SAFECOMPOSE: PREDICTING UNSAFE LORA COMPOSITIONS WITHOUT EXHAUSTIVE TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapter hubs let practitioners compose several LoRA modules on one open-weight base, yet each adapter is safety-screened only in isolation. We formalise composition screening as learning a set function over an adapter pool, separating a *threshold crossing* (safe parts, unsafe whole), its *mechanism* (distinct-adapter interaction versus summed merge dose), and genuine *third-and-higher-order Mobius mass*, which alone defeats low-degree predictors (Theorem 1, a conditional result). We pair this with a screener that scores candidate compositions from adapter fingerprints and a calibrated accept/test/reject decision with a distribution-free false-negative guarantee (Theorem 3). Empirically, the picture is largely negative for interaction: across 500 compositions of individually screened public adapters on a 7B base, none crossed the deployment threshold (95% upper bound 0.76% under the sampled law), and sub-threshold hazard is node-predictable with no resolved gain from interaction features; a deliberately constructed collusion is explained by summed merge *dose*, not a distinct-adapter mechanism. On a constructed benchmark whose unsafe compositions have every constituent individually safe at its coefficient, a coefficient-aware screener with a nonlinear dose term reaches AUROC 0.99 (node-only 0.87) and transfers to a held-out risky pair (0.93–0.95), while interaction architectures add nothing; an exploratory calibrated accept region holds out of sample at $\alpha{=}0.10$. The lesson is that merge configuration, not only the adapter file, is part of the safety object. Code and measured records: `https://anonymous.4open.science/r/SafeCompose-F316`.

1 INTRODUCTION

Low-rank adaptation (Hu et al., 2022) made specialisation cheap enough that adapters became artifacts in their own right: published, versioned, downloaded, and combined. Public hubs distribute them by the thousand (Pfeiffer et al., 2020), and operators compose several into one model by weight averaging (Wortsman et al., 2022; Matena & Raffel, 2022), task arithmetic (Ilharco et al., 2023), interference-aware merging (Yadav et al., 2023; Yu et al., 2024), or learned routing (Huang et al., 2024a; Wu et al., 2024). The safety practice alongside is per-artifact: an adapter is tested, or its weights screened, on its own (Hsu et al., 2024; Puertolas Merenciano et al., 2026).

That practice assumes safety is compositional. It is not. Merging carries misalignment from one contributed model into the result (Hammoud et al., 2024) through interference between the parameters contributors rely on (Ma et al., 2025), and can also destroy a backdoor reliably present before the merge (Arora et al., 2024). Most directly, Ding (2026) construct LoRA pairs that score no attack successes in isolation but whose linear composition suppresses refusal almost entirely across three open-weight families — a combinatorial blind spot for unit-centric defences, with exhaustive composition testing intractable. That paper establishes the hazard; it does not say which compositions to worry about.

Running them all is what the arithmetic forbids: a pool of n adapters admits $\sum_{k=2}^{K}\binom{n}{k}$ compositions of size at most K, above 10^5 already at $n{=}40, K{=}4$, each needing a merge and a full safety sweep, and uniform sampling does not help when the unsafe ones are rare. The cheap surrogate, screening adapters one at a time, measures the wrong object when the quantity of interest is by construction the part of the hazard that no single adapter carries.

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.

This paper. We treat composition screening as learning and calibrating a set function $r(S)$, the hazard of the model composing the adapters in S, with Mobius representation $r(S) = \sum_{T \subseteq S} a(T)$ (Rota, 1964; Grabisch & Roubens, 1999). Its starting point is to *not* equate the practitioner's worry — a threshold crossed though each adapter passed alone — with the third-and-higher-order coefficients that alone defeat a degree-2 predictor (Theorem 1): a crossing can come from summed merge dose with no distinct-adapter mechanism, and an additive dose is not an additive hazard (Remark 1). Whether higher-order features help is then *empirical*, and because a screener errs, the deployable object is a calibrated decision, not a score.

SAFECOMPOSE is the system this implies (Section 5): it fingerprints each adapter in weight- and activation-space (projections onto the alignment structure Hsu et al. (2024) and Wei et al. (2024) show carries safety, and displacement along the refusal direction of Arditi et al. (2024)), scores candidate hyperedges of an interaction hypergraph with a degree-d model, spends a testing budget on the candidates it is least sure about (Sawade et al., 2010; Kossen et al., 2021), and emits a calibrated accept/test/reject decision (Angelopoulos et al., 2025). Active acquisition breaks the exchangeability calibration needs; we prove it and give two repairs.

Contributions. **(C1) Formulation**: composition safety screening as learning a set function over an adapter pool (Section 3), separating threshold crossing, non-additivity, and higher-order Mobius mass so that "what defeats per-adapter screening" becomes a measurable rather than an assumed quantity. **(C2) Theory (conditional)**: an exact decomposition (Lemma 1) and a quantitative separation — *if* the hazard carries third-and-higher-order Fourier mass, any degree-≤ 2 predictor suffers squared error at least that mass, with a mis-estimation-frequency corollary (Theorem 1) — plus a truncation bound and a degree-d representation statement (Proposition 1, Theorem 2). **(C3) Calibrated screening**: a three-way policy with a finite-sample high-probability false-negative guarantee (Theorem 3), a proof that active acquisition breaks the exchangeability it needs with two repairs (Proposition 2), and a ranking-to-recall bound (Proposition 3), exercised end to end on a constructed benchmark (E6) with risk-ranked testing; the active uncertainty/diversity acquisition rule and repair remain unvalidated. **(C4) Benchmark and measurement**: a screening benchmark and a rigorous *negative* finding — ordinary composition is rarely emergently unsafe and interaction features give no resolved gain over node ones (E1–E4; a non-resolution, not equivalence); a constructed collusion needs no distinct-adapter mechanism (E2); and the discriminating feature is nonlinear merge *dose*, not higher-order interaction (E6).

2 RELATED WORK

Appendix L gives the full version, with the exact reason each nearby line of work does not already solve the problem.

Composing adapters, and safety under adaptation. Weight-space composition spans averaging (Wortsman et al., 2022; Matena & Raffel, 2022), task arithmetic (Ilharco et al., 2023), interference-aware merging (Yadav et al., 2023; Yu et al., 2024) and learned routing (Huang et al., 2024a; Wu et al., 2024); Ortiz-Jimenez et al. (2023) show it behaves additively only under weight disentanglement — the condition our higher-order coefficients measure the violation of — but evaluate task accuracy, not safety. Fine-tuning removes alignment cheaply, including through LoRA (Qi et al., 2024; Lermen et al., 2023), and alignment is shallow (Qi et al., 2025), localised in a small band of weights and layers (Wei et al., 2024; Li et al., 2025), collapses in a step under perturbation (Peng et al., 2024), and routes through a single refusal direction (Arditi et al., 2024; Zou et al., 2023a). All of it concerns one adapted model.

Safety of merged models. Merging propagates a contributor's misalignment (Hammoud et al., 2024), induces cross-task collapse (Ma et al., 2025), and can be repaired with a safety vector (Bhardwaj et al., 2024); a poisoned contributor can survive merging (Zhang et al., 2024; Yin et al., 2024; Liu et al., 2025; Pawlak et al., 2025; Yang et al., 2025) or be erased by it (Arora et al., 2024). In each, one ingredient is deliberately malicious. Ding (2026) is the exception and closest work — adapters benign in isolation whose composition suppresses refusal — but it is an attack demonstration with no predictor, no calibration, and no way to choose which compositions to execute; SAFECOMPOSE addresses the defensive problem it names.

Screening artifacts; sets and interactions; budgeted decisions. Trojan meta-classifiers (Xu et al., 2021), weight-space spectral signatures for poisoned LoRA (Puertolas Merenciano et al., 2026), and alignment-subspace projection (Hsu et al., 2024) each screen *one* artifact; composite backdoors firing only on co-occurring triggers (Lin et al., 2020; Huang et al., 2024b) are the right analogy but live in a single model — the analogue of Theorem 1. Permutation-invariant and hypergraph encoders (Zaheer et al., 2017; Lee et al., 2019; Feng et al., 2019; Chien et al., 2022; Wagstaff et al., 2019) supply the architecture; the Möbius transform and interaction indices (Rota, 1964; Grabisch & Roubens, 1999; Tsai et al., 2023; Amrollahi et al., 2019; Hamilton et al., 2017; Xu et al., 2019; Morris et al., 2019) the decomposition; active testing (Sawade et al., 2010; Kossen et al., 2021; Kirsch et al., 2019; Baptista & Poloczek, 2018) and automated red-teaming (Perez et al., 2022; Zou et al., 2023b) the search; and distribution-free risk control with selective prediction (Bates et al., 2021; Angelopoulos et al., 2025; 2024; Geifman & El-Yaniv, 2017) the decision layer.

3 Problem Setting

3.1 System model

A base model has weights $W_0 \in \mathbb{R}^p$. An adapter pool $\mathcal{P} = \{a_1, \ldots, a_n\}$ contains LoRA-style modules (Hu et al., 2022; Houlsby et al., 2019): each a_i carries factors $U_i^{(\ell)} \in \mathbb{R}^{d_\ell \times r_i}$ and $V_i^{(\ell)} \in \mathbb{R}^{r_i \times d'_\ell}$ on a set of layers $\ell \in \mathcal{L}_i$, inducing the update $\Delta_i^{(\ell)} = (\alpha_i / r_i)\, U_i^{(\ell)} V_i^{(\ell)}$ and $\Delta_i = 0$ elsewhere. A *composition* is a triple $(S, c, \oplus)$ with $S \subseteq [n]$, coefficients $c \in \mathbb{R}^{|S|}$, and an operator $\oplus \in \mathcal{O}$ producing weights

$$W(S, c, \oplus) \;=\; \oplus\big(W_0, \{(\Delta_i, c_i)\}_{i \in S}\big). \tag{1}$$

$\mathcal{O}$ holds six operators: weighted addition, linear merge, task arithmetic (Ilharco et al., 2023), sequential loading, routing (Wu et al., 2024), and layer-selective merge. Weighted addition, $W = W_0 + \sum_{i \in S} c_i \Delta_i$, is the running example; nothing below assumes it.

An evaluation suite $\mathcal{T}$ contains prompts and scored checks across six dimensions: refusal and policy compliance, benign utility, privacy extraction, backdoor probes, tool-use safety, and robustness to prompt variation (Appendix G). It defines a bounded hazard $\mathrm{haz}(W) \in [0, 1]$ and a utility $u(W) \in [0, 1]$. Writing $r(S) \coloneqq \mathrm{haz}(W(S, c, \oplus))$ for fixed $c, \oplus$ gives a set function $r : 2^{[n]} \to [0, 1]$ with $r(\emptyset) = \mathrm{haz}(W_0)$; the dependence on c and $\oplus$ is suppressed only where held fixed, and Section 6 varies both. The label is $Y(S) = \mathbb{1}[r(S) > \tau]$ for a deployment threshold τ.

Definition 1 (Individually screened pool; emergent unsafety)**.** $\mathcal{P}$ is *individually screened at level* $\tau_0 < \tau$ if $r(\{i\}) \le \tau_0$ for every $i \in [n]$: every adapter passes the same suite alone, with margin. A composition S, $|S| \ge 2$, is *emergently unsafe* if $Y(S) = 1$ while $Y(\{i\}) = 0$ for all $i \in S$; in an individually screened pool every unsafe composition is emergently unsafe.

This is the regime the paper is about, and the regime in which per-adapter screening has by construction no signal at a *fixed* coefficient. It is enforced in the benchmark rather than assumed of the world (Appendix F).

Remark 1 (Threshold emergence, mechanism, and interaction order)**.** Keep apart: (a) *threshold emergence* (Definition 1), $Y(S) = 1$ with all singletons safe; (b) *non-additivity*, $a(T) \neq 0$ for some $|T| \ge 2$; and (c) *third-and-higher-order mass*, $a(T) \neq 0$ for some $|T| \ge 3$. Only (c) obstructs a degree-2 predictor (Theorem 1) — a degree-2 model already fits any pairwise term, so (b) alone does not — and (a) implies neither ($r(S) = 0.02 + 0.08|S|$ makes every size-4 set emergently unsafe with all $a(T)$, $|T| \ge 2$, zero). A last caution used in Section 6.1: an additive merge *dose* is not an additive *hazard*. A nonlinear $r(S) = F(\sum_{i \in S} c_i v_i)$ is a function of summed dose yet can carry large coefficients at every order — $r(S) = \mathbb{1}[|S| \ge k]$ is a pure dose rule with order-k Mobius coefficient 1. So a control showing "a single adapter at matched dose reproduces the crossing" identifies the *mechanism* as dose, not a distinct-adapter pairing; it does not show (b) absent or (c) negligible, which needs an estimator of the coefficients.

3.2 Threat and failure model

Four settings, in increasing adversarial strength. (i) *Accidental interaction*: adapters trained independently, unsafety arising from interference alone, the regime Ma et al. (2025) diagnose for

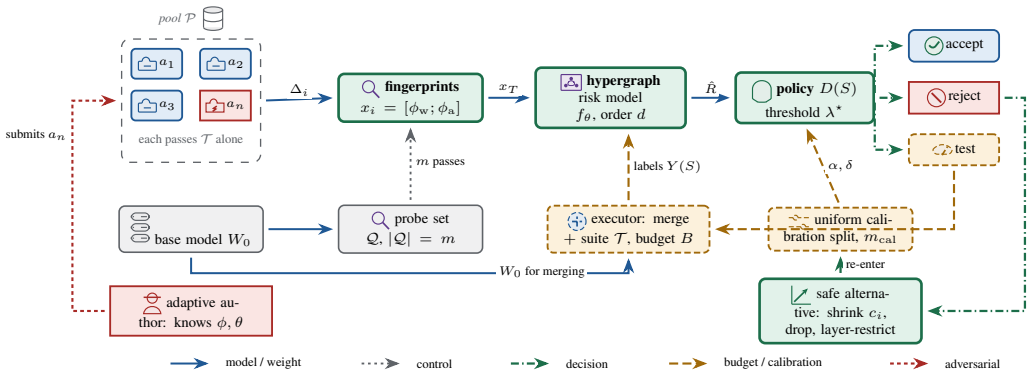


Figure 1: SAFECOMPOSE screening architecture, with one lane per flow as named in the legend. Red is threat setting (iv), where the author submits an adapter optimised against ϕ and f_θ. The screener never executes the compositions on the accept or reject branches.

full models. (ii) *One hidden malicious adapter* among benign ones, the merging-backdoor setting (Zhang et al., 2024; Yin et al., 2024; Liu et al., 2025), made plausible by trojaned plugins (Dong et al., 2025), poisoning (Wan et al., 2023) and deferred activation (Hubinger et al., 2024). (iii) *All adapters individually pass*, that is (ii) under Definition 1, the setting of Ding (2026). (iv) *Adaptive author*, who knows ϕ, θ and optimises an adapter to keep utility while lowering its predicted interaction risk. The system is evaluated on (iii); the adaptive setting (iv) is analysed (Appendix I) but not yet run (E7). Settings (i) and (ii) give context.

The screener holds W_0, the pool, the operator family, a small probe set $\mathcal{Q}$ with $|\mathcal{Q}| = m$, and a labelling oracle returning $Y(S)$ by composing and running $\mathcal{T}$ at a cost of one *execution*, with an acquisition budget of $B \ll |\mathcal{C}|$ executions; the calibration draws m_{cal} are counted separately, so total executions are $B + m_{\text{cal}}$ (Section 5.4). It does not hold the adapters' training data.

3.3 THE SCREENING PROBLEM

The candidate set is $\mathcal{C} = \{S \subseteq [n] : 2 \le |S| \le K\}$, of size

$$|\mathcal{C}| = \textstyle\sum_{k=2}^{K} \binom{n}{k}, \tag{2}$$

which is 102,050 at $n = 40, K = 4$. The screener returns a decision $D(S) \in \{\textsf{accept}, \textsf{test}, \textsf{reject}\}$ for every $S \in \mathcal{C}$ within its execution budget, and is judged primarily on the false-negative rate among accepted compositions at a stated safe-coverage level or testing budget:

$$\text{FNR}_{\text{acc}} = \mathbb{P}\big[Y(S) = 1 \mid D(S) = \textsf{accept}\big], \quad \text{coverage} = \mathbb{P}\big[D(S) = \textsf{accept}\big]. \tag{3}$$

This is an *accepted-set risk* (the unsafe fraction among accepts), not the recall $\mathbb{P}[D = \textsf{accept} \mid Y = 1]$ of unsafe candidates; the two differ, and a low FNR_{acc} can coexist with poor recall when unsafe compositions are rare, so we report coverage alongside it. Accepting everything gives full coverage at the base rate of unsafe compositions; testing everything gives $\text{FNR}_{\text{acc}} = 0$ at cost $|\mathcal{C}|$. Everything interesting lies between.

4 COMPOSITION HAZARD AS A SET FUNCTION

This section fixes what a screening model must represent and what it can be guaranteed. Proofs are in Appendix B and notation in Appendix A; throughout, $z_i(S) = \mathbb{1}[i \in S]$ and r is the hazard set function of Section 3 at a fixed operator and coefficient vector.

Assumption 1 (Bounded, deterministic hazard)**.** $r : 2^{[n]} \to [0, 1]$, and re-evaluating $\mathcal{T}$ on the same composed weights returns the same value. The label-noise ablation relaxes determinism.

Assumption 2 (Candidate measure)**.** Candidates are drawn from the p-biased product measure μ_p on $2^{[n]}$, each adapter included independently with probability p. With $p = \kappa/n$ the mass sits on sets of size $O(\kappa)$, so μ_p idealises a user drawing a handful of adapters. It is not the law on $\mathcal{C}$: conditioning on $2 \le |S| \le K$ breaks independence, so Theorem 1 is claimed under μ_p only.

Assumption 3 (Clean calibration split)**.** The calibration sample $\{(S_t, Y_t)\}_{t=1}^{m}$ is i.i.d. from the deployment candidate distribution and independent of everything used to fit the score $\hat{R}$, which is therefore fixed when calibration begins.

Lemma 1 (Mobius representation; Rota, 1964; Grabisch & Roubens, 1999; Tsai et al., 2023)**.** *For every $r : 2^{[n]} \to \mathbb{R}$ there is a unique a with $r(S) = \sum_{T \subseteq S} a(T)$, given by $a(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} r(U)$; in particular $a(\{i,j\}) = r(\{i,j\}) - r(\{i\}) - r(\{j\}) + r(\emptyset)$.*

Call r *order-d additive* if $a(T) = 0$ for $|T| > d$, and write r_d for its degree-d truncation. Lemma 1 names the coefficients: $a(T)$ with $|T| \geq 2$ is the part of the hazard that appears only when every adapter in T is present, i.e. the non-additive part. An individually screened pool bounds the first-order terms, $a(\{i\}) \leq \tau_0$, and says nothing about the rest. Two cautions carried from Remark 1: this higher-order mass is what the next result concerns, and it is *not* the same as the threshold event of Definition 1, which an order-1 (additive) hazard can also trigger; and per-adapter screening at a fixed coefficient is blind to *both* the higher-order mass and the additive dose contributed at merge time, which is a distinct blind spot our measurements find dominant.

4.1 What pairwise models cannot represent

Let $\sigma = \sqrt{p(1-p)}$ and $\chi_T(S) = \prod_{i \in T} (z_i(S) - p)/\sigma$, an orthonormal basis of $L^2(\mu_p)$, so $r = \sum_T \hat{r}(T)\chi_T$ and Parseval holds. The two bases grade alike: $\max\{|T| : a(T) \neq 0\} = \max\{|T| : \hat{r}(T) \neq 0\}$, both being the degree of the multilinear extension (Lemma 2). Let $\mathcal{H}_d$ be the set functions of degree at most d, which by Lemma 1 is exactly what a model with free scalars on hyperedges up to order d expresses. What is restricted is the class, not the feature set; Appendix B makes the distinction precise. Let

$$\Delta^2_{\geq 3}(r) \;=\; \sum_{|T| \geq 3} \hat{r}(T)^2 \tag{4}$$

be the higher-order mass.

Theorem 1 (Pairwise insufficiency)**.** *Under Assumption 2, every $h \in \mathcal{H}_2$ has $\|r - h\|^2_{L^2(\mu_p)} \geq \Delta^2_{\geq 3}(r)$, with equality exactly at the orthogonal projection $r^{\leq 2} = \sum_{|T| \leq 2} \hat{r}(T)\chi_T$, which in general is* not *the Mobius truncation r_2. If in addition h takes values in $[0,1]$, then for every $\tau \in (0,1)$*

$$\mu_p\big(|r(S) - h(S)| > \tau\big) \;\geq\; \Delta^2_{\geq 3}(r) - \tau^2. \tag{5}$$

Equation (5) turns a representation gap into a frequency of *mis-estimation*: a pairwise model's hazard estimate is off by more than τ on at least a $\Delta^2_{\geq 3} - \tau^2$ fraction of candidates. This bounds estimation error, not the accept/reject decision — a model can misestimate a hazard yet still classify it correctly (Appendix B), so the result is a necessary-signal statement, not a proof that a low-degree screener must produce false negatives. It is the formal counterpart of the observation that scanners tuned to atomic units can miss compositional failures (Lin et al., 2020), and it bites only when $\Delta^2_{\geq 3}$ is non-negligible.

Remark 2 (The signal is real but faint)**.** For the order-k conjunction $r(S) = \prod_{i \in T_0} z_i(S)$, $|T_0| = k$, one gets $\hat{r}(T) = p^{k-|T|}\sigma^{|T|}$ on $T \subseteq T_0$ and hence $\Delta^2_{\geq 3} = \sum_{j=3}^{k} \binom{k}{j} p^{2(k-j)}(p(1-p))^j$, equal to $(p(1-p))^3$ at $k = 3$ and $O(p^k)$ under Assumption 2 with $p = \kappa/n$. The separation is genuine but the signal decays geometrically in the order, so estimation needs many labels or a structural prior (Amrollahi et al., 2019); hence the pair-before-triple curriculum in acquisition.

