurate. The decline is consistent with the role of omitted contextual terms, but it does not independently estimate the uniform derivative bound in Theorem 8. The relative reduction of $45.1\%$ is computed from the means at scale $0.25$, not from an average of per-seed percentage changes.

**Test reporting and uncertainty.** Digit accuracy is averaged over the six task accuracies on the 270 held-out identities. Their transformed copies are not treated as 1,620 independent identities for uncertainty calculations. Standard deviations are over the five train/split seeds. The selected method can have lower clipped loss without higher accuracy because those are different objectives. Table 30 reports losses without standard deviations only to reduce width; the full means, standard deviations and worst-task accuracies are in `digits_summary.json`. Random search achieves $46.16 \pm 1.98\%$ dense accuracy and $45.68 \pm 2.29\%$ after compression; the base achieves $42.67 \pm 1.99\%$. No significance or state-of-the-art claim is made from these small differences.

**Stored evidence.** `digits_seedS.pt` stores the base, six adapter state dictionaries and all split identities for seed S. `digits_seedS_observations.npz` stores the complete diagnostic discovery tensor, the frozen rank-4 candidates' per-example audit losses and the random query design. The decision-making code restricts each method to its declared discovery subset. `digits_raw.csv`, `digits_diagnostics.csv`, `digits_metadata.json` and `digits_specialists.json` contain the reported measurements. The full tensor is released to make oracle comparisons inspectable, not to suggest that low-order methods need all its observations in deployment.

## O ALGORITHMIC PSEUDOCODE AND CLAIM BOUNDARIES

**Scope of the results.** Theorem 2 gives the sharp cardinality cost of downward closure in single-layer rank-one LoRA, with Corollaries 2–5 describing recovery, realization, and robustness. Theorem 5 gives the minimax loss-oracle ratio across degree classes, a common-base rank-one realization, and the minimum depth for a nonzero zero-transcript squared loss. It requires the stated information and low-order feasibility assumptions. Applying either theorem to an empirical rescue still requires measuring the relevant finite differences. The remaining results derive conditions for representing literal conflicts, bounding omitted terms, and certifying exclusions or final-model risk.

**CH-MERGE: context-lifted selection with a deployment audit**

**Inputs:** frozen base and adapters, scales, order $q$, task caps, utility, selection/final-rank budgets, discovery data, independent audit data, and optional certified approximation radii.
1. Evaluate the declared discovery coalitions. Compute taskwise signed coefficients by inclusion–exclusion; do not remove negative terms solely because they are negative.
2. If valid radii exist, test candidate faces with local-moment dual lower bounds and retain only certified literal cuts. Shorten a cut only by rechecking its bound. Otherwise label structural decisions heuristic.
3. Solve a sufficient-risk constrained integer program when certification is possible, or use signed surrogate minimization to propose a shortlist otherwise. Include baselines explicitly rather than assuming the full merge is useful.
4. Fuse realized products, apply the declared rank compression, and freeze every shortlisted model. Record its complete scale, mask, rank and numerical representation.
5. Evaluate the frozen candidates on independent audit examples. Apply the simultaneous bound for the entire candidate list. Return the highest-utility passing model, or abstain.

The analysis uses established conflict hypergraphs, no-good cuts, Möbius inversion, Sherali–Adams relaxations, LoRA gauge invariance, SVD compression, and Hoeffding bounds, with attribution where each is used. Related Hodge/cohomological diagnostics are discussed in Section 2. Our experiments compare selection rules under a fixed merge operator; they do not reproduce the large-model merging baselines or establish a general benchmark advantage.

# P RELEASED-CLIP PROTOCOL AND EVIDENCE PROVENANCE

## P.1 MODELS, DATA, AND NUMERICAL EVALUATION

The released adapters come from the KnOTS setup at source revision `baa35e85768b8841820d35edc3e41b25304248c5`. The workspace records the base-model and adapter revisions, hashes, prompts, and saved classifier heads. The four adapter/task names are MNIST, SVHN, GTSRB, and EuroSAT, in this mask-bit order. Dense increments are checked against PEFT's merged products in FP32, including identity checks on unaffected weights. The initial and fresh-example scans use the same FP16 image encoder and FP32 normalized features, heads, logits, and cross-entropy, with FP64 averaging. Dense increments and coalition sums are accumulated in FP32 and then cast once. Inference uses a single NVIDIA GeForce RTX 4090 D, batch size 64. The recorded initial scan peaks at approximately 0.85 GiB allocated GPU memory.

