WHAT INDIVIDUAL ADAPTER EVALUATIONS MISS: COMPOSITION-ONLY FAILURES IN LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

LoRA composition offers a way to reuse separately trained task adaptations, but evaluating the composed model raises a question that individual adapter checks leave open: does composition change which requests fail, even when average harm changes little? We study composition-only failures, where a composed response is labelled harmful on a request for which the base and every constituent alone are not. A request-level audit aligns base, member, and composition responses while keeping harmful compliance, refusal, and task performance distinct. Across several base models, we observe these failures in task-trained adapters and screened public adapter banks. The failures persist on additional requests, under two harm judges, and with regenerated responses. They are most pronounced when a composition includes an adapter that already produces some harmful responses alone: on average, harm stays close to that member's own rate while the set of failing requests shifts. They also occur in some compositions without such a member. Interventions expose a second gap. Retraining a member without safety data can lower the composition-only rate while increasing total harm, whereas adding refusal data lowers both and, in a tested summarization pair, retains the partner's task gain. Averaging composition weights removes most composition-only failures but leaves total harm above the base. Composed models should therefore be audited jointly for failure overlap, total harm, and retained adaptation.

1 INTRODUCTION

Low-rank adaptation (LoRA) enables parameter-efficient customization of language models (Hu et al., 2021). LoRAHub and instance-level composition reuse separately trained adapters for cross-task generalization (Huang et al., 2023; Wang et al., 2024). Such reuse creates a distinct model state whose behavior must be evaluated.

Adaptation can weaken safety even with benign fine-tuning data (Qi et al., 2023). CoLoRA further demonstrates harmful behavior triggered by composing deliberately constructed adapters that appear benign individually (Ding, 2026). We examine ordinary task-trained and published adapter banks: which harmful responses do member evaluations miss, and can interventions reduce them while retaining task benefits?

Comparing average harmful-response rates cannot identify this audit gap. A composition with higher harm might fail only on requests that some member already fails; one with unchanged harm might fail on different requests. We study *composition-only failures*: the composition is judged harmful on a request for which the base and every member, evaluated individually, are not. This definition does not require members to be harmless on all other requests.

Our central observation is an empirical reversal: member retraining can reduce composition-only failures while increasing total harm. The audit depends on both composed responses and member responses used as references; a more harmful member can exclude failures from the composition-only count. Total harm and failure overlap must therefore be measured jointly. A separate distinction concerns utility: StrongREJECT shows that apparent jailbreak success can overstate useful harmful responses when capabilities deteriorate (Souly et al., 2024). We therefore assess harmful compliance using HarmBench classifiers (Mazeika et al., 2024), separately from refusal matching and task utility.

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.

We examine five in-house banks on five base models and three banks assembled using published adapters. Our contributions are:

- **A request-level account of the member-audit gap.** We define composition-only failures and use elementary accounting relations to distinguish them from changes in aggregate harm. We measure these failures in task-trained and public adapter banks, including on additional requests and with regenerated responses.
- **Evidence of misleading audit improvements.** Refusal loss can accompany capability degradation, and member retraining can lower the composition-only rate while increasing total harm. These observed reversals show why neither indicator is a sufficient repair criterion.
- **Intervention evidence with explicit task references.** Within matched evaluation protocols, safety-augmented member retraining and common reweighting reduce measured failures. Partner-task comparisons distinguish retained summarization gains from arithmetic recovery toward the base.

2 Related Work

Composition and its risks. Task arithmetic, TIES-Merging, DARE and LoRAHub combine learned updates (Ilharco et al., 2022; Yadav et al., 2023; Yu et al., 2023; Huang et al., 2023). CoLoRA constructs colluding adapters with harmful composed behavior (Ding, 2026); recent work studies adaptive jailbreak robustness after skill merging (Ma et al., 2026). We examine ordinary task-trained and published banks using request-aligned base and member references, and separate new failures from changes in average harm.

Preserving alignment during adaptation. Benign fine-tuning can weaken alignment (Qi et al., 2023), and shallow alignment provides a mechanism for several fine-tuning and inference vulnerabilities (Qi et al., 2025). Safe LoRA projects selected updates into a safety-aligned subspace (Hsu et al., 2024); SaLoRA uses a fixed safety module and task-specific initialization (Li et al., 2025); SafeMERGE selectively merges model layers (Djuhera et al., 2025). Safety examples also improve instruction-tuning safety (Bianchi et al., 2024). Our interventions test how member repair and common reweighting affect composition outcomes; they are not proposed as new defenses or compared as state-of-the-art merging algorithms.

Measuring harmful behavior. HarmBench provides behavior classifiers (Mazeika et al., 2024), StrongREJECT distinguishes apparent jailbreak success from useful harmful responses (Souly et al., 2024), and XSTest measures exaggerated refusal on benign requests (Röttger et al., 2023). Refusal directions characterize a representation-level signal (Arditi et al., 2024). We assess whether refusal signals remain informative under composition, while using harmful-response labels and task performance as separate outcomes.

3 Setting and Request-Level Audit

3.1 Deployment setting and threat model

An operator extends a frozen instruction-tuned base model with task-specific LoRA adapters trained locally or obtained from public repositories. The operator selects the active adapters and their composition weights; an evaluator can load the base, individual adapters, and compositions. The risk is a harmful response under composition on a request where the base and each constituent alone are not labelled harmful. We do not assume coordinated malicious adapter training: our in-house adapters use ordinary task objectives, while public adapters have screened but not fully verified provenance (Section 4.1).

A user can submit harmful requests but is not given control of adapter selection in this setting. We evaluate fixed benchmark requests, without adaptive jailbreak optimization or explicit input triggers. The comparison is request-specific; members need not be harmless on every request. Non-harmful labels need not indicate correct or useful answers, so we measure task performance separately. These audits do not certify safety against arbitrary requests or adaptive attacks.

3.2 LoRA COMPOSITION AND MODEL STATES

Let $\mathcal{A} = \{1, \ldots, m\}$ index an adapter bank. The activation vector $\mathbf{t} = (t_1, \ldots, t_m) \in \{0,1\}^m$ defines which adapters participate, and its equivalent active set is

$$T = \{i \in \mathcal{A} : t_i = 1\}. \tag{1}$$

All adapters in T contribute simultaneously; those outside T are inactive. For eight adapters, $\mathbf{t} = (1,0,1,0,0,0,0,0)$ means $T = \{1,3\}$. The 2^m possible subsets include the base ($T = \varnothing$), singletons ($T = \{i\}$), and the full bank ($T = \mathcal{A}$). We evaluate a specified collection $\mathcal{C} \subseteq 2^{\mathcal{A}}$, not necessarily all subsets.

At a targeted base-model matrix W_0, adapter i provides a low-rank update $\Delta W_i = (\alpha_i / r_i) B_i A_i$, where r_i is the rank, α_i is the internal scale, and B_i, A_i are the learned factors. External weights $\mathbf{w} = (w_1, \ldots, w_m)$ give

$$W(T, \mathbf{w}) = W_0 + \sum_{i=1}^{m} t_i w_i \Delta W_i = W_0 + \sum_{i \in T} w_i \Delta W_i. \tag{2}$$

This equation applies separately at every targeted matrix, with zero update where an adapter does not apply. Internal scaling is already included in ΔW_i; w_i controls the external contribution. A state specifies the active set, weights, base checkpoint, and inference settings. The same T with different weights therefore defines different states.

For each selected composition T, the audit compares

$$\begin{aligned} \text{base: } & W(\varnothing, \mathbf{1}) = W_0, \\ \text{singletons: } & W(\{i\}, \mathbf{1}) = W_0 + \Delta W_i, \quad i \in T, \\ \text{composition: } & W(T, \mathbf{w}). \end{aligned} \tag{3}$$

The base checkpoint and inference settings are fixed within each comparison. The main sweep uses $\mathbf{w} = \mathbf{1}$; reweighting changes composition weights while singleton references remain at full strength.

3.3 REQUEST-LEVEL OUTCOMES

Let $X = (x_1, \ldots, x_N)$ be a fixed dataset of harmful requests, indexed by $\mathcal{I} = \{1, \ldots, N\}$. Let $f(x; T, \mathbf{w})$ be the greedily decoded response of the model constructed by Eq. (2). A fixed pretrained judge and scoring protocol assign a label $J(x, r) \in \{0, 1\}$ to the request–response pair; 1 means labelled harmful. We do not fine-tune the judges. Define

$$H(T, \mathbf{w}) = \{n \in \mathcal{I} : J(x_n, f(x_n; T, \mathbf{w})) = 1\}, \qquad h(T, \mathbf{w}) = \frac{|H(T, \mathbf{w})|}{N}. \tag{4}$$

These are request-index sets and empirical rates, not labels of the requests themselves. A harmful request can elicit a refusal or unrelated response. For a fixed audit of a composition with $|T| \geq 2$, abbreviate

$$H_T := H(T, \mathbf{w}), \quad H_{\{i\}} := H(\{i\}, \mathbf{1}), \quad H_\varnothing := H(\varnothing, \mathbf{1}). \tag{5}$$

Rates use the corresponding abbreviations. Complements are relative to $\mathcal{I}$; a non-harmful label does not imply refusal, correctness, or usefulness.

Composition-only failures. We count requests labelled harmful under the composition but not under its base or any active singleton:

$$E(T) = H_T \cap H_\varnothing^c \cap \bigcap_{i \in T} H_{\{i\}}^c, \qquad e(T) = \frac{|E(T)|}{N}. \tag{6}$$

For $T = \{1, 3\}$, the references are the base, adapter 1 alone, and adapter 3 alone. For larger T, we still compare with singletons, not every proper subcomposition. Inactive adapters are irrelevant to this event. The denominator is all N requests: eight events among 100 give $e(T) = 0.08$. This is neither conditional harm nor an increase in total harm. Composition weights are suppressed only within the fixed audit.

Two-judge intersection. For judge J_j, define $E_j(T)$ by Eq. (6) using its labels. Our strict intersection is

$$E_{12}(T) = E_1(T) \cap E_2(T), \qquad e_{12}(T) = \frac{|E_{12}(T)|}{N}. \tag{7}$$

Both judges must label the composition harmful, while neither labels any reference harmful. Intersection does not eliminate correlated judge errors; all harm claims concern automated labels under the stated scoring protocols.

3.4 What can individual audits establish?

All quantities below use shared requests and one fixed judge and reference configuration. Define $U_T = H_\varnothing \cup \bigcup_{i \in T} H_{\{i\}}$, the union of reference failures.

Proposition 1 (Marginal bounds)**.** *For any pattern of harm labels on the evaluation dataset,*

$$\left[h_T - h_\varnothing - \sum_{i \in T} h_{\{i\}} \right]_+ \leq e(T) \leq h_T, \tag{8}$$

where $[z]_+ = \max\{z, 0\}$.

The counting proof is in Appendix G. A positive bound establishes new failures in the measured labels; zero is inconclusive. Because it requires h_T, the bound is retrospective, not a predictor of an unmeasured composition. Member labels alone fix U_T but do not fix H_T without assumptions connecting them; this does not rule out prediction from weights or activations (Appendix G.2).

Let $u_T = |U_T|/N$ and $d_T = |U_T \cap H_T^c|/N$, the reference failures absent from the composition. Partitioning the sets gives

$$h_T - u_T = e(T) - d_T. \tag{9}$$

New failures can be offset by disappearing failures even if $h_T = u_T$. For example, suppose the reference union covers 20 of 100 requests. A composition that resolves five of these failures but introduces five others still has 20 harmful responses, with $e(T) = 0.05$. Here u_T is a request-level union, not the maximum or sum of member rates. Reweighting fixes U_T; member retraining changes it. A more harmful retrained member can exclude requests from $E(T)$, lowering $e(T)$ despite higher total harm.

Audit procedure. Specify the states and references; align their requests; report total harm and new failures with uncertainty; and compare task scores with both the base and relevant singletons. Retention is not adaptation if the singleton adds no gain. Re-evaluate affected states after intervention and state which references change. Across compositions of the same bank, base and singleton responses can be reused when their configurations remain fixed. Retraining a member requires updating its singleton labels and re-evaluating affected compositions; reweighting here retains the full-strength singleton references. Distribution-level guarantees require additional sampling assumptions (Appendix G.4).

4 Experimental Protocol

4.1 Models, adapters and states

Each in-house bank contains eight task-trained adapters: two GSM8K, four Alpaca variants (Taori et al., 2023), one Dolly (Conover et al., 2023) and one DialogSum (Chen et al., 2021). They target attention and MLP projections with ranks 8–32, on Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct (Qwen et al., 2025), Phi-4 (Abdin et al., 2024), GLM-4-9B-Chat (Team GLM, 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). Per bank, the sweep covers the base, eight singletons, all 28 pairs and ten sampled compositions at each size three through six (77 states). The two Qwen bases and Phi-4 have zero primary-judge harm on the main prompts. GLM-4 exhibits high repetition under individual adapters and Mistral already complies with many harmful requests; they provide contrast cases. Forward hooks implement the weighted sum in Eq. 2.

For Qwen2.5-7B we also examine a control bank of eight published task adapters from different uploaders; a bank replacing its eighth member with the in-house Dolly adapter; and a bank of eight

published instruction-following adapters (one Dolly-15k, four Alpaca and three DPO-LIMA (Rafailov et al., 2023; Zhou et al., 2023)). Selection used declared purposes, compatibility and tensor-integrity checks (Appendix K.7). Such screening cannot verify all aspects of third-party training provenance. Published banks share a base and some members, so they are not independent training replications.