Proposition 1 (Truncation error under coefficient decay)**.** *If $|a(T)| \leq C\gamma^{|T|}$ for all $|T| \geq d+1$ then, for every S with $|S| = k$,*

$$|r(S) - r_d(S)| \;\leq\; C \sum_{j=d+1}^{k} \binom{k}{j} \gamma^j \;\leq\; C\,\frac{(k\gamma)^{d+1}}{(d+1)!}\, e^{k\gamma}. \tag{6}$$

The decay rate γ is an assumption about the world, not a finding; E3 tests it and Figure 6(b) plots the bound at three assumed rates.

Theorem 2 (Capacity and width). *(i) A model $h(S) = b_\emptyset + \sum_{T\subseteq S, 1\le|T|\le d} b(T)$ with free scalars realises exactly $\mathcal{H}_d$, so it represents r exactly if and only if r is order-d additive. (ii) If the scalars are tied through a feature map, $b(T) = g_\theta(x_T)$, exact representation additionally needs that map to separate hyperedges: if $x_T = x_{T'}$ while $a(T) \neq a(T')$, no g_θ suffices. (iii) For the hyperedge-free alternative $h(S) = \rho(\sum_{i\in S}\phi(x_i))$ with continuous ρ, ϕ and latent width L, representing all continuous permutation-invariant functions of sets of size at most K requires $L \ge K$ (Wagstaff et al., 2019, Thms. 4.3–4.4).*

Part (ii) is why the benchmark contains matched compositions with identical node fingerprints and different outcomes (Appendix F); Appendix B restates the hypotheses of the imported result in part (iii).

4.2 Calibrated acceptance

Fix targets α, δ, a finite grid Λ of accept thresholds, and $\mathcal{A}(\lambda) = \{S : \hat{R}(S) \le \lambda\}$. Define the bounded loss

$$L_\lambda(S, Y) \;=\; \alpha + (Y - \alpha)\mathbb{1}[S \in \mathcal{A}(\lambda)] \;\in\; [0, 1], \tag{7}$$

so that, whenever $\mathbb{P}[S \in \mathcal{A}(\lambda)] > 0$, $\mathbb{E}[L_\lambda] \le \alpha$ holds if and only if the selective risk $\mathrm{FNR}_{\mathrm{acc}}(\lambda)$ of Eq. (3) is at most α.

Theorem 3 (Finite-sample control of the accepted false-negative rate). *Let Assumption 3 hold, let p_λ be the Hoeffding–Bentkus p-value (Bentkus, 2004; Bates et al., 2021) for $H_\lambda : \mathbb{E}[L_\lambda] > \alpha$ on the calibration split, and let $\hat{\Lambda} \subseteq \Lambda$ be returned by any procedure controlling the family-wise error rate at level δ. Then*

$$\mathbb{P}\big[\forall\lambda \in \hat{\Lambda} :\ \mathrm{FNR}_{\mathrm{acc}}(\lambda) \le \alpha \ \text{ or } \ \mathbb{P}[S \in \mathcal{A}(\lambda)] = 0\big] \;\ge\; 1 - \delta. \tag{8}$$

Remark 3 (Why this machinery). $\mathrm{FNR}_{\mathrm{acc}}(\lambda)$ is a ratio of expectations and need not be monotone in λ, so the monotone-risk hypotheses of RCPS (Bates et al., 2021) are unavailable and conformal risk control (Angelopoulos et al., 2024) would give only expectation-level control over the joint draw, whereas a screener is calibrated once and then run. Appendix B gives the comparison in full. Only the accept branch is calibrated: test and reject cannot create a false negative, so they trade coverage rather than safety. Appendix N states the scope of the certificate in full.

Proposition 2 (Active acquisition breaks Assumption 3, and two repairs). *(i) If calibration labels are the ones chosen by an acquisition rule depending on $\hat{R}$, the sample is in general neither i.i.d. nor exchangeable with a fresh candidate and Eq. (8) fails; Appendix B exhibits a rule and a label configuration under which $\hat{\Lambda}$ contains a λ with $\mathrm{FNR}_{\mathrm{acc}}(\lambda) > \alpha$ with probability tending to one. (ii) A calibration sample drawn i.i.d. (uniformly with replacement) from $\mathcal{C}$ independently of acquisition restores Theorem 3 verbatim, conditional on $\hat{R}$, at a cost of m executions. (iii) If acquired labels must be reused, let acquisition be randomised with known propensities $\pi(S) \ge \pi_{\min} > 0$ under Poisson sampling over $\mathcal{C}$, $N = |\mathcal{C}|$, with $\hat{R}$ fixed before that round so that $\mathcal{A}(\lambda)$ does not depend on which candidates are sampled. Then $\hat{L}_\lambda = N^{-1}\sum_{t\,\mathrm{sampled}} L_\lambda(S_t, Y_t)/\pi(S_t)$ is unbiased for the pool mean and deviates from it by at most $\pi_{\min}^{-1}\{\log(2/\delta')/(2N)\}^{1/2}$ with probability $1 - \delta'$, which supplies a valid p-value and hence Theorem 3 with that variance inflation.*

Proposition 3 (Recall at a fixed testing budget). *Let $\hat{R}$ have empirical ROC area A on $\mathcal{C}$, let unsafe prevalence be π, and execute the top βN candidates by score with $\beta > \pi$. Then with $\beta' = (\beta - \pi)/(1 - \pi)$,*

$$\mathrm{Recall}(\beta) \;\ge\; (\mathrm{A} - 1 + \beta')/\beta'. \tag{9}$$

At $\mathrm{A} = 0.95$, $\pi = 0.02$, $\beta = 0.20$ this gives $\mathrm{Recall} \ge 0.72$. It is worst-case and deterministic, and its role is to say what a target recall demands of A before any model is trained.

5 SafeCompose

5.1 Stage 1: adapter fingerprints

Each adapter is summarised by $x_i = \phi(a_i) = [\phi_{\mathrm{w}}(a_i); \phi_{\mathrm{a}}(a_i)]$. Appendix D has the full schema and Figure 3 the five stages with their costs; the design rationale is that both halves are known carriers of safety-relevant structure.

Algorithm 1 SAFECOMPOSE screening. Steps 2–5 are the four steps of the specification, in order.

Require: W_0; pool $\mathcal{P}$; operator $\oplus$; candidates $\mathcal{C}$; probes $\mathcal{Q}$; budget B; calibration size m_{cal}; targets α, δ; grid Λ

1: $x_i \leftarrow \phi(a_i)$ for all $i \in [n]$ ▷ n reads, nm probe passes
2: $\mathcal{D}_{\text{cal}} \leftarrow$ execute m_{cal} candidates drawn uniformly from $\mathcal{C}$ ▷ reserved split, Prop. 2(ii)
3: $\mathcal{D} \leftarrow$ seed labels; $b \leftarrow |\mathcal{D}|$ ▷ seeds count toward B
4: **while** $b < B$ **and** stopping rule not met **do**
5: fit θ on $\mathcal{D}$; score all $S \in \mathcal{C}$ ▷ step 1: score
6: $\mathcal{B} \leftarrow$ batch acquired by uncertainty with hypergraph diversity ▷ step 2: select
7: execute $\mathcal{B}$, append to $\mathcal{D}$, $b \leftarrow b + |\mathcal{B}|$ ▷ step 3: update
8: **end while**
9: $\hat{\Lambda} \leftarrow$ CALIBRATE($\hat{R}, \mathcal{D}_{\text{cal}}, \alpha, \delta, \Lambda$); $\lambda^\star \leftarrow \max \hat{\Lambda}$ if $\hat{\Lambda} \neq \emptyset$ else abstain (empty accept region) ▷ step 4
10: **for** $S \in \mathcal{C}$: $D(S) \leftarrow$ accept if $\hat{R}(S) \leq \lambda^\star$; reject if $\hat{R}(S) \geq \lambda^+$; else test; report SAFEALTERNATIVE(S) on rejections
11: **return** D, $\hat{\Lambda}$, and the certificate $(\alpha, \delta, m_{\text{cal}})$

The *weight-space* block records per-layer norms of Δ_i, leading singular values and the rank concentration $\sum_{j \leq 3} s_j^2 / \sum_j s_j^2$, the layer-localisation profile over $\mathcal{L}_i$ where Li et al. (2025) find the discriminative safety layers, and the projection of Δ_i onto the alignment subspace Safe LoRA operates on (Hsu et al., 2024): spectral structure rather than magnitude, because spectral statistics alone already separate benign from poisoned *single* adapters (Puertolas Merenciano et al., 2026) and safety is localised at the rank level (Wei et al., 2024; Shuttleworth et al., 2025). The *activation-space* block, measured on a probe set $\mathcal{Q}$ of m prompts spanning harmful, borderline, benign and tool-invoking requests, records refusal-token logit margins, per-layer hidden-state displacement and its component along the refusal direction of Arditi et al. (2024), linear-probe readouts (Zou et al., 2023a), tool-invocation rates, and predictive entropy, at m cached forward passes per adapter.

5.2 STAGE 2: INTERACTION HYPERGRAPH AND THE RISK MODEL

Nodes are adapters, hyperedges are candidate compositions. For $2 \leq |T| \leq d$ we build x_T from the constituent fingerprints without composing the model: principal angles between the Δ's, layer-support overlap, sign agreement along the alignment and refusal directions, and interference proxies of the kind Yadav et al. (2023) use to explain merge degradation, with third-order features adding the triple overlap and the residual of the pairwise terms. By Theorem 2(i) the model is a degree-d hyperedge sum with tied parameters,

$$\hat{R}(S) \;=\; \varsigma\Big(b_\emptyset + \sum_{i \in S} g_\theta^{(1)}(x_i) + \sum_{\{i,j\} \subseteq S} g_\theta^{(2)}(x_{ij}) + \sum_{\{i,j,k\} \subseteq S} g_\theta^{(3)}(x_{ijk}) \Big), \tag{10}$$

with ς the logistic link and $g^{(1)}, g^{(2)}, g^{(3)}$ small shared MLPs, so the parameter count is independent of n and the model applies unchanged to adapters never seen in training, the inductive setting of Hamilton et al. (2017). Ablations switch terms off; Appendix D gives the hypergraph-convolution variant (Feng et al., 2019; Chien et al., 2022). Because parameters are tied across hyperedges of one order, a composition of size $k > d$ is scored by summing its sub-hyperedges of order at most d, which is what makes E3 well posed at all.

5.3 STAGES 3 AND 4: BUDGETED ACQUISITION AND THE CALIBRATED DECISION

Acquisition follows active testing rather than active learning: the goal is an accurate estimate of where the decision boundary sits, not a uniformly good model (Sawade et al., 2010; Kossen et al., 2021). Batches are scored by predictive uncertainty with a diversity penalty over shared adapters (Kirsch et al., 2019; Baptista & Poloczek, 2018), and a pair-before-triple curriculum makes the higher-order terms of Eq. (10) fit residuals rather than raw labels, the concrete answer to the sparsity Remark 2 quantifies. Whether it beats random acquisition at an equal execution count is what the ablations decide. Calibration is Theorem 3 on the reserved split of line 2, with fixed-sequence testing over Λ; line 2 exists because of Proposition 2(i) and is the condition under which the certificate

means anything. The reuse variant, Proposition 2(iii), is an ablation, since it pays $1/\pi_{\min}$ in width for the executions it saves.

5.4 Stage 5: safe alternatives

For a rejected composition, SafeAlternative (Algorithm 3, Appendix D) searches its neighbourhood — shrink the most-implicated coefficient, drop that adapter, or merge outside the safety band of Li et al. (2025) — and re-screens each survivor by the same score and threshold. The certificate of Theorem 3 does *not* transfer to this selected repair sub-population, so its accepted-set risk must be validated on the repair distribution itself; this stage is specified but unmeasured (E8, Appendix D).

Cost. Fingerprinting costs $O(n)$ adapter reads and nm forward passes, paid once; features are $O(|\mathcal{C}|d^2)$ and scoring $O(|\mathcal{C}|)$, neither executing the model. The executions are the B acquisitions (seeds included, Algorithm 1 line 3) and the m_{cal} calibration draws, so the up-front screened fraction is $(B+m_{\text{cal}})/|\mathcal{C}|$ against the oracle's $|\mathcal{C}|$: under two percent at n=40, K=4, $B+m_{\text{cal}}$=2000, arithmetic from Eq. (2) rather than a measured speed-up. Candidates routed to test and executed repairs add to this ledger only when run — a test is not a completed evaluation (Appendix N). Appendix M gives the full model, including the regime below roughly 2000 candidates where exhaustive testing is simply the right method.

6 Experiments

Pool, suite, baselines. The evaluation follows a pre-registered protocol. The pool mixes domain, capability, style, instruction and safety-oriented adapters on distinct corpora plus a controlled risky subset, all filtered to Definition 1; the suite covers the six dimensions of Section 3, scored by a trained classifier (Mazeika et al., 2024), the StrongREJECT rubric (Souly et al., 2024), XSTest (Röttger et al., 2024) and held-out backdoor triggers, with a second classifier (Inan et al., 2023) labelling independently. Appendix G records the eight baselines, the base, judge and adapter identities of the realised run, and which scorer is used where. **E1**–**E4** and **E6** (Section 6.1) are measured; E3 (train $\{2,3\}$/test 4) tests Theorem 1 and Proposition 1, while **E5** (cross-base transfer), **E7** (adaptive evasion) and **E8** (SafeAlternative utility) remain protocol.

6.1 Results

E1 (emergent unsafe rate). On an open-weight 7B base with 37 public adapters each passing Definition 1 at τ_0=0.10 (τ=0.30; Appendix G.8), across 500 compositions (stratified $150/200/150$ over sizes 2/3/4) none crossed τ: an emergent unsafe rate of 0 under that sampling law, 95% Wilson upper bound 0.76% (both the 13B reporting and 7B in-loop classifiers give the same zero). Composition raises the hazard above the single-adapter floor of 0.05 (median 0.09, max 0.235) but stays below τ.

Robustness and baselines (E3, E4). The zero result is stable across merge operators (linear, layer-selective; $0/200$ each) and rises only gently with size (max hazard 0.235 at size four). Predicting hazard from fingerprints by ridge regression, node-only features do as well as pairwise or higher-order ones (rank correlation 0.71, 0.71, 0.66), and a 2000-sample paired bootstrap puts every node-minus-interaction difference across zero: no measured gain from interaction features (evidence about prediction, not Mobius mass), while pairs-to-quadruples transfer degrades somewhat (Appendix K).

E2 (constructed collusion, and what it really is). Two adapters each carry a harmless backdoor (a rare trigger to a benign marker). On 240 *held-out* trigger probes over two independently trained pairs, each adapter alone fires on $\sim 0\%$ at merge weights $c \in [0.30, 0.40]$ while their composition already fires $\sim 100\%$ at c=0.30: unsafe where per-adapter screening at that weight sees nothing. A pre-registered dose-matched control isolates the mechanism — a single adapter at the matched coefficient-sum dose $2c$ also fires — so the crossing needs *no distinct-adapter mechanism*, though a saturating rate speaks to *mechanism*, not to the hazard being additive or its interaction mass small (Remark 1). Fixed-coefficient per-adapter screening misses a *dose*-dependent crossing.

Screening pipeline (E6). On an expanded benchmark of 1177 compositions at eight dose levels (461 positive), the hazard is *non-monotone* in dose — a full-weight merge often destroys the backdoor

(Arora et al., 2024). Applying the singleton screen of Definition 1 to *both* classes gives an eligible population of 462 (199 unsafe, 263 safe); on it a *coefficient-aware* model with a *quadratic* dose term reaches out-of-fold AUROC 0.993 (0.994 subset-grouped, node-only 0.87) and transfers to a *held-out* backdoor pair at 0.93–0.95. With i.i.d. (with-replacement) calibration, Learn-then-Test certifies an accept region at α=0.10 in $8/8$ seeds (full) and $6/8$ (eligible; FNR 0.03–0.05), borderline at α=0.05. E6 validates from-weights screening *mechanics*, not execution savings — a direct 20-probe test wins on this sub-2000 bank (Table 6).

7 Discussion and conclusion

What the measurements say. Two findings stand together. First, on a screened pool of public 7B adapters, emergent unsafety at the deployment threshold is rare and what hazard exists below it is node-predictable, with no resolved gain from pairwise or higher-order features (E1–E4): under the sampled composition law, ordinary composition is not, on this evidence, a common source of emergent harm — a *non-resolution*, not a proof of equivalence. Second, where emergent unsafety *is* constructed, it is real and screenable: on the 199 genuinely emergent positives whose constituents are each safe at their own coefficient (Definition 1), a coefficient-aware screener reaches out-of-fold AUROC 0.993 and transfers to a held-out risky pair at 0.93–0.95, against 0.87 for a node-only screener.

The blind spot is merge *dose*, not higher-order interaction. The decisive signal is the merge coefficient entered *nonlinearly*: the hazard is non-monotone in dose (a full-weight merge can destroy a backdoor), so a quadratic dose term lifts AUROC from 0.89 to 0.97 while pairwise and triple features, a pairwise MLP, and a set transformer add nothing once the same dose summary is supplied. Because LoRA merging is additive in weight space, this is what one should expect, and it leaves Theorem 1 — which bites only under genuine third-and-higher-order mass — neither confirmed nor refuted here. The actionable lesson is a shift in the unit of screening: *a LoRA's safety status is not portable across merge configurations*; the coefficient and operator are part of the security object, and a fixed-coefficient per-adapter screen is blind to a dose-dependent crossing.

Limitations. The positive evidence is a constructed, mechanism-specific benchmark: two independently trained canary pairs from one recipe, so the transfer result is across training instances of one mechanism, not across qualitatively different emergent failures. E6 is a proof of concept for the screener's discrimination and calibration *mechanics*, not for execution savings — it labels $\sim 85\%$ of a 1177-candidate bank, which sits below the $\gtrsim$ 2000-candidate regime where a from-weights screen beats exhaustive testing, and a direct 20-probe composition test is near-perfect when executing every candidate is affordable. The calibration certificate is exploratory (α=0.10, m_{cal}=647), theorem-backed only for the reserved calibration pool's empirical law, and non-empty in 6–8 of 8 draws; the guarantee does not survive distribution shift or an adaptive author, and repair utility is unmeasured. Results are for one 7B base and weight-space fingerprints at τ=0.30.

Conclusion. We recast composition safety screening as learning and calibrating a set function over an adapter pool, separated threshold emergence from non-additivity and higher-order mass, and measured where emergent unsafety does and does not appear. The contribution is the formulation, a conditional separation theorem with a distribution-free calibrated decision, a benchmark, and the finding that the object a deployable screen must encode is nonlinear merge dose. Cross-base transfer (E5), an adaptive author (E7), repair (E8), and a qualitatively different emergent mechanism are the natural next steps.

Use of large language models

This statement is provided under the ICLR 2027 policy on the use of large language models. The research idea, the conceptual framing that separates a threshold crossing from non-additivity of the hazard and from genuine higher-order Möbius mass, the structure of the paper, and every design decision and interpretation are the authors' own. The authors directed the work throughout and take full responsibility for all claims and results.

Under that direction, an LLM-based assistant was used as a tool to help carry the work out. It assisted in formalising the mathematical statements of Section 4 and drafting their proofs in Appendix B; every such statement was checked independently against its assumptions by the authors before inclusion, and the one result imported rather than proved here is labelled as such in Appendix B. It also implemented and ran, on the authors' compute and to the authors' specification, the experimental pipeline behind Section 6.1 — training the harmless-canary adapters, the vLLM generation and HarmBench-classifier judging, and the analyses, including the node-versus-interaction baseline and the dose-matched control; the authors specified the design, inspected the raw generations and the resulting hazards, and the per-composition records are in the anonymised repository with the code so the analyses can be re-executed independently. Finally it helped search and summarise related literature, produce the TikZ and PGFPlots figures and the arithmetic tables, format references, and edit the manuscript for readability. All 78 references were verified against an authoritative record. No LLM is listed as an author or contributor, and no result was included without the authors' specification and verification.

ETHICS STATEMENT

Dual use. This paper is defensive: it predicts which adapter compositions are unsafe so that they are not deployed. The dual-use hazard is that the same interaction features that rank compositions by risk could be used to search for a dangerous one. Two considerations bound that risk. The hazard is already public: Ding (2026) publishes a working construction, and single-adapter routes to the same end are older and cheaper (Lermen et al., 2023; Dong et al., 2025). And the screening features are diagnostic rather than generative; they rank a candidate set, they do not synthesise an adapter. Threat setting (iv) and experiment E7 study an adaptive author precisely so that the limits of the defence are published together with it rather than discovered later.

Release. The anonymised repository contains the fingerprint extractor, the risk models, the calibration code, the evaluation harness, and the measured records: adapter provenance, composition identifiers, and the per-composition labels. The positive class measured here is a harmless canary (a trigger mapped to a benign marker string), and no harmful-behaviour completion is included. For the deliberately risky adapters of the full benchmark design, we do not intend to release weights or the completions they produce; reproduction of that positive class would be supported through a documented construction recipe and an access-controlled request process rather than an open download, which keeps the benchmark reproducible for a screening researcher without shipping a ready-made harmful artifact.

Human subjects, data, and licensing. No human subjects and no personal data are involved. The privacy-extraction probes follow the published protocol of Carlini et al. (2021) and target only memorised content from public pretraining corpora; any extracted string is reported as an aggregate rate and never quoted. The licence each public adapter declares is recorded in the adapter manifest; four of the 37 declare none, and no adapter weights are redistributed.