Classifier construction uses the official KnOTS class names and prompts, with saved frozen heads of norm 100. Evaluation applies the original CLIP logit scale; the additional factor of 100 used in the KnOTS training loop is not added at evaluation. There is no temperature fitting or head selection. Consequently high accuracy need not correspond to a very small cross-entropy for these checkpoints. Every original mean loss and accuracy was recomputed from saved logits independently of the inference aggregation.

MNIST, SVHN, and GTSRB use checksum-verified canonical test pools. EuroSAT uses the public 2,700-example test mirror from `tanganke/eurosat`. Exact membership relative to the KnOTS authors' local split is not verifiable from the available repository. Neither discovery nor fresh-example resampling resolves possible overlap with adapter training data. The fresh validation supports consistency on disjoint evaluation images, conditional on this common source limitation.

## P.2 FOUR DISTINCT EVALUATION PROTOCOLS

**Descriptive discovery scan.** All 16 coalitions at two scales and four tasks give 128 task/scale/coalition losses on 128 images per task, seed 42. The 96 pair–rescuer candidates reuse these values. The conflict threshold is $C > 0.02$, yielding 44 pairs and 88 rescuer opportunities; 13 opportunities satisfy $R > 0.02$. One conflicting pair has two strong rescuers, hence 12 rescued pairs. Contextual contrasts are descriptive recomputations of the same losses.

**Strict pair-only extrapolation.** Masks $0, 1, 2, 3, 4, 5, 6, 8, 9, 10, 12$ form the training design; masks $7, 11, 13, 14, 15$ are unseen targets. Separate task/scale interpolation retains every coefficient's sign. All fitting completes before higher-order targets are read. The graph uses $C > 0.02$; an allow is a predicted strict decrease relative to at least one contained conflict pair. This is 11 observations per task/scale, or 88 task-specific observations overall; the complete 128-loss scan is retained as an evaluation oracle. Primary metrics count 40 distinct higher coalitions, not 132 overlapping pair–superset comparisons. Cases shown as examples are selected for exposition after evaluation and do not tune the policy.

**Fresh examples with frozen decisions.** The original 40 cases, 30 positive cases, 28 allows, predictions, and discovery-selected witness pairs are frozen in the protocol before inference. Each task uses 512 examples sampled from its original test pool with seed 43, excluding all discovery indices and RGB hashes and enforcing unique fresh hashes. Checkpoints and heads are hash-verified. A fresh margin is the fixed witness's fresh pair loss minus the coalition's fresh loss. No new witness is chosen even if another pair would yield a larger gain. Cutoffs $0, 0.01, 0.02, 0.05$ report this margin at different sizes without altering the original decision. The full-policy precision denominator remains 28. These are empirical replication counts, not the bounded population-risk audit of Theorem 6; unmodified cross-entropy does not satisfy its $[0, 1]$ premise.

**Coalition holdout and scale scan.** In leave-one-coalition-out fitting, each model is independently fit on the other 15 masks, using an intercept and every monomial through the declared degree with unrestricted coefficients and equal weights. Cubic fits have 15 parameters; quadratic fits have 11. No held-out loss enters fitting. On the balanced 16-point Boolean design, leverage is $15/16$ and $11/16$, respectively; PRESS checks multiply full-fit residuals by 16 and 3.2 and reproduce the explicit refits. This identity is a verification of the folds, not the fitting procedure. The scale scan instead uses the same 512 fresh images and frozen heads in FP32 for encoder and weights, with all 16 coalitions at five scales (320 task-specific values). Its coefficients are full Möbius differences, not globally refit regression coefficients. Exact reconstruction is checked to tolerance $10^{-12}$; median, mean absolute, and RMS magnitudes are computed by order. Pooled summaries concatenate coefficients across tasks before aggregation. Log–log slopes use all five scales without deleting nonmonotonic observations. There is only one fourth-order coefficient per task/scale, and no slope confidence claim is made.