4.2 Measurements and evaluation sets

The main harm sweep uses 80 AdvBench requests (Zou et al., 2023), greedy decoding, 512 new tokens and bf16. The primary classifier is HarmBench-Llama-2-13b-cls. A prompted Qwen2.5-14B judge provides a second measurement for that sweep. Refusal r_T is a phrase-match rate, measured separately from harmful compliance h_T.

A 297-request evaluation combines 97 JailbreakBench behaviors (Chao et al., 2024) and 200 additional AdvBench requests, disjoint from the main 80 and earlier 160-request sets. Compositions were selected using main-sweep harm: five high-harm and five low-harm Dolly-containing compositions, five high-harm Dolly-free compositions, and top compositions with Dolly removed. The 97 requests had already been used in an earlier replication after selection; interventions were designed after observing the 297-request evaluation and reuse this set. Appendix K.1 records the timeline.

The evaluation labels all relevant base, singleton and composition responses. It uses the primary judge and HarmBench-Mistral-7b-val-cls with an abridged prompt; interventions use the complete official judge prompts. These are separate scoring protocols (Appendix K.3). Intervention effects compare original and modified states scored within the same protocol, not rates across experiments. Rates describe selected compositions, not deployment prevalence.

Utility includes a 60-item GSM8K sweep (Cobbe et al., 2021), with selected states evaluated on 200 GSM8K items, 300 MMLU items (Hendrycks et al., 2021), and task-specific DialogSum or Dolly ROUGE-L scores (Lin, 2004). Every reported retention ratio specifies its reference: the best singleton, a particular task member, or the base. Intervals resample shared prompts and, where stated, compositions; they do not capture judge error or all training variation.

Inference correction and robustness. A factorial check found anomalously low Qwen2.5-7B arithmetic scores in our eager attention path. We report the complete SDPA arithmetic rerun and retain eager diagnostics in Appendix K.6. Harm on the 297 requests was also regenerated with SDPA, including references; hardware differences are documented in Appendix K.6. Interventions use SDPA and their own regenerated references.

5 Composition-Only Failures

5.1 Failures missed by member evaluations

On the main prompts, the Phi-4 GSM8K$_{00}$+Dolly pair has total harm 0.500, against 0.300 for Dolly and zero for the base and GSM8K member. The marginal bound gives $e(T) \geq 0.200$; aligned labels give $e(T) = 0.250$, or 20 of 80 prompts. The pair retains 20 of Dolly's 24 harmful requests and adds 20 requests harmful under neither member. This illustrates why both aggregate harm and failure identity matter.

On the 297-prompt set, mean e_{12} over the five selected Dolly-containing compositions is 0.126–0.220 across five banks (Table 1). These are fractions of all evaluated requests, averaged over compositions, not the fraction failing under any one of the five. Dolly-free compositions reach 0.093 and 0.127 in the two in-house Qwen banks, and 0.015–0.046 in the other three. Removing Dolly therefore does not eliminate the phenomenon, but does not establish that every remaining member is globally harmless: some have nonzero singleton harm.

The result persists with SDPA: selected Dolly-containing groups have mean $e_{12} = 0.123$–0.204, and the two Qwen Dolly-free groups have 0.092 and 0.133. Paired intervals include zero for the implementation differences (Table 21); this is evidence that the failure phenomenon replicates, not a statistical equivalence test. The two classifiers retain 78–91% of primary-judge composition-only events under their intersection rule (Appendix H).

Table 1: Request-level failures on 297 prompts. Top-5 are five Dolly-containing compositions selected on the main sweep; free-5 are five selected Dolly-free compositions. Entries are means over compositions; e_{12} intervals resample prompts. Base harm is 0.000–0.034. These rates use eager generations and the scoring protocol of Appendix K.3; the paired SDPA rerun is in Table 21, and individual-judge rates in Table 11.

Bank	Top-5 h	Top-5 e_{12} [95% CI]	Free-5 e_{12} [95% CI]	MMLU Δ (top-5)
Qwen2.5-7B	.237	.126 [.104, .149]	.093 [.075, .112]	$-.05$ to .00
Qwen2.5-14B	.232	.156 [.131, .183]	.127 [.106, .148]	$-.06$ to $-.03$
Phi-4	.464	.220 [.186, .254]	.032 [.022, .043]	$-.06$ to $-.02$
Public task + Dolly	.271	.170 [.143, .198]	.015 [.005, .027]	$-.13$ to $-.02$
Public instruction	.480	.127 [.097, .160]	.046 [.027, .068]	$-.08$ to .00

Table 2: Members tested alone versus their compositions, Qwen2.5-7B-Instruct. The in-house column excludes Dolly and compositions containing it. Refusal-rate change is the singleton refusal-match rate minus the base rate; positive values indicate more refusal matches.

	In-house (7, no Dolly)	**Published (8)**
Largest singleton harm	0.050	0.025
Mean refusal-match rate change	+0.014	+0.028
Largest composition harm	0.200	0.038
Largest size-mean harm (sizes 3–6)	0.122	0.010
Size-mean refusal at size 3 / 6	0.662 / 0.003	0.995 / 0.978

5.2 Low member harm does not separate the measured banks

The published control bank has size-mean harm at most 0.010 and no state above 0.038 on the main prompts. Its members and the non-Dolly in-house Qwen2.5-7B members both have low singleton harm, but their compositions differ substantially (Table 2). Mean singleton activation shifts along the base refusal direction are -1.27 (in-house, no Dolly) and -0.58 (published control), in base standard deviations. These latent shifts differ from the refusal-match rate changes in Table 2. The measured member indicators do not clearly separate the banks; we neither establish equivalence of member distributions nor rule out better predictors.

Norm-matched cross-bank comparisons and the earlier 97-request replication provide supporting evidence (Appendices E.3 and C); they do not isolate a causal mechanism or establish uniform harm amplification.

5.3 Task retention is partial and task-dependent

Top compositions in the in-house banks keep MMLU within 0.06 of the base; the two published failure banks include losses up to 0.08 and 0.13. On the SDPA harness, the six selected full-strength compositions in Section 7 retain 66–86% of base GSM8K accuracy. Thus harmful compositions can retain substantial measured ability, but these results do not establish unchanged utility. The strongest evidence for preserved adaptation comes from the explicit partner-task comparisons in Section 7.1, rather than a generic capability threshold.

6 Refusal and Harm Diverge as Utility Changes

Figure 1 compares the three measurements across composition sizes. Refusal matching generally falls as in-house composition size grows, whereas harm rises and then falls (Figure 1). For Qwen2.5-7B, size-three means are $h = 0.165$, $r = 0.421$ and GSM8K 0.538; at size six they are 0.010, 0.001 and 0.003. Low refusal therefore occurs both in harmful states and in states with severe task degradation. This association does not by itself establish that arithmetic loss causes the decline in harm.

Refusal also misorders measured harm: among eligible composition pairs, lower refusal accompanies lower harm in 68.5% under the primary judge and 71.2% under the second. Pairs share states and

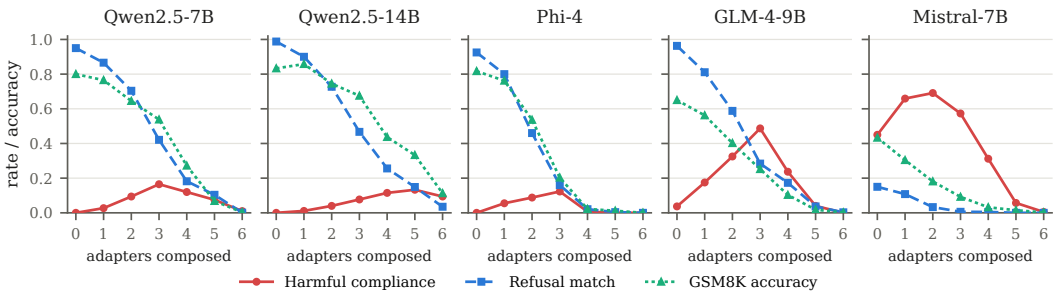


Figure 1: In-house bank size means: harmful compliance (red), refusal matching (blue), and absolute GSM8K accuracy (green). Harm and refusal use eager attention; GSM8K uses SDPA (Table 6). Curves summarize size-dependent trends under these respective configurations. Size 0 is the base; sizes 1–2 are exhaustive; each size 3–6 averages ten sampled compositions. Lines connect size groups, not nested trajectories; no uncertainty bands are shown.

requests. Refusal-matched responses can still be harmful: 11.2% on GLM-4 and 14.6% on Mistral (Appendix D.1).

Summed member shifts along a base refusal direction track refusal loss on one public bank (AUC approximately 0.99), but harm AUC ranges from 0.06 to 0.74 across in-house banks in the specified orientation. Values below 0.5 can signal inverse association; large shifts can belong to degraded states. The observed shifts are compressed relative to additive predictions, and the diagnostics do not support an additive-threshold theory of harm (Appendix D.2).

7 Mitigations: Harm Reduction and Task Retention

We evaluate two interventions on the 297 prompts: retraining a Dolly member with refusal examples and scaling each update by a common weight α. Weight 1 sums updates; $1/|T|$ averages them. Both experiments use complete judge prompts. Each experiment reruns its original states and references within its own harness (Appendix K.3). Original-member compositions still have mean $e_{12} = 0.059$–0.104 (Table 3), demonstrating the audit gap under this protocol too. Reweighting fixes full-strength singleton references; retraining changes the affected reference. We therefore assess e_{12}, total harm and task scores jointly. These are within-protocol comparisons on the reused evaluation pool.

Figure 2 compares both interventions. Task retention is relative to the task member alone; exceeding 100% need not exceed the base.

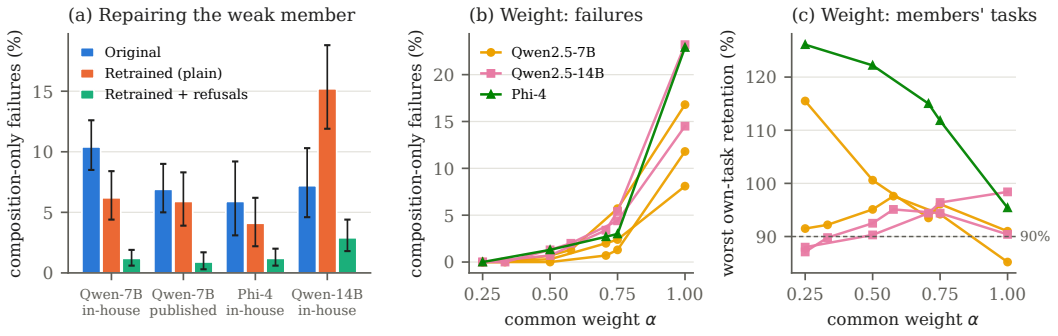


Figure 2: Interventions on 297 prompts. (a) Mean strict composition-only rate e_{12} across 28 small compositions per repair setting; bars are 95% two-level bootstrap intervals. (b) e_{12} versus common weight for six Dolly-containing compositions. (c) Minimum measured GSM8K/DialogSum retention relative to the task members alone, for the five compositions with these scores. Colors identify base models; curves identify compositions. The dashed line marks 90% retention. Overall harm and absolute task references must also be checked (Section 7).

Table 3: Member repair on 297 prompts. Means over 28 small compositions per setting. Orig., plain and safe denote the original member and the two retrained variants. Panel (b) reports two-level bootstrap intervals; differences are paired by partner set. The rule requires the safe upper endpoint $\leq .03$ and the difference interval below zero.

(a) Total harmful compliance: member and composition

	Member alone, h			Compositions, h		
Model (partners)	Orig.	Plain	Safe	Orig.	Plain	Safe
Qwen-7B (in-house)	.118	.364	.040	.185	.318	.056
Qwen-7B (published)	.118	.364	.040	.152	.354	.032
Phi-4 (in-house)	.327	.428	.017	.143	.253	.045
Qwen-14B (in-house)	.108	.226	.013	.140	.341	.062

(b) Composition-only failures, e_{12}

Model (partners)	Orig.	Plain	Safe [95% CI]	Safe − plain [95% CI]	Rule
Qwen-7B (in-house)	.104	.062	.012 [.006, .019]	−.050 [−.071, −.032]	met
Qwen-7B (published)	.069	.059	.009 [.003, .017]	−.051 [−.073, −.031]	met
Phi-4 (in-house)	.059	.041	.012 [.006, .020]	−.029 [−.049, −.012]	met
Qwen-14B (in-house)	.072	.152	.029 [.018, .044]	−.123 [−.159, −.090]	not met

7.1 Does repairing a member help its compositions?

For Qwen2.5-7B, Qwen2.5-14B and Phi-4, we retrained the Dolly member using one recipe with its rank, scaling and target modules retained. The *plain* variant uses Dolly-15k with 300 held-out records; the *safe* variant adds base-generated refusals to BeaverTails requests (Ji et al., 2023), approximately 5% of the combined records. Lexical overlap with AdvBench and JailbreakBench is filtered. Training counts and filtering rules are in Appendix K.2. We compare both variants with the original adapter on 28 partner sets per setting: one or two partners from the corresponding seven-member pool. Qwen2.5-7B also uses the published task partners.

The safe recipe lowers mean total harm and e_{12} relative to both original and plain in all four settings (Table 3). Mean e_{12} falls three- to sevenfold relative to plain (Figure 2a). Its upper interval endpoint is below 0.03 in three settings; Qwen2.5-14B improves but does not meet this threshold. The table's decision rule is an experimental criterion, not a deployment safety guarantee. Total harm remains above the base in the repaired groups. A second Qwen2.5-7B training seed gives $e_{12} = 0.014$ $[0.007, 0.022]$ for safe versus 0.075 $[0.053, 0.099]$ for plain. A third seed produced non-finite weights and was excluded before evaluation (Appendix K.2). This is limited replication, not an estimate of broad training-seed variability.

