ERASING THE NON-ADDITIVE RESIDUE IS INSUFFICIENT TO REPAIR OVER-SCALED MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging by task arithmetic can fail under over-scaling, but an internal non-additivity measure need not identify the cause of that failure. We isolate the layerwise non-additive residue between a merged model and the additive reference built from its single-task paths. In two-task merges, a numerically audited factorial ledger shows that this residue is transported, regenerated after early erasure, and coupled to output non-additivity through its direction. The same audit shows that naive bfloat16 estimates of local generation are dominated by rounding error. We then erase the residue in two tested over-scaled configurations, at 1.5B and 7B. Complete erasure at the latest tested boundary does not reliably restore the audited code scores, and persistent erasure at three late boundaries across every generated position also fails to restore the 1.5B merge. In a fixed 12-item code instrument, partial persistent erasure ($\lambda = 0.25$) raises the score in all three seeds while worsening expert-reference negative log-likelihood in all three. That dose was exploratory. Complete residue removal is therefore insufficient as a repair in these configurations. Without same-coefficient single-task controls we cannot tell whether these failures are specific to merging, and the partial-dose response precludes calling the residue behaviourally irrelevant. At 1.5B the output-side interaction ratio loses method-ranking information under a shared counterfactual, while corresponding state-space measures rank four audited methods at both scales.

1 INTRODUCTION

Task arithmetic adds finetuned weight displacements to a base model, sometimes combining skills and sometimes degrading them (Ilharco et al., 2023). Activation and parameter diagnostics describe different objects: SurgeryV2's representation bias contains marginal as well as interaction effects (Yang et al., 2024a), CTL measures departures from cross-task linearity (Zhou et al., 2024), and TSV measures parameter-subspace overlap (Gargiulo et al., 2025). Here we test a narrower candidate—the non-additive activation residue—as both a measurable structure and a repair target. Its removal can be evaluated directly, without assuming that every kind of task interference is a second-order interaction.

Does removing the measured residue repair an over-scaled merge?

Answering that question, rather than correlating with it, dictates the method. It requires an exact handle on one candidate non-additive quantity and an intervention that removes it while keeping

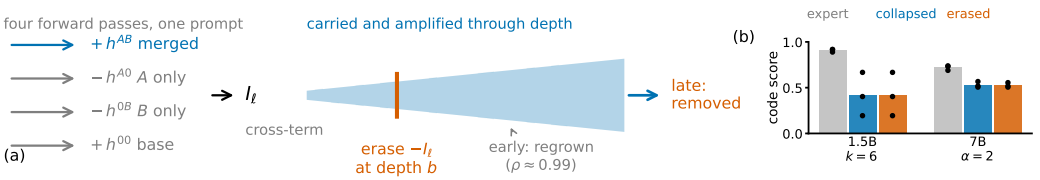


Figure 1: **Complete residue erasure does not reliably repair the tested failures.** (a) Four forward passes isolate I_ℓ. (b) At the latest tested boundary, full erasure leaves the code score unchanged in five of six seed–configuration cells (bars: seed means; dots: seeds); partial persistent erasure has a different, endpoint-dependent effect (Table 2).

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.

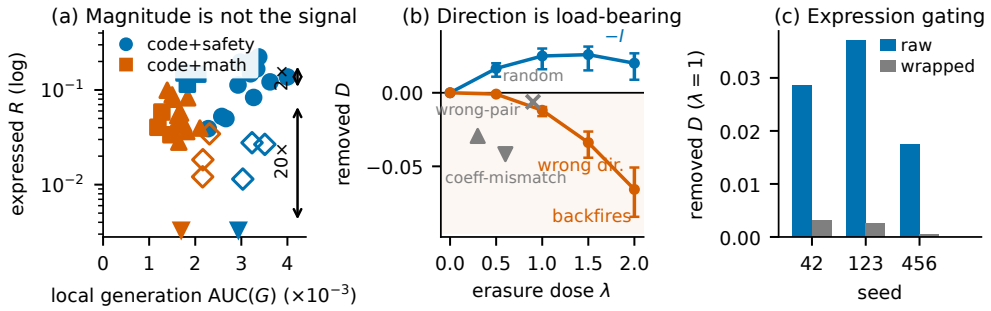


Figure 2: **Direction, magnitude, and format affect different measurements.** (a) Across 36 cells, local generation magnitude separates task pairs only coarsely and reverses across prompt formats. Colour gives the pair, marker the stratum (filled: controlled; square: raw public; open: wrapped). (b) Erasure along $-I_\ell$ saturates near $\lambda = 1$; a norm-matched cross-prompt direction backfires (error bars: seed range; grey: single-dose controls). (c) The Alpaca wrapper increases internal generation while suppressing expressed interference and the measured reduction after identical erasure 13-fold; this is not an independent identification of the format feature.

the merging weights fixed. So we take the most direct candidate the ledger can isolate exactly: the layerwise cross-term $I_\ell = h_\ell^{AB} - h_\ell^{A0} - h_\ell^{0B} + h_\ell^{00}$, the exact non-additive residue of merging, obtained from four forward passes over one prompt (Figure 1a). We track it through every block, separate what each block *generates* from what it *transports*, and then erase it surgically against norm-matched controls that hold the displacement fixed and destroy one structural property at a time.