### P.3 Source map and reproduction scope

Paths below are relative to the workspace root; the supplementary archive's README gives the packaged evidence map and reproduction commands. The main original loss table has SHA256

```
52d26481b6d517a2abdee976bf8a0b043512653bdc68124f8ed8e20246fc2fea
```

- `real_lora_rescue_scan/results/`: coalition losses, raw logits, prompts, model metadata, sample manifest, and validation records for the initial scan.
- `real_lora_rescue_scan/pair_only_experiment/`: training masks, coefficients, held-out predictions, pair/superset decisions, and aggregate metrics for Table 26.
- `real_lora_rescue_scan/contextual_analysis/`: $D_0, D_k, Q$ and excess-rescue contrasts; `signed_polynomial_analysis/` contains the in-sample regression comparison under the same parent directory.
- `real_lora_rescue_scan/leave_one_coalition_out/`: independent refits, predictions, and full confusion matrices for Table 28.
- `real_lora_rescue_scan/independent_data_validation/`: frozen protocol, fresh sample hashes, losses, and full-policy metrics for Table 27.
- `real_lora_rescue_scan/conjecture_experiments/`: full synthetic trial records, exact-search validation, and `interaction_scaling/` observations for the scale analysis.

The corresponding analysis scripts are retained in `real_lora_rescue_scan/`; GPU preparation and evaluation scripts are also in `workflow/`. The manuscript-local `run_all.sh` runs the original CPU controls only. Rebuilding the PDF requires neither downloading models nor rerunning inference.

Table 26: Pair-only prediction on 40 unseen larger coalitions. Both rules use the same 11 losses per task and scale. A positive case improves on at least one contained conflicting pair; 30 cases meet this criterion.

| Policy | Correct allows | False allows | False exclusions | Correct exclusions |
|---|---|---|---|---|
| Conflict graph | 0 | 0 | 30 | 10 |
| Signed quadratic | 27 | 1 | 3 | 9 |

Public checkpoint/data acquisition and the recorded GPU environment are required to regenerate the released-CLIP observations from scratch. The existing results are evidence, not a claim that all benchmark dependencies are bundled in this manuscript folder.

# Q Released-adapter extrapolation and interaction-order controls

## Q.1 Released LoRAs and the pair-only test

We use four public KnOTS rank-16 LoRA adapters for CLIP ViT-B/32, trained on MNIST, SVHN, GTSRB, and EuroSAT (Stoica et al., 2025). This four-adapter study uses no new adapter training. The fixed merge is $W_0 + \lambda \sum_{i \in S} D_i$, with $\lambda \in \{0.5, 1\}$ and all 16 coalitions. Discovery uses 128 examples per task, sampled before inference with seed 42. Loss is unmodified mean cross-entropy in nats. Encoder inference is FP16; dense update accumulation, classifier heads, logits, and cross-entropy use FP32. These dense sums are not compressed to a common final rank. Appendix P records the precision, classifier, split, and source limitations.

The initial scan finds 44 conflicting task–pair–scale combinations and 13 strong pair–rescuer cases among 88 opportunities containing a conflicting pair. Twelve of the 44 pairs have at least one strong rescuer. These counts establish empirical rescue on the scanned samples. They do not establish its interaction order: only 2 of the 13 cases have $Q > 0.02$, none halves an initially positive $D_0 > 0.02$, and 12 have $R - G_0 > 0.02$. Equation (40) explains why this last observation alone is insufficient to identify a cubic mechanism.