Limits of the guarantee. The false-negative certificate of Theorem 3 is conditional on assumptions that a deployment can violate. Presenting a calibrated screener as a safety guarantee when its calibration distribution no longer matches the deployment is itself a hazard, and Sections 6.1 and 7 state the boundary in the terms a deployer would need.

REPRODUCIBILITY STATEMENT

Code, measured records and audits are available in an anonymised repository, `https://anonymous.4open.science/r/SafeCompose-F316`. The theoretical results are stated with their assumptions in Section 4 and proved in full in Appendix B; Appendix A tabulates notation and Figure 2 records which result depends on which. The fingerprint schema is enumerated feature by feature in Appendix D, the benchmark construction including the matched-group control in Appendix F, and the evaluation protocol including scoring rules, thresholds, splits and seeds in Appendix G. Appendix M gives the cost model behind every count in Section 5.4. The measurements reported in Section 6.1 — E1–E4, the node-versus-interaction baseline, and the E2

construction with its dose-matched control — were run on a single open-weight 7B base with a pool of public LoRA adapters and trained safety classifiers; the base and classifier checkpoints, the composition sampling law, generation and scoring settings, splits and seeds are in Appendix G, and the adapter manifest, the per-composition hazards and the control tables are in the repository (`outputs/`) with the scripts that produced them. The risk model and calibrated accept region (E6) were run on the expanded constructed benchmark and are reported in Section 6.1 and Table 6 (`pipeline_results_v2_n1177.json`): out-of-fold discrimination by feature tier, the Learn-then-Test accept region with its out-of-sample coverage and FNR, and risk-ranked versus random sequential testing (the active uncertainty/diversity acquisition rule and repair stage remain unrun). The analysis steps re-run on a CPU from the released records; generation and judging need the base model and a GPU. The remaining experiments (E5, E7, E8) are *not* yet run and are marked as future work.

REFERENCES

Andisheh Amrollahi, Amir Zandieh, Michael Kapralov, and Andreas Krause. Efficiently learning Fourier sparse set functions. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c77331e51c5555f8f935d3344c964bd5-Abstract.html.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2208.02814.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2), 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Ansh Arora, Xuanli He, Maximilian Mozes, Srinibas Swain, Mark Dras, and Qiongkai Xu. Here's a free lunch: Sanitizing backdoored models with model merge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15059–15075. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.894. URL https://aclanthology.org/2024.findings-acl.894/.

Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 462–471. PMLR, 2018. URL https://proceedings.mlr.press/v80/baptista18a.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 2023. doi: 10.1214/23-AOS2276. URL https://arxiv.org/abs/2202.13415.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535. URL https://arxiv.org/abs/2101.02703.

Vidmantas Bentkus. On hoeffding's inequalities. *The Annals of Probability*, 32(2), 2004. doi: 10.1214/009117904000000360. URL https://arxiv.org/abs/math/0410159.

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language models are Homer Simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14138–14149. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.762. URL https://aclanthology.org/2024.acl-long.762/.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html. 100 behaviours; success scored by an LLM judge (Llama-3-70B with a custom prompt in the v1.0 release).

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *10th International Conference on Learning Representations (ICLR)*, 2022. URL https://iclr.cc/virtual/2022/poster/6302.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Sihao Ding. Colluding LoRA: A compositional vulnerability in LLM safety alignment, 2026. URL https://arxiv.org/abs/2603.12681.

Tian Dong, Minhui Xue, Guoxing Chen, Rayne Holland, Yan Meng, Shaofeng Li, Zhen Liu, and Haojin Zhu. The philosopher's stone: Trojaning plugins of large language models. In *Proceedings of the 32nd Network and Distributed System Security Symposium (NDSS)*, 2025. URL https://www.ndss-symposium.org/ndss-paper/the-philosophers-stone-trojaning-plugins-of-large-language-models/.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://ojs.aaai.org/index.php/AAAI/article/view/4235.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999. doi: 10.1007/s001820050125. URL https://link.springer.com/article/10.1007/s001820050125.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 1024–1034, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

Hasan Abed Al Kader Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, Bernard Ghanem, and Mete Ozay. Model merging and safety alignment: One bad model spoils the bunch. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13033–13046. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.762. URL https://aclanthology.org/2024.findings-emnlp.762/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *The Ninth International Conference on Learning Representations (ICLR 2021)*, 2021. URL https://arxiv.org/abs/2009.03300.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when finetuning large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/77baa7c2a3a675823e89131698fd6e19-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2106.09685. URL https://arxiv.org/abs/2106.09685.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *Conference on Language Modeling (COLM)*, 2024a. URL https://arxiv.org/abs/2307.13269.

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. Composite backdoor attacks against large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1459–1472. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-naacl.94. URL https://aclanthology.org/2024.findings-naacl.94/.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566. arXiv preprint, not peer reviewed.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://iclr.cc/virtual/2023/poster/12254.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for human-AI conversations, 2023. URL https://arxiv.org/abs/2312.06674. arXiv preprint, not peer reviewed.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://arxiv.org/abs/1604.04173.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. LoRA fine-tuning efficiently undoes safety training in Llama 2-chat 70B, 2023. URL https://arxiv.org/abs/2310.20624.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to LLM security. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f3bfbd65743e60c685a3845bd61ce15f-Abstract-Conference.html.

Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 113–131. ACM, 2020. doi: 10.1145/3372297.3423362. URL https://dblp.org/rec/conf/ccs/LinXL020.html.

Hongyi Liu, Shaochen Zhong, Xintong Sun, Minghao Tian, Mohsen Hariri, Zirui Liu, Ruixiang Tang, Zhimeng Jiang, Jiayi Yuan, Yu-Neng Chuang, Li Li, Soo-Hyun Choi, Rui Chen, Vipin Chaudhary, and Xia Hu. LoRATK: LoRA once, backdoor everywhere in the share-and-play ecosystem. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23009–23047. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1253. URL https://aclanthology.org/2025.findings-emnlp.1253/. Earlier preprint (arXiv:2403.00108) was titled "LoRA-as-an-Attack! Piercing LLM Safety Under The Share-and-Play Scenario".

Qianli Ma, Dongrui Liu, Qian Chen, Linfeng Zhang, and Jing Shao. LED-Merging: Mitigating safety-utility conflicts in model merging with location-election-disjoint. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21749–21767. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1055. URL https://aclanthology.org/2025.acl-long.1055/.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html. Success scored by a fine-tuned Llama 2 13B Chat classifier (hashing-based classifier for copyright behaviours).

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4602–4609, 2019. doi: 10.1609/aaai.v33i01.33014602. URL https://ojs.aaai.org/index.php/AAAI/article/view/4384.

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 36 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.12827.

Stanislaw Pawlak, Jan Dubinski, Daniel Marczak, and Bartlomiej Twardowski. Backdoor vectors: a task arithmetic view on backdoor attacks and defenses, 2025. URL https://dblp.org/rec/journals/corr/abs-2510-08016.html. arXiv preprint, not peer reviewed.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ada93fa6643735f294be51dc31eebbd4-Abstract-Conference.html.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.225. URL https://aclanthology.org/2022.emnlp-main.225/.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pp. 46–54. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.7. URL https://aclanthology.org/2020.emnlp-demos.7/.

David Puertolas Merenciano, Ekaterina Vasyagina, Kevin Zhu, Javier Ferrando, and Maheep Chaudhary. Weight space detection of backdoors in LoRA adapters, 2026. URL https://arxiv.org/abs/2602.15195. arXiv preprint, not peer reviewed.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/83b7da3ed13f06c13ce82235c8eedf35-Abstract-Conference.html.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

Gian-Carlo Rota. On the foundations of combinatorial theory I. theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. doi: 10.1007/BF00531932. URL https://link.springer.com/article/10.1007/BF00531932.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024), Volume 1: Long Papers*, pp. 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/. 250 safe + 200 unsafe prompts; responses hand-labelled by the authors as full compliance / full refusal / partial refusal.

Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 725–732, Haifa, Israel, 2010. URL https://icml.cc/Conferences/2010/papers/285.pdf. Derives the optimal sampling distribution over an unlabelled test pool for estimating a fixed model's risk at minimal labelling cost.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL https://www.jmlr.org/papers/v9/shafer08a.html.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ff541950d1e885af90f523571564a401-Abstract-Conference.html.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html. 313 forbidden prompts; graded 0–1 by a rubric-based LLM judge (refused, specific, convincing) or a fine-tuned Gemma 2B grader.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023. URL https://www.jmlr.org/papers/v24/22-0202.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Cham, 2 edition, 2022. ISBN 978-3-031-06648-1. doi: 10.1007/978-3-031-06649-8. URL https://link.springer.com/book/10.1007/978-3-031-06649-8. XXVI+476 pages.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the limitations of representing functions on sets. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6487–6494. PMLR, 2019. URL https://proceedings.mlr.press/v97/wagstaff19a.html.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35413–35425. PMLR, 2023. URL https://proceedings.mlr.press/v202/wan23b.html.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://arxiv.org/abs/2010.09686.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52588–52610. PMLR, 2024. URL https://proceedings.mlr.press/v235/wei24f.html.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2404.13628.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1810.00826.

Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A. Gunter, and Bo Li. Detecting AI trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (S&P)*, pp. 103–120. IEEE, 2021. doi: 10.1109/SP40001.2021.00034. URL https://dblp.org/rec/conf/sp/XuWLBGL21.html.

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

Jinluan Yang, Anke Tang, Didi Zhu, Zhengyu Chen, Li Shen, and Fei Wu. Mitigating the backdoor effect for multi-task model merging via safety-aware subspace. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/28963.

Ming Yin, Jingyang Zhang, Jingwei Sun, Minghong Fang, Hai Li, and Yiran Chen. LoBAM: LoRA-based backdoor attack on model merging, 2024. URL https://arxiv.org/abs/2411.16746. Poster at the 2nd Workshop on Navigating and Addressing Data Problems for Foundation Models (DATA-FM), ICLR 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 3391–3401, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Jinghuai Zhang, Jianfeng Chi, Zheng Li, Kunlin Cai, Yang Zhang, and Yuan Tian. BadMerging: Backdoor attacks against model merging. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 4450–4464. ACM, 2024. doi: 10.1145/3658644.3690284. URL https://arxiv.org/abs/2408.07362.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b. URL https://arxiv.org/abs/2307.15043. Introduces GCG and the AdvBench dataset. arXiv preprint, not peer reviewed. The v1 author list omitted Carlini and Nasr.

A NOTATION

Symbols are listed in the order they first appear. Four are reused, and context disambiguates each. p is the base-model weight dimension in $W_0 \in \mathbb{R}^p$ and the inclusion probability of μ_p everywhere else; the p-value of Theorem 3 is always written p_λ. α is the LoRA scaling only when subscripted as α_i, and the target false-negative rate when bare. m is the probe-set size in Section 3, and the calibration-split size in Assumption 3, Theorem 3 and Appendix J, where it denotes the same quantity as m_{cal}. γ is the coefficient-decay constant of Proposition 1 except in Appendix J, which redefines it locally as coverage.

Symbol	Meaning
$W_0 \in \mathbb{R}^p$	base-model weights
$\mathcal{P} = \{a_1, \ldots, a_n\}$	adapter pool; $n = \lvert\mathcal{P}\rvert$
$U_i^{(\ell)}, V_i^{(\ell)}$	LoRA factors of adapter a_i at layer ℓ
$r_i, \alpha_i, \mathcal{L}_i$	rank, scaling and adapted layers of a_i
Δ_i	weight update induced by a_i, zero outside $\mathcal{L}_i$
$S \subseteq [n]$	a composition, i.e. a set of adapters
$c \in \mathbb{R}^{\lvert S\rvert}$	composition coefficients
$\oplus \in \mathcal{O}$	composition operator; $\lvert\mathcal{O}\rvert = 6$
$W(S, c, \oplus)$	composed weights, Eq. (1)
$\mathcal{T}, \mathcal{Q}$	evaluation suite; probe set with $\lvert\mathcal{Q}\rvert = m$
$\mathrm{haz}(\cdot), u(\cdot)$	hazard and utility functionals, both in $[0, 1]$
$r(S)$	composition hazard set function, $r(S) = \mathrm{haz}(W(S, c, \oplus))$
τ, τ_0	deployment threshold; per-adapter screening level, $\tau_0 < \tau$
$Y(S) = \mathbb{1}[r(S) > \tau]$	unsafe label
$z_i(S) = \mathbb{1}[i \in S]$	membership indicator
$a(T)$	Mobius (interaction) coefficient of r at T, Lemma 1
r_d	Mobius truncation of r at order d
$\mu_p,\ p,\ \sigma$	p-biased candidate measure; inclusion probability; $\sigma = \sqrt{p(1-p)}$
χ_T	p-biased Fourier basis function, $\chi_T = \prod_{i \in T}(z_i - p)/\sigma$
$\hat{r}(T)$	Fourier coefficient of r at T
$r^{\leq d}$	degree-d orthogonal projection, $\sum_{\lvert T\rvert \leq d} \hat{r}(T)\chi_T$
$\mathcal{H}_d$	set functions of degree at most d
$\Delta^2_{\geq 3}(r)$	third-and-higher-order Fourier mass, Eq. (4)
C, γ	coefficient-decay constants, Proposition 1
K	largest composition size screened
$\mathcal{C}$	candidate set $\{S : 2 \leq \lvert S\rvert \leq K\}$; $N = \lvert\mathcal{C}\rvert$
x_i, x_{ij}, x_{ijk}	node, pairwise and third-order fingerprints
$\phi_{\mathrm{w}}, \phi_{\mathrm{a}}$	weight-space and activation-space fingerprint maps
d	interaction order of the risk model
θ	parameters of the risk model f_θ; $g^{(1)}, g^{(2)}, g^{(3)}$ its order-wise heads
ς	logistic link
$\hat{R}(S)$	predicted risk, Eq. (10)
$\lambda, \Lambda, \hat{\Lambda}$	accept threshold; its grid; the FWER-validated subset
$\lambda^\star, \lambda^+$	deployed accept and reject thresholds
$\mathcal{A}(\lambda)$	accept region $\{S : \hat{R}(S) \leq \lambda\}$
$D(S)$	decision in $\{\mathsf{accept}, \mathsf{test}, \mathsf{reject}\}$
α, δ	target false-negative rate; failure probability of the certificate
$L_\lambda(S, Y)$	bounded calibration loss, Eq. (7)
$\mathrm{FNR}_{\mathrm{acc}}$	false-negative rate on the accepted set, Eq. (3)
$B,\ m_{\mathrm{cal}}$	execution budget; reserved calibration executions
π	prevalence of unsafe compositions in $\mathcal{C}$
$\pi(S), \pi_{\min}$	acquisition propensity and its floor, Proposition 2(iii)
β, β'	executed fraction; prevalence-corrected false-positive fraction
$\mathrm{A}, \mathrm{TPR}(\cdot)$	area under the empirical ROC; true-positive rate

B PROOFS

Every result stated in Section 4 is proved here in full. Nothing in this appendix is a sketch. Where a step imports an external theorem, that theorem's hypotheses are restated and checked against the present model before use. Figure 2 shows which result depends on which.

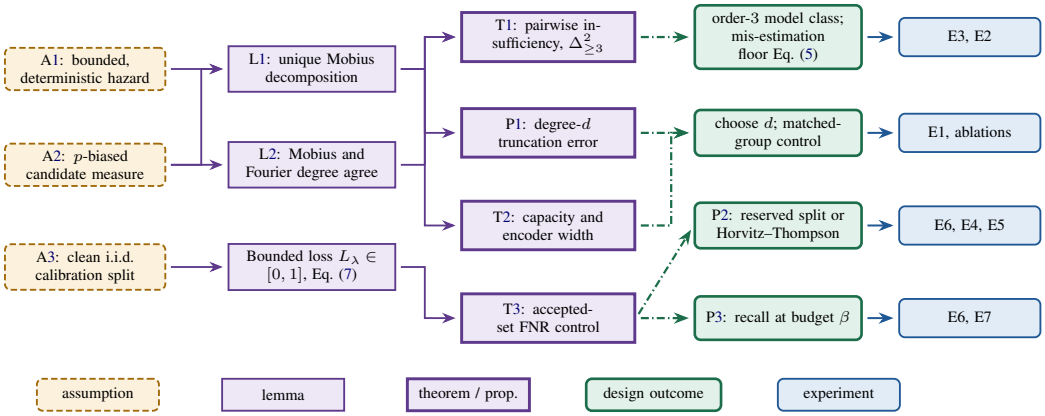


Figure 2: Dependency structure of the formal results. Arrows run from a statement to the statements that use it, and from a design consequence to the experiment that tests it. Assumptions are drawn with a dashed amber rule, lemmas with a thin purple rule, theorems and propositions with a thick purple rule, design consequences in green and experiments in blue, so the four kinds remain distinguishable in grayscale. Theorem 2(iii) is imported from Wagstaff et al. (2019) and Zaheer et al. (2017) and is the only node with an external source.

B.1 THE TWO BASES

Proof of Lemma 1. Regard r and a as vectors in $\mathbb{R}^{2^{[n]}}$ and the relation $r(S) = \sum_{T \subseteq S} a(T)$ as $r = Za$ with the zeta matrix $Z_{S,T} = \mathbb{1}[T \subseteq S]$. Order the subsets by any linear extension of inclusion; then Z is lower triangular with unit diagonal, so it is invertible and a is unique. It remains to identify Z^{-1} with the matrix $M_{S,T} = (-1)^{|S|-|T|}\mathbb{1}[T \subseteq S]$. For $U \subseteq S$,

$$(MZ)_{S,U} = \sum_{T:\, U \subseteq T \subseteq S} (-1)^{|S|-|T|} = \sum_{j=0}^{|S \setminus U|} \binom{|S \setminus U|}{j} (-1)^{|S \setminus U|-j} = (1-1)^{|S \setminus U|} = \mathbb{1}[U = S]\,,$$

where $j = |T \setminus U|$ ranges over subsets of $S \setminus U$, and the sum is empty unless $U \subseteq S$. Hence $MZ = I$ and $a = Mr$, which is the stated formula. Taking $S = \{i, j\}$ gives $a(\{i,j\}) = r(\{i,j\}) - r(\{i\}) - r(\{j\}) + r(\emptyset)$. This is the Mobius inversion of Rota (1964) specialised to the Boolean lattice, in the form used by Grabisch & Roubens (1999) and Tsai et al. (2023). □

Lemma 2 (The two gradings agree)**.** *Let $\tilde{r}(z) = \sum_T a(T) \prod_{i \in T} z_i$ be the multilinear extension of r to $\mathbb{R}^n$. For any $p \in (0,1)$ the functions $\{\chi_T\}_{T \subseteq [n]}$ with $\chi_T = \prod_{i \in T}(z_i - p)/\sigma$, $\sigma = \sqrt{p(1-p)}$, form an orthonormal basis of $L^2(\mu_p)$, and*

$$\max\{|T| : a(T) \neq 0\} \;=\; \max\{|T| : \hat{r}(T) \neq 0\} \;=\; \deg \tilde{r}.$$

In particular $r \in \mathcal{H}_d$ if and only if $a(T) = 0$ for all $|T| > d$, if and only if $\hat{r}(T) = 0$ for all $|T| > d$.

Proof. The space of multilinear polynomials in $z_1, \ldots, z_n$ has dimension 2^n, and restriction to $\{0,1\}^n$ is injective on it, hence bijective onto $\mathbb{R}^{2^{[n]}}$; so $\tilde{r}$ is the unique multilinear polynomial agreeing with r on the cube, and Lemma 1 exhibits its monomial coefficients.

Orthonormality: under μ_p the coordinates z_i are independent Bernoulli(p), so $\mathbb{E}[(z_i - p)/\sigma] = 0$ and $\mathbb{E}[((z_i - p)/\sigma)^2] = \mathrm{Var}(z_i)/\sigma^2 = 1$. For $T \neq T'$ pick $i \in T \triangle T'$; conditioning on the other coordinates and using independence, $\mathbb{E}[\chi_T \chi_{T'}]$ contains the factor $\mathbb{E}[(z_i - p)/\sigma] = 0$ if i lies in

exactly one of T, T', so the product vanishes. For $T = T'$ the expectation is 1. Since there are 2^n such functions and they are orthonormal in a 2^n-dimensional space, they form a basis.

Grading: each χ_T is a multilinear polynomial of degree exactly $|T|$, and expanding $\prod_{i\in T}(z_i - p)$ shows that χ_T lies in the span of $\{\prod_{i\in T'} z_i : T' \subseteq T\}$ with the leading coefficient $\sigma^{-|T|} \neq 0$ on $T' = T$. The change of basis between $\{\prod_{i\in T} z_i\}$ and $\{\chi_T\}$ is therefore triangular with respect to the partial order by inclusion and has nonzero diagonal, so it maps the span of monomials of degree at most d onto the span of $\{\chi_T : |T| \leq d\}$, for every d. Both maxima above therefore equal $\deg \tilde{r}$. □

B.2 Pairwise insufficiency

What the class does and does not contain. $\mathcal{H}_2$ is a linear subspace: the set functions of multilinear degree at most two. It is *not* the set of predictors built from node and pairwise features. A predictor that applies a nonlinearity after aggregating such features — a link function outside the hyperedge sum, an attention pooling of node embeddings, or a multilayer perceptron on $\{x_i\}, \{x_{ij}\}$ — is in general of full degree as a set function and Theorem 1 does not bind it. Three consequences are worth stating. Baseline 5 is the natural empirical stand-in for the ceiling of the theorem but is not literally inside it, and the same is true of baseline 6, so the two are not separated by class membership. SafeCompose's own model, Eq. (10), applies its link outside the hyperedge sum and is likewise not a member of $\mathcal{H}_3$; what the degree-d statement governs is the span its hyperedge terms range over, not the composed predictor. And the "no third-order hyperedges" ablation is a restriction of the feature set, not a projection onto $\mathcal{H}_2$. The theorem is therefore a statement about what no degree-two *function* can do, and the empirical baselines approximate that ceiling rather than realise it.