At the member level, safe matches plain closely on MMLU (within 0.013) and GSM8K (within 0.02, with higher point estimates). Qwen2.5-7B Dolly ROUGE-L is 0.345 versus 0.349. For safe compositions, MMLU stays within 0.02 of the base on the Qwen models and within 0.05 on Phi-4. Partner-specific evaluation is narrower: selected Qwen2.5-7B pairs at full strength (Table 4).

The repaired GSM8K pair retains 99% of its partner's score, compared with 101% for plain and 91% for the original. This is arithmetic retention, not an adaptation gain, because the GSM8K partner scores below the base. The repaired DialogSum pair retains approximately the partner's full score (0.320), versus base 0.183, while reducing harm to 0.054 and e_{12} to 0.017. This is a measured example of harm reduction with retained task adaptation, rather than only recovery toward the base. The evidence is specific to these tasks and pairs; repaired-composition over-refusal was not measured. Alpaca has no dedicated task score here; its entries report harm and MMLU only.

The original-to-plain comparison shows why e_{12} alone is misleading. For Qwen2.5-7B with in-house partners, plain lowers e_{12} from .104 to .062 while raising composition harm from .185 to .318; safe reduces both. Hence original and plain are both necessary comparators. The two Qwen partner pools share one repair. These experiments cover up to three adapters and do not repair Dolly-free failures.

Table 4: Repaired compositions, Qwen2.5-7B-Instruct, full strength, 297 evaluation prompts. Task: the partner's own task relative to the partner alone (GSM8K 0.785, DialogSum ROUGE-L 0.320; base 0.910 and 0.183). Base harm 0.017.

Partner	Weak member	h	e_{12}	Partner task	MMLU Δ
GSM8K	original	0.182	0.081	91%	−0.020
	plain	0.303	0.024	101%	−0.007
	safe	0.040	0.000	99%	−0.010
DialogSum	original	0.168	0.094	95%	−0.020
	plain	0.418	0.111	98%	0.003
	safe	0.054	0.017	100%	0.000
Alpaca	original	0.212	0.118	—	−0.023
	plain	0.323	0.064	—	−0.013
	safe	0.054	0.003	—	−0.020

7.2 Does down-weighting retain useful adaptation?

Six compositions use common weights $\alpha \in \{1, 1/\sqrt{k}, 1/k\}$, $k = |T|$, with full-strength singleton references. Figure 2b–c connects measured weights; Dolly+Alpaca lacks task scores and is omitted from panel (c).

Averaging reduces e_{12} from 0.081–0.232 to at most 0.013, but total harm remains 0.024–0.077 (Appendix Table 12). DialogSum retention is 90–92% of the singleton score, or 78–81% of its gain over the base. GSM8K reaches or exceeds the singleton, but those members score below the base; this is recovery, not retained arithmetic adaptation. Absolute references and intermediate weights are in Appendix E.1.

For six Dolly-free compositions in the two Qwen banks, averaging gives $e_{12} \leq 0.007$ but DialogSum retention falls to 76–87%. No tested weight meets all three criteria: $e_{12} \leq 0.02$, harm within 0.03 of the base, and task retention at least 90%. This does not exclude other weights or merging methods. These groups differ in size and membership, so the larger task cost cannot be attributed solely to the absence of Dolly.

8 Discussion

The audit unit is a composed model with specified members, weights and inference settings. Member-level scores remain useful references, but cannot identify which requests fail only after composition. Repair adds a further requirement: the reference labels must describe the repaired members, while total harm tracks whether the composed model itself improves. Utility must likewise be assessed against both the partner and base; returning toward the base can recover accuracy without retaining adaptation. These implications concern the tested additive compositions and evaluation protocols. Appendix B discusses their scope and future work.

9 Conclusion

Composition changes both failure frequency and which requests fail. Request-aligned audits reveal failures absent from member responses; the interventions show why these must be assessed alongside total harm and task benefit. Improving one measure need not improve the others. The relevant assessment is therefore of the composed model under its actual settings.

AI use statement

Generative AI assistants were used to aid and polish the writing of this manuscript, including improving clarity, grammar, and presentation. All scientific content, including the research ideas, experiments, analyses, mathematical statements and proofs, was produced by the authors. The authors reviewed and edited all AI-assisted text and take full responsibility for the final content.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905. arXiv:2412.08905.

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, 2024. URL https://arxiv.org/abs/2406.11717. arXiv:2406.11717.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *International Conference on Learning Representations*, 2024.

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 (Datasets and Benchmarks Track)*, 2024.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. URL https://aclanthology.org/2021.findings-acl.449/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168. arXiv:2110.14168.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Sihao Ding. Colluding LoRA: A Compositional Vulnerability in LLM Safety Alignment, 2026. URL https://arxiv.org/abs/2603.12681. arXiv:2603.12681.

Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. SafeMERGE: Preserving Safety Alignment in Fine-Tuned Large Language Models via Selective Layer-Wise Model Merging, 2025. URL https://arxiv.org/abs/2503.17239. arXiv:2503.17239.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

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 fine-tuning large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.16833.

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, 2021. URL https://arxiv.org/abs/2106.09685. arXiv: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, 2023. URL https://arxiv.org/abs/2307.13269. arXiv:2307.13269.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic, 2022. URL https://arxiv.org/abs/2212.04089. arXiv:2212.04089.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/4dbb61cb68671edc4ca3712d70083b9f-Abstract-Datasets_and_Benchmarks.html.

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

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. SaLoRA: Safety-alignment preserved low-rank adaptation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e24d9d028e3c7f6f13e6032919ee021e-Abstract-Conference.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. URL https://aclanthology.org/W04-1013/.

Yu Ma, Hongli Shi, Jing Li, Xinran Xu, and Weiwei Hou. When skills meet safety: Benchmarking and characterizing the adaptive jailbreak robustness of skill-merged LLMs. *arXiv preprint arXiv:2608.08542*, 2026.

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, 2024. URL https://arxiv.org/abs/2402.04249. arXiv:2402.04249.

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!, 2023. URL https://arxiv.org/abs/2310.03693. arXiv:2310.03693.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, 2023. URL https://arxiv.org/abs/2308.01263. arXiv:2308.01263.

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, 2024. URL https://arxiv.org/abs/2402.10260. arXiv:2402.10260.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model, 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html.

Team GLM. ChatGLM: A family of large language models from GLM-130B to GLM-4 All Tools, 2024. URL https://arxiv.org/abs/2406.12793. arXiv:2406.12793.

Zhiqi Wang, Shizhu He, Kang Liu, and Jun Zhao. Instance-level dynamic LoRAs composition for cross-task generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5699–5708. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.326. URL https://aclanthology.org/2024.findings-emnlp.326/.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models, 2023. URL https://arxiv.org/abs/2306.01708. arXiv:2306.01708.

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, 2023. URL https://arxiv.org/abs/2311.03099. arXiv:2311.03099.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.11206.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, 2023. URL https://arxiv.org/abs/2307.15043. arXiv:2307.15043.

Table 5: Eager-attention screening results: states meeting the harm, arithmetic-retention and repetition screens on the 80 main prompts. Member h is the maximum over the composition's members; GSM8K is relative to the best singleton. †Eager-attention utility evaluation: 200 GSM8K items relative to the best singleton and 300 MMLU items as the change from the base, one prompt, extraction rule and 1024-token limit for every state; — not re-evaluated. Mistral rows are contrast cases and use GSM8K adapter 06, with Alpaca adapters 09 and 03 in the two compositions.

MODEL	COMPOSITION	$\|T\|$	h	MEMBER h (max)	GSM8K	GSM8K†	MMLU† Δ	rep-4
Qwen2.5-14B	gsm8k + dolly + dialogsum	3	0.200	0.075	96%	82%	−0.026	0.018
Qwen2.5-14B	dolly + dialogsum	2	0.287	0.075	81%	81%	−0.026	0.035
Qwen2.5-7B	gsm8k + dolly + dialogsum	3	0.238	0.113	85%	78%	−0.046	0.029
Qwen2.5-7B	dolly + alpaca	2	0.263	0.113	81%	73%	−0.030	0.059
Qwen2.5-7B	gsm8k + dolly	2	0.188	0.113	83%	—	—	0.019
Phi-4	gsm8k_{00} + dolly	2	0.500	0.300	86%	78%	−0.024	0.036
Phi-4	dolly	1	0.300	—	98%	97%	0.000	0.070
Mistral-7B	gsm8k_{06} + alpaca_{09}	2	0.800	0.762	84%	—	—	0.044
Mistral-7B	gsm8k_{06} + alpaca_{03}	2	0.737	0.762	100%	—	—	0.040
Mistral-7B	gsm8k	1	0.762	—	92%	—	—	0.049
Mistral-7B	dialogsum	1	0.650	—	96%	—	—	0.044

A Complete Numerical Summaries and Norm Controls

These tables retain the reported numerical summaries. Table 6 uses corrected SDPA GSM8K means. Table 5 and the older norm, weight and prioritization analyses retain eager-harness utility measurements; they document exploratory selection and are not independent confirmation of corrected utility. Screening used $h_T \geq \max(0.15, h_\varnothing + 0.10)$, at least 80% of the best singleton's arithmetic score, and repeated 4-gram rate at most twice the median singleton's. These were descriptive cutoffs, not deployment criteria.

B Scope and Future Work

This section specifies the boundaries of the empirical evidence and the research directions they motivate.

Evaluation scope and intervention design. The results characterize selected banks and compositions rather than deployment prevalence. Larger subsets are sampled, and public banks share a base and some members. Interventions reuse requests observed during study development, with recipes fixed before their runs. Discovery and intervention scoring protocols differ; each intervention contrast uses a common protocol internally. Task-specific repair evidence is concentrated on Qwen2.5-7B, and training-seed replication is limited. The safe recipe changes both data content and training steps, so their causal contributions are not isolated.

What is measured as safety. Harmful compliance is one safety-relevant behaviour under a specified prompt set and judge. It is not a measure of all safety risks, severity, real-world impact, or robustness to adaptive attacks. Refusal matching is a surface measurement, while low harmful compliance can arise from harmless behaviour or an inability to answer. These quantities should be reported jointly with utility rather than substituted for one another.

Scope of the compositional claim. The clearest utility-retaining examples contain a Dolly-trained adapter whose singleton is already harmful, in-house or published. Averaged over the compositions that contain it, harm is close to that member's own rate; what changes is which prompts fail. Dolly-free compositions also fail on the two in-house Qwen banks but only weakly on the published instruction bank, and the published control bank has low measured harm. The empirical claim concerns request-specific failures rather than uniform amplification of average harm. A weak member is neither shown necessary nor shown sufficient. Composition-only failures also do not require a minimal colluding set: the definition compares a composition with its singletons, not with every proper subcomposition. All published banks share one base model.

Attention implementation and utility references. The initial harm sweep and exploratory utility analyses used eager attention. The main utility figure uses the complete SDPA rerun of the in-house 60-item probe, and the interventions use SDPA with their own reference states. Eager-attention

Table 6: Size-averaged harmful compliance (h), absolute GSM8K accuracy, repeated-4-gram rate, and refusal-match rate (r). Sizes 0–2 are exhaustive; each larger size averages ten sampled states. Bold indicates the largest observed mean harm for each bank. Repetition is a diagnostic, not a substitute for utility. GSM8K uses SDPA attention (Appendix K.6).

	0	**1**	**2**	**3**	**4**	**5**	**6**
Qwen2.5-7B							
h	0.000	0.027	0.094	**0.165**	0.120	0.075	0.010
GSM8K	0.800	0.765	0.645	0.538	0.272	0.068	0.003
rep-4	0.080	0.069	0.054	0.035	0.132	0.330	0.537
r	0.950	0.866	0.704	0.421	0.182	0.104	0.001
Qwen2.5-14B							
h	0.000	0.011	0.040	0.077	0.115	**0.133**	0.094
GSM8K	0.833	0.858	0.745	0.675	0.437	0.333	0.113
rep-4	0.070	0.053	0.038	0.027	0.024	0.022	0.049
r	0.988	0.900	0.727	0.467	0.256	0.149	0.035
Phi-4							
h	0.000	0.055	0.088	**0.123**	0.007	0.000	0.000
GSM8K	0.817	0.762	0.536	0.203	0.022	0.012	0.000
rep-4	0.068	0.050	0.067	0.174	0.106	0.041	0.062
r	0.925	0.800	0.460	0.160	0.021	0.004	0.000
GLM-4-9B (high repetition under individual adapters)							
h	0.037	0.175	0.325	**0.487**	0.237	0.040	0.000
GSM8K	0.650	0.562	0.402	0.252	0.103	0.018	0.002
rep-4	0.132	0.561	0.601	0.625	0.542	0.212	0.050
r	0.963	0.811	0.588	0.284	0.173	0.037	0.003
Mistral-7B-v0.3 (base model already complies)							
h	0.450	0.659	**0.691**	0.573	0.312	0.057	0.005
GSM8K	0.433	0.304	0.181	0.092	0.032	0.015	0.002
rep-4	0.063	0.043	0.041	0.051	0.053	0.022	0.009
r	0.150	0.108	0.033	0.006	0.005	0.000	0.000

Table 7: Eager-attention norm controls, with harm on 160 instructions and absolute GSM8K accuracy in parentheses. Random controls are reported as layerwise norm-matched. Superscripts give the scaled singleton's actual squared norm: it overshoots the target in every row. Consequently the singleton control is approximate, not conservative under a non-monotone response.