**Matched information, unseen coalitions.** For each task and scale, the pair-only model interpolates the empty coalition, four singletons, and six pairs: 11 observations. All four triples and the four-adapter coalition are held out. The graph and signed model share these observations; no held-out loss selects a coefficient, threshold, or model. There are 40 distinct test cases across tasks and scales. Let $E_t(S)$ be the conflicting pairs contained in $S$. For this comparison, a rescued superset satisfies

$$F_t(S) < \max_{e \in E_t(S)} F_t(e). \tag{67}$$

The signed policy substitutes its predicted loss on the left. This existential criterion requires improvement over at least one conflicting pair, with strictly positive gain; it differs from the scan's 0.02 strong-rescue cutoff. The graph forbids every $S$ containing any edge. "False exclusion" below is relative to (67), not a claim of global feasibility or optimality.

The quadratic has held-out MAE 0.0990 and RMSE 0.1804 nats; mean within-task/scale Spearman correlation is 0.700. It anticipates 27 of 30 rescued supersets (Table 26). However, only 3 of the 40 coalitions outperform *every* contained conflicting pair. At the stricter unit of each pair–superset comparison, it detects 33 of 37 positive gains but makes 21 false allows across 132 comparisons. Preserving a promising superset is therefore weaker than correctly predicting every pairwise rescue relation.

## Q.2 Frozen replication on disjoint examples

Before fresh inference, we freeze the original predictions, 28 allows, and one discovery-selected witness pair for each graph-forbidden coalition. We evaluate 512 examples per task with seed 43, excluding discovery indices and RGB hashes and deduplicating the fresh examples by RGB hash. No coefficients, graph edges, witnesses, or decisions are refitted. The numerical inference protocol

Table 27: Complete frozen 40-case policy on fresh examples. Margins compare each coalition with its fixed discovery witness. All 28 original allows remain in the precision denominator at every reporting cutoff; decisions are never retuned.

| Fresh margin $> \gamma$ | Positive cases | Correct allows | Precision | Recall |
|---|---|---|---|---|
| 0 | 29 | 26 | $26/28 = 92.9\%$ | $26/29 = 89.7\%$ |
| 0.01 | 21 | 19 | $19/28 = 67.9\%$ | $19/21 = 90.5\%$ |
| 0.02 | 17 | 15 | $15/28 = 53.6\%$ | $15/17 = 88.2\%$ |
| 0.05 | 10 | 9 | $9/28 = 32.1\%$ | $9/10 = 90.0\%$ |

Table 28: Strict leave-one-coalition-out comparison. Loss errors use 128 distinct task/scale/coalition holdouts. Strong-rescue classification uses all 96 pair–rescuer candidates, including negatives, and the fixed $C, R > 0.02$ rule.

| Model | Loss MAE | Loss RMSE | TP | FP | FN | TN |
|---|---|---|---|---|---|---|
| Signed quadratic | 0.06793 | 0.10062 | 11 | 6 | 2 | 77 |
| Signed cubic | 0.06054 | 0.08397 | 11 | 18 | 2 | 65 |

matches discovery. This is replication within the same test pools, with a new set of examples, not transfer to new tasks or adapters.

Of the 30 original rescued supersets, 29 retain a positive margin and 26 of the original 27 hits replicate. Table 27 also includes discovery negatives: full-policy precision is $26/28$, whereas $26/27$ describes only the selected discovery hits. At margin $0.02$, 17 cases remain positive and 15 are anticipated. Thus much of the directional evidence survives resampling, but the evidence for a substantial improvement is weaker. The largest frozen EuroSAT example improves by $0.4567$ nats on fresh data. Shared examples and coalitions make case outcomes dependent; we do not interpret these fractions as independent-trial significance tests.

### Q.3 What does increasing interaction order buy?

Global in-sample least squares on all 16 coalitions recovers 9 of 13 strong rescues with a signed quadratic and all 13 with a signed cubic. That comparison fits the outcomes it evaluates. We therefore independently refit each model on the other 15 coalitions for every held-out target. For rescue classification only the triple is held out; the observed pair and singleton losses determine $C$ and the pair term of predicted $R$. These folds use higher coalitions and are a separate protocol from the strict pair-only test.