Proof of Theorem 1. By Lemma 2, $\mathcal{H}_2 = \mathrm{span}\{\chi_T : |T| \leq 2\}$, a linear subspace of $L^2(\mu_p)$. Write $h = \sum_{|T|\leq 2} \hat{h}(T)\chi_T$. By Parseval in the orthonormal basis $\{\chi_T\}$,

$$\|r - h\|^2_{L^2(\mu_p)} = \sum_{|T|\leq 2} \big(\hat{r}(T) - \hat{h}(T)\big)^2 + \sum_{|T|\geq 3} \hat{r}(T)^2 \;\geq\; \Delta^2_{\geq 3}(r),$$

with equality if and only if $\hat{h}(T) = \hat{r}(T)$ for every $|T| \leq 2$, that is, if and only if $h = r^{\leq 2}$. This is the orthogonal projection onto $\mathcal{H}_2$ and is in general different from the Mobius truncation r_2, because a monomial $\prod_{i\in T} z_i$ with $|T| = 3$ has nonzero $\chi_{T'}$ components for every $T' \subseteq T$, so deleting high-order monomials is not the same operation as deleting high-order orthogonal components.

For the second claim, suppose additionally that h takes values in $[0, 1]$. Since $r \in [0, 1]$ by Assumption 1, $|r - h| \leq 1$ pointwise. Splitting on the event $\{|r - h| > \tau\}$,

$$\Delta^2_{\geq 3} \;\leq\; \mathbb{E}_{\mu_p}\big[(r-h)^2\big] \;\leq\; \tau^2\,\mu_p\big(|r-h| \leq \tau\big) + 1 \cdot \mu_p\big(|r-h| > \tau\big) \;\leq\; \tau^2 + \mu_p\big(|r-h| > \tau\big),$$

which rearranges to Eq. (5). The bound is vacuous when $\tau^2 \geq \Delta^2_{\geq 3}$, which is the correct behaviour: a pairwise model is only forced to err by more than the interaction mass it cannot represent. □

Computation in Remark 2. Let $r(S) = \prod_{i\in T_0} z_i(S)$ with $|T_0| = k$. Substituting $z_i = p + \sigma\chi_{\{i\}}$ and expanding the product over the independent coordinates,

$$r = \prod_{i\in T_0} \big(p + \sigma\chi_{\{i\}}\big) = \sum_{T\subseteq T_0} p^{\,k-|T|}\sigma^{|T|}\chi_T,$$

using $\prod_{i\in T} \chi_{\{i\}} = \chi_T$. Hence $\hat{r}(T) = p^{k-|T|}\sigma^{|T|}$ for $T \subseteq T_0$ and $\hat{r}(T) = 0$ otherwise, so $\Delta^2_{\geq 3} = \sum_{j=3}^{k} \binom{k}{j} p^{2(k-j)} \big(p(1-p)\big)^j$, which at $k = 3$ is $(p(1-p))^3$. Every term carries a factor p^{2k-j} with $j \leq k$, so with $p = \kappa/n$ the whole sum is $O(p^k)$ and decays geometrically in the interaction order. □

Proof of Proposition 1. Fix S with $|S| = k$. By Lemma 1, $r(S) - r_d(S) = \sum_{T\subseteq S,\,|T|>d} a(T)$, a sum of $\sum_{j=d+1}^{k} \binom{k}{j}$ terms grouped by $|T| = j$. The triangle inequality and the hypothesis $|a(T)| \leq$

$C\gamma^{|T|}$ give the first inequality of Eq. (6). For the second, use $\binom{k}{j} \le k^j/j!$ and set $x = k\gamma$:

$$\sum_{j=d+1}^{k} \binom{k}{j}\gamma^j \;\le\; \sum_{j\ge d+1} \frac{x^j}{j!} \;=\; \sum_{i\ge 0} \frac{x^{\,d+1+i}}{(d+1+i)!} \;\le\; \frac{x^{d+1}}{(d+1)!}\sum_{i\ge 0}\frac{x^i}{i!} \;=\; \frac{x^{d+1}}{(d+1)!}\,e^x,$$

where the middle inequality uses $(d+1+i)! \ge (d+1)!\,i!$, valid for all integers $d, i \ge 0$. □

B.3 Capacity and width

Proof of Theorem 2. (i) Let $\mathcal{G}_d = \{h : h(S) = b_\emptyset + \sum_{T\subseteq S, 1\le|T|\le d} b(T)\}$ as b ranges over all real assignments to subsets of size at most d. Any such h has Mobius coefficients $a_h(T) = b(T)$ for $|T| \le d$ and $a_h(T) = 0$ otherwise, by the uniqueness in Lemma 1; hence $\mathcal{G}_d \subseteq \mathcal{H}_d$. Conversely any $h \in \mathcal{H}_d$ is of that form with $b = a_h$. So $\mathcal{G}_d = \mathcal{H}_d$. By uniqueness again, $h = r$ forces $b(T) = a(T)$ for $|T| \le d$ and $a(T) = 0$ for $|T| > d$, so exact representation holds if and only if r is order-d additive.

(ii) Suppose the assignment is tied through a feature map, $b(T) = g_\theta(x_T)$ for a single function g_θ per order. If $x_T = x_{T'}$ with $|T| = |T'|$ then $b(T) = b(T')$ for every θ. If r has $a(T) \ne a(T')$, part (i) requires $b(T) = a(T) \ne a(T') = b(T')$, a contradiction. So no θ represents r exactly, whatever the capacity of g_θ. The obstruction is in the feature map, not the head.

(iii) This part is imported. Zaheer et al. (2017, Thm. 2) prove that a permutation-invariant function of sets drawn from a *countable* universe admits an exact decomposition $\rho(\sum_{x\in S}\phi(x))$; the ϕ their proof constructs is a discontinuous injective encoding, and the uncountable-domain statements in their appendix are either restricted to a fixed cardinality or are approximation rather than exact representation. Wagstaff et al. (2019, Thms. 4.3–4.4) close the gap: for continuous ρ and ϕ with latent width L, representing all continuous permutation-invariant functions of sets of size at most K requires $L \ge K$. They state this for sets of real scalars, $f : \mathbb{R}^K \to \mathbb{R}$, whereas our fingerprints lie in $\mathbb{R}^D$ with $D > 1$. The necessity direction transfers: composing a width-L sum-decomposition over $\mathbb{R}^D$ with the continuous injection $t \mapsto (t, 0, \ldots, 0)$ yields one over $\mathbb{R}$, so $L < K$ already fails on the scalar slice. We do not claim their converse for $D > 1$. The remaining hypotheses hold in our setting: sets have size at most K, and the baseline encoder is continuous. Neither result is used anywhere else in this paper. □

B.4 The calibrated accept region

Lemma 3 (The loss is equivalent to the selective false-negative rate)**.** *With L_λ as in Eq. (7), $L_\lambda \in [0,1]$ pointwise, and*

$$\mathbb{E}[L_\lambda] \le \alpha \quad\Longleftrightarrow\quad \mathbb{E}\big[Y\mathbb{1}[S \in \mathcal{A}(\lambda)]\big] \le \alpha\,\mathbb{P}[S \in \mathcal{A}(\lambda)],$$

which, when $\mathbb{P}[S \in \mathcal{A}(\lambda)] > 0$, is exactly $\mathrm{FNR}_{\mathrm{acc}}(\lambda) \le \alpha$, and which holds trivially when $\mathbb{P}[S \in \mathcal{A}(\lambda)] = 0$.

Proof. $Y \in \{0,1\}$ and $\mathbb{1}[S \in \mathcal{A}(\lambda)] \in \{0,1\}$, so $(Y-\alpha)\mathbb{1}[\cdot] \in [-\alpha, 1-\alpha]$ and $L_\lambda = \alpha + (Y-\alpha)\mathbb{1}[\cdot] \in [0,1]$. Taking expectations, $\mathbb{E}[L_\lambda] = \alpha + \mathbb{E}[Y\mathbb{1}[\cdot]] - \alpha\mathbb{P}[\mathcal{A}(\lambda)]$, so $\mathbb{E}[L_\lambda] \le \alpha$ is equivalent to $\mathbb{E}[Y\mathbb{1}[\cdot]] \le \alpha\mathbb{P}[\mathcal{A}(\lambda)]$. Dividing by $\mathbb{P}[\mathcal{A}(\lambda)] > 0$ gives $\mathbb{E}[Y \mid S \in \mathcal{A}(\lambda)] \le \alpha$, which is Eq. (3). If $\mathbb{P}[\mathcal{A}(\lambda)] = 0$ both sides are zero and $\mathbb{E}[L_\lambda] = \alpha$. □

Proof of Theorem 3. Fix $\lambda \in \Lambda$ and consider $H_\lambda : \mathbb{E}[L_\lambda] > \alpha$. Under Assumption 3 the calibration points are i.i.d. and $\hat{R}$ is fixed, so $L_\lambda(S_1, Y_1), \ldots, L_\lambda(S_m, Y_m)$ are i.i.d. random variables in $[0,1]$ with mean $\mathbb{E}[L_\lambda]$. The Hoeffding–Bentkus p-value

$$p_\lambda = \min\Big\{\exp\big(-2m\,(\alpha - \hat{L}_\lambda)_+^2\big),\; e\,\mathbb{P}\big(\mathrm{Bin}(m,\alpha) \le \lceil m\hat{L}_\lambda\rceil\big)\Big\}, \qquad \hat{L}_\lambda = \tfrac{1}{m}\textstyle\sum_t L_\lambda(S_t, Y_t),$$

is super-uniform under H_λ, that is $\mathbb{P}[p_\lambda \le u] \le u$ for all $u \in [0,1]$ whenever $\mathbb{E}[L_\lambda] > \alpha$; this is the bounded-loss p-value of Bates et al. (2021, Thm. 3), built from Hoeffding's inequality and the Bentkus inequality (Bentkus, 2004), and used in exactly this role by Angelopoulos et al. (2025, Sec. 3). Its hypotheses are i.i.d. observations bounded in $[0,1]$, both established above; no monotonicity in λ is required, which matters here because $\lambda \mapsto \mathrm{FNR}_{\mathrm{acc}}(\lambda)$ need not be monotone.

Let $\hat{\Lambda}$ be returned by any procedure with family-wise error rate at most δ applied to $\{p_\lambda\}_{\lambda\in\Lambda}$ over the finite grid, for instance Bonferroni or fixed-sequence testing. By definition of FWER control, $\mathbb{P}[\exists\lambda \in \hat{\Lambda} \text{ with } H_\lambda \text{ true}] \le \delta$. On the complement, every $\lambda \in \hat{\Lambda}$ satisfies $\mathbb{E}[L_\lambda] \le \alpha$, which by Lemma 3 is $\mathrm{FNR}_{\mathrm{acc}}(\lambda) \le \alpha$ when the accept region has positive probability and is vacuous otherwise. That is Eq. (8).

Tighter p-values may be substituted without changing the conclusion, provided they are super-uniform under the same null; the betting bounds of Waudby-Smith & Ramdas (2024) are the usual choice, and Bates et al. (2021, Thm. 4) states them in this form. The statement is conditional on $\hat{R}$; since Assumption 3 makes the calibration sample independent of $\hat{R}$, the tower property gives the unconditional statement as well. □

Proof of Proposition 2. **(i)** Let $\mathcal{Q}$ be an acquisition rule that queries the m candidates with the smallest $\hat{R}$, and let the labels satisfy: $Y = 0$ for every candidate in the lowest $\hat{R}$ decile, and $\mathbb{P}[Y = 1] = \eta$ for candidates in the remainder of $\mathcal{A}(\lambda)$, with $\eta > \alpha$ and the remainder carrying at least half the mass of $\mathcal{A}(\lambda)$. Such a configuration is consistent with any $\hat{R}$ that is confidently correct at the very bottom of its range and poorly calibrated just above it, which is the ordinary behaviour of a model fitted on few labels. Every queried point then lies in the lowest decile, so $Y_t = 0$ and $L_\lambda(S_t, Y_t) = \alpha - \alpha = 0$ for all t, giving $\hat{L}_\lambda = 0$ and $p_\lambda \le \exp(-2m\alpha^2) \to 0$. Thus $\lambda \in \hat{\Lambda}$ with probability tending to one. Meanwhile the true rate is $\mathrm{FNR}_{\mathrm{acc}}(\lambda) = \mathbb{E}[Y \mid \mathcal{A}(\lambda)] \ge \eta/2 > \alpha$ for $\eta > 2\alpha$, so Eq. (8) fails. The mechanism is that the sample is no longer exchangeable with a fresh candidate: acquisition conditions on $\hat{R}$, and $\hat{R}$ is correlated with Y, so $\hat{L}_\lambda$ is not an unbiased estimate of $\mathbb{E}[L_\lambda]$ and p_λ loses super-uniformity. An uncertainty-based rule biases the estimate in the opposite direction and is merely conservative in this example, but it is invalid for the same reason and is conservative only by accident, not by construction.

(ii) Draw $\mathcal{D}_{\mathrm{cal}}$ by i.i.d. uniform sampling with replacement from $\mathcal{C}$ before acquisition begins, and fit $\hat{R}$ only on the actively acquired labels $\mathcal{D}$. Then $\mathcal{D}_{\mathrm{cal}} \perp \hat{R}$, and conditional on $\hat{R}$ the calibration points are i.i.d. from the uniform candidate distribution, so Assumption 3 holds and Theorem 3 applies verbatim. The cost is m_{cal} executions, which are charged against B in line 2 of Algorithm 1 rather than being obtained for free.

(iii) Enumerate $\mathcal{C} = \{S^{(1)}, \dots, S^{(N)}\}$ and let each be sampled independently with known probability $\pi(S) \ge \pi_{\min} > 0$ (Poisson sampling). Write $\bar{L}_\lambda = N^{-1}\sum_{S\in\mathcal{C}} L_\lambda(S, Y(S))$, which is well defined because $Y(S)$ is deterministic given S under Assumption 1. The Horvitz–Thompson estimator is $\hat{L}_\lambda = N^{-1}\sum_{S\in\mathcal{C}} \mathbb{1}[S \text{ sampled}]\; L_\lambda(S, Y(S))/\pi(S)$, and

$$\mathbb{E}[\hat{L}_\lambda] = \frac{1}{N}\sum_{S\in\mathcal{C}} \frac{\mathbb{P}[S \text{ sampled}]}{\pi(S)} L_\lambda(S, Y(S)) = \frac{1}{N}\sum_{S\in\mathcal{C}} L_\lambda(S, Y(S)) = \bar{L}_\lambda.$$

The N summands are independent, and the S-th lies in $[0,\, 1/(N\pi(S))] \subseteq [0,\, 1/(N\pi_{\min})]$ because $L_\lambda \in [0, 1]$. Hoeffding's inequality for independent bounded variables gives, for any $t > 0$,

$$\mathbb{P}\big[|\hat{L}_\lambda - \bar{L}_\lambda| \ge t\big] \le 2\exp\Big(-\frac{2t^2}{\sum_S \big(1/(N\pi_{\min})\big)^2}\Big) = 2\exp\big(-2t^2 N\pi_{\min}^2\big),$$

and setting the right-hand side to δ' yields $|\hat{L}_\lambda - \bar{L}_\lambda| \le \pi_{\min}^{-1}\{\log(2/\delta')/(2N)\}^{1/2}$ with probability at least $1 - \delta'$. The one-sided version supplies the p-value $p_\lambda = \exp(-2N\pi_{\min}^2(\alpha - \hat{L}_\lambda)_+^2)$, which is super-uniform under $H_\lambda : \bar{L}_\lambda > \alpha$: writing $s = \big(\log(1/u)/(2N\pi_{\min}^2)\big)^{1/2}$,

$$\mathbb{P}[p_\lambda \le u] = \mathbb{P}[\hat{L}_\lambda \le \alpha - s] \le \mathbb{P}[\hat{L}_\lambda \le \bar{L}_\lambda - s] \le \exp(-2N\pi_{\min}^2 s^2) = u.$$

Substituting this p-value into the proof of Theorem 3 gives the same conclusion with the target now the pool mean $\bar{L}_\lambda$ over the enumerated $\mathcal{C}$ rather than an expectation over a hypothetical candidate distribution. The width of the resulting interval carries the factor $1/\pi_{\min}$, which is the price of reusing acquired labels. This is the design-based analogue of the weighted-conformal correction under covariate shift (Tibshirani et al., 2019; Barber et al., 2023), and it is the easy case, since the screener chose the propensities and therefore knows them exactly rather than estimating them. □

Proof of Proposition 3. Work on the finite pool $\mathcal{C}$ with $N = |\mathcal{C}|$, $N_+ = \pi N$ unsafe and $N_- = (1-\pi)N$ safe candidates, and let ties in $\hat{R}$ be broken arbitrarily but fixed. Executing the top βN candidates corresponds to a threshold t_β with

$$\beta N = \mathrm{TPR}(t_\beta) N_+ + \mathrm{FPR}(t_\beta) N_-.$$

Since $\mathrm{TPR}(t_\beta) \le 1$, we get $\mathrm{FPR}(t_\beta) N_- \ge \beta N - N_+ = (\beta - \pi)N$, hence $\mathrm{FPR}(t_\beta) \ge (\beta - \pi)/(1-\pi) = \beta'$, which is positive because $\beta > \pi$. The empirical ROC curve $u \mapsto \mathrm{TPR}_{\mathrm{ROC}}(u)$ is non-decreasing, so $\mathrm{Recall}(\beta) = \mathrm{TPR}(t_\beta) \ge \mathrm{TPR}_{\mathrm{ROC}}(\beta')$.

For the second step, split the area under that curve at β' and bound each piece using monotonicity on the left and $\mathrm{TPR}_{\mathrm{ROC}} \le 1$ on the right:

$$\mathrm{A} = \int_0^1 \mathrm{TPR}_{\mathrm{ROC}}(u)\, du = \int_0^{\beta'} + \int_{\beta'}^1 \;\le\; \beta' \,\mathrm{TPR}_{\mathrm{ROC}}(\beta') + (1 - \beta').$$

Rearranging gives $\mathrm{TPR}_{\mathrm{ROC}}(\beta') \ge (\mathrm{A} - 1 + \beta')/\beta'$, which combined with the previous display is Eq. (9). The argument is deterministic and concerns the empirical ROC of one pool; it is not a generalisation bound, and it is tight only for the ROC that is flat at $\mathrm{TPR}_{\mathrm{ROC}}(\beta')$ on $[0, \beta']$ and equal to 1 afterwards, so it is loose for scores with a well separated tail. □

B.5 Why Learn-then-Test rather than the neighbouring machinery

Remark 3 compresses a comparison worth stating in full, because the three frameworks control genuinely different objects and substituting one for another would invalidate Theorem 3. RCPS (Bates et al., 2021) controls $\mathbb{P}[R(\lambda) \le \alpha] \ge 1 - \delta$ for a risk that is an expected loss, but its Theorem 1 requires a *nested* family of set-valued predictors and a loss monotone non-increasing in the set, so that a single scalar λ traces a monotone risk curve. Our $\lambda \mapsto \mathrm{FNR}_{\mathrm{acc}}(\lambda)$ is a ratio of two expectations whose numerator and denominator both grow as λ increases, and it need not be monotone: adding a batch of safe compositions to the accept region lowers it, adding an unsafe one raises it, and which happens depends on the ordering the score induces. Learn-then-Test (Angelopoulos et al., 2025) needs only super-uniform p-values and a family-wise error procedure, which is why the reduction in Lemma 3 is routed through it: monotonicity is not required and Λ may even be multidimensional, which matters if a deployment wants to calibrate the accept and reject thresholds jointly. Conformal risk control (Angelopoulos et al., 2024) is the third option and is genuinely simpler, but its Theorem 1 gives $\mathbb{E}[L_{n+1}(\hat{\lambda})] \le \alpha$, an expectation over the joint draw of the calibration set and the test point, under exchangeability and a loss that is non-increasing and right-continuous in λ and bounded. A screening system is calibrated once and then used for many decisions on that one calibration draw, so a statement that holds on average over hypothetical re-calibrations is the wrong object; the high-probability statement of Eq. (8) is what a deployer can act on. Reporting a conformal-risk-control guarantee as though it were a high-probability one, or the converse, would be the single easiest way to misstate what this system provides.

B.6 What is not proved here

Three statements in the paper are not theorems and are not presented as such. The claim that weight-space and activation-space fingerprints *do* carry the interaction signal is an empirical hypothesis, tested by RQ1 and the fingerprint ablations. The decay hypothesis $|a(T)| \le C\gamma^{|T|}$ behind Proposition 1 is an assumption about real adapter pools, estimated by E3 rather than derived. And no result here says anything about threat setting (iv): an adaptive author chooses the candidate distribution, which is precisely the object Assumptions 2 and 3 hold fixed, so the certificate of Theorem 3 is void under that setting by construction and E7 measures the damage rather than bounding it.

C A worked three-adapter instance

The theory of Section 4 is short enough to be checked by hand on a small instance, and doing so exposes a distinction that the aggregate statement of Theorem 1 hides. This appendix works one instance through completely.

C.1 THE INSTANCE

Take $n = 3$ adapters on a base model with $\mathrm{haz}(W_0) = r(\emptyset) = 0.02$, a per-adapter screening level $\tau_0 = 0.10$ and a deployment threshold $\tau = 0.30$. Suppose the suite returns

S	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$r(S)$	0.02	0.05	0.04	0.06	0.07	0.09	0.08	**0.62**
$Y(S)$	0	0	0	0	0	0	0	**1**

Every adapter passes alone with margin, so the pool is individually screened in the sense of Definition 1; every pair also passes; the triple is unsafe. By Definition 1 the triple is emergently unsafe, and it is the only positive.