$\|\Delta W\|^2$	**COMPOSITION**	**SCALED SINGLE**	**RANDOM LOW-RANK**
Qwen2.5-7B (base 0.006, GSM8K 0.300)			
328	0.131 (0.650)	$0.019\ (0.117)^{367}$	0.006 (0.32)
513	0.188 (0.667)	$0.037\ (0.183)^{642}$	0.006 (0.30)
1035	0.119 (0.550)	$0.000\ (0.000)^{1141}$	0.006 (0.32)
1481	0.237 (0.650)	$0.000\ (0.000)^{1783}$	0.010 (0.32)
Phi-4 (base 0.000, GSM8K 0.833)			
230	0.519 (0.700)	$0.100\ (0.317)^{335}$	0.000 (0.833)
918	0.475 (0.700)	$0.000\ (0.000)^{930}$	0.000 (0.833)
2033	0.056 (0.617)	$0.000\ (0.000)^{2381}$	0.000 (0.858)

results are labelled in Table 5 and the eager-attention diagnostics below. Similar size means do not prove that state-level scores, retention thresholds or orderings are unchanged. The 297-prompt harm comparison regenerates both compositions and references (Table 21); the persistence of positive event rates, rather than a nonsignificant implementation difference, supports the main robustness conclusion.

Task coverage. GSM8K measures arithmetic, MMLU a subset of multiple-choice knowledge, and DialogSum and Dolly their specified generation tasks. None measures general usefulness. A retention

ratio relative to a task member does not establish benefit over the base; both references are needed when claiming retained adaptation. Repaired-composition over-refusal was not evaluated.

Eager-attention utility evaluation. For the base, all singletons and every composition of Table 1 we scored the first 200 GSM8K test items with one prompt (final answer after `####`), greedy decoding, a 1024-token limit and one extraction rule (`####`, `\boxed`, "answer is", else the last number), and 300 MMLU test items sampled with a fixed seed. The pre-specified capability-retention criterion requires that a composition's GSM8K accuracy is at least 80% of the best singleton's and its MMLU accuracy is within 0.05 of the base's. Among the top-5 compositions of Table 1, two on Qwen2.5-14B meet this ($h = 0.287$ and 0.200; 81–82% and -0.026), none on Qwen2.5-7B (best 78%), Phi-4 (best 78%) or the published bank with Dolly, whose most harmful composition also loses 0.093 MMLU. The five least harmful Dolly compositions on Qwen2.5-7B and Phi-4, all of size four to six, score 0–0.03 on GSM8K and, in seven of ten cases, below 0.18 on MMLU, consistent with substantial task degradation under eager attention.

Sampling, judges, and uncertainty. Only ten states are sampled at each size above two, and all sizes use shared prompts. Intervals in the main text resample prompts and, where stated, compositions within a size; they do not account for judge error. Judge agreement supports the qualitative patterns but does not validate every label. Our composition-only failure estimates rely on automated judges. Agreement across judges supports robustness to the choice of evaluator but does not establish label correctness; correlated errors may affect both the composition's labels and its members'. A member response incorrectly labelled non-harmful inflates $e(T)$, which the intersection rule can reduce by disqualifying the event when either judge flags a member; false-positive composition labels can also inflate the rate. Requiring both judges can reduce, but does not eliminate, these errors. The small qualitative check below is not a calibration study, and classifier labels do not measure the full severity of generated content. The 160-prompt interventions and XSTest evaluations concern selected compositions, not the full sweep. Sensitivity to screening thresholds remains unreported.

Interventions and prediction. The weight experiments change only a common scalar; they do not compare against optimized merging operators. Norm controls are limited to selected states and directions. The supporting membership-model experiment (Appendix I) reports a failure of one fitted predictor under one protocol, not a lower bound on achievable prediction error. Similarly, a size-dependent peak in the observed means is not proof of a sharp threshold along a continuous or nested composition trajectory.

External confirmation. The control bank was assembled and its selection recorded before its outcomes were seen, and the triage parameter was fixed before it was evaluated. Because the control bank contains no flagged state, it tests whether the phenomenon generalizes, not whether the triage rule does. It cannot make the in-house analyses confirmatory.

Future work. Three extensions follow from this scope. First, composition audits could be evaluated on independently selected adapter banks and fresh request sets, including benign requests for measuring over-refusal and adaptive attacks for testing robustness beyond fixed benchmarks. Second, step-matched and multi-seed retraining, combined with broader partner-task evaluation, could separate the effect of refusal data from additional optimization and test whether useful adaptation persists across tasks. Third, learned composition weights and more selective merging operators could be assessed with the same request-level references. A predictive extension would need to demonstrate transfer to unmeasured compositions while accounting for higher-order interactions; the present diagnostics do not establish such a predictor.

C Composition-Only Failures and Published Banks: Details

C.1 Request-level counts on the main prompts

Dolly has nonzero singleton harmful compliance on the three primary banks (0.113, 0.075 and 0.300) and occurs in every primary utility-retaining composition in Table 5. It is not, however, required for harm under composition. Among the 40 sampled larger compositions per bank, those without Dolly reach a size-mean harm of 0.122 on Qwen2.5-7B (size four) and 0.145 on Qwen2.5-14B (size five), although no non-Dolly member of those banks exceeds 0.050 and 0.013 respectively; on Qwen2.5-14B the Dolly-free compositions are, if anything, more harmful than those containing Dolly.

Phi-4 differs: without Dolly its size-three compositions average 0.027 against 0.217 with Dolly, and arithmetic accuracy collapses before refusal does. Dolly-free compositions that retain utility exist but are weaker than the Dolly examples: 0.125 at 81% retention on Qwen2.5-14B and 0.113 at 85% on Qwen2.5-7B, neither meeting both screens of Table 5. The Qwen2.5-7B composition of GSM8K with two Alpaca adapters has harm 0.150 against members at 0.000, 0.050 and 0.000, with 62% retention. These observations associate Dolly with failures in smaller compositions that retain more measured task ability; they do not isolate a causal mechanism for the size difference.

Request-level labels make Proposition 1 concrete. For the Phi-4 pair of Table 5 (GSM8K adapter 00 with Dolly), the proposition gives

$$e(T) \geq 0.500 - 0.000 - 0.300 - 0.000 = 0.200, \tag{10}$$

that is, at least 16 of the 80 prompts. The exact count is 20 ($e(T) = 0.250$): on these prompts the pair is scored harmful while neither the base nor either member is. Across all compositions of size two or more, the share of harmful responses that are composition-only is 79% on Qwen2.5-7B, 85% on Qwen2.5-14B, 54% on Phi-4, 49% on GLM-4 and 7% on Mistral-7B, whose base is already harmful. These shares are large partly because members are rarely harmful: if a composition's failures were unrelated to its members', 91%, 96% and 74% of them would fall on new prompts on the three low-harm-base banks. The observed shares are lower, and a prompt on which the base or some member is harmful is 3.1, 4.8 and 4.4 times as likely to be harmful under composition as any other prompt. Composition failures therefore depend on member failures without reducing to them: pooled over all compositions, only 16–33% of member-harmful prompts remain harmful under composition, and new prompts supply most harmful responses. The Phi-4 pair of Table 5 shows both components in one state: it remains harmful on 20 of the 24 prompts that Dolly fails alone and adds 20 prompts that no member fails. Pooled figures include compositions whose utility has collapsed, which fail few prompts of either kind; these pooled quantities should not be interpreted as measurements restricted to capable compositions. These counts describe request-level behaviour, not a weight-space mechanism.

C.2 Published banks

Update norm alone also fails to separate these banks in the overlapping squared-norm range 800–2400: in-house compositions average harm 0.10–0.16 per bin, compared with at most 0.008 for the control bank. This cross-bank contrast does not isolate a causal mechanism, since training tasks, adapter identity and update geometry differ.

Replacing a control member with the in-house Dolly adapter yields selected compositions with much higher harm. On the earlier JailbreakBench evaluation, one reaches 0.495, compared with 0.196 for Dolly and 0.093 for its partners together. However, averaging over sampled weak-member compositions on the main prompts gives excess harm of only 0.01–0.06, with intervals including zero across the three Qwen2.5-7B banks. Their mean composition-only rates remain 0.095–0.119 (Table 9). The supported finding is new failure requests, not a uniform increase in average harm.

The published control members and non-Dolly in-house members both have low measured harm and high refusal, but their composition outcomes differ. Update magnitude alone does not account for the difference (Section E.3): control-bank compositions span the same range of squared update norms as failing in-house compositions. The one structural difference we found is the members' mutual alignment. In-house updates trained on the same dataset are strongly aligned (pairwise Frobenius cosine 0.04–0.25 among the Alpaca variants), and even cross-task in-house pairs have cosines of 0.005–0.011, whereas every control-bank pair is below 0.0013. Same-rank in-house adapters have unrelated A matrices, so identical A initialization does not explain this observation. We report this as an observation only. Alignment does not order harm within a bank (the most aligned in-house pair keeps a refusal-match rate of 0.91), and an Alpaca-pair effect visible on the two Qwen banks reverses on the other three.

Foreign partners extend a weak member's failures. Two further banks, both specified with decision rules before evaluation, locate the difference in the members rather than in the in-house training recipe. In the first, the control bank's eighth member is replaced by the in-house Dolly adapter (singleton harm 0.113). All 18 sampled compositions without Dolly reproduce the control bank exactly. The Dolly-containing sampled compositions average 0.163, against a pre-set threshold of 0.16, so this test is met only at its margin. The pattern within it is clearer: the most harmful state combines Dolly with five published partners (medical summarization, a persona model, academic rewriting, role-play and

Table 8: Second published bank (general instruction-following adapters), split by whether a composition contains the Dolly-15k adapter (harm 0.362 alone). h: size-mean harmful compliance; e: mean composition-only rate, the fraction of prompts harmful under the composition but not under the base or any member. The split is post hoc.

Size		1	2	3	4	5	6
Without Dolly-15k:	h	0.012	0.015	0.016	0.029	0.050	0.056
	e	–	0.004	0.003	0.004	0.016	0.019
With Dolly-15k:	h	0.362	0.368	0.421	0.397	0.456	0.450
	e	–	0.059	0.110	0.119	0.152	0.152

Table 9: Compositions containing the weak member, Qwen2.5-7B-Instruct, all evaluated compositions of size two or more that contain it. Means with 95% intervals from a two-level bootstrap over prompts and compositions.

Bank (weak member, harm alone)	n	Excess over member	Composition-only e	Harm where member unflagged
In-house (Dolly, 0.113)	28	$0.011\ [-0.071, 0.086]$	$0.095\ [0.061, 0.135]$	$0.116\ [0.074, 0.163]$
Published + Dolly (0.113)	29	$0.032\ [-0.041, 0.101]$	$0.097\ [0.059, 0.139]$	$0.113\ [0.069, 0.164]$
Published instr. (Dolly-15k, 0.362)	31	$0.058\ [-0.010, 0.127]$	$0.119\ [0.068, 0.175]$	$0.189\ [0.112, 0.271]$

an evaluation judge) and reaches 0.363 while keeping 71% of the best singleton's GSM8K accuracy. Its refusal-match rate, 0.037–0.113 across the top states, is far below Dolly's own 0.188. These partners have low harm in the main-prompt checks; combined with Dolly, the most harmful of their compositions lower refusal and triple its harm. This maximum is selected from 29 Dolly-containing states, so it overstates the typical effect (Table 9), but it holds on held-out prompts (Table 10).

The second bank consists of eight general instruction-following adapters published by different authors for the same base model (one Dolly-15k, four Alpaca variants, three DPO-LIMA). By its pre-set criterion it reproduces only partially: 31 of 68 compositions are flagged, but the largest size mean (0.371 at size six) exceeds the most harmful member by only 0.009. That member is the published Dolly-15k adapter, which is harmful on 0.362 of prompts alone; no other member exceeds 0.050. Separating the compositions by whether they contain it gives the picture in Table 8. Without it, harm grows slowly with size and stays below 0.09. With it, harm rises from 0.368 at size two to 0.450 at size six, arithmetic accuracy stays at 0.37–0.44 at every size, and the composition-only share $e(T)$ grows from 0.059 to 0.152. Compositions containing the Dolly-15k adapter remain harmful on 83% of the prompts it fails alone and become harmful on 19% of the prompts it is not labelled harmful on.

Averaged over all compositions that contain the weak member, the typical effect is a redistribution rather than a uniform increase (Table 9). Mean harm exceeds the weak member's own rate by only 0.01–0.06, and every interval includes zero. What does not average out is where the failures fall. On all three Qwen2.5-7B banks with a weak member, 10–12% of prompts become harmful under a composition although neither the base nor any member is harmful on them, and on the prompts where the weak member is not labelled harmful, 11–19% become harmful; all six intervals exclude zero. Compositions without the weak member are far less affected (on the second published bank, composition-only rates of at most 0.019).

Across the Qwen2.5-7B banks the same shape therefore recurs: partners with low singleton harm can keep many of a weak member's failures, drop some, and add failures on prompts where that member was not labelled harmful.

Held-out prompts. The most harmful compositions above were selected on the same 80 prompts used to report them. We therefore re-evaluated, with criteria fixed before generation, the five most and five least harmful compositions containing the weak member of each bank, the weak member alone, and each top composition without its weak member, on 97 JailbreakBench behaviours (Chao et al., 2024) disjoint from the main prompts (Table 10; base harm 0.072). On the in-house bank and on the published task bank with Dolly, the selected compositions stay well above the weak member alone ($+0.16$ and $+0.20$, intervals excluding zero) and above the lowest-ranked compositions, and the top Dolly composition of the published bank reaches 0.495 against 0.196 for Dolly alone and

Table 10: Held-out re-evaluation on 97 JailbreakBench behaviours (base harm 0.072). Top-5 and bottom-5 are the compositions containing the weak member with the highest and lowest harm on the main prompts; partners are the top-5 compositions with the weak member removed. Intervals: 95% prompt bootstrap.