The anatomy comes first, because a null result about a quantity is only informative once the quantity is known to be real. We therefore establish that the cross-term is structured and causally live, being transported, regenerated and direction-selective, with a two-assumption model deriving its angle law (Figure 2, §3–4). These measurements make the residue a substantive intervention target, without establishing it as the source of behavioural failure.

We next test erasure in two over-scaled configurations: six adapters at α=1.5 for 1.5B and three at α=2 for 7B. At the latest tested boundary (b=25 of 28 blocks), complete erasure does not recover the audited code score in the former and recovers at most 4% of the expert gap in the latter. At 1.5B, clamping three late boundaries throughout generation also fails to restore the score. Yet partial persistent erasure (λ=0.25) improves code scores in 3/3 seeds while worsening expert-reference NLL in 3/3. Our conclusion is therefore about an intervention: *complete removal of the measured residue is insufficient to repair these failures*. It does not identify which marginal displacement causes them, nor imply that the residue has no behavioural effect at other doses.

A separate diagnostic audit shows why the choice of readout matters. At 1.5B, an output-side interaction ratio reverses method ranking under method-specific counterfactuals; under a shared counterfactual the reversal disappears but the signal nearly vanishes. At 7B the reversal does not reproduce. State-space quantities rank the four audited methods in both scale checks (§6.1). These are convention- and regime-dependent empirical findings, not a general inversion law.

Contributions. We make three bounded empirical contributions. **First,** a numerically audited factorial ledger isolates the activation residue and separates local generation, transport and marginal-state mismatch (§2–3); the bf16 audit exposes a large rounding floor. **Second,** norm-matched interventions show that erasure changes output non-additivity in a direction- and depth-dependent way, while complete erasure does not reliably repair the tested over-scaled behavioural failures; partial persistent erasure produces an endpoint-dependent improvement (§4–5). **Third,** output- and state-space diagnostics respond differently to prompt format and merging method, with the output ratio losing ranking information under a shared reference at 1.5B (§6). These results establish a limitation of residue removal as a repair, not a unique causal explanation of collapse.

2 ISOLATING THE NON-ADDITIVE RESIDUE: A FACTORIAL LEDGER AND ERASURE

The intuition first. Merging two task vectors and running a prompt gives four natural models, namely base, each single task, and the merge. The *interference* at any point in the computation is exactly what the merge has that the two single tasks do not add up to. Our method is nothing more than taking this four-way difference seriously at every layer, which makes the candidate exact rather than a proxy, and then removing it surgically, which turns "where does collapse live" into an experiment rather than a correlation.

2.1 SETUP

We follow the training and measurement protocol of Zhu et al. (2026) exactly. Base model Qwen2.5-1.5B (Qwen Team, 2024) (weights pinned to a single commit), six tasks (math, code, instruction, safety, summarization, translation (Cobbe et al., 2021; Taori et al., 2023; Ji et al., 2024; Narayan et al., 2018; Tiedemann, 2012)) finetuned as rank-16 LoRA (Hu et al., 2022) task vectors with response-only loss, three training seeds each. Task vectors are norm-matched to the global core-median norm before composition, and merged models $\theta_0 + \alpha\Delta_A + \beta\Delta_B$ are evaluated on a 6×6 interior grid of $(\alpha, \beta) \in (0, 1.2]^2$. Each prompt stratum contains 60 fixed prompts, and the principal hidden-state quantities are read at the final prompt token. The coefficient grid has 7×7 points including axes, of which 6×6 are interior. Our running example throughout is the code+safety merge evaluated on code prompts. Every confirmatory run executes on a single A100 with deterministic attention kernels and TF32 disabled, for the reason §2.4 makes concrete.

2.2 LAYERWISE LEDGER: CARRIED VS. GENERATED

For a prompt x, let $h_\ell^{00}, h_\ell^{A0}, h_\ell^{0B}, h_\ell^{AB}$ be the residual state entering block ℓ under the four settings of Figure 1. The **cumulative cross-term**

$$I_\ell \;=\; h_\ell^{AB} - h_\ell^{A0} - h_\ell^{0B} + h_\ell^{00} \tag{1}$$

is everything non-additive accumulated before block ℓ. To separate what block ℓ *adds* from what it *transports*, we evaluate all four block functions at one **common reference state** $\bar{h}_\ell = h_\ell^{A0} + h_\ell^{0B} - h_\ell^{00}$ (the additive reconstruction) and take the same four-corner difference:

$$G_\ell \;=\; F_\ell^{AB}(\bar{h}_\ell) - F_\ell^{A0}(\bar{h}_\ell) - F_\ell^{0B}(\bar{h}_\ell) + F_\ell^{00}(\bar{h}_\ell). \tag{2}$$

Intuitively, G_ℓ is the cross-term block ℓ would create from scratch on interference-free input. The exact ledger $I_{\ell+1} = G_\ell + P_\ell$ then attributes the remainder P_ℓ to transport of what was already carried; §3.1 splits P_ℓ exactly. Two structural identities hold exactly and serve as built-in sanity checks. Here $I_1 = 0$ (LoRA leaves the embedding untouched) and $I_2 = G_1$, both verified to machine zero in every run.