C.2 THE MOBIUS COEFFICIENTS

Applying Lemma 1 term by term,

$$a(\emptyset) = 0.02, \qquad a(\{1\}) = 0.03, \qquad a(\{2\}) = 0.02, \qquad a(\{3\}) = 0.04,$$
$$a(\{1,2\}) = 0, \qquad a(\{1,3\}) = 0, \qquad a(\{2,3\}) = 0,$$

since for instance $a(\{1,2\}) = 0.07 - 0.05 - 0.04 + 0.02 = 0$ and $a(\{1,3\}) = 0.09 - 0.05 - 0.06 + 0.02 = 0$, while the third-order coefficient is

$$a(\{1,2,3\}) = 0.62 - (0.07 + 0.09 + 0.08) + (0.05 + 0.04 + 0.06) - 0.02 = 0.51.$$

So the entire hazard above the per-adapter baseline sits in a single third-order coefficient. All pairwise interaction coefficients are exactly zero: the three pairs are not merely safe, they are *exactly additive*. This is the regime the paper is about, in its purest form.

C.3 WHAT EACH SCREENING FAMILY PREDICTS

Per-adapter screening (baseline 1) sees $\max_i r(\{i\}) = 0.06 < \tau_0$ and reports the pool clean. It is not wrong about the adapters; it is answering a different question.

The Mobius truncation at order two is $r_2(\{1,2,3\}) = a(\emptyset) + a(\{1\}) + a(\{2\}) + a(\{3\}) = 0.11$, comfortably below $\tau = 0.30$. Any predictor that fits node and pair terms on the observed pairs and then *extrapolates* to the triple, which is what baselines 3 and 5 do and what E3 tests, therefore accepts an unsafe composition. Its error on that one candidate is $0.62 - 0.11 = 0.51$, exactly $a(\{1,2,3\})$.

The degree-two orthogonal projection is a different object, and it is worth being precise because conflating the two would overstate Theorem 1. Under $\mu_{1/2}$ one has $\chi_T(S) = \prod_{i \in T}(2z_i - 1)$ and, expanding each monomial as $\prod_{i \in U} z_i = 2^{-|U|} \sum_{T \subseteq U} \chi_T$, the Fourier coefficients are $\hat{r}(T) = \sum_{U \supseteq T} 2^{-|U|} a(U)$, giving

$$\hat{r}(\emptyset) = 0.12875, \qquad \hat{r}(\{1\}) = 0.07875, \qquad \hat{r}(\{2\}) = 0.07375,$$
$$\hat{r}(\{3\}) = 0.08375, \qquad \hat{r}(\{i,j\}) = 0.06375, \qquad \hat{r}(\{1,2,3\}) = 0.06375.$$

As a check, $\hat{r}(\emptyset)$ must equal $\mathbb{E}_{\mu_{1/2}}[r]$, and the eight values above average to $1.03/8 = 0.12875$. Hence $\Delta^2_{\geq 3} = \hat{r}(\{1,2,3\})^2 = 0.06375^2 \approx 4.06 \times 10^{-3}$, and Theorem 1 says no degree-two function achieves mean squared error below that, so its root mean squared error is at least 0.0637 across the eight subsets.

Now the nuance. The optimal degree-two approximation is $r^{\leq 2} = r - \hat{r}(\{1,2,3\})\chi_{\{1,2,3\}}$, and at $S = \{1,2,3\}$, where $\chi_{\{1,2,3\}} = 1$, it evaluates to $0.62 - 0.06375 = 0.55625$, which is far above τ. On this instance the L^2-optimal pairwise model *does* flag the composition. There is no contradiction with Theorem 1, which bounds squared error and not decision quality, but there is a lesson: $r^{\leq 2}$ can only be formed by someone who already knows r on all eight subsets, including the triple. A screener does not. What a screener can form from pairs alone is the Mobius truncation r_2, and that is the object that fails, by exactly $a(\{1,2,3\})$.

Remark 4 (The operative failure is order extrapolation)**.** Theorem 1 bounds how well a degree-two function can fit a hazard whose values are all available. The failure a screening system actually suffers is stronger and more specific, but it holds only under a stated fit. *If* a model exactly interpolates all Mobius coefficients up to order d (as a degree-d polynomial with enough evaluated lower-order points does), then predicting at order $> d$ returns the Mobius truncation, whose error at S is $\sum_{T \subseteq S, |T|>d} a(T)$ with no cancellation available. An arbitrary regularised or nonlinear fit on *sampled* pairs need not return that truncation — it may do better or worse — so this is a statement about the truncation object, not a guarantee about any fitted baseline. Proposition 1 bounds the truncation error under a decay assumption, and E3 measures the ranking degradation empirically. Reporting Theorem 1 as though it directly implied the extrapolation failure would be an overstatement, and we avoid it.

C.4 What the instance says about the design

Three things, each of which appears elsewhere in the paper as a design decision rather than as an observation. First, the pairwise features on this instance carry no signal at all, since every $a(\{i,j\}) = 0$; a system that scores triples must therefore have a term that is not a sum over pairs, which is Theorem 2(i). Second, the fingerprint map has to distinguish the triple from a matched triple with the same three nodes but a different outcome, or Theorem 2(ii) says no head can fit both; this is the matched-group construction of Appendix F. Third, with $\Delta^2_{\geq 3} \approx 4 \times 10^{-3}$ on a three-element pool the signal is small in absolute terms even when it is the whole story, which is Remark 2 at $n = 3$: the bound of Eq. (5) at $\tau = 0.05$ gives only $\mu_p(|r - h| > 0.05) \geq 1.6 \times 10^{-3}$, a statement too weak to be useful on an instance this small. The aggregate bound is an asymptotic instrument; the decision-level failure is visible directly, and the paper leans on the second rather than the first.

D Fingerprint schema, model variants, and the remaining algorithms

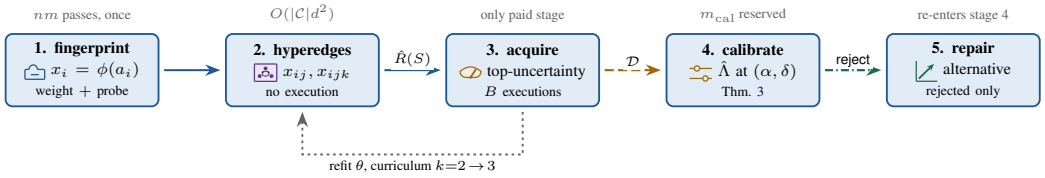


Figure 3: The five stages of SafeCompose. Stage 1 costs n adapter reads and nm probe forward passes and is paid once per pool. Stages 3 and 4 are the only stages that consume the execution budget. Stage 4 is the calibrated policy of Theorem 3; stage 5 runs only on rejected candidates. Interaction features (stage 2) are formed from stage-1 fingerprints without touching the base model again, and the dotted loop is the refit between acquisition rounds.

D.1 Node features $x_i = \phi(a_i)$

Every feature is computable from the adapter file and at most m forward passes of the base model with that adapter attached. Nothing here requires composing two adapters, which is what keeps stage 1 linear in n.

D.2 Interaction features

For $\{i,j\}$: cosine similarity between Δ_i and Δ_j per layer and pooled; principal angles between their leading singular subspaces; the overlap of their layer supports; the signed agreement of their alignment projections and of their refusal-direction components; a sign-conflict count in the style of the interference diagnosis of Yadav et al. (2023); and the difference between the sum of their individual activation displacements and the displacement of the merged pair measured on a 16-prompt sub-probe, which is the only interaction feature that touches the model and is optional (the no-execution variant drops it, and the ablation reports the cost).

Table 1: Weight-space block $\phi_{\mathrm{w}}(a_i)$. Dimensions are per adapted layer ℓ and are then summarised over $\mathcal{L}_i$ by mean, maximum, and the vector of per-layer values on a fixed layer grid, so adapters that adapt different layer sets remain comparable.

Feature	Dim. / layer	Why it is here
$\|\Delta_i^{(\ell)}\|_F$, $\|\Delta_i^{(\ell)}\|_2$	2	scale of the update; baseline 2 uses these alone
top-8 singular values of $\Delta_i^{(\ell)}$	8	spectral shape separates benign from poisoned single adapters (Puertolas Merenciano et al., 2026)
rank concentration $\sum_{j\le 3} s_j^2 / \sum_j s_j^2$	1	safety is localised at the rank level (Wei et al., 2024)
principal left/right singular directions	2×8	the geometry pairwise features are built from
alignment projection onto $W_0^{\mathrm{aligned}} - W_0^{\mathrm{base}}$	1	the quantity Safe LoRA screens (Hsu et al., 2024)
residual outside the alignment subspace	1	what the projection discards
new-direction (intruder) mass	1	low-rank updates introduce directions absent from the base spectrum (Shuttleworth et al., 2025)
layer-localisation profile over $\mathcal{L}_i$	$\|\mathcal{L}\|$	overlap with the safety layers (Li et al., 2025)
rank r_i, scaling α_i, adapted-module mask	3	configuration, controlled for in the rank ablation

Table 2: Activation-space block $\phi_{\mathrm{a}}(a_i)$, measured on the probe set $\mathcal{Q}$. All statistics are aggregated over $\mathcal{Q}$ by mean and by the 90th percentile, and activations are cached so the cost is paid once per adapter.

Feature	Dim.	Why it is here
refusal-token logit margin	4	direct read of the behaviour the label is about
mean hidden-state displacement per layer	$\|\mathcal{L}\|$	where in depth the adapter moves the model
component along the refusal direction	$\|\mathcal{L}\|$	refusal is mediated by one direction (Arditi et al., 2024)
linear-probe readouts (harmfulness, compliance)	4	representation-level readout (Zou et al., 2023a)
representational similarity to the base model	2	how far the adapter moves representations at all
tool-invocation rate on tool-shaped prompts	2	tool-use safety is one of the six suite dimensions
predictive entropy and margin	2	uncertainty, used by the acquisition rule

For $\{i, j, k\}$: the triple subspace overlap, the volume spanned by the three leading directions, and the residual of the three pairwise terms, so the third-order head sees what the pairwise heads cannot already explain. This mirrors the residual structure of the Mobius decomposition and is what makes the curriculum of Algorithm 1 coherent.

D.3 Model variants

The default is Eq. (10) with $d = 3$ and three shared MLP heads of two hidden layers. Two alternatives are evaluated as ablations rather than as the method. *Hypergraph convolution* replaces the fixed sum by message passing between node and hyperedge representations (Feng et al., 2019; Chien et al., 2022), which adds depth at the cost of an inductive bias that is harder to read off. *Set encoder* drops hyperedges entirely and pools node embeddings with attention (Lee et al., 2019), which is baseline 6 and which Theorem 2(iii) says needs latent width at least K.

Algorithm 2 CALIBRATE: the accept region of Theorem 3.

Require: score $\hat{R}$; reserved split $\mathcal{D}_{\text{cal}} = \{(S_t, Y_t)\}_{t=1}^{m}$; targets α, δ; grid $\Lambda = \{\lambda_1 < \cdots < \lambda_G\}$ placed on the quantiles of $\hat{R}$ over the *training pool* that fit $\hat{R}$, fixed before $\mathcal{D}_{\text{cal}}$ is drawn

1: drop from Λ every λ whose coverage is below $n^{\text{crit}}(m, \alpha, \delta)/m$, the smallest accepted fraction that could ever give $p \le \delta$; renumber $\lambda_1 < \cdots < \lambda_G$ ▷ grid trimming: no score on $\mathcal{D}_{\text{cal}}$ and no label enters

2: **for** $g = 1, \ldots, G$ **do**

3: $\quad \hat{L}_g \leftarrow \frac{1}{m} \sum_t \big[\alpha + (Y_t - \alpha) \mathbb{1}\big[\hat{R}(S_t) \le \lambda_g\big]\big]$

4: $\quad p_g \leftarrow \min\{\exp(-2m(\alpha - \hat{L}_g)_+^2),\ e\,\mathbb{P}[\text{Bin}(m, \alpha) \le \lceil m\hat{L}_g \rceil]\}$

5: **end for**

6: $\hat{\Lambda} \leftarrow$ fixed-sequence test: accept $\lambda_1, \lambda_2, \ldots$ while $p_g \le \delta$, stop at the first failure ▷ FWER $\le \delta$ without a Bonferroni penalty

7: **if** $\hat{\Lambda} = \emptyset$ **return** $\emptyset$ and coverage 0 (empty accept region; the policy abstains) ▷ matches Algorithm 1

8: **return** $\hat{\Lambda}$ and the empirical coverage $\frac{1}{m} \sum_t \mathbb{1}\big[\hat{R}(S_t) \le \max \hat{\Lambda}\big]$

D.4 ALGORITHMS 2 AND 3

Fixed-sequence testing controls the family-wise error rate for *any* pre-specified ordering, with no multiplicity correction, which is all Theorem 3 needs; a prefix structure of the valid region is not required for validity. What the ordering affects is *power* — whether the sequence reaches the thresholds we care about before it stops. We order the grid so that this power is spent where it matters, as follows.

Where the grid starts, and why it matters. Ordering is about power, not validity. By Eq. (7) a candidate that is *not* accepted still contributes α to the loss, so at the most conservative grid point, where almost nothing is accepted, $\hat{L}_g \approx \alpha$ and $p_g \approx 1$. A fixed sequence begun there halts at its first step and returns $\hat{\Lambda} = \emptyset$ *however well* $\hat{R}$ *separates*, which is not conservatism but an absence of power. Concretely, the smallest number of accepted calibration points that could ever yield $p_g \le \delta$, achieved when none of them is unsafe, solves $\alpha(1 - n/m) \le \hat{L}^{\text{crit}}$; below it no grid point is rejectable whatever the labels are. Line 1 of Algorithm 2 therefore removes every grid point below that coverage before any p-value is formed. The bound is a function of (m, α, δ) alone, so the trimmed grid is exactly as pre-specified as the untrimmed one and Theorem 3 is untouched; Bonferroni over the full grid is the alternative, valid with no ordering assumption at the cost of the multiplicity penalty. For the same reason Λ is placed on the quantiles of $\hat{R}$ over the *training pool* that fit $\hat{R}$ and is fixed before $\mathcal{D}_{\text{cal}}$ is drawn, so the grid depends on neither the identities nor the labels of the calibration sample.

Algorithm 3 SAFEALTERNATIVE: propose screened repairs for a rejected composition.

Require: rejected S; coefficients c; risk model $\hat{R}$; utility model $\hat{u}$; policy thresholds $\lambda^\star, \lambda^+$; search width w

1: attribute risk: $\rho_i \leftarrow g_\theta^{(2)}$ and $g_\theta^{(3)}$ terms of Eq. (10) containing i, summed ▷ interaction-head magnitude, a ranking signal, not a Mobius estimate

2: $\mathcal{N} \leftarrow \emptyset$

3: **for** i in the top-w adapters by ρ_i **do**

4: $\quad \mathcal{N} \leftarrow \mathcal{N} \cup \{(S, c^{(i,\varrho)})\}$ for $\varrho \in \{0.75, 0.5, 0.25\}$ ▷ shrink c_i

5: $\quad \mathcal{N} \leftarrow \mathcal{N} \cup \{(S \setminus \{i\}, c_{-i})\}$ ▷ drop a_i

6: $\quad \mathcal{N} \leftarrow \mathcal{N} \cup \{(S, c, \oplus_{\mathcal{L} \setminus \mathcal{L}_{\text{safety}}})\}$ ▷ restrict the merge away from the safety layers

7: **end for**

8: keep $\mathcal{N}' \leftarrow \{S' \in \mathcal{N} : \hat{u}(S') \ge \zeta\, \hat{u}(S)\}$ ▷ utility floor ζ, reported not tuned

9: **return** the elements of $\mathcal{N}'$ re-submitted to the policy of Algorithm 1, ranked by $\hat{u}$

Line 8 is the part that matters for honesty: an alternative is not returned as safe, it is returned as a new candidate that must clear the same accept threshold. It does *not* inherit the certificate: Theorem 3 bounds the false-negative rate over the calibration distribution, whereas repairs are a selected sub-population, so the accepted-set guarantee does not transfer to them per candidate and must be validated on the repair distribution itself (E8). If none clears the threshold, the answer to the user is that no repair was found.

Two moves the model cannot screen. The feature map of Appendix D is built from adapter fingerprints alone: neither the coefficient vector c nor the operator $\oplus$ appears in it, so $\hat{R}$ of Eq. (10) is *constant* in both. Two of the three moves are therefore invisible to the model that proposes them. Line 5 shrinks c_i and line 7 restricts the merge away from the safety layers; each leaves every feature, and hence the predicted risk, exactly as it was, so re-submitting the result at line 8 returns the decision that rejected it. Only line 6, which drops an adapter and so changes S, is screened in the ordinary way. Proposals of the first two kinds are meaningful only once *executed*, and an implementation should mark them so that a caller cannot mistake one for a screened repair. Admitting c and $\oplus$ to the feature map would remove the restriction, but it is a change to the model rather than to the algorithm and we do not make it here; the consequences of leaving them out are collected in Section 6.1.

E PROBE-SET DESIGN AND ITS SENSITIVITY

The probe set $\mathcal{Q}$ is the one component that sits between the two halves of the fingerprint and determines the cost of stage 1. It is also the component most likely to be quietly overfitted, since the same prompts that build the features could be reused to build the labels. This appendix fixes its construction, its separation from the labelling suite, and the two conditions that stress it.

E.1 CONSTRUCTION

$\mathcal{Q}$ contains m prompts in five strata, with the target proportions below. The strata exist so that a fingerprint records how an adapter moves the model on each kind of input rather than only on the harmful ones, which is what makes the features informative about adapters that are not harmful at all.

Stratum	Share	Purpose
Harmful requests	25%	measures refusal-direction displacement where refusal is the correct behaviour
Borderline requests	20%	the region where over-refusal and under-refusal separate; drawn from the safe-but-refusable style of Röttger et al. (2024)
Benign task requests	25%	measures capability displacement, so that a capability adapter is not scored as a safety event
Tool-shaped requests	15%	tool-use safety is one of the six suite dimensions and has its own activation signature
Neutral continuations	15%	a low-variance reference for representational similarity to the base model

E.2 SEPARATION FROM THE LABELLING SUITE

$\mathcal{Q}$ and $\mathcal{T}$ are disjoint in prompts and, where possible, in source. A prompt that appears in $\mathcal{Q}$ never appears in $\mathcal{T}$, and the harmful stratum of $\mathcal{Q}$ is drawn from a different release of the harmful-behaviour corpora than the one the labelling suite uses. The reason is not tidiness: if the two overlap, an activation feature can encode the eventual label directly, the predictor learns a lookup rather than an interaction, and E2 will not catch it because the leak travels with the prompt rather than with the adapter. The overlap is checked by exact and near-duplicate matching before any run, and the check is reported.

E.3 Size

Stage-1 cost is nm forward passes, so m is the parameter that decides whether fingerprinting is cheap. The design target is m in the low hundreds, with the $m/4$ ablation reporting the cost of a quarter-size probe. Two things make small m plausible. The activation features are aggregates, means and high quantiles over $\mathcal{Q}$, whose standard error falls as $m^{-1/2}$, and the refusal direction they project onto is one-dimensional (Arditi et al., 2024), so the quantity being estimated is low-dimensional even though the representation is not. Against that, the borderline stratum is where adapters differ most and also where per-prompt variance is highest, so the ablation reports per-stratum sensitivity rather than a single curve.

E.4 Two stress conditions

Probe distribution shift. The features are recomputed on a probe set drawn with different stratum proportions, 40% harmful and 10% borderline, and the model trained on the original mix is evaluated without refitting. This is the condition that most directly threatens deployment, because a hub operator will not use our probe set. Being precise about the guarantee: recomputing fingerprints on a shifted probe set changes the feature map and therefore the *score function* $\hat{R}$ itself (the score is head $\circ$ features, not the head alone), so Theorem 3's fixed-score certificate no longer applies to the shifted score — it must be recalibrated on fresh exchangeable data. What the condition measures is how much a screener *calibrated on the original probes* then loses under shift: accuracy and hence coverage at fixed α degrade, and the honest report is that the old certificate is void until recalibration, with coverage collapse the expected symptom.

Removal of the alignment and refusal projections. The two features with the strongest prior justification (Hsu et al., 2024; Wei et al., 2024; Arditi et al., 2024) are deleted and the model refitted. If performance is unchanged, the prior is decorative and the model is working from generic representational statistics; if it collapses, the fingerprint is doing what the design claims and the safety-geometry literature is load-bearing rather than ornamental. Either outcome is reportable and the second is the one the design predicts.

E.5 Caching and determinism

Activations for $(a_i, \mathcal{Q})$ are computed once and cached, keyed by the adapter hash, the probe-set version, the base checkpoint hash and the decoding configuration. Decoding is greedy wherever the feature does not require sampling, so that Assumption 1's determinism holds for the fingerprint half exactly and for the label half up to the judge variability that the label-noise condition measures.

F Benchmark construction

The benchmark is the first contribution the specification asks for (C1) and the one most able to mislead if built carelessly. This appendix fixes its construction before any data is collected.

F.1 Adapter pool

Six groups, sized so that the pair and triple candidate space is enumerable while quadruples remain sampled.

Every adapter, including the risky ones, must satisfy Definition 1: it passes the full suite alone at level τ_0 with margin. An adapter that fails alone is removed from the pool rather than kept as an easy positive, because keeping it would let a node-only baseline succeed for the wrong reason. This filter is applied before any composition is formed and its pass rate is reported.

F.2 Constructing the positive class

Emergent unsafety must be present often enough to measure and must not all come from one mechanism. Four generators, reported separately so that a result driven by one of them cannot be presented as a result about all four.

Table 3: Pool composition. Counts are the design target, not a measurement. *Risky* adapters populate the positive class and are the only ones whose weights are withheld from release (ethics statement).