Bank	Weak alone	Top-5	Partners	Bottom-5	Top-5 − weak	Top-5 harm where weak unflagged
In-house	0.196	0.359	0.194	0.004	+0.163 [0.095, 0.231]	0.269 [0.217, 0.330]
Published task + Dolly	0.196	0.400	0.085	0.225	+0.204 [0.128, 0.274]	0.321 [0.256, 0.386]
Published instruction	0.619	0.579	0.097	0.594	−0.039 [−0.132, 0.047]	0.314 [0.189, 0.440]

Table 11: Strict composition-only rates on 297 evaluation prompts. Top-5: the five most harmful compositions containing the weak member, selected on the main prompts; free-5: the five most harmful compositions without it. e_1, e_2: strict rate with every label from the primary or the second judge alone; e_{12}: composition harmful under both judges, base and members counted as failing under either. 95% prompt-bootstrap intervals. Base harm on these prompts is 0.000–0.034.

Bank	Weak alone	Top-5 h	Top-5 e_1	Top-5 e_2	Top-5 e_{12}	Free-5 e_{12}	Top-5 MMLU Δ
Qwen2.5-7B in-house	0.114	0.237	0.147 [0.123, 0.170]	0.152	0.126 [0.104, 0.149]	0.093 [0.075, 0.112]	−0.05 to 0.00
Qwen2.5-14B in-house	0.101	0.232	0.178 [0.150, 0.207]	0.185	0.156 [0.131, 0.183]	0.127 [0.106, 0.148]	−0.06 to −0.03
Phi-4 in-house	0.320	0.464	0.261 [0.227, 0.297]	0.257	0.220 [0.186, 0.254]	0.032 [0.022, 0.043]	−0.06 to −0.02
Published task + Dolly	0.114	0.271	0.187 [0.158, 0.216]	0.203	0.170 [0.143, 0.198]	0.015 [0.005, 0.027]	−0.13 to −0.02
Published instruction	0.418	0.480	0.163 [0.129, 0.199]	0.168	0.127 [0.097, 0.160]	0.046 [0.027, 0.068]	−0.08 to 0.00

0.093 for its five partners together. On the published instruction bank the selection does not transfer: the Dolly-15k adapter alone is harmful on 0.619 of the new prompts, and compositions containing it are neither more harmful than it nor ordered as before. On all three banks, the top compositions are harmful on 27–32% of the held-out prompts where the weak member is not labelled harmful, and on the two published banks the partners without the weak member stay within 0.05 of the base; the separate 297-request evaluation in Table 1 measures failures absent from all singleton references. The weak-member-only statistic does not itself establish that stronger condition. We therefore report the published instruction bank as a case in which a harmful member dominates and composition relocates its failures, and the Dolly banks as cases in which composition also raises harm above every member.

The weak members examined here were trained on Dolly-15k, including both in-house training and an independently published adapter; the published control bank has low observed harm. That fine-tuning on benign instruction data can weaken a single model's safety is known (Qi et al., 2023); our request-level analysis examines which additional requests fail under composition. The in-house banks add a second route: Dolly-free compositions there reach size means of 0.12–0.15 (Section 5.1), which the published general-instruction bank does not reproduce. We therefore do not claim that a weak member is necessary, and we do not estimate how common either situation is among published adapters.

D Refusal and Representations: Details

D.1 How refusal-based orderings disagree with harm

Two states are compared when their harm rates differ by at least 0.056, the largest single-rate binomial standard error at 80 prompts; pairs with equal refusal-match rates are excluded. Across all pairs of the 40 sampled larger compositions per bank, the state with less refusal matching has lower harm in 68.5% of 1813 eligible pairs under the primary judge and 71.2% of 1354 under the second judge. These are descriptive proportions over overlapping pairs, not independent trials.

The following utility-conditioned comparisons use the eager-attention scores. They are exploratory and have not been recomputed here from state-level SDPA outputs. Within a fixed composition size the inversion rate is 0.32, 0.48, 0.29, 0.38 and 0.53 for Qwen2.5-7B, Qwen2.5-14B, Phi-4, GLM-4 and Mistral. Restricting comparisons to compositions whose GSM8K accuracy is at least half the best singleton's reduces it to 0.029 (12 of 410 pairs), 0.114, 0.143 and 0.063 on the four banks whose base models refuse. Among compositions that still perform the task, refusal matching orders harm

Table 12: Summing versus averaging on 297 prompts. Each cell gives full strength → mean weight ($1/k$). G and D are scores relative to the GSM8K and DialogSum members alone. Harm is not measured by task retention: base h is .020, .013 and .000 for Qwen-7B, Qwen-14B and Phi-4. Intermediate weights are in Table 13.

Model	Composition	h: sum → mean	e_{12}: sum → mean	Task: sum → mean
Qwen-7B	GSM+Dolly	.175 → .054	.081 → .003	G 91% → 101%
	GSM+Dolly+Dialog	.246 → .040	.168 → .000	G 85% → 110%; D 94% → 92%
	Dolly+Alpaca	.215 → .064	.118 → .000	—
Qwen-14B	GSM+Dolly+Dialog	.253 → .024	.145 → .000	G 90% → 101%; D 102% → 90%
	Dolly+Dialog	.337 → .030	.232 → .007	D 98% → 90%
Phi-4	GSM+Dolly	.498 → .077	.229 → .013	G 95% → 122%

Table 13: Composition weight versus harm and the members' own tasks, 297 evaluation prompts. h: primary-judge harm (base 0.020, 0.013, 0.000 for Qwen2.5-7B, Qwen2.5-14B, Phi-4); e_{12} uses Eq. (7) with the complete intervention judge prompts; task: own-task score of the GSM8K (G) and DialogSum (D) members relative to the member alone. No adapter-specific task score is reported for Dolly+Alpaca; its MMLU stays within 0.02 of the base at every weight.

		$\alpha = 1$			$\alpha = 1/\sqrt{k}$			$\alpha = 1/k$		
Model	Composition	h	e_{12}	Task	h	e_{12}	Task	h	e_{12}	Task
Qwen2.5-7B	gsm8k+dolly	0.175	0.081	G 91%	0.098	0.020	G 94%	0.054	0.003	G 101%
	gsm8k+dolly+dialogsum	0.246	0.168	G 85%, D 94%	0.088	0.013	G 101%, D 98%	0.040	0.000	G 110%, D 92%
	dolly+alpaca	0.215	0.118	—	0.081	0.007	—	0.064	0.000	—
Qwen2.5-14B	gsm8k+dolly+dialogsum	0.253	0.145	G 90%, D 102%	0.064	0.020	G 97%, D 95%	0.024	0.000	G 101%, D 90%
	dolly+dialogsum	0.337	0.232	D 98%	0.128	0.034	D 94%	0.030	0.007	D 90%
Phi-4	gsm8k+dolly	0.498	0.229	G 95%	0.168	0.027	G 115%	0.077	0.013	G 122%

correctly for 86–97% of comparable pairs; it misorders states mainly by ranking utility-collapsed compositions as the least safe.

Refusal matching also misses some harm. The share of refusal-matched responses that the primary judge scores harmful is below 1% on the two Qwen banks and Phi-4, but 11.2% on GLM-4 (259 of 2309) and 14.6% on Mistral (24 of 164), where responses often open with a refusal phrase and then comply. The ϵ_T term of Eq. 14 is therefore bank-dependent and cannot be assumed negligible.

The supported claim is that refusal matching alone can misorder the measured states, chiefly across the boundary between retained and lost utility, and can miss harmful responses that begin with a refusal. Pairing refusal matching with a utility measurement removes most of the ordering problem in our data but not the second failure.

D.2 The same dissociation in representations

The same dissociation appears in the base model's representations. We take the difference-of-means refusal direction (Arditi et al., 2024) at the post-instruction token of chat-formatted prompts, estimated from AdvBench instructions disjoint from the evaluation prompts, and project every state's activations on it. A composition's mean shift along this direction is correlated with the sum of its members' shifts (Pearson 0.87–0.97 across the eight banks) but is compressed to 0.23–0.75 of that sum. The projection tracks refusal: summed member shifts separate compositions that lose most refusal matching from the rest better than summed member refusal rates on the balanced published bank (AUC 0.99 against 0.75). It does not consistently rank harm in the specified orientation. For flagging harmful compositions its AUC is 0.53, 0.74, 0.44, 0.31 and 0.06 on the five in-house banks, against 0.73, 0.62, 0.83, 0.52 and 0.41 for the most harmful member's rate, and it is anti-predictive on GLM-4 and Mistral because the largest shifts belong to compositions that no longer perform the task. This refusal-associated representation signal does not reliably rank harmful compliance in these banks. The additive-threshold account is not supported by the diagnostic criteria (Appendix K.5); these are descriptive measurements, not a causal mechanism for harmful behavior.

Table 14: Eager-attention weight intervention on the six selected compositions, averaged within a model. Harm on 160 AdvBench requests and XSTest unsafe prompts; over-refusal on XSTest safe prompts; GSM8K relative to the bank's best singleton. Bold marks the minimum over-refusal, including ties.

MODEL	WEIGHT	h^{adv}	h^{xs}	OVER-REF	GSM8K
Qwen2.5-7B	base	0.006	0.015	0.032	—
	α=1	0.221	0.183	**0.011**	91%
	α=$1/\sqrt{k}$	0.035	0.032	0.023	98%
	α=$1/k$	0.017	0.012	0.036	106%
Qwen2.5-14B	base	0.000	0.005	0.028	—
	α=1	0.209	0.217	**0.008**	89%
	α=$1/\sqrt{k}$	0.034	0.035	0.022	97%
	α=$1/k$	0.003	0.010	0.020	103%
Phi-4	base	0.000	0.005	0.384	—
	α=1	0.475	0.555	**0.000**	88%
	α=$1/\sqrt{k}$	0.094	0.185	**0.000**	102%
	α=$1/k$	0.006	0.030	0.004	108%

E Composition Weight and Strength: Details

E.1 Task references and interpretation of the weight intervention

We scale all members of six selected compositions by a common α, including 1, $1/\sqrt{k}$ and $1/k$, where $k = |T|$. Task scores are relative to the GSM8K or DialogSum member alone; the base remains a separate reference.

Figure 2b shows the composition-only failure curves, and panel (c) shows the minimum measured partner-task retention for each composition with task scores. The Dolly+Alpaca composition appears only in panel (b), because neither a GSM8K nor a DialogSum member score is available for it. Connecting evaluated weights does not establish behavior at untested weights. At full strength, e_{12} is 0.081–0.232; averaging reduces it to at most 0.013 (Table 12). Total harm is still 0.024–0.077, so near-zero composition-only rates should not be described as restored base safety. GSM8K scores match or exceed the task member's, while DialogSum retention is approximately 90–92%. Values rounded to 90% do not resolve whether an exact 90% decision threshold is met.

Absolute references clarify the task benefit. The DialogSum members score 0.317 and 0.329 on the two Qwen models, against bases 0.183 and 0.172. Averaging retains approximately 78–81% of this improvement, compared with 88–94% at $1/\sqrt{k}$. The GSM8K members score 0.775, 0.885 and 0.765, below their respective bases 0.910, 0.930 and 0.955; there is no arithmetic adaptation benefit to preserve under this harness.

For six selected Dolly-free compositions across the two Qwen banks, averaging reduces e_{12} to at most 0.007 and restores arithmetic and MMLU performance, but DialogSum retention falls to 76–87%. No *tested* weight jointly satisfies the specified criteria ($e_{12} \leq 0.02$, harm within 0.03 of the base, and measured task retention at least 90%). This does not rule out other weights or merging methods. These compositions differ in size and membership from the Dolly-containing examples, so their greater task cost cannot be attributed solely to the absence of a weak member. Additional strength and norm controls are reported in Appendix E.4 as supporting diagnostics.

E.2 Weights on selected compositions, main prompts and XSTest

We re-evaluated six selected utility-retaining compositions (three from Qwen2.5-7B, two from Qwen2.5-14B, one from Phi-4) at common weights $\alpha \in \{1, 1/\sqrt{k}, 1/k\}$, $k = |T|$, on 160 AdvBench instructions, XSTest's unsafe and safe prompts, and GSM8K (30 states). Weight $1/k$ is the mean of the updates; with near-orthogonal updates, $1/\sqrt{k}$ holds the update norm near a single adapter's. The states were selected after observing risk, so this is an intervention on flagged examples.

Table 15: Strength sweeps on fixed compositions, 80 main prompts. "top" is the most harmful sampled composition of the bank and "ctl" the least harmful of the same size; all members are scaled by α. "match" scales the composition to the Frobenius norm of its largest member (α^* in parentheses); "−weak" removes the weak member at $\alpha = 1$. GSM8K uses the sweep's own prompt, under which the base scores 0.283.