Output-side estimands. Following Zhu et al. (2026), expressed interference on a prompt batch is the interaction ratio R: the mean Jensen–Shannon divergence between the merged model's next-token distribution and its no-interaction additive prediction (logit-space $\ell^{A0} + \ell^{0B} - \ell^{00}$), normalized by the mean main-effect JSD. This is a ratio of means, never a mean of ratios. The causal effect of an intervention is

$$D \;=\; R_{\text{original}} - R_{\text{patched}}, \tag{3}$$

with the *unpatched* main-effect denominator frozen in both terms, so D moves for exactly one reason. Contrasts are paired differences of these summaries across conditions sharing prompts, evaluated with prompt-paired bootstrap CIs and 3/3 seed-direction agreement, seeds being the outer replication unit.

2.3 CROSS-TERM ERASURE AND ITS CONTROLS

The intervention replaces the merged model's residual state at a fixed boundary with $h_\ell^{AB} - \lambda I_\ell$: at $\lambda = 1$, the carried cross-term is exactly erased and the state is restored to the additive reconstruction. The ladder $\lambda \in \{0, 0.5, 1, 1.5, 2\}$ probes dose dependence. The claim "the *direction* of I_ℓ is what

matters" is only as strong as its controls. Every control therefore preserves $\|I_\ell\|$ per prompt while destroying exactly one structural property. Each control's overlap $\cos(\delta, -I_\ell)$ is reported, so direction-mediation is checked rather than assumed:

- **wrong-pair**: the additive state borrowing the second axis from the *other* task pair (kills: "any task-structured direction works");
- **coefficient-mismatch**: the $(+1, -1, +1)$ combination in the same affine span (kills: "any span-consistent displacement works");
- **cross-prompt**: the additive state of a *donor* prompt (fixed length-binned derangement), applied as a full dose ladder (kills: "prompt identity is irrelevant");
- **distance-matched random** (kills: "perturbation per se helps").

Interventions run in float32. Damage is audited both at the first token (top-1 agreement, entropy shift, JSD from unpatched) and over prefill-patched greedy generation on both endpoint strata.

2.4 The estimator is a catastrophic cancellation, and the gates caught three real errors

The four-corner difference is a catastrophic cancellation, since the four paths differ at $O(\alpha)$ while I_ℓ is $O(\alpha\beta)$. Every estimator in this paper therefore carries its own noise floor and gates: (i) a *determinism null* (four forwards at identical coefficients must agree bitwise, enforced at run time); (ii) a *propagated-ULP floor* (half-ULP representation jitter on the four stored states, propagated in float64 through the cancellation; any layer below its floor is marked uninterpretable); (iii) finite-difference probes carry step-halving convergence checks; (iv) the exact identities of §2.2. The gates earned their place before any conclusion depended on them, catching an all-bf16 mixed-derivative estimate that was pure noise, a floor computation silently zeroed by float32 rounding, and ultimately the bf16 precision finding itself.

Every hypothesis, criterion, layer position, dose ladder and control construction was fixed before the corresponding data existed, in a versioned log with a full change record; falsified predictions, including two of three headline expectations, are reported as falsified with their prior confidences. Verdicts, immutable per-cell results and the full chain are released.

3 The candidate is real (I): transport dominates, and magnitude is not the signal

Before interpreting erasure, the residue has to be established as a numerically valid object. We start where the field's metrics do, with size. How much cross-term does each layer produce, and does that quantity carry the functional differences?

Measured naively at training precision, local generation looks like a universal constant. The result dissolves under audit, and instructively so. Across all twelve (pair $\times$ stratum) conditions of our main sweep (1,296 cells spanning two task pairs, six prompt strata, three seeds, and the full composition grid), the bf16 full-depth generation summary AUC(G) occupies a band of 0.0101–0.0116 (per-pair ratios 1.02–1.15$\times$), including the geometrically most conflicting and most orthogonal pairs. The precision audit below attributes this uniformity to a task-independent quantization floor. Recomputed with float32 blocks at the single $(1, 1)$ coefficient point (36 pair–stratum–seed cells, all 28 blocks), the picture is different. The fp32 band spans 3.4$\times$ (0.0012–0.0040, falsifying our own $<2\times$ expectation), the pair ratio is 1.81–1.90$\times$, *coarsely tracking* the 2$\times$ output difference as every magnitude quantity in this paper turns out to (§6), and generation under instruction wrappers is 1.7–1.8$\times$ *larger* than raw, 3/3 seeds.

The output side sets the standard these numbers must meet. The two pairs differ by roughly 2$\times$ in expressed interference (family scores 10.33% vs. 5.31%) and, causally by erasure, 14$\times$–337$\times$ in removable interference (§4); the same prompts differ by roughly 20$\times$ across formats ($+6.91$pp raw