Group	Count	Trained on	Role
Domain / capability	10–14	disjoint task corpora	realistic user adapters
Style / persona	6–8	style transfer corpora	known to shift refusal phrasing
Instruction following	4–6	public instruction sets	the common stacking partner
Safety oriented	3–5	refusal and policy data	tests whether safety adapters help or interfere
Controlled risky	6–10	constructed, see below	positive class
Third-party public	4–8	unknown, as published	external validity

1. **Found.** Compositions of independently trained benign adapters that happen to be unsafe. This is the E1 quantity and the only generator that requires nothing constructed; if it is empty, the paper says so (Section 6.1).
2. **Interference driven.** Pairs selected to maximise a weight-space conflict statistic without any safety-directed training, following the interference diagnosis of Yadav et al. (2023) and the neuron-level account of Ma et al. (2025).
3. **Collusion.** Pairs and triples trained jointly so that harm appears only in composition, reproducing the construction of Ding (2026). These are the hardest positives and the ones an adversary would actually build.
4. **Merge-activated backdoor.** One contributor carries a trigger that is inert alone and active after merging, the setting of Zhang et al. (2024) and Yin et al. (2024). Reported separately because it violates the spirit of Definition 1 even when it satisfies the letter: the node is malicious, only undetectably so.

F.3 Matched groups: the control that decides the claim

For each positive composition S^+ we construct one or more *matched* negatives S^- with the same multiset of node fingerprints, or as close as the pool allows, and a different outcome. Two regimes must be kept apart. Under an *exact* node-fingerprint collision Theorem 2(ii) applies: no model whose prediction is a function of the node fingerprints alone can separate the group. The subtlety is that any interaction feature that is itself a deterministic function of those fingerprints collides too, so an interaction-aware model separates the group only through information the node summaries do not carry — here the composition's own small-probe activation fingerprint (and raw-weight interaction terms), not a re-encoding of $\{x_i\}$. Exact collisions are therefore where the claim is decided, and we report which feature supplies the separating signal. A sibling swap (one adapter replaced by a same-corpus, different-seed twin) gives only a *near* match: proximity is not an impossibility, since a classifier can separate two close points absent a regularity constraint, so these are reported as an empirical robustness probe with the within-group fingerprint distance, not as an instance of Theorem 2(ii); groups above a distance threshold are excluded.

A second construction holds S fixed and changes only the coefficient vector c or the operator $\oplus$. This leaves $\{x_i\}$ exactly unchanged and changes r, but — as E2 makes concrete — the score $\hat{R}$ of Section 5 is itself constant in c and $\oplus$, so it cannot separate these pairs either. They are therefore *not* a node-versus-interaction control; we report them instead as a probe of the score's blindness to merge dose and operator, a limitation the coefficient-aware extension of Section 6.1 is meant to address.

F.4 Composition sampling

All $\binom{n}{2}$ pairs are enumerated. Triples are sampled with probability proportional to a mixture of uniform and a diversity term over groups, so that no group dominates. Quadruples are sampled uniformly at a fixed count for E3 and are never used for training in the pair-and-triple condition. The reported first-measurements run (Section 6.1) drew from a 37-adapter pool and analysed 500 compositions with complete fingerprints and labels, split across sizes $2/3/4$ as $150/200/150$; the per-composition records are released so the interval and rankings can be recomputed against this

exact sample rather than an ecosystem-wide population. Coefficients are drawn from a fixed grid, and each $(S, \oplus)$ appears at more than one coefficient setting so that the coefficient-shift ablation is possible without new executions. Figure 4 summarises which threat setting each generator instantiates.

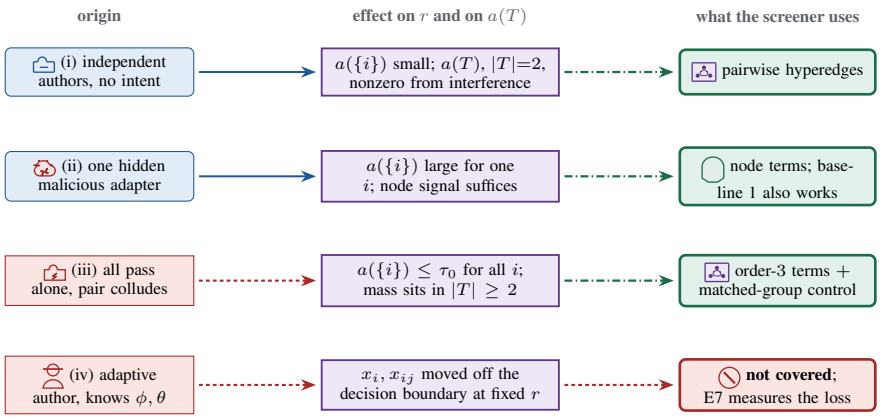


Figure 4: Threat settings, their effect on the interaction coefficients of the hazard set function, and the part of the screener responsible for each. Rows (iii) and (iv) are the evaluation targets. Row (iv) is drawn in red on both sides because the certificate of Theorem 3 does not cover it, and E7 measures the resulting loss rather than bounding it.

F.5 Splits

Splits are at the level of adapters, not compositions, because two compositions sharing an adapter are dependent and a composition-level split would leak. E2 partitions the pool; E3 partitions by size; E4 partitions by operator; E5 partitions by base-model family. The calibration split of Algorithm 1 line 2 is drawn uniformly from the test-side candidate set and is disjoint from everything used for fitting, as Proposition 2(ii) requires.

G Evaluation protocol

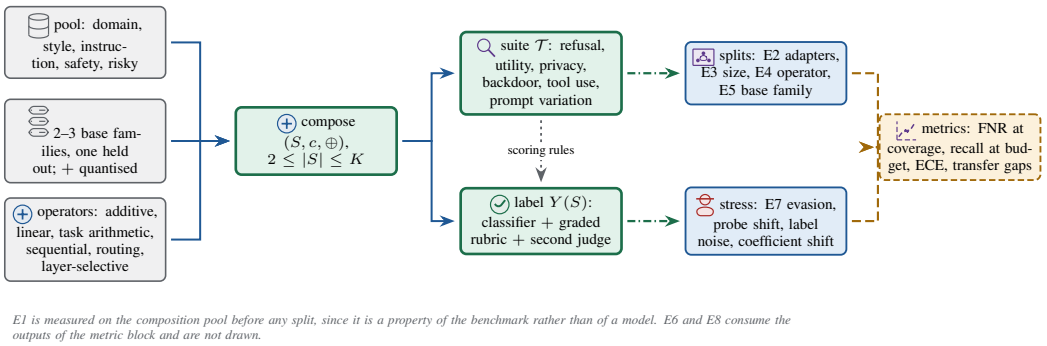


Figure 5: Protocol dependencies. The pool, base models and operator family determine the candidate set; each executed candidate passes through the suite and the labelling stage; splits and stress conditions consume labelled candidates; metrics consume both. Scoring rules differ across suite components and are not pooled.

G.1 The six suite dimensions and how each is scored

Scoring mechanisms differ, and the difference matters: a fine-tuned classifier, a graded rubric applied by a model, and a human-annotated set are not interchangeable, so each dimension records which it uses.

Table 4: Suite composition. *Scoring* names the mechanism actually used by the cited source.

Dimension	Prompts	Scoring
Refusal and policy	behaviours from Mazeika et al. (2024) and Chao et al. (2024); attack strings from Zou et al. (2023b)	HarmBench's fine-tuned classifier, plus the graded 0–1 rubric of Souly et al. (2024); the two are reported separately, never averaged
Over-refusal	Röttger et al. (2024)	the benchmark's labels are human annotations, so an automated stand-in is reported as an approximation and its agreement with the released labels is stated
Benign utility	Hendrycks et al. (2021)	exact match, no judge
Privacy extraction	protocol of Carlini et al. (2021)	extraction rate, aggregate only; no extracted string is stored or quoted
Backdoor probes	held-out triggers from the constructed adapters	target-output match
Tool-use safety	tool-shaped prompts with a scripted sandbox	scripted checks on the invoked call

The hazard $\mathrm{haz}(W)$ is a fixed weighted combination of the dimension scores with weights declared before collection, and the sensitivity of every conclusion to those weights is an ablation. A second classifier (Inan et al., 2023) labels every executed composition independently; the label used is the primary one, and disagreement between the two is reported as a rate and used to drive the label-noise stress condition, since judge disagreement is a documented error source (Zheng et al., 2023).

G.2 Metrics

Eleven, with the estimator fixed for each. AUROC and AUPRC for ranking. Accepted-set false-negative rate $\mathrm{FNR}_{\mathrm{acc}}$ and safe coverage, always reported as a pair at a stated operating point, since either alone is trivially optimisable. Test reduction, $1-(B+m_{\mathrm{cal}})/|\mathcal{C}|$. Recall at testing budget, with Eq. (9) drawn as the analytic floor. Expected calibration error under the equal-mass binned estimator (Guo et al., 2017) and under the debiased estimator of Kumar et al. (2019), reported together because the binned estimate is biased downward and the two can disagree materially. Empirical coverage of the calibrated policy over seeds against nominal $1-\delta$. Utility retention, the ratio of the composed model's suite utility to the best single constituent's. Cross-adapter and cross-operator transfer gaps, each the drop from the matched in-distribution condition. Screening compute in forward passes and in executions, reported separately because they differ by orders of magnitude in cost. The primary deployment metric is $\mathrm{FNR}_{\mathrm{acc}}$ at a stated safe-coverage level or testing budget.

G.3 Ablations and stress tests

Sixteen, each removing or perturbing one component: weight-only fingerprints; activation-only fingerprints; no pairwise features; no third-order hyperedges; set encoder in place of the hypergraph; probe set reduced to $m/4$; probe distribution shifted away from the training probe mix; alignment and refusal projections removed; adapter rank varied across the pool; quantised base model (Dettmers et al., 2023); merge coefficients shifted off the training grid; calibration reusing actively acquired labels under Proposition 2(iii) instead of a reserved split; and safety labels corrupted at the observed judge-disagreement rate. Three of the sixteen perturb the acquisition rule of Stage 3 rather than the risk model, and they are the control for RQ4: uniform random acquisition; the pair-before-triple curriculum removed; the batch diversity penalty removed. Each runs on the same trained ranker to a matched execution count, so the comparison isolates the selection rule. They are necessary because baselines 5, 6 and 8 all acquire under the same rule at the same budget, so a budgeted comparison between them confounds ranking quality with acquisition quality, and Eq. (9)

cannot stand in: it is a property of the ranker's ROC area, not of the selection rule, and is vacuous at $\mathrm{A} = 1/2$. If uniform random acquisition matches the full rule on recall at the matched execution count, the acquisition rule is dropped, Stage 3 is reported as uniform sampling under the same budget, and the reduction reported for RQ4 is credited to screening rather than to active testing. E7, adaptive fingerprint evasion, is reported as an experiment rather than an ablation because it changes the data-generating process rather than the model.

G.4 Thresholds

τ is set from the base model's own hazard plus a margin declared in advance, so that $\mathrm{haz}(W_0) \leq \tau_0 < \tau$. τ_0 is the per-adapter screening level of Definition 1. Neither is tuned on results. The policy targets are $\alpha \in \{0.01, 0.05\}$ and $\delta = 0.05$, and the grid Λ has $G = 200$ points on the score quantiles of the *training pool* that fit $\hat{R}$ — fixed before the calibration sample is drawn, so the threshold family depends on neither the identities nor the labels of the calibration split, as Theorem 3 requires. Using quantiles of candidates in (or selected around) the calibration sample would make the family data-dependent and is not what is run.

G.5 Budgets and repetitions

B is the acquisition budget (seeds included) and is swept over $\{0.5, 1, 2, 5, 10, 20\}$ percent of $|\mathcal{C}|$. m_{cal} is *not* a fixed constant: it is set from Table 5 at the intended operating point, with a floor of 1000, and is counted *separately* from B, so total executions are $B + m_{\mathrm{cal}}$ (plus any test-branch and executed-repair labels), matching Section 5.4 and Algorithm 1. Appendix J shows that a fixed 1000 is an order of magnitude too small at $\alpha = 0.01$, which is why the protocol states a rule rather than a number. Five seeds control adapter initialisation, composition sampling, acquisition, and the calibration draw. Confidence intervals come from the seed spread and are reported as ranges, not as standard errors of a single run.

G.6 Pre-registered decision rules

Three rules are fixed now so that they cannot be chosen after the fact. **(P1)** The interaction claim stands only if SafeCompose beats baseline 1 on the matched groups of Appendix F; a win on the unmatched set alone is reported as evidence about adapters, not about interactions. **(P2)** The higher-order claim stands only if removing the third-order head changes the size-4 result in E3; if it does not, the paper reports that $d = 2$ suffices at the sizes tested and Theorem 1 is reported as a possibility not realised in this pool.

What P2 can and cannot see. A margin is only a decision rule if the comparison can resolve it. Treating the two AUROCs as paired on the same candidates and taking the Hanley–McNeil variance, a 0.01 margin at the E3 operating point — $4{,}000$ held-out candidates at a prevalence of 0.15, base AUROC 0.90 — reaches power 0.47 when the two models' AUROCs correlate at 0.8, and the smallest difference detectable at power 0.80 is 0.0152, larger than the margin itself. Power is strongly sensitive to that correlation, which is not identified by the AUROCs alone: 0.14 at $r{=}0$, 0.47 at $r{=}0.8$, 0.96 at $r{=}0.95$. Two models sharing a fingerprint map and a training split are correlated, so the favourable end is plausible, but it is an assumption rather than a measurement. We therefore report P2 asymmetrically: passing it supports the higher-order claim, whereas failing it withdraws the claim *without* establishing that $d = 2$ suffices, unless the pool is large enough that the margin exceeds the minimum detectable difference at the realised prevalence. On a pool of $20{,}000$ candidates the same margin reaches power 0.93.

(P3) The calibration claim stands only if empirical coverage over seeds contains the nominal $1 - \delta$; under-coverage is reported as a failure of the assumptions, and the first suspect is Proposition 2.

G.7 What would falsify the paper

If E1 returns an emergent unsafe rate indistinguishable from zero on independently trained adapters, the screening problem as posed is not a live one outside adversarial construction, and that is the finding. If baseline 5, the pairwise model, matches SafeCompose at size 4, the higher-order machinery

is unnecessary at these sizes. If E7 removes the advantage entirely, the method is not deployable against an informed author and the contribution reduces to the accidental and one-malicious-adapter settings. Each of these outcomes is reportable, and the protocol is arranged so that none of them can be quietly avoided.

G.8 Realized first-measurements run

The measurements of Section 6.1 were produced by the following concrete configuration; the per-composition records and adapter manifest are in the anonymised repository (`outputs/`) with the scripts that produced them.

Base and judges. Base model `Qwen/Qwen2.5-7B-Instruct` served with vLLM at temperature 0 (greedy). Labels come from the HarmBench classifiers (Mazeika et al., 2024). The reporting judge, which labels the 500-composition cohort, is `cais/HarmBench-Llama-2-13b-cls`; the second classifier, `cais/HarmBench-Mistral-7b-val-cls`, is the single-adapter screen and was the in-loop judge for 484 of the cohort's compositions. The reported zero holds under both ($0/500$ and $0/484$), and a separate 13B re-scoring of 512 saved generations with a prompt-length guard also gives $0/512$.

Pool. 37 distinct public LoRA adapters for this base, each passing Definition 1 at $\tau_0{=}0.10$ on the HarmBench standard behaviours (median solo hazard 0.05 under the 7B screen, against a median of 0.08 over the 484 compositions it scored). The manifest lists each adapter's source repository, Hub revision, licence, rank and LoRA targets, and the training configuration its model card declares (30 of 37 declare one). The deployment threshold is $\tau{=}0.30$.

Hazard measured. The headline E1/E4 rate is the *harm/compliance* dimension — HarmBench standard-behaviour attack-success — not the full six-dimensional weighted hazard of Appendix G; the backdoor dimension is measured separately (E2), and the two classifiers agree only at the composition-level zero, not on every response. Per-dimension weights and the other four dimensions are part of the protocol design, not this realised run; conclusions at thresholds below $\tau{=}0.30$ are not determined by the zero-crossing count (the maximum observed harm hazard is 0.235).

Compositions and operators. 500 compositions with complete weight-space fingerprints and labels, as unique subsets S at the default operator, split $150/200/150$ over sizes $2/3/4$ (E1, under `weighted_addition`). E4 re-runs 200 pairs from the same seeded sampler (150 of them the cohort's pairs) under `linear_merge` and `layer_selective`. Generation uses the harm suite at ≤ 192 new tokens; the backdoor probes use ≤ 32.

Baseline regression. Ridge regression with 5-fold cross-validation and $\lambda \in \{1, 10\}$; features are the node, pairwise and triple aggregates of the per-module weight-space fingerprints of Appendix D. Reported rank correlations and top-quintile AUROC are the $\lambda{=}1$ values on the 500 compositions; the 2000-sample paired bootstrap (`baseline_bootstrap.json`) resamples compositions with replacement and refits each tier on the same resample, giving the node-minus-interaction difference intervals reported. We do *not* claim the per-tier tie-induced optimism cancels in the difference (the bias of a difference is the difference of biases, not necessarily zero); we report the paired-difference intervals directly, and they straddle zero, which we read as *no resolved difference* on this pool — not as proven equivalence, for which we set no margin. The resampling code is released and recomputes the out-of-fold predictions and the intervals exactly from the released records.

E2 construction and control. Two backdoor adapters trained on the closed-QA corpus with a 6% trigger→marker poison fraction (trigger `zq7-diagnostic-marker`, harmless marker `BANANA-CANARY-7Z ok`), rank 16, 250 steps, learning rate 10^{-4}, batch 1, gradient accumulation 8, sequence length 512, seeds 0 and 1; a second independent pair uses seeds 2 and 3. The dose-matched control (`control_dose.json`) sweeps the merge coefficient and compares the distinct pair at c against a single adapter at $2c$ on 40 trigger probes. The dose-response (`dose_response.json`) repeats this over a fine grid $c \in [0.30, 0.90]$, 240 held-out probes and both pairs, and additionally reports the merged-update Frobenius norm of each comparator, so "matched dose" can be read as either coefficient sum or update norm; the two give different verdicts (equal under coefficient-sum matching, the pair lower-norm under norm matching), which is why we claim only that the crossing needs no distinct-adapter mechanism, not that it is a pure concentrated dose.

Cohort note. The reported cohort is 500 compositions, all with complete fingerprints and labels (150/200/150 over sizes 2/3/4); an earlier draft quoted 512 and a mid-run pull 484, both superseded by this complete set. The released records keep the denominator auditable: the 500-composition cohort, the 484 of its compositions labelled in-loop, and the 512 saved generations re-scored by the 13B classifier (484 of them in the cohort). The unit of analysis is the subset S at the default operator (E4 varies the operator on a 200-subset).

H WHY EACH BASELINE FAILS, AND WHAT WOULD MAKE IT SUCCEED

A baseline table is only informative if the reasons for the ordering are stated before the numbers arrive. This appendix does that, and in three cases it also states the condition under which the baseline would win, which is the outcome that would force us to withdraw a claim.

1. Independent adapter risk (max or mean). Predicts $\hat{R}(S) = \max_{i \in S} r(\{i\})$ or the mean. By Lemma 1 this is a function of the first-order coefficients only, so on an individually screened pool it is bounded above by τ_0 for every candidate (Definition 1). It is uninformative by construction on the matched groups of Appendix F, where every candidate in a group shares the same node fingerprints and hence the same score; on the unmatched pool any advantage it shows is evidence about adapters, not about interactions. *It would win* if the pool were not individually screened, that is, if unsafe compositions were mostly compositions containing an already-unsafe adapter. That is threat setting (ii), it is the setting most of the merging-backdoor literature studies (Zhang et al., 2024; Yin et al., 2024), and E1 reports what fraction of positives fall into it. If that fraction is near one, the interesting problem is not present in this benchmark and we say so.

2. Weight-norm or spectral threshold. A richer node-only statistic: it can detect an adapter whose update is unusually large or spectrally unusual, which is what separates poisoned from benign single adapters in Puertolas Merenciano et al. (2026) and what Safe LoRA's alignment projection measures (Hsu et al., 2024). It remains a function of $\{x_i\}$ alone, so Theorem 2(i) with $d = 1$ applies and the analysis of baseline 1 carries over. Ding (2026) report empirically that a Safe-LoRA-style weight-space test does not flag their colluding pairs, which is the expected behaviour. *It would win* if adapters that participate in unsafe compositions were individually identifiable from their spectra even while passing the behavioural suite. That is a real possibility and it is exactly what the weight-only-fingerprint ablation isolates: if weight-only features match the full system, the activation half is not earning its cost.

3. Pairwise cosine similarity. The first baseline with any interaction content, and the cheapest possible one: a single scalar per pair, aggregated over the pairs inside S. Aggregated by a sum over the pairs it sits in $\mathcal{H}_2$ and inherits the bound of Theorem 1, and on the instance of Appendix C it has literally nothing to work with because every second-order coefficient is zero. *It would win* if composition risk were driven by gross geometric overlap between updates, which is the hypothesis that merge interference and safety loss are the same phenomenon. Ma et al. (2025) give the closest evidence for that hypothesis at the neuron level, so it deserves a fair test rather than a rhetorical dismissal; if baseline 3 matches the full system, the paper's contribution reduces to a better interference statistic.

4. Small-probe safety test of the composition. The only baseline that executes something for every candidate: it merges each candidate and runs a short probe rather than the full suite. It has no representation limit at all, so Theorem 1 says nothing about it. Its cost is what disqualifies it: $|\mathcal{C}|$ merges, which is the quantity Eq. (2) makes prohibitive, and its recall is bounded by how well a short probe approximates the full suite. It is included because it is what a careful practitioner would actually do today, and because the gap between it and the exhaustive oracle measures how much of the labelling signal a cheap probe recovers. *It would win* on a small pool: Table 9 gives the crossover, below roughly 2000 candidates, and we do not claim an advantage there.