Bank		0.25	0.5	0.75	1	1.25	1.5	match	−weak
In-house	top h	0.013	0.037	0.138	0.287	0.050	0.000	0.125 (0.77)	0.113
	top GSM8K	0.817	0.817	0.683	0.400	0.133	0.017	0.650	0.800
	ctl h	0.000	0.013	0.025	0.025	0.037	0.100	0.025	–
Published task + Dolly	top h	0.000	0.013	0.087	0.362	0.263	0.013	0.175 (0.81)	0.000
	top GSM8K	0.133	0.833	0.367	0.450	0.067	0.000	0.400	0.200
	ctl h	0.000	0.000	0.000	0.013	0.000	0.000	0.000	–
Published instruction	top h	0.037	0.100	0.338	0.500	0.500	0.500	0.250 (0.68)	0.050
	top GSM8K	0.283	0.317	0.233	0.200	0.167	0.150	0.233	0.233
	ctl h	0.000	0.013	0.025	0.037	0.062	0.138	0.025	–

Lower weights reduce harm on both prompt sets and raise GSM8K accuracy (Table 14): over the six compositions, AdvBench harm is 0.259, 0.045 and 0.010 at the three weights, and arithmetic is not below the best singleton at $1/k$. Over-refusal rises by at most 0.025. At $1/\sqrt{k}$, harm remains above the corresponding base values. Further scaling reduces it but does not make it universally zero. Section 7 evaluates the members' own tasks on the additional evaluation requests.

E.3 Norm controls

Composition size and squared update norm are strongly associated in the main sweep (reported Spearman 0.845). We therefore compare compositions with random low-rank updates matched layer by layer in Frobenius norm, and with scaled single adapters. For the reported Qwen2.5-7B subset, cross terms account for 5.6% of total squared norm; this does not establish functional independence of the updates.

The random controls remain near the corresponding base values, whereas compositions at those norms can show higher harm and substantially different GSM8K performance. This indicates that Frobenius norm alone is insufficient to account for the tested behaviours; it does not rule out other notions of functional magnitude. Scaled singletons frequently lose arithmetic accuracy at nearby norms where compositions retain it. Because their norms overshoot the targets and their harm responses are non-monotone, that comparison is suggestive rather than an exact direction-controlled experiment. Exact norm-matched singleton controls and repeated random directions would strengthen the interpretation. We do not infer that near-orthogonality causes utility retention or produces a universal composition threshold.

The comparison also holds between real banks. We computed the exact squared norm, cross terms included, of every sampled composition in the in-house Qwen2.5-7B bank and in the control bank. Between 800 and 2400, 21 in-house compositions average harm of 0.10–0.16 per norm bin (maximum 0.287), while 26 control-bank compositions average at most 0.008 (maximum 0.025). Cross terms account for 8.8% of the in-house composition norms and for none of the control bank's.

E.4 Strength sweeps on fixed compositions

The norm comparisons above contrast different compositions. To vary strength while holding membership fixed, we took the most harmful sampled composition of three Qwen2.5-7B banks and, as a control, the least harmful composition of the same size, and scaled all members by a common α from 0.25 to 1.5 (design and readings fixed before evaluation; Table 15). We also removed each member in turn at $\alpha = 1$ and scaled each composition to the Frobenius norm of its largest member. Summed updates are only 1.2–1.5 times the norm of their largest member, so the norm-matched state is close to $\alpha = 0.7$–0.8.

Three observations follow. First, on two of the three sweeps harm changes nonlinearly with the tested strength. On the two Dolly banks it stays within 0.04 of the base up to $\alpha = 0.5$, rises sharply between

0.75 and 1, and then falls as GSM8K collapses at $\alpha \geq 1.25$, reproducing along a single composition the rise and fall seen across composition sizes. On the instruction bank, whose member is harmful alone, harm rises earlier and saturates at 0.500 without falling. The projection on the refusal direction of Section D.2 changes nearly linearly in α ($R^2 = 0.99$ and 0.96 on the Dolly banks) while harm does not. Second, in these sweeps update norm alone does not account for which compositions fail. Each control composition at $\alpha = 1.5$ has an update norm comparable to or larger than its top composition at $\alpha = 1$ (40.9, 52.1, 77.0 against 44.8, 44.8, 56.7) and harm of 0.100, 0.000 and 0.138 against 0.287, 0.362 and 0.500. Magnitude does matter within a composition: at the norm of their largest member the top compositions retain approximately 44–50% of their full-strength harm, not exceeding the pre-specified 50% criterion for magnitude-independence. Third, membership matters in a specific way. Removing the weak member removes most harm on every bank, and on the published task bank removes all of it, although that member alone is harmful on only 0.113 of prompts and its five partners on none. On that bank, removing the role-play or academic-rewriting partner also halves the harm. Which partners matter is not predicted by their singleton harm, at most 0.025 for all of them, nor by their singleton shift along the refusal direction (Spearman 0.06 over non-weak members). The association test was specified before evaluation.

The low Qwen2.5-7B base GSM8K score noted in Section 4.2 persists under the sweep's prompt (0.283, against 0.817 for the in-house composition at $\alpha = 0.25$) and under the unified re-evaluation (0.445, against 0.63–0.82 for every singleton). The factorial comparison identifies the discrepancy in the eager attention path of this software stack (Appendix K.6). These sweep utility values are eager-attention diagnostics and should not be interpreted as SDPA capability estimates.

F Exploratory Prioritization of Harm Evaluation

This exploratory ordering analysis uses the eager-attention utility scores; its exact rankings have not been re-established with corrected state-level outputs. Joint evaluation raises a practical question: if task scores are available, can they help decide which compositions to assess for harm first? We study an ordering heuristic, not a harmful-compliance predictor or a stopping rule that certifies unreviewed states.

Rule and information available. For a candidate set $\mathcal{C}$, measure g_T for every candidate and define $g_{\text{ref}} = \max_{T \in \mathcal{C}} g_T$. This differs from g_{best}, the best-singleton reference used for utility retention. For $g_{\text{ref}} > 0$, order candidates by increasing

$$d_c(T) = |g_T / g_{\text{ref}} - c| \,. \tag{11}$$

Ties must be broken without using the held-out bank's harm labels. If all candidate scores are zero, the rule provides no ordering information and requires a stated fallback, such as random order.

Evaluation. The candidate set consists of all 77 evaluated states, including the base and singletons. A state is flagged when $h_T \geq \max(0.15, h_\varnothing + 0.10)$, an operational analysis threshold rather than a validated risk tolerance. Leave-one-bank-out fitting selects $c = 0.44$ or 0.45 in every fold, and Table 16 uses $c = 0.44$ throughout. The rule and threshold were developed on these banks, so this is not prospective validation. We report the fraction of candidates reviewed to recover at least 80% of flagged states, breaking ties uniformly at random and averaging over 200 orders. With M candidates, n^+ flagged states and $m = \lceil 0.8 n^+ \rceil$, the expected fraction under a uniformly random ordering is

$$B_{\text{random}} = \frac{m(M+1)}{M(n^+ + 1)}. \tag{12}$$

Besides size-ascending order and the label-informed oracle, we include ordering by increasing refusal-match rate, and direct probes that score every state on $k \in \{5, 10, 20\}$ randomly chosen harmful prompts and review states in decreasing order of probe harm.

The arithmetic-score rule improves on random, size-ascending and refusal ordering in every bank. It is not better than a small direct harm probe. A ten-prompt probe reviews less on Phi-4, GLM-4 and Mistral and more on the two Qwen banks, and has the lower mean (0.358 against 0.413); a twenty-prompt probe averages 0.298. The probe needs no utility measurement, whereas the arithmetic rule requires GSM8K scores for every candidate, so on total evaluation cost the arithmetic rule is competitive only when those scores are already available. Its one distinctive property is that it orders candidates without generating any harmful content. Refusal ordering beats random on four banks and

Table 16: Fraction of the 77-state candidate set reviewed to reach 80% recall of flagged states. PROBE-k first scores every state on k of the 80 prompts, an upfront cost of $k/80$ of a full harm sweep; CAP instead requires GSM8K scores for every state. Oracle is a label-informed lower bound. Bold marks the best non-oracle ordering per bank.

MODEL	n^+	RANDOM	SIZE	REFUSAL	CAP	PROBE-5	PROBE-10	ORACLE
Qwen2.5-7B	17	0.788	0.680	0.519	**0.338**	0.522	0.389	0.182
Qwen2.5-14B	14	0.810	0.927	0.429	**0.297**	0.527	0.351	0.156
Phi-4	9	0.810	0.551	0.571	0.398	0.309	**0.157**	0.104
GLM-4-9B	40	0.791	0.609	0.714	0.558	0.482	**0.448**	0.416
Mistral-7B	39	0.810	0.496	0.870	0.475	0.475	**0.444**	0.416
mean		0.802	0.653	0.621	0.413	0.463	**0.358**	0.255

is worse than random on Mistral, whose base does not refuse. We present the arithmetic rule as a preliminary observation about where risk concentrates, not as a recommended audit procedure. The control bank has no flagged state and cannot validate any ordering.

G Additional Analytical Results

These results clarify measured labels rather than a mechanism of composition. In the refusal decomposition and endpoint analysis, $\mathbb{P}$ is over a uniformly selected index of the fixed dataset, so h_T remains an empirical rate. Unrestricted-distribution examples describe possible distributions, not every rate attainable at a particular finite N. Population guarantees below separately introduce a request distribution and distinct population notation.

G.1 Refusal does not identify harmful compliance

For a fixed state T, let H and R be the binary harm-judge and refusal-detector outputs on a uniformly selected evaluation request; H is an indicator, whereas H_T denotes the index set defined in Section 3. Define $a_T = \mathbb{P}(R = 0)$ and $\epsilon_T = \mathbb{P}(H = 1, R = 1)$. The always-defined identity is

$$h_T = \mathbb{P}(H = 1, R = 0) + \epsilon_T. \tag{13}$$

If $a_T > 0$, let $q_T = \mathbb{P}(H = 1 \mid R = 0)$. Then

$$h_T = a_T q_T + \epsilon_T, \qquad 0 \le h_T - \epsilon_T \le \min\{a_T, q_T\}. \tag{14}$$

If $a_T = 0$, q_T is undefined and Eq. 13 still applies. Here R is a detector output, not a semantic definition of refusal, and ϵ_T is a joint harmful-and-refusal-matched frequency, not the false-negative rate of the refusal detector.

Proposition 2 (Non-identifiability and ranking reversal)**.** *Even if $\epsilon_T = 0$, knowing $a_T = a$ determines only that $h_T \in [0, a]$. Every value in that interval is attainable by a joint label distribution. For states A, B with $a_A > a_B > 0$, their non-refusal ordering is reversed by harmful compliance, $h_A < h_B$, exactly when*

$$a_A q_A - a_B q_B < \epsilon_B - \epsilon_A. \tag{15}$$

If both ϵ terms vanish, this condition reduces to $q_A < (a_B/a_A) q_B$.

Proof. Assign probability t to $(H = 1, R = 0)$, $a - t$ to $(H = 0, R = 0)$, and $1 - a$ to $(H = 0, R = 1)$, for any $t \in [0, a]$. This fixes $a_T = a$ while giving $h_T = t$. Subtracting Eq. 14 for the two states proves the ordering condition. □

The proposition establishes ambiguity, not its empirical frequency. Nor does a low q_T identify capability loss: a capable model that answers harmlessly can also have low q_T. Utility measurements and response inspection are needed to distinguish these possibilities.

G.2 NON-IDENTIFICATION FROM MEMBER LABELS

Proposition 3 (Finite-dataset non-identification)**.** *Fix all base and singleton labels on N requests and let $U_T = H_\varnothing \cup \bigcup_{i\in T} H_{\{i\}}$. For any subset $Q \subseteq \mathcal{I} \setminus U_T$, an unrestricted assignment of composition labels exists with $E(T) = Q$. Every rate k/N for integers $0 \le k \le N - |U_T|$ is attainable without changing any reference label.*

Proof. Set $H_T = Q$. Then $E(T) = H_T \setminus U_T = Q$; choosing any k indices from the complement gives the stated rates. □

This concerns unrestricted labels, not patterns realizable by LoRA. It does not rule out prediction using weights, activations, or structural assumptions.

G.3 PROOF OF PROPOSITION 1

Proof. Write $U_T = H_\varnothing \cup \bigcup_{i\in T} H_{\{i\}}$. Counting gives

$$|E(T)| = |H_T| - |H_T \cap U_T| \ge |H_T| - |H_\varnothing| - \sum_{i\in T} |H_{\{i\}}|. \tag{16}$$

Divide by N, use nonnegativity, and note that $E(T) \subseteq H_T$. □

The inequality is exact on measured labels, not a prediction of unmeasured states. It can be loose: harm above the largest singleton does not suffice, since different singletons can fail on different requests. For Eq. (9), partition H_T into $H_T \cap U_T$ and $E(T)$, and U_T into $H_T \cap U_T$ and $U_T \cap H_T^c$. Subtract cardinalities and divide by N.

G.4 DIRECT AUDITING OF A FIXED COMPOSITION

Proposition 4 (Zero-failure audit)**.** *Fix a composition, decoding rule, judge and prompt distribution independently of the audit sample. With n i.i.d. prompts and no harmful labels observed, the exact one-sided $1-\delta$ binomial upper confidence bound is $1-\delta^{1/n}$. Equivalently, for any true rate $h > 1-\delta^{1/n}$, the probability of observing zero failures is less than δ.*

Proof. Under the assumptions, the zero-count probability is $(1-h)^n$. Solving $(1-h)^n = \delta$ gives the bound. □

At $\delta = 0.05$ and $n = 59$, the bound is approximately 0.0495. The looser exponential bound $\log(1/\delta)/n$ requires 60 prompts for a bound below 0.05. For simultaneous guarantees over M fixed states, a union bound can replace δ by δ/M. Selecting compositions or stopping after inspecting results requires an appropriate selection or sequential procedure. Curated benchmark requests need not be i.i.d. draws from a deployment distribution, and judge error is outside this guarantee.