Cubic loss prediction improves, but strong-rescue precision falls from $64.7\%$ to $37.9\%$ at the same $84.6\%$ recall. The cubic has 15 parameters and 15 training observations per fold, with zero residual degrees of freedom. This near-interpolation regime is one possible reason for the poorer threshold decisions. Candidates sharing a triple also share its held-out prediction. Here, lower average reconstruction error does not yield better rescue decisions.

### Q.4 Scaling and controlled checks

The stored scaling experiment uses the boundary case $G_0$ ($\gamma = 0$), not the new positive-margin case. It verifies the noiseless 2-versus-$n$ separation for $n \in \{4, 8, 16, 32, 64, 128\}$: signed utility is always $n$, while graph utility remains 2. With independent Gaussian noise per coalition, 100 trials for each $n$ and $\sigma \in \{0.01, 0.05, 0.1\}$ give signed feasibility rates of 73–80%. At $n = 128, \sigma = 0.1$, nominal signed utility averages $124.06$ against true optimum $127.49$, but only $73\%$ of outputs are feasible. Assigning zero utility to infeasible outputs gives a mean ratio $0.7010$, versus $0.01569$ for the graph. Nominal utility alone is not an approximation guarantee. Appendix H derives the noise amplification and records the exact-search checks.

On the same 512 examples per task, an additional FP32 scan evaluates all coalitions at $\lambda \in \{0.125, 0.25, 0.5, 0.75, 1\}$. Pooled median absolute Möbius magnitudes have fitted log–log slopes

Table 29: Planted third-order losses, 20 seeds. Retained utility is a fraction of the total available utility; high utility with conflicts is not a success. Regret is true loss minus the exhaustive minimum.

| Method | Probes | Retained utility | Conflicts | Loss regret |
|---|---|---|---|---|
| Anchored order 2 | 137 | $1.000 \pm 0.000$ | $16.0 \pm 0.0$ | $267.05 \pm 13.84$ |
| Global quadratic | 697 | $0.357 \pm 0.077$ | $0.0 \pm 0.0$ | $5.70 \pm 1.39$ |
| Random search | 697 | $0.624 \pm 0.031$ | $0.0 \pm 0.0$ | $1.43 \pm 0.53$ |
| Oracle clique expansion | privileged | $0.411 \pm 0.036$ | $0.0 \pm 0.0$ | $4.84 \pm 0.61$ |
| Signed order 3 | 697 | $0.714 \pm 0.026$ | $0.0 \pm 0.0$ | $0.00 \pm 0.00$ |
| Exhaustive oracle | 65,536 | $0.714 \pm 0.026$ | $0.0 \pm 0.0$ | $0.00 \pm 0.00$ |

1.164, 1.399, 2.144, and 2.826 for orders one through four. These finite-scale slopes do not follow a universal $1, 2, 3, 4$ law. They also do not refute a local $O(\lambda^k)$ bound under bounded mixed derivatives; an upper bound does not assert a nonzero leading coefficient or a finite-interval exponent.

The original controlled experiments are retained in Appendix R: exact cubic recovery on 20 planted seeds, a rank-one rescue construction retaining 15 versus 10 adapters, and five learned-digit seeds. In the latter, cubic prediction improves but compressed accuracy does not, and the independent bounded-loss audit accepts no candidate. These controls establish mechanisms and limitations; the released-CLIP study does not run the full certified CH-MERGE deployment pipeline.

## R CONTROLLED MECHANISM AND COMPRESSION RESULTS

All reported results were generated by the accompanying CPU code. Means and standard deviations are across seeds, not confidence intervals. These are controlled mechanism studies, not pretrained-transformer benchmark results. Appendix N supplies all protocols, raw-result filenames and additional tables.

**Pure third-order conflicts.** For each of 20 seeds, we sample 16 distinct triples on 16 adapters with maximum incidence degree at most 6. Utilities are uniform on $[0.7, 1.3]$ and each conflict costs $1 + \sum_i a_i$. Theorem 7 therefore embeds the experiment in per-layer rank-at-most-6 LoRA updates. We enumerate all 65,536 coalitions as an oracle. The order-2 and order-3 methods see only their 137 and 697 anchored probes. A globally fitted quadratic receives 697 uniformly sampled coalitions, including the base; random search uses the same query set. Clique expansion is given the true hyperedges and is therefore a privileged, not data-budget-matched, baseline.