5. Pairwise MLP on $\{x_i\}, \{x_{ij}\}$. The strongest baseline restricted to node and pair *features*: an unrestricted learned function of them. The restriction is on what it sees, not on the class of set functions it realises, and as a nonlinear map it need not lie in $\mathcal{H}_2$; Theorem 1 therefore does not

literally bind it. What binds it is weaker and empirical: any third-order structure it recovers has to be a function of node and pair features alone. It is the practical stand-in for the ceiling of Theorem 1, and the comparison between it and the full system at size 4 is the single most informative number in the evaluation, because both models see the same node and pairwise features and differ in whether third-order features are available to them (not in a strict multilinear degree, since the linked model need not be low-degree, Appendix B). Pre-registered rule P2 of Appendix G turns that comparison into a decision: if baseline 5 matches SAFECOMPOSE at size 4, we report that $d = 2$ suffices at the sizes tested and that Theorem 1 describes a possibility not realised in this pool.

6. Set transformer without hyperedges. Permutation-invariant attention pooling over node embeddings (Lee et al., 2019). Like baseline 5, and unlike the linear class $\mathcal{H}_2$ of Theorem 1, it is not confined to degree two: attention can in principle express higher-order dependence between the pooled elements. What separates it from baseline 5 is that it sees no pair feature at all, so any higher-order structure must be recovered from node features alone. Theorem 2(iii)'s latent-width bound is stated for *sum-decomposition* architectures $\rho(\sum_i \phi(x_i))$ (Wagstaff et al., 2019); the attention-pooled set transformer is not of that form, so we do not claim the bound constrains it directly — we report it as a strong learned opponent whose result we are least able to predict. *It would win* if the interaction information is already present in the node fingerprints and only needs to be combined nonlinearly, in which case explicit hyperedge features are an inductive bias rather than a necessity, and the honest report is that the hypergraph helps sample efficiency rather than expressivity.

7. Exhaustive oracle. Executes every candidate and is correct by construction, at cost $|\mathcal{C}|$ merges and suite runs. It exists to bound the recall column and to make the cost comparison concrete, and it is computed on a sub-pool small enough to enumerate; the full-pool oracle is left undefined rather than extrapolated.

8. SAFECOMPOSE. The full system. Its advantage over baseline 1 measures whether interactions matter at all; over baseline 3, whether learned interaction features beat one hand-chosen statistic; over baseline 5, whether order > 2 matters; over baseline 6, whether explicit hyperedges beat implicit pooling; and against baseline 7, how much execution is saved. Each of those four comparisons is a separate claim with a separate way of failing, and the paper reports them separately rather than as a single win.

What would make the whole comparison uninformative. If E1 returns an emergent unsafe rate near zero, every baseline and SAFECOMPOSE alike will have almost no positives to rank, the AUROC column will be dominated by variance, and the correct report is that the benchmark did not exhibit the phenomenon. That is stated as a falsification condition in Appendix G rather than handled by enlarging the constructed positive class until the numbers become interesting.

I THE ADAPTIVE AUTHOR: THREAT SETTING (IV) AND EXPERIMENT E7

Threat setting (iv) is the one the certificate of Theorem 3 does not cover, and the honest way to present a defence that has an uncovered setting is to specify the attack against it precisely enough that the result is falsifiable. This appendix does that. It specifies an evaluation adversary, not a deployable attack, and no artifact from it is released; the ethics statement gives the release conditions.

I.1 CAPABILITIES

The adaptive author knows the fingerprint map ϕ, the trained parameters θ, the decision thresholds $\lambda^\star, \lambda^+$, the operator family $\mathcal{O}$, and the base model W_0. This is the strongest reasonable assumption: an open screening tool is exactly what we propose to release, so Kerckhoffs's position is the right one. The author does *not* know the calibration split, the labels of executed compositions, or the hidden portion of the suite $\mathcal{T}$. The author controls one adapter $a_\star$ and knows which pool it will join.

I.2 Objective

The author wants a composition $S \ni \star$ that is genuinely unsafe but predicted safe, while $a_\star$ retains utility so that it attracts adoption. Writing $u(\cdot)$ for suite utility and $r(\cdot)$ for the hazard,

$$\max_{a_\star} \; r\big(S\big) \quad \text{subject to} \quad \hat{R}_\theta(S) \le \lambda^\star, \qquad r(\{\star\}) \le \tau_0, \qquad u(\{\star\}) \ge u_{\min}, \tag{11}$$

where the first constraint is evasion, the second keeps $a_\star$ admissible under Definition 1, and the third keeps it useful. In practice the author optimises a Lagrangian of a differentiable surrogate for r (a harmful-compliance loss on a held-out harmful set), the exact $\hat{R}_\theta$, which is differentiable in $\phi(a_\star)$ by construction, a refusal-preservation loss on $\{\star\}$ alone, and a task loss.

I.3 Two attack families

Fingerprint-space evasion moves $\phi(a_\star)$ toward the region the model scores as safe while holding $r(S)$ fixed. It exploits Theorem 2(ii) from the attacker's side: if two hyperedges share a feature vector, the model must assign them the same score, so an author who can match the fingerprint of a benign configuration inherits its score. The natural defence is a feature map that is expensive to invert, and we do not claim to have one; what we claim is that the failure mode is a property of the feature map rather than of the head, which tells a defender where to look.

Order-shifting evasion moves the hazard from order 3, which the model represents, to order 4, which it does not. Under Proposition 1 this succeeds to the extent that the decay assumption fails, so an author who can concentrate $a(T)$ at $|T| = d + 1$ defeats a degree-d model by construction. The defence is not a better head at fixed d, it is a larger d, whose cost Table 10 gives, or a detector for the order-shifting itself. This is the cleaner of the two attacks and the one we expect to be more effective.

I.4 What E7 reports

Four quantities, none of them aggregated into a single robustness score. (i) The recall of SafeCompose on evasive positives at the operating point calibrated on non-evasive data. (ii) The utility cost to the author, $u_{\min}$ versus achieved $u(\{\star\})$, since an attack that destroys the adapter's usefulness does not get adopted and is not a live threat. (iii) The number of queries to $\hat{R}_\theta$ the attack needs, which distinguishes a white-box optimisation from something a hub operator could rate-limit. (iv) Whether $\mathrm{FNR}_{\mathrm{acc}}$ measured on the evasive candidates exceeds α. Theorem 3 does not cover this setting — an adaptive author chooses the candidate distribution, precisely what Assumptions 2 and 3 hold fixed — so the guarantee is inapplicable here by construction. We therefore do *not* predict the outcome in advance; E7 measures how much an adaptive author actually gains, and reporting that measured loss is more useful than either claiming a guarantee that does not apply or asserting an attack must succeed before running it.

I.5 What would count as an adequate defence, and why we do not claim one

An adequate defence would either make Eq. (11) infeasible at acceptable utility, or make the certificate hold under a bounded-shift model of the adaptive candidate distribution. The second route is the one the literature offers a handle on: beyond-exchangeability analysis (Barber et al., 2023) degrades coverage by a term controlled by the total-variation distance between calibration and deployment distributions, so a certificate could be stated conditional on a bound for that distance. We do not state one, because we have no way to estimate the distance an adaptive author can achieve without first running E7. That is the correct order of operations, and it is why E7 is an experiment in this paper rather than a theorem.

J How large the calibration split has to be

Theorem 3 says the certificate is valid; it does not say it is attainable. Attainability is a question about m_{cal}, and the answer turns out to depend sharply on which concentration bound supplies the p-value. This appendix works it out, because the conclusion changes the protocol.

J.1 THE SIZING RULE

Write $\gamma = \mathbb{P}[S \in \mathcal{A}(\lambda)]$ for coverage at the operating point and $\alpha_0 = \mathrm{FNR}_{\mathrm{acc}}(\lambda)$ for the true selective false-negative rate, with $\alpha_0 < \alpha$. From Eq. (7), L_λ takes exactly three values: α when the candidate is not accepted, 0 when it is accepted and safe, and 1 when it is accepted and unsafe. Hence

$$\bar{L}_\lambda \;=\; \mathbb{E}[L_\lambda] \;=\; (1-\gamma)\alpha + \gamma\alpha_0, \qquad \alpha - \bar{L}_\lambda \;=\; \gamma\,(\alpha - \alpha_0), \tag{12}$$

so the margin the test must detect is the product of coverage and the gap between the target and the truth. Halving coverage halves the margin, which is the sense in which a conservative accept region is harder to certify rather than easier.

Corollary 1 (Hoeffding sizing)**.** *Under Assumption 3, if the empirical mean equals its expectation, the Hoeffding component of the p-value in Theorem 3 satisfies $p_\lambda \le \delta$ if and only if*

$$m \;\ge\; \frac{\log(1/\delta)}{2\,\gamma^2(\alpha - \alpha_0)^2}. \tag{13}$$

Proof. The Hoeffding p-value is $\exp(-2m(\alpha - \hat{L}_\lambda)_+^2)$. Substituting $\hat{L}_\lambda = \bar{L}_\lambda$ and Eq. (12) gives $\exp(-2m\gamma^2(\alpha-\alpha_0)^2)$, and requiring this to be at most δ rearranges to Eq. (13). □

Equation (13) is a feasibility statement, not a guarantee for a particular draw: it assumes the empirical mean lands exactly on its expectation, so a real deployment needs headroom above it. Its value is that the scaling is explicit, $m = \Theta(\gamma^{-2}(\alpha-\alpha_0)^{-2})$, and the scaling is bad. Tightening the target from $\alpha = 0.05$ to $\alpha = 0.01$ at fixed relative gap multiplies the requirement by twenty-five.

J.2 WHY THE BOUND ONE USES MATTERS MORE THAN THE CONSTANT

Hoeffding treats L_λ as an arbitrary variable in $[0,1]$ and therefore prices in a variance it does not have: by Eq. (12) the loss is α or 0 almost always and 1 only on the rare accepted positives. The Bentkus inequality (Bentkus, 2004) compares against a binomial tail instead and is far tighter in exactly this regime, which is why Theorem 3 uses the minimum of the two, following Bates et al. (2021, Thm. 3). Table 5 gives the exact minimal m for both, computed in `scripts/calibration_size.py`. One subtlety governs how: the Hoeffding–Bentkus p-value is *not* monotone in m, because the Bentkus term evaluates the binomial tail at $\lceil m\hat{L}_\lambda \rceil$ and that ceiling is a step function of m. The first m whose p-value falls below δ can therefore be followed by larger m whose p-value rises back above it: at $\alpha = 0.05$, $\alpha_0 = 0.01$, $\gamma = 0.8$ the first crossing is $m = 211$, sizes 223–239 fail again, and the smallest size from which the certificate is available for every larger m is 240. A binary search, or a scan that stops at the first crossing, returns a size at which the certificate is not in fact attainable. The column reports the *stabilised* minimum: the smallest m from which the p-value stays at or below δ.

Table 5: Minimal calibration size at $\delta = 0.05$ under the noiseless sizing rule, computed exactly. γ is coverage, α_0 the true selective false-negative rate at the operating point. These are arithmetic consequences of Eq. (12) and the two p-value formulas, not measurements.

α	α_0	γ	$\bar{L}_\lambda$	m (Hoeffding)	m (Hoeffding–Bentkus)
0.05	0.010	0.80	0.0180	1,463	240
0.05	0.010	0.50	0.0300	3,745	603
0.05	0.025	0.80	0.0300	3,745	603
0.05	0.025	0.50	0.0375	9,587	1,469
0.01	0.002	0.80	0.0036	36,569	1,213
0.01	0.002	0.50	0.0060	93,617	3,037
0.01	0.005	0.80	0.0060	93,617	3,037
0.01	0.005	0.50	0.0075	239,659	7,605

The ratio between the two columns is about 6 at $\alpha = 0.05$ and about 30 at $\alpha = 0.01$. At the strict target the Hoeffding requirement is tens of thousands of executed compositions, which would consume the entire budget and defeat the purpose of screening; the Bentkus requirement is in the

low thousands, which is affordable. The betting bounds of Waudby-Smith & Ramdas (2024), which Bates et al. (2021, Thm. 4) state in the same form, are tighter again and are the obvious next step if the numbers in Table 5 prove binding in practice.

J.3 Consequence for the protocol

The value $m_{\text{cal}} = 1000$ that Appendix G states as a default is adequate for $\alpha = 0.05$ at coverage 0.8 with a comfortable true rate, and is *not* adequate for $\alpha = 0.01$ in any row of Table 5, nor for $\alpha = 0.05$ once coverage drops to 0.5 with $\alpha_0 = \alpha/2$. The protocol therefore sets m_{cal} from Table 5 at the intended operating point rather than fixing it in advance, with a floor of 1000 and a ceiling set by the budget, and reports the value used alongside every certificate. The protocol carries this change before any run, and it illustrates why the sizing calculation belongs in the paper: a plausible-looking default was wrong by an order of magnitude at the tighter target.

J.4 What this does not cover

Three things. The rule assumes the empirical mean equals its expectation, so it sizes feasibility and not power; a proper power calculation would fix a probability of successfully certifying and is straightforward but longer. It assumes a single λ; the family-wise correction over the grid Λ costs a factor of at most $|\Lambda|$ under Bonferroni and nothing under fixed-sequence testing for any pre-specified order (a prefix structure affects only power, not validity), which is why Algorithm 2 uses the latter, subject to the grid-placement caveat of Appendix D: the sequence must not begin at a coverage no amount of clean labels could certify. And it assumes Assumption 3: under the reuse variant of Proposition 2(iii) every entry in Table 5 inflates by $1/\pi_{\min}^2$, which is the concrete price of skipping the reserved split.

K E6 results, baselines, and analytic panels

The measured results are in Section 6.1; the full-pipeline run (E6) is in Table 6, its baseline comparison in Table 7 and its feature ablation in Table 8; cross-base transfer (E5) remains future work. The column headings, units and comparands of the panels below are fixed in advance so that the reported version cannot quietly change what is being shown. The plots in Figure 6 are analytic curves evaluated from formulas proved in Appendix B; Figure 7 plots the measured E6 acquisition and calibration.

Table 6: E6 full-pipeline run on the expanded constructed benchmark (1177 compositions over the same 4 backdoor $+ 4$ benign adapters at 8 merge-dose levels, 461 (39%) positive on the backdoor dimension; hazard $=$ marker rate over 240 held-out probes, τ=0.30). The hazard is *non-monotone* in dose — positive rate 0.09 at c=1 (full-weight merge destroys the backdoor) rising to ~ 0.60 at $c \in [0.3, 0.5]$ — so the coefficient-aware score adds a *quadratic* dose term. AUROC is out-of-fold (5-fold logistic, 5 seeds); the Definition-1 eligible-population and held-out-pair results, and the Learn-then-Test accept region (i.i.d. with-replacement calibration), are in the "Stratification" paragraph below. `pipeline_results_v2_n1177.json`, `round6_results.json`.

Feature tier	out-of-fold AUROC
node (fingerprint only)	0.81 ± 0.00
dose-only (linear)	0.82 ± 0.01
node + dose (linear)	0.89 ± 0.00
node + dose^2 (coefficient-aware)	$\mathbf{0.97 \pm 0.00}$
+ pairwise	0.98 ± 0.00 (+0.002)

Acquisition (sequential, charging seed+acquired labels), recall at label budget $0.30/0.45/0.60$: risk-ranked $0.54/0.86/0.98$ vs. random $0.30/0.45/0.60$. Discrimination on the Definition-1 eligible population and the Learn-then-Test certificate are reported in the "Stratification" paragraph below.

Stratification, generalisation, and scope. We measured each backdoor adapter's singleton canary rate at every dose level (`singleton_rates.json`): all four fire on $\leq 5\%$ of probes at $c \leq 0.4$ and saturate by $c \geq 0.6$. To evaluate the Definition-1 phenomenon — *safe* constituents composing into

an *unsafe* model — we apply the singleton-safety screen to **both** classes, keeping only configurations in which every constituent is individually safe at that exact coefficient. This *eligible population* has 462 configurations: 199 emergent-unsafe (all at $c \leq 0.4$) and 263 safe (backdoor-carrying at $c \leq 0.4$ that do not cross, plus benign-only). We do *not* report an asymmetric filter that would keep all 716 negatives of the mixed bank, since the high-dose negatives ($c \geq 0.5$) contain already-unsafe singletons and lie outside Definition 1. On the eligible population the coefficient-aware node+dose^2 score reaches out-of-fold AUROC $\mathbf{0.993}$ (configuration-random) and $\mathbf{0.994}$ under a *subset-grouped* split (no adapter subset shared across train/test), against node-only 0.87; and it transfers to a *completely held-out backdoor pair* (train on one independently-trained canary pair, test on the other) at AUROC 0.93–0.95. So the compositional-emergence phenomenon is real, screenable, and generalises to unseen subsets and unseen risky pairs — not an artefact of separating a low-dose band from an ineligible high-dose one. On the full mixed benchmark AUROC is 0.975 (configuration-random), 0.965 subset-grouped, and the frozen train-only (30%) deployed score reaches 0.977 on held-out test. A limitation: both risky pairs use the same harmless-canary/backdoor recipe, so the held-out-pair result shows transfer across independently trained instances of *one* mechanism family, not across a qualitatively different emergent-unsafety mechanism; establishing the latter needs a second constructed mechanism and is left to future work.

Calibration protocol. The bank is split into three *disjoint* pools — 353 train, 647 calibration, 177 test (30/55/15) — with the score fit on the train pool only. The threshold grid Λ is fixed from the *train* scores (independent of the calibration draw), and calibration observations are drawn *with replacement* from the reserved calibration pool, so the draws are i.i.d. (Assumption 3) and the certificate covers that pool's empirical configuration law; the statistical unit is $(S, c, \oplus)$, not S. The 647 pool configurations are each executed once and their labels cached; the with-replacement draw is a *statistical* resample of those fixed labels, so once the pool is labelled its accepted-set FNR is directly computable and Learn-then-Test here demonstrates the finite-sample certificate mechanics rather than a deployment necessity at this scale. This is a proof-of-concept scale: m_{cal}=647 is below the deployment floor of 1000 in Appendix G, and we report $\alpha \in \{0.05, 0.10\}$ (the protocol targets $\{0.01, 0.05\}$); both choices are fixed before examining outcomes and logged as an exploratory small-scale run. *Attainability* (per seed, one authoritative ledger — Table 6, Figure 7, `round7_results.json`): at α=0.10 a non-empty accept region certifies in $\mathbf{8/8}$ seeds on the full configuration distribution (coverage 0.55, held-out FNR 0.045) and $\mathbf{6/8}$ on the eligible distribution (coverage 0.51 conditional on certifying, 0.38 counting the 2 empty draws as zero; FNR 0.026); at α=0.05 it certifies in $1/8$ (full) and $0/8$ (eligible) — borderline, consistent with the sample-size floor of Proposition 3. Empty acceptance is a valid (abstaining) outcome, not a violation. E6 executes $\sim 85\%$ of the bank; at $1177 < 2000$ candidates it sits below the regime where a from-weights screen beats exhaustive testing (Section 5.4) — indeed a direct 20-probe composition test reaches AUROC ≈ 1.0 here but must *execute* every candidate, whereas the screener predicts from weights alone. E6 therefore validates the screener's discrimination and calibration *mechanics*, not execution savings at deployment scale. Finally, the measured E6 predictor is the coefficient-aware node+dose^2 logistic; the full interaction hypergraph of Section 5 is a broader design not separately validated here. The learned interaction baselines confirm the reading: given the explicit dose scalars a set transformer matches it (0.984) and a pairwise MLP nearly does (0.96), while without them they fall to 0.87/0.91 and interaction features add $+0.002$ (Table 7). The advantage is thus the explicit nonlinear-dose *representation*, not interaction-specific architecture.

The dose-aware SAFECOMPOSE score dominates every from-weights baseline. Weight-norm ranks *below* random (0.35): a larger merge update is not more dangerous, because a full-weight merge tends to *destroy* the backdoor (Section 6.1), so norm and hazard are anti-correlated over this pool. Pairwise cosine is near chance. Crucially, when the learned models are given the *same explicit dose scalars* (c-sum, c-sum^2, c-mean, c-mean^2), they close the gap — a set transformer reaches AUROC 0.984 and a pairwise MLP 0.96, versus 0.87 and 0.91 without them (`matched_dose_baselines.json`) — so the decisive ingredient is the *explicit nonlinear dose feature*, not interaction-specific architecture: given fingerprints and dose, model class barely matters, and dose scalars alone (no fingerprints) reach only 0.81. Separately, a direct spot-test that *executes* each composition on 20 trigger probes attains AUROC ≈ 1.0 — the strongest option when execution is affordable, which is exactly the small-candidate regime where the from-weights screener is not yet economically motivated.

Table 7: Baseline comparison *measured* on the E6 benchmark (1177 compositions, 461 positive; hazard = canary rate $> \tau$). AUROC is out-of-fold for learned scores and direct for the unsupervised rankers; Recall@B executes the top B=0.02$|\mathcal{C}|$=23 candidates, Recall@0.3 the top 30%. Learned rows (pairwise MLP, set transformer) train on the same per-member (fingerprint, coefficient) inputs (`learned_baselines.py`); all rows here screen *from weights* without executing the composition. `baselines_e6.py`.

#	Method	AUROC	Recall@B	Recall@0.3
1	independent risk (node)	0.81	0.03	0.51
2	weight-norm threshold	0.35	0.00	0.08
3	pairwise cosine	0.49	0.03	0.30
4	pairwise MLP (learned)	0.91	0.04	0.62
5	set transformer (learned)	0.87	0.04	0.58
6	random	0.50	0.02	0.30
7	exhaustive oracle	1.00	0.05	0.77
8	**SAFECOMPOSE (node+dose^2)**	**0.98**	**0.05**	**0.74**

Table 8: Feature ablation of the E6 coefficient-aware score (out-of-fold AUROC on the 1177-composition benchmark, 5 seeds; full model = node+dose^2, AUROC 0.975 ± 0.001). Component ablations of the full hypergraph/activation-space method (activation features, hypergraph vs. set encoder, alignment/ refusal projections, probe-shift, quantisation, reused-label calibration) require that method's run and remain protocol.