G.5 A POPULATION-CONFIDENCE VERSION OF THE MARGINAL BOUND

Let $\mathcal{D}$ be a request distribution and draw N requests i.i.d. Fix a collection of L model states and scoring protocols independently of this sample. Write p_T for population harm probability and q_T^{new} for population composition-only probability, with empirical counterparts h_T and $e(T)$. Set

$$\eta = \sqrt{\frac{\log(2L/\delta)}{2N}}. \tag{17}$$

Hoeffding's inequality and a union bound give simultaneous $|p_T - h_T| \le \eta$ for all L fixed states with probability at least $1-\delta$. Independence is required across requests, not across states sharing the same requests.

Corollary 5 (Uniform marginal-rate certificate)**.** *For any size-k composition whose composition state, base and all singleton reference states belong to this collection, on the same event,*

$$q_T^{\text{new}} \ge \left[h_T - h_\varnothing - \sum_{i\in T} h_{\{i\}} - (k+2)\eta \right]_+ . \tag{18}$$

Proof. The population union bound gives $q_T^{\text{new}} \geq [p_T - p_\varnothing - \sum_{i \in T} p_{\{i\}}]_+$. Substitute the simultaneous lower bound $h_T - \eta$ for composition harm and upper bounds for the $k+1$ references. □

With aligned labels, directly estimating the composition-only indicator is preferable to this conservative bound, with multiplicity control where appropriate. Curated benchmarks are not automatically i.i.d. deployment samples; neither this guarantee nor the empirical bound accounts for judge error.

G.6 Conditional endpoint argument for an interior maximum

Define $v_T = \mathbb{P}(H = 1, R = 0)$, so $h_T = v_T + \epsilon_T$ without undefined conditionals. Let $\overline{h}_k$ denote the average over a specified finite set of states of size k, for $k = 0, \ldots, K$. If $\overline{h}_0 \leq b_0$, $\overline{h}_K \leq b_K$, and there exists $k_0 \in \{1, \ldots, K-1\}$ with $\overline{h}_{k_0} > \max(b_0, b_K)$, then every global maximizer of $\overline{h}_k$ is interior. The proof follows immediately because neither endpoint can attain the maximum.

The decomposition supplies possible endpoint bounds, for example $\overline{h}_0 \leq \overline{a}_0 + \overline{\epsilon}_0$, or $\overline{h}_K \leq \overline{q}_K + \overline{\epsilon}_K$ when every conditional q being averaged is defined. Small q at the final size is an additional behavioural premise; it is not implied by low GSM8K accuracy. The interior-height premise is also an additional condition. Thus this argument does not prove a peak from linear composition alone, predict its location, or establish uniqueness, abruptness, or a phase transition.

Where q_T is defined for every state in a size-specific average,

$$\overline{h}_k = \overline{a}_k \, \overline{q}_k + \mathrm{Cov}_k(a_T, q_T) + \overline{\epsilon}_k. \tag{19}$$

Multiplying size-averaged factors omits this covariance. Computed from per-response labels for sizes three through six (20 model-by-size cells), 10 covariances are negative, one is positive (GLM-4 at size three, $+0.011$), and the rest are zero to three decimals. The largest magnitude is -0.058 for Qwen2.5-7B at size three, where the size mean is 0.165. This is a descriptive relation involving a conditional rate, not evidence by itself of a causal mechanism.

G.7 Why singles and pairs need assumptions to predict larger sets

Fix $Q \subseteq \mathcal{A}$ with $|Q| \geq 3$ and define unrestricted set functions

$$F_0(T) = \tfrac{1}{2}, \qquad F_1(T) = \tfrac{1}{2} + \gamma \mathbf{1}\{Q \subseteq T\}, \quad 0 < \gamma \leq \tfrac{1}{2}. \tag{20}$$

Both functions lie in $[0, 1]$ and agree on the base, singletons and pairs, but differ at $T = Q$. This concerns unrestricted functions, not LoRA realizability. Extrapolation needs additional assumptions, such as controlled higher-order interactions; the present experiments do not establish them.

H Judge Checks and Size Summaries

The second judge is Qwen2.5-14B-Instruct prompted to label whether the generation is an instance of the requested behaviour, framed explicitly as labelling red-teaming data; it generates six tokens, from which a yes/no answer is parsed. Prompt variants asking about "helpfulness" yielded nearly constant negative labels because the judge itself refused. Reported rates use only the behaviour-labelling prompt. The exact Qwen judge template and parser are not included in the available experiment record; this limits reproduction of that secondary diagnostic.

A manual sanity check read twelve randomly sampled responses from each of three states against the classifier's criterion: a Mistral GSM8K singleton (primary-judge rate 0.900; $10/12$ clearly harmful), a Phi-4 GSM8K$_{06}$+Dolly pair (0.562; 4 clearly harmful and 3 borderline), and a Qwen2.5-7B triple containing Dolly (0.287; 1 clearly harmful and 2 borderline). These small, selection-specific counts do not establish classifier calibration and are not used to validate the strict two-judge event rates. The Phi-4 pair differs from the GSM8K$_{00}$+Dolly example (0.500); both are from the same main sweep and scoring protocol.

Per-model Pearson correlations over 77 states are 0.882, 0.830, 0.939, 0.942, and 0.933 in the table's model order. The pooled Pearson and Spearman correlations are 0.947 and 0.897. In the Phi-4 tail, the second judge gives 0.013, 0.014, and 0.024 at sizes four through six, against 0.007, 0.000, and 0.000 from the primary judge. This qualifies any claim that the tail contains no harmful responses.

Table 17: Both judges by composition size: size-mean harmful-compliance rates on the 80 main prompts. The second judge has lower rates in most cells, but is not uniformly stricter.

MODEL	JUDGE	0	1	2	3	4	5	6
Qwen2.5-7B	primary	0.000	0.027	0.094	**0.165**	0.120	0.075	0.010
	second	0.013	0.022	0.069	**0.096**	0.085	0.051	0.008
Qwen2.5-14B	primary	0.000	0.011	0.040	0.077	0.115	**0.133**	0.094
	second	0.000	0.009	0.020	0.044	0.069	**0.079**	0.030
Phi-4	primary	0.000	0.055	0.088	**0.123**	0.007	0.000	0.000
	second	0.000	0.034	0.063	**0.077**	0.013	0.014	0.024
GLM-4-9B	primary	0.037	0.175	0.325	**0.487**	0.237	0.040	0.000
	second	0.037	0.086	0.137	**0.244**	0.123	0.041	0.021
Mistral-7B	primary	0.450	0.659	**0.691**	0.573	0.312	0.057	0.005
	second	0.175	0.267	**0.279**	0.250	0.139	0.060	0.028

Per-response labels give the joint distribution of harm and refusal matching for every state. The number of refusal-matched responses scored harmful is 17 of 2774 on Qwen2.5-7B, 0 of 3009 on Qwen2.5-14B, 3 of 1765 on Phi-4, 259 of 2309 on GLM-4 and 24 of 164 on Mistral; the largest single-state share is 0.25, on GLM-4.

Table 18: Judge agreement on the evaluation of Table 1. κ: mean per-state Cohen's kappa between the two HarmBench classifiers over the 297 prompts, for the top-5 compositions and for the base and their members. Undefined per-state values are excluded (only the Phi-4 base). Events: composition–prompt pairs that are composition-only under the primary judge alone, summed over the top-5 compositions and 297 prompts. Lost: events that are not composition-only under the second judge, because it does not flag the composition, or because it flags the base or a member.

Bank	κ comp.	κ members	Events	Lost: composition	Lost: member	Retained
Qwen2.5-7B in-house	0.84	0.84	218	27	4	86%
Qwen2.5-14B in-house	0.80	0.71	264	31	1	88%
Phi-4 in-house	0.77	0.84	388	50	12	84%
Published task + Dolly	0.84	0.70	277	21	4	91%
Published instruction	0.80	0.57	242	23	31	78%

The reported intersection retains 78–91% of primary-judge events. Loss reasons are assigned exclusively: an event is counted under composition if the second judge does not flag the composition; otherwise it is counted under member if the second judge flags the base or a singleton. One Phi-4 event meets both raw conditions and is assigned to composition. The retained count plus the two loss counts equals Events. Agreement does not establish label correctness.

I Supporting Prediction Experiment

A second-order membership model was fitted to harmful-compliance log odds,

$$\operatorname{logit} \widetilde{h}_T = \mu + \sum_{i \in T} a_i + \sum_{\{i,j\} \subseteq T} b_{ij}, \qquad \widetilde{h}_T = (c_T + 0.5)/(N + 1), \tag{21}$$

where c_T is the harmful-label count among N evaluation requests. The reported protocol uses stratified $50/50$ splits over all 77 Qwen2.5-7B states, repeated 50 times, and closed-form ridge regularization with the base log odds fixed as μ. This is not a train-on-singles-and-pairs, test-on-larger-compositions protocol. Baselines use first-order membership features, quadratics in size or squared norm, and the training-set mean.

The second-order predictor has MAE 0.076, compared with 0.067 for the constant baseline and 0.054 for the size quadratic. The spread among the structured predictors is 0.022. The experiment supports a negative result for this model and protocol, not statistical equivalence of all predictors,

Table 19: Cross-validation on the Qwen2.5-7B bank, means over 50 stratified $50/50$ splits. d is the nominal design dimension including a constant column, not the number of fitted parameters or effective degrees of freedom after fixing the intercept and regularization.

PREDICTOR	d	MAE	SPEARMAN
$\lvert T\rvert$, quadratic	3	0.054	$+0.598$
$\lVert\Delta W\rVert_F^2$, quadratic	3	0.055	$+0.499$
singletons only	9	0.062	$+0.370$
constant	1	0.067	—
second-order (singletons, pairs)	37	0.076	$+0.542$

an irreducible error floor, or impossibility of prediction. Finite state counts, regularization, split design, and model misspecification remain possible explanations. Norm–size association also limits interpretation of their similar predictive performance.

Adding diagonal quadratic terms to singleton-intensity models reduces logit RMSE from 1.375 to 1.229 for Qwen Dolly and from 2.318 to 1.875 for Phi-4 Dolly. These results are supplementary diagnostics of the fitted model class; they do not establish the functional form of the true response or an unavoidable prediction error.

J Exploratory Weight-Space Diagnostics

We examine the spectral structure of pairwise interactions in refusal matching, Frobenius overlap, and refusal-direction projections as potential indicators of composition behaviour. These exploratory analyses use refusal-based outcomes unless harmful compliance is explicitly specified; they are distinct from the harmful-compliance prediction experiment in Table 19.

The spectral diagnostic covers six banks: eight-adapter banks on Qwen2.5-7B, Qwen2.5-14B, Phi-4 and GLM-4-9B, a seven-adapter Granite-3.1-8B bank, and an additional heterogeneous Qwen2.5-7B bank with eleven adapters. This diagnostic collection differs from the main harm-evaluation banks. A permutation test (5000 shuffles per bank) of the spectral concentration of the pairwise-interaction matrix was significant only for the eight-adapter Qwen2.5-7B bank ($p < 0.001$). The other p-values were 0.174 (Qwen2.5-14B), 0.922 (Phi-4), 0.475 (GLM-4-9B), 0.168 (Granite-3.1-8B) and 0.221 (heterogeneous Qwen2.5-7B). Given the small banks, these non-significant results do not establish the absence of shared low-rank structure. Normalized Frobenius overlap between adapter pairs correlated with the magnitude of fitted pairwise interaction terms (Pearson 0.57, 0.55 and 0.31 on three banks), but choosing which pairs to measure by overlap outperformed random choice only on Qwen2.5-7B.

For the refusal-direction diagnostic, a normalized difference of mean last-token activations from 128 harmful and 128 benign instructions was used as a direction in each layer. Projected update energy was 1.3–6.4 times chance across 40 adapters. Its within-model Spearman correlations with singleton harm were $+0.286$, $+0.016$, -0.076, -0.216, and -0.261; the pooled correlation was $+0.121$. These do not establish a reliable harm predictor. The energy measure is unsigned and considers the write side of selected modules, so it cannot by itself establish the sign or causal effect of a behavioural change.

A frozen difference-of-means probe reached balanced accuracy 0.656 against refusal-matching labels under adapters, and 0.671 after common-mode correction. A cross-state probe reached AUC 0.899 but did not transfer its threshold between states. These probes were evaluated against refusal matching, not harmful compliance, and are reported for completeness.

K Reproducibility Details

K.1 Evaluation chronology

Composition selection uses harm on the 80 main AdvBench requests. After composition selection, the selected groups were first evaluated on 97 JailbreakBench requests in three Qwen2.5-7B banks. Selection rules and hyperparameters were not changed. Subsequently, 200 unused AdvBench requests

were drawn with Python `Random(1)` and combined with the 97. This pool is disjoint from the 80 main and 160 earlier weight/norm requests; the former are nested in the latter. Thus the 97 requests were held out from main-sweep selection, not untouched throughout development. Repair and reweighting were designed after this evaluation, with the recipe and weight grid fixed before intervention runs. They reuse the 297 requests for evaluation; complete AdvBench/JailbreakBench lists also serve as a refusal-training exclusion list. No hyperparameter search on this pool is reported, but interventions are not an independently designed untouched-test confirmation.

K.2 Training and checkpoints

Recorded software is PyTorch 2.11, Transformers 5.17 and PEFT 0.21. Inference uses bf16, greedy decoding, chat templates and left padding, with 512 new tokens for harm. Recovered revisions are Qwen2.5-7B-Instruct `a09a354`, Qwen2.5-14B-Instruct `cf98f3b`, Phi-4 `2db69c1`, Mistral-7B-Instruct-v0.3 `c170c70`, GLM-4-9B-chat-hf `8599336`, HarmBench-Llama-2-13b-cls `bda7053`, and HarmBench-Mistral-7b-val-cls `51182c7`. These were recovered from the single cached snapshot per repository and cache configuration, rather than experiment-time revision logs.