The cubic reconstruction is exact by design, so Table 29 tests the mechanism and implementation rather than generalization. Equal-budget global quadratic fitting is safer than the anchored pairwise fit but retains only about half of the oracle utility. Hypergraph selection preserves combinations that clique expansion discards.

**Rescue, not just triples.** Five disjoint copies of (41), with unit penalties and 15 adapters, are realized by a depth-4 network with rank-one updates. Maximizing retained utility under zero interference keeps all 15 adapters with literal conflicts, versus 10 after incorrectly forbidding each positive pair. Both outputs are conflict-free, but the unsigned policy needlessly loses one third of the utility. On a face fixing one harmful pair present and its rescuer absent, the independent-term bound is $-3$, while the local-moment bound is 1 (the exact minimum). The verified-dual implementation returns a downward-rounded value just below 1. This tests the complete signed-constraint/MILP component, not only polynomial fitting.

**Learned LoRA under domain shift.** Five seeds use six shifted/rotated domains of the bundled digit dataset (scikit-learn developers, 2026), with independent rank-4 adapters and identity-disjoint training, discovery, audit, and test splits. Appendix N gives the full architecture and split protocol. Base accuracy on original test images averages $96.52\%$, and specialists average $93.40\%$ on their own domains, confirming learned specialization.

Loss is $\min(\mathrm{CE}, 5)/5$. All methods share the scale grid $\{0.25, 0.5, 0.75, 1\}$. A surrogate proposes one minimizer per scale; the four proposals are validated on discovery data, costing up to four extra

Table 30: Learned-LoRA audit, five seeds. Test accuracy is averaged over six shifted/rotated domains. Dense sums have variable rank; rank-4 columns use the same final rank in every layer. These are proposal outputs, *not* certified deployments.

| | Dense fused update | | Rank-4 compressed update | |
|---|---|---|---|---|
| Method | Clipped loss | Accuracy (%) | Clipped loss | Accuracy (%) |
| All-adapter sum | $0.4628$ | $41.58 \pm 2.76$ | $0.4626$ | $42.09 \pm 2.86$ |
| Best single | $0.4282$ | $46.37 \pm 2.51$ | $0.4282$ | $46.37 \pm 2.51$ |
| Anchored order 2 | $0.4196$ | $46.31 \pm 2.17$ | $0.4214$ | $46.16 \pm 1.61$ |
| Global quadratic | $0.4187$ | $46.35 \pm 1.85$ | $0.4224$ | $45.99 \pm 1.99$ |
| Signed order 3 | $0.4185$ | $46.25 \pm 1.91$ | $0.4246$ | $45.57 \pm 2.44$ |
| Exhaustive discovery | $0.4185$ | $46.25 \pm 1.91$ | $0.4246$ | $45.57 \pm 2.44$ |

coalition evaluations. Thus the order-2 and order-3 budgets are at most 92 and 172 calls, respectively, versus 256 for exhaustive discovery. The global quadratic receives 42 random masks per scale and the same four-proposal allowance. No test metric chooses a model.

At scale $0.25$, third-order prediction reduces mean absolute loss error on unseen coalitions of size at least four from $0.00775 \pm 0.00210$ to $0.00426 \pm 0.00049$, a $45.1\%$ reduction. It selects the same scale and coalition as exhaustive discovery on all five seeds. Nevertheless, improved surrogate reconstruction does not produce a convincing accuracy advantage; rank compression can reverse the small pre-compression loss benefit.

For the predeclared per-task audit cap $0.30$, $K = 8$ frozen rank-4 candidates and $\delta_{\mathrm{aud}} = 0.05$, the radius is $0.1381$. The audit abstains on every seed because no candidate passes. The clipped-loss experiment does not establish the smooth derivative bound; its coefficients and errors remain empirical diagnostics. The independent audit has the stated conditional risk interpretation.