Ablation (from full node+dose^2)	AUROC	ΔAUROC
− quadratic dose (node + linear dose)	0.888	−0.087
− all dose (node fingerprints only)	0.806	−0.169
− fingerprints (dose features only)	0.820	−0.155
+ pairwise features	0.977	+0.002

The single decisive component is the *quadratic* dose term; removing it costs 0.09 AUROC and removing dose entirely costs 0.17, while adding pairwise interaction features changes nothing (+0.002) — the same node-versus-interaction pattern as E1–E4, now with the non-monotone dose response made explicit.

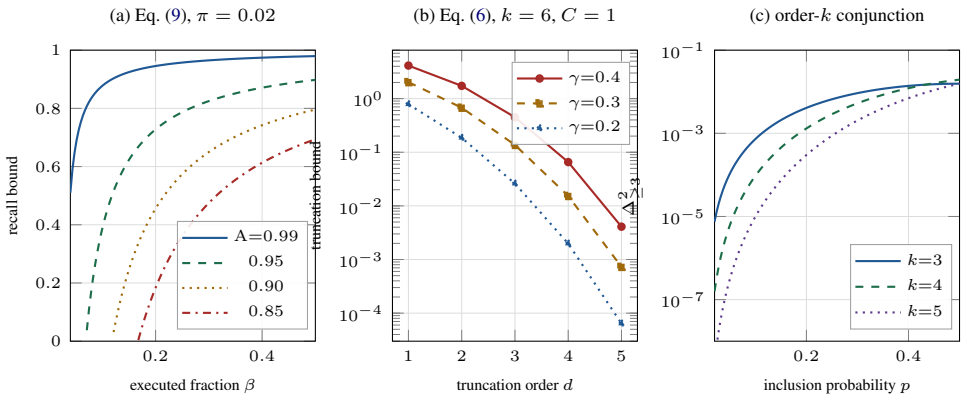


Figure 6: Analytic panels, not experimental results. (a) The recall floor of Proposition 3 as a function of the executed fraction, at four values of the ranking quality A and prevalence $\pi = 0.02$; the curve says what A a target recall demands before any model is trained. (b) The truncation bound of Proposition 1 at $k = 6$ and $C = 1$ for three decay rates, on a logarithmic axis. (c) The higher-order Fourier mass $\Delta^2_{\geq 3}$ of a pure order-k conjunction as a function of the inclusion probability (Remark 2); the decay in k is why the estimation problem is harder than the representation problem.

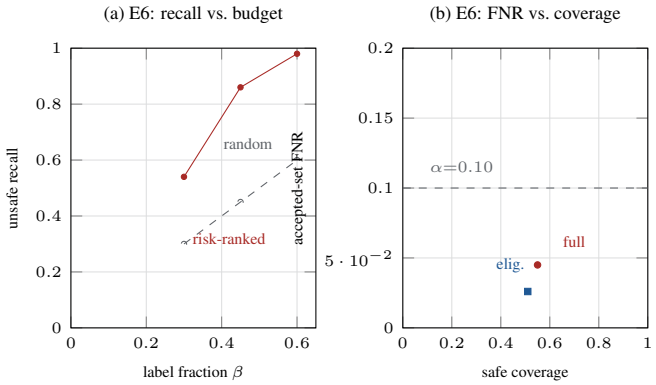


Figure 7: Measured E6 (1177-composition benchmark, coefficient-aware node+dose^2 score). (a) Sequential risk-ranked acquisition recall versus label budget, against the random diagonal. (b) The Learn-then-Test α=0.10 operating points on the full configuration distribution (coverage 0.55, FNR 0.045) and the eligible distribution (coverage 0.51, FNR 0.026), with the target line.

K.1 Reporting rules

Every number arrives with the seed spread, the executed fraction that produced it, and the split it came from. The exhaustive oracle is computed on a sub-pool small enough to enumerate and is left undefined on the full pool rather than extrapolated. Any metric that depends on the judge is reported under both judges. No aggregate is reported over the four positive-class generators of Appendix F; they are reported separately, because a system that only catches merge-activated backdoors and a system that only catches collusion are different systems with the same average.

L Extended related work

Section 2 is compressed to fit the page limit. This appendix restores the detail, and in particular records for each nearby line of work the exact reason it does not already solve the problem.

L.1 Parameter-efficient adaptation and its composition

LoRA (Hu et al., 2022) and the adapter family before it (Houlsby et al., 2019) made specialisation cheap; quantised variants (Dettmers et al., 2023) made it cheap on one device, which is the economic precondition for a pool large enough to have a combinatorial candidate space. Pfeiffer et al. (2020) built the distribution channel. Shuttleworth et al. (2025) report that low-rank updates introduce directions absent from the base model's spectrum rather than merely reweighting existing ones, which is the reason our weight-space fingerprint records spectral shape rather than magnitude.

On the composition side, weight averaging (Wortsman et al., 2022) and Fisher weighting (Matena & Raffel, 2022) treat merging as an estimation problem; task arithmetic (Ilharco et al., 2023) treats task-specific updates as vectors that can be added and subtracted; TIES (Yadav et al., 2023) and DARE (Yu et al., 2024) address the interference that arises when several such vectors are combined. Ortiz-Jimenez et al. (2023) explain when task arithmetic works at all, through weight disentanglement in the tangent space; read through Lemma 1, their condition is the statement that the higher-order coefficients of the relevant functional vanish, and our $\Delta^2_{\geq 3}$ measures how badly it fails for the hazard rather than for the loss. Routing approaches (Huang et al., 2024a; Wu et al., 2024) compose without a single merged weight vector, which is why the operator family $\mathcal{O}$ includes routing and why E4 tests transfer across operators. None of this literature measures safety; that is not a criticism of it, it is the gap.

L.2 Where safety lives, and what breaks it

The result that fine-tuning removes alignment cheaply (Qi et al., 2024) is the origin of the current concern, and Lermen et al. (2023) give the LoRA-specific version at scale: refusal rates driven to about one percent on a 70B chat model for a few hundred dollars. Qi et al. (2025) argue alignment is shallow in the depth of the output-token distribution, a claim about tokens and not about weights; we cite it for what it says and not for the weight-subspace claim it is often enlisted for. For the weight side, Wei et al. (2024) find safety-critical regions covering roughly three percent of parameters and two and a half percent at the rank level and, importantly, that freezing them is not sufficient; Li et al. (2025) locate a contiguous band of middle layers that discriminates malicious from benign queries. Peng et al. (2024) give the geometric picture our whole premise leans on: safety is preserved under random weight perturbation inside a basin and then collapses in a step, unlike capability, which degrades smoothly. A step function of the displacement is exactly a function with large higher-order structure once the displacement is a sum of several adapter updates.

On the activation side, Arditi et al. (2024) show refusal is mediated by a single residual stream direction across thirteen models, which is what makes a one-dimensional activation feature meaningful rather than arbitrary, and Zou et al. (2023a) supply the general read-and-steer machinery. Repair by weight arithmetic (Bhardwaj et al., 2024) is evidence that the safe alternative search of Section 5.4 is searching a non-empty space, at least for single models.

L.3 Safety of composed models, and why the gap remains

Four papers study safety of merged models directly. Hammoud et al. (2024) show merging propagates misalignment and propose data-aware merging that treats alignment as a skill to maximise; the framing is that one bad ingredient spoils the merge, which is propagation of an existing defect rather than emergence from safe parts. Ma et al. (2025) attribute merge-induced safety collapse to neuron misidentification and cross-task interference and isolate the conflicting parameters; it is the closest published mechanism for our non-additivity, but it merges full task-expert models and is a mitigation rather than a predictor. Arora et al. (2024) report the opposite polarity, merging removing an implanted backdoor by around three quarters of its attack success rate, which is the cleanest available evidence that a safety property measured on one model does not transfer to the composition in either direction. Ding (2026) is the one that studies our exact object: LoRA pairs benign in isolation whose linear composition suppresses refusal on three open-weight families, with the observation that unit-centric defences have a combinatorial blind spot and exhaustive verification is intractable. It is a single-author preprint without peer review, its adapters are deliberately co-designed to collude, and it supplies no predictor, no calibration and no acquisition rule. We treat it as establishing that the hazard exists in the adversarial case and as leaving open both whether it occurs accidentally, which is E1, and what to do about it, which is the rest of the paper.

The merging-backdoor line (Zhang et al., 2024; Yin et al., 2024; Liu et al., 2025; Pawlak et al., 2025) and its defences (Yang et al., 2025) all assume a deliberately backdoored contributor. That setting is threat setting (ii) and is reported for context; it is easier than ours precisely because the node carries signal. Supply-chain capabilities that make the malicious node plausible are documented separately (Dong et al., 2025; Wan et al., 2023; Hubinger et al., 2024).

L.4 Screening artifacts rather than behaviours

Xu et al. (2021) train a meta-classifier over a population of shadow models and detect trojans from query access with generalisation to unseen attack types; it is the methodological ancestor of learning a model over model artifacts, and it classifies one model at a time. Puertolas Merenciano et al. (2026) do the low-rank version without any execution, extracting spectral statistics from LoRA updates and separating benign from poisoned adapters on held-out data; it is the closest existing thing to our node fingerprint, and it is a per-adapter test. Hsu et al. (2024) project an adapter's update onto the alignment subspace and patch it, which is both a defence and, for us, a feature; notably Ding (2026) report that this weight-space test does not flag their colluding pairs, which is a concrete instance of Theorem 1's message. Composite backdoors that fire only when several trigger components co-occur (Lin et al., 2020; Huang et al., 2024b) are the structural analogue of our failure mode inside a single model, and Lin et al. (2020) explicitly report that scanners built for atomic triggers fail on them.

L.5 Set functions, hypergraphs, and interaction indices

Deep Sets (Zaheer et al., 2017) gives the sum-decomposition result whose hypotheses Theorem 2(iii) restates carefully; Wagstaff et al. (2019) supply the width requirement that turns it into a design constraint. Set Transformer (Lee et al., 2019) is the attention-pooling variant and our baseline 6. Hypergraph neural networks (Feng et al., 2019) and the multiset framework of Chien et al. (2022) give the deeper alternative in Appendix D. On the expressivity side, Xu et al. (2019) and Morris et al. (2019) establish the Weisfeiler–Leman ceiling for message passing and the strict gain from higher-order variants; our Theorem 1 is the set-function analogue, proved directly rather than by a colour-refinement argument. The interaction algebra is Rota (1964) for the Mobius inversion, Grabisch & Roubens (1999) for the axiomatic treatment of interaction indices, and Tsai et al. (2023) for the faithful Shapley interaction index and its explicit Boolean-lattice formulas. Amrollahi et al. (2019) give the sample-complexity picture for learning Fourier-sparse set functions, which is the right yardstick for Remark 2. Generalising to adapters not seen during training is the inductive setting of Hamilton et al. (2017).

L.6 Budgeted evaluation and distribution-free decisions

Active testing chooses which points to label in order to estimate a metric efficiently, which is our problem; active learning chooses points to improve a model, which is not. Sawade et al. (2010) give the optimal sampling distribution and debiasing weights for risk estimation and report matching passive evaluation with roughly a third of the labels; Kossen et al. (2021) name the framework and show that testing acquisitions differ from learning acquisitions. Kirsch et al. (2019) supply diverse batch acquisition, and Baptista & Poloczek (2018) the combinatorial-domain machinery. Automated red teaming (Perez et al., 2022; Zou et al., 2023b) searches inputs against a fixed model; our search is over models against a fixed input suite, which is why their acquisition strategies do not transfer directly.

For the decision layer, conformal prediction (Vovk et al., 2022; Shafer & Vovk, 2008; Lei et al., 2018) gives marginal coverage under exchangeability. Beyond coverage, RCPS (Bates et al., 2021) controls a bounded monotone risk with high probability over the calibration draw; Learn-then-Test (Angelopoulos et al., 2025) removes the monotonicity requirement by reducing risk control to multiple testing, which is why we use it; conformal risk control (Angelopoulos et al., 2024) controls a monotone bounded loss in expectation over the joint calibration and test draw, which is a different object and, for a screener calibrated once, the wrong one. Selective prediction (Geifman & El-Yaniv, 2017) supplies the abstention framing of the three-way decision. Under shift, weighted conformal methods (Tibshirani et al., 2019) and the beyond-exchangeability analysis of Barber et al. (2023) bound the damage when the shift is bounded, and Proposition 2(iii) is the design-based special case where the shift is one the screener created and therefore knows. Calibration is measured with the standard estimator (Guo et al., 2017) and its debiased counterpart (Kumar et al., 2019), reported together because the naive estimate is biased.

M Cost model

Section 5.4 states the headline counts. This appendix gives the terms behind them, so that the comparison against exhaustive testing is auditable rather than rhetorical. Every figure is arithmetic from the counting argument; none is a measurement.

M.1 Candidate space

From Eq. (2), $|\mathcal{C}| = \sum_{k=2}^{K} \binom{n}{k}$.

The first row is included because it is the honest boundary: below roughly 2000 candidates there is nothing to screen for, and the exhaustive oracle is the right method. A screening system is only worth building where the candidate space is large, and the table says where that begins.

Table 9: Candidate counts. The last column is the fraction of candidates a screener executes when $B + m_{\text{cal}} = 2000$.

n	K	pairs	triples	$\lvert\mathcal{C}\rvert$	executed fraction
20	3	190	1,140	1,330	$> 100\%$ (exhaustive is cheaper)
40	3	780	9,880	10,660	18.8%
40	4	780	9,880	102,050	2.0%
100	3	4,950	161,700	166,650	1.2%
100	4	4,950	161,700	4,087,875	0.05%

Table 10: Cost of each stage. m is the probe-set size, F the cost of one forward pass, M the cost of one merge, E the cost of one full suite run, and $|\mathcal{T}| \gg m$.

Stage	Count	Unit	Model executions
1. fingerprints	n adapter reads + nm passes	F	nm
2. hyperedge features	$O(\lvert\mathcal{C}\rvert\, d^2)$ vector ops	—	0
2. optional pair probe	$\binom{n}{2} \cdot 16$ passes	F	optional
3. scoring	$O(\lvert\mathcal{C}\rvert)$	—	0
3. acquisition	B compositions	$M + E$	B
4. calibration	m_{cal} compositions	$M + E$	m_{cal}
5. repair	$\leq w \cdot 5$ per rejection	—	0 until re-submitted
Exhaustive oracle	$\lvert\mathcal{C}\rvert$ compositions	$M + E$	$\lvert\mathcal{C}\rvert$

M.2 Where the work goes

The ratio of interest is therefore $(B + m_{\text{cal}})/|\mathcal{C}|$ in units of $(M + E)$, plus a fixed nm forward passes that do not scale with the candidate space. The fixed term dominates only in the small-n regime where screening is not worth doing anyway. The optional pair probe is the one term that grows quadratically in n while still touching the model; the no-execution variant drops it, and the ablation in Appendix K reports what that costs.

M.3 What this count leaves out

Four things a real deployment pays that the table does not: loading and unloading adapter weights between merges, which is I/O rather than compute but is not free at 10^5 candidates; retraining θ once per acquisition round, which is small but not zero; the human cost of adjudicating disagreement between the two judges; and the cost of constructing the probe set and the suite in the first place, which is paid once per deployment and is not amortised across pools. The claimed saving is therefore a saving in model executions, which is the dominant term, and is stated as such rather than as a wall-clock speed-up.

N Deployment: what a hub operator would have to record

The output of SafeCompose is a decision plus a certificate, and a certificate is only meaningful if the conditions under which it was issued travel with it. This appendix states what those conditions are, in the form of what an adapter hub would have to store, because several of them are consequences of the theory rather than engineering preferences.

What the certificate does and does not say. Eq. (8) bounds the false-negative rate *among accepted compositions*, at the stated α and δ, for candidates drawn from the distribution the calibration split came from, and with respect to the safety suite that produced the labels. It is not a statement about a different pool, a different operator family, a different probe distribution, an adaptive author, or a hazard the suite does not test. Every reported certificate should therefore carry $(\alpha, \delta, m_{\text{cal}})$ and the identity of the suite alongside it, and a certificate quoted without them is not interpretable.

The certificate artifact. Each screening run emits, for every candidate, the decision $D(S)$, the score $\hat{R}(S)$, and a reference to a run record containing: the base checkpoint hash; the pool manifest with an adapter hash each; the operator $\oplus$ and the coefficient grid; the probe-set version; the suite and judge versions; the targets (α, δ); the realised m_{cal} and the sizing row of Table 5 it was drawn from; the validated set $\hat{\Lambda}$ and the deployed $\lambda^\star$; and the empirical coverage. A decision without that record is not a certified decision, because Theorem 3 is a statement about a particular calibration draw against a particular candidate distribution.

When the certificate expires. Three events invalidate it, and each follows from an assumption rather than from caution. Adding an adapter to the pool changes $\mathcal{C}$ and therefore the candidate distribution, so Assumption 3 no longer holds for the stored split and a fresh uniform draw is required; note that this is cheap relative to rescoring, since the fingerprints of existing adapters are cached and only the new adapter needs stage 1. Changing the operator family or the coefficient grid changes r itself, since Section 3 defines the hazard set function at a fixed $(c, \oplus)$. And changing the suite or a judge version changes the labels, hence Y, hence everything downstream. A hub that silently upgrades its safety classifier has silently invalidated every certificate it has issued, and the run record is what makes that detectable.

Rate limiting the score. Threat setting (iv) assumes white-box access to ϕ and θ, which is the right assumption for an open release. A hub that exposes $\hat{R}$ as a queryable service faces a weaker adversary, and Appendix I reports query counts precisely so that a hub can decide whether rate limiting closes the gap. We do not claim it does; we claim the number is worth measuring before assuming either way.

What the end user is told. An accept should be reported with its α and δ and with the coverage of the region it fell in, not as a bare assurance. A test should be reported as unresolved rather than as a soft accept, because the whole design of the three-way policy rests on the test branch absorbing the candidates the model cannot rank confidently. A reject should carry the attributed interaction terms from Eq. (10) and the alternatives from Algorithm 3, since a rejection with no explanation and no alternative is the outcome most likely to be ignored. And every one of the three should carry the statement that the certificate covers the calibrated candidate distribution and not an adaptive author.

Failure modes an operator should watch. Coverage drifting downward at fixed α usually means the score has degraded relative to the pool, most often because the pool has grown into regions the model was not trained on; that is E2's condition arriving in production. Coverage holding while the executed positives in the test band rise means the model is still ranking well but the base rate has shifted, which changes π in Proposition 3 and therefore the recall a given budget buys. Empirical coverage on the reserved split falling outside its nominal range is the signal that Assumption 3 has broken, and the first thing to check is whether the split was drawn independently of acquisition, which is exactly the failure Proposition 2(i) constructs.

O Reproducibility details

What exists today. The formal results and their proofs (Appendix B), the fingerprint schema feature by feature (Appendix D), the benchmark construction (Appendix F), the protocol including scoring rules, thresholds, budgets, seeds and the three pre-registered decision rules (Appendix G), and the cost model (Appendix M). The measurements of Section 6.1 were run, and the anonymised repository (`https://anonymous.4open.science/r/SafeCompose-F316`) holds their records under `outputs/`: the adapter screen and manifest, the fingerprints, the executed compositions and their hazards for E1–E4 under both classifiers, the node-versus-interaction baseline with its bootstrap, the E2 control and dose-response tables with the held-out probe set, and the E6 benchmark with every pipeline and baseline result, each with the script that produced it. The analytic panels of Figure 6 are reproducible from the formulas in the captions, and Table 5 from `scripts/calibration_size.py`.

Screening pipeline (E6), measured. The fitted risk model, the calibrated accepted-set risk, and *risk-ranked* sequential testing are run on the expanded constructed benchmark of 1177 composi-

tions — the active uncertainty/diversity acquisition rule and the repair stage remain unrun — (Section 6.1, `pipeline_results_v2_n1177.json`): out-of-fold AUROC by feature tier (including the coefficient-aware node+dose^2 score the non-monotone dose response requires), the Learn-then-Test accept region on a calibration-heavy clean split with its out-of-sample coverage/FNR over 8 seeds at $\alpha \in \{0.05, 0.10\}$, and risk-ranked versus random sequential-testing recall with a label ledger. The benchmark builder and pipeline runner are released (`build_benchmark_v2.py`, `run_full_pipeline_v2.py`).

What does not exist yet. Repair utility (E8), cross-base-model transfer (E5), the adaptive author (E7), and the same pipeline on the 37-adapter ordinary pool (which has no positives, so its accepted-set risk is trivial); cross-base transfer (E5) remains future work.

Release. The repository contains the SAFECOMPOSE package (fingerprint extractor, risk models, calibration, acquisition and repair code, execution and evaluation harness) with its unit tests, the GPU runners and CPU analysis scripts behind every measured number, the records listed above, and audits of the implementation. Adapter weights and raw generations are not included: the public adapters are identified by repository and revision in the manifest, and the canary construction is specified in Appendix G. For the deliberately risky adapters of the full benchmark design, weights and harmful completions would be withheld; the ethics statement gives the reasoning.

Environment and determinism. Every executed composition is identified by the tuple (W_0 checkpoint hash, adapter hashes, c, $\oplus$, suite version, judge version, seed), and that tuple is stored with the label. Decoding is greedy where the suite permits it and otherwise uses a fixed temperature and seed, so that Assumption 1's determinism holds up to the documented judge variability, whose rate the label-noise condition reports rather than assumes away.

Anonymity. No institution, funder, or thanks section appears anywhere in this PDF. The code and records are linked only through the anonymised mirror above, which serves the repository's files without its history or author metadata.