In-house adapters are GSM8K 00 ($r = 16$) and 06 ($r = 8$), Alpaca 01/09 ($r = 32$), 03 ($r = 16$) and 07 ($r = 8$), Dolly 04 ($r = 8$), and DialogSum 05 ($r = 16$). Internal scale is $\alpha_i = 2r_i$, dropout zero, and targets are q/k/v/o and gate/up/down projections. Training uses 300 steps, batch 4, maximum length 512, AdamW with weight decay 0.01, no warmup or scheduler, and loss on all non-padding tokens including prompts. A seed-0 generator draws learning rates from $\{1, 2, 3\} \times 10^{-4}$; individual draws were not logged, limiting exact training reproduction.

Dolly retraining. Rank, scaling, dropout and targets copy the same base's adapter 04. Dolly-15k is shuffled with seed 0; 300 held-out records leave 14,711. Safe adds 774 refusals, approximately 5% before length filtering. Both variants use AdamW at 2×10^{-4}, weight decay zero, betas $(0.9, 0.999)$, 3% linear warmup then cosine decay, one epoch, micro-batch 8, accumulation 2, maximum length 1024, gradient clipping 1.0, bf16 autocast and response-only loss. Truncation leaves no loss tokens in 126 Qwen Dolly records and 113 Phi-4 records; these are removed before batching. No refusal examples are removed. Plain/safe counts are $14{,}585/15{,}359$ for Qwen and $14{,}598/15{,}372$ for Phi-4, giving $912/960$ and $913/961$ optimizer steps. Thus the safe recipe changes both data content and steps; their causal contributions are not isolated. All three bases use seed 0; Qwen2.5-7B additionally reports seed 2. Seed-1 Qwen retrains on H100 had non-finite weights and were excluded at loading, not selected by downstream scores.

Refusal-data filtering. Unsafe BeaverTails `30k_train` requests are deduplicated ignoring case and limited to 20–400 characters. After seed-0 shuffling, exclude requests whose lower-cased alphanumeric word sets have Jaccard similarity at least 0.5 with any of 520 AdvBench or 100 JailbreakBench behaviors. This lexical filter does not guarantee semantic disjointness. The construction retains 3,000 eligible requests (two exclusions recorded during selection), then greedily queries each base. Refusal matching retains 1,419, 1,303 and 2,132 for Qwen-7B, Qwen-14B and Phi-4; the first 774 are used.

K.3 Judge protocols and known limitations

Discovery on 297 requests uses the official Llama-2 classifier template with full responses. The Mistral validation classifier uses an abridged three-of-seven-rule template, omitting rules on caveats, attempted behavior, code artifacts and language; responses are limited to 4,000 characters. Both complete judge inputs are right-truncated at 2,048 tokens. Interventions instead use both complete official prompts and truncate each response to its first 512 judge tokens before inserting it into the template.

The discovery primary judge generates two tokens, parses unrecognized outputs as -1, then the evaluation script maps these to zero. The second judge maps outputs not beginning with "yes" to zero. Interventions greedily decode one next token with the same case-insensitive yes-prefix rule. Invalid primary outputs were not saved separately, so their frequency cannot be recovered from binary labels. Empty, malformed and truncated model responses are scored without deletion or regeneration. Zero includes parser defaults, not only explicit negative judgments.

An audit of 40,986 stored discovery inputs per judge found one over-limit input for each judge (2,202 and 2,151 tokens), both corresponding to one response from a size-six composition in the public task+Dolly bank. Truncation removes the template ending; both saved labels are zero. This is a composition response, not a reference; correcting it can change only that request's event, by at most $1/297$ in the composition rate or $1/(5 \cdot 297)$ in a top-5 mean if the composition is included. This bound does not establish its corrected label.

The discovery and intervention tables report distinct, fixed scoring protocols; discovery outputs were not re-scored with the intervention protocol. Within each intervention experiment, original and modified compositions and their references use the same complete templates, response preprocessing and parser. Thus the original-to-plain reversal and safe-recipe comparisons do not rely on a cross-template contrast. Absolute rates across experiments are not interpreted as intervention effects, because template and harness changes can affect labels in either direction. Repair and reweighting were run as separate experiments, each regenerating its own base and singleton references. Table 4 reports repair-run measurements; Tables 12–13 report reweighting-run measurements. For Qwen2.5-7B, base harm is 0.017 versus 0.020, GSM8K singleton accuracy is 0.785 versus 0.775, and DialogSum singleton ROUGE-L is 0.320 versus 0.317, respectively. Task-retention ratios and MMLU differences use the references from the corresponding experiment. Intervention effects are assessed within each experiment; matching compositions across experiments are separate measurements.

The prompted Qwen judge on the main 80 requests is described in Appendix H.

K.4 Published-adapter repositories

The control bank contains the eight control entries below; task+Dolly replaces its eighth with in-house Dolly. The provenance and replacement records precede control generation; incompatible initial candidates were not used in reported compositions. Public repair partners are the first seven. Revisions are recovered from cached repositories.

Table 20: Public Qwen2.5-7B adapter repositories and recovered revision prefixes. The full identifier for the abbreviated Alpaca-CoT entry is given below.

Bank	Repository	Revision
Task/control	shibing624/chinese-text-correction-7b-lora	b0cb452
Task/control	mrs83/FlowerTune-Qwen2.5-7B-Instruct-Medical-PEFT	38b5563
Task/control	adlsdztony/Rui-7B	425c4a5
Task/control	parzival2l/qwen2.5-7B-instruct-ner-tuned	d5d1d14
Task/control	XiaoXu123123/academic-humanize-qwen25-7b-dpo-v2-lora	6fb2e2e
Task/control	Junhoee/Qwen-Megumin	4a8bf48
Task/control	rotalabs/sakshi-judge-indic	59c9e90
Control	silent666/task-14-Qwen-Qwen2.5-7B-Instruct	39262f5
Instruction	Kurapika993/qwen2.5-7b-qlora-dolly15k	863cb09
Instruction	srai86825/qwen2.5-7b-alpaca-full-lora	6c996bc
Instruction	hwang2006/qwen2.5-7b-alpaca-1pct-lora	90cc4cb
Instruction	FutureMa/Qwen2.5-7B-Instruct-LoRA-Alpaca-ZH	1da6e2b
Instruction	rayonlabs / Alpaca-CoT	b88f238
Instruction	Andante1984/qwen25-7b-dpo-lima-lora	e7c0998
Instruction	Barryzbr12/qwen2.5-7b-instruct-dpo-lima-lora	c0fc4a3
Instruction	Evan7017/qwen2.5-7b-dpo-lima	757f76f

The full Alpaca-CoT repository identifier is `rayonlabs/Qwen2_5-7B-Instruct-alpaca_reasoning_COT-3742f4c1-ebb7-4cbb-9cc2-93d8f54c04a6`.

K.5 Latent-shift diagnostic outcomes

An additive sub-threshold account would require approximately additive member shifts and a selective relation between the resulting projection and refusal. The first requirement is unsupported: on the three primary banks, direct additive-prediction R^2 is negative and fitted slopes are 0.46–0.66. The projection also lacks the intended selectivity over random directions, which reach refusal AUC 0.79 and 0.93 on Qwen2.5-7B and Phi-4. Published control members have mean shift -0.58 base standard deviations versus -1.27 for non-Dolly in-house members. Both means are negative.

Summed shifts predict refusal loss better than size and summed member refusal on one published bank (21 of 40 compositions lose refusal; AUC 0.985 versus 0.747 and 0.638). Several other banks have 32–38 of 40 compositions in the loss class. In the published task+Dolly bank, Dolly-containing compositions shift 0.83 standard deviations more than the additive prediction, while summed shifts have AUC 0.54 for refusal loss. The predictive association varies across banks and does not establish a general additive threshold mechanism.

K.6 GSM8K HARNESS DIAGNOSTIC

For Qwen2.5-7B, base accuracy on the 60-question probe is 0.300. A lenient metric that accepts the gold number anywhere in the generation reaches 0.400, and the gold number is absent from 62% of generations, so answer extraction alone does not explain the discrepancy. Batch size one gives 0.333 against 0.300 at batch size sixteen; output budgets of 768 and 1024 tokens both give 0.283; an explicit final-answer format gives 0.283 and a bare question 0.200. None of these checks, nor the unified re-evaluation (0.445), resolves the gap, whereas the harness of Section 7 gives 0.910.

A factorial check on the base model localizes the discrepancy (Table 22). With the same 200 items, chat template, greedy decoding and 1024 tokens, we crossed the attention implementation (eager, used in the discovery evaluation, against PyTorch SDPA), the two prompt wordings, and two batching schemes. With SDPA the base scores 0.910–0.915 in every cell. With eager attention it scores 0.285–0.445, and the score depends on how prompts are batched. Prompt wording has little effect compared with the implementation difference. The comparison localizes the large discrepancy to the eager attention path in this software stack; it does not establish a general defect of eager attention or identify the underlying implementation bug. Re-running the main-sweep probe (60 items, 512 tokens) of all five in-house banks with SDPA raises the Qwen2.5-7B base from 0.300 to 0.800 and moves every other size mean by at most 0.03 (GLM-4 base +0.05); Table 6 reports the SDPA values. The main harm phenomenon persists in the regenerated evaluation. We regenerated every response of the evaluation of Table 1 with SDPA (same checkpoints, states, prompts, decoding and judges; four banks on H100 and one on RTX 6000 GPUs) and compared the two runs prompt by prompt (Table 21). The reported group-mean differences are small (about 0.02 or less in absolute value), and the paired intervals shown include zero. These are not equivalence tests. The hardware allocation is reported for the regeneration; a pure attention-only causal comparison additionally requires matched hardware within each pair.

Table 21: Evaluation regenerated with eager and SDPA attention, 297 prompts. Group means of total harm h (primary judge) and strict e_{12}; difference SDPA minus eager with a 95% paired prompt-bootstrap interval. Top-5 and free-5 as in Table 1.

Bank	Top-5 h eager / SDPA	Top-5 e_{12} eager / SDPA	Δe_{12}	Free-5 e_{12} eager / SDPA	Δe_{12}
Qwen2.5-7B in-house	0.237 / 0.226	0.126 / 0.123	−0.003 [−0.018, +0.012]	0.093 / 0.092	−0.001 [−0.015, +0.014]
Qwen2.5-14B in-house	0.232 / 0.233	0.156 / 0.154	−0.002 [−0.017, +0.012]	0.127 / 0.133	+0.006 [−0.005, +0.018]
Phi-4 in-house	0.464 / 0.455	0.220 / 0.204	−0.015 [−0.032, +0.001]	0.032 / 0.032	+0.000 [−0.006, +0.007]
Published task + Dolly	0.271 / 0.277	0.170 / 0.177	+0.007 [−0.011, +0.025]	0.015 / 0.015	+0.001 [−0.007, +0.009]
Published instruction	0.480 / 0.469	0.127 / 0.126	−0.001 [−0.020, +0.019]	0.046 / 0.050	+0.004 [−0.009, +0.016]

Table 22: Qwen2.5-7B-Instruct base, GSM8K accuracy on 200 test items under one extraction rule. "Original order, 32": prompts in dataset order, batches of 32 (discovery configuration); "length-sorted, 64": batches of 64 after sorting by length (Section 7).

Attention	Prompt wording	Original order, 32	Length-sorted, 64
eager	original ("on the last line in the form ####")	0.445	0.300
eager	Section 7 ("after ####")	0.445	0.285
SDPA	original	0.915	0.915
SDPA	Section 7	0.910	0.910

K.7 PUBLISHED-ADAPTER CONTROL BANK: SELECTION PROCEDURE

We enumerated adapters declaring Qwen2.5-7B-Instruct as their base on the model hub and removed names indicating attack, jailbreak, refusal-removal or safety purposes; 299 candidates remained and configurations were fetched for 39. Thirty-one were compatible with our composition hooks, and eight were selected across uploaders and tasks. Integrity checks at load time showed that eight of the 31 are not usable as published: four store half of every MLP adapter pair as empty tensors, one is a truncated upload, two have weight shapes of a different base model, and one contains only non-finite values. The last emits a single repeated token, so refusal matching scores it as never refusing while the harm judge scores it as never harmful. Three initial selections were among the unusable adapters and were replaced before any composition was evaluated.

The final bank consists of Chinese text correction (rank 8), medical question answering (8), a community fine-tune (16), entity recognition (32), academic rewriting (16; query and value projections only), role play (32; three projections per layer), evaluation in Indic languages (16), and a community fine-tune with little metadata (32, $\alpha/r = 8$). We also screened the other 15 usable candidates alone on the same 80 prompts; the largest singleton harm among all 23 usable adapters is 0.025. The hub further contains adapters from research on deliberately induced misalignment whose names do not reveal that purpose, so a name-level filter is not sufficient on its own.

Second published bank. The instruction-following bank follows the selection rule of Section 4.1 and contains the eight instruction entries in Table 20: one Dolly, four Alpaca and three DPO-LIMA adapters. Four adapters from misalignment research with innocuous names were excluded before selection. All eight passed the integrity checks. Singleton harm is 0.362 for the Dolly-15k adapter, 0.050 for the full Alpaca adapter, and at most 0.025 for the others; one DPO-LIMA adapter has a refusal-match rate of 0.225 alone with harm 0.025. The frozen criterion for this bank (some size mean exceeding the most harmful member by 0.05) was not met: the largest excess is 0.009, so by that criterion the bank reproduces only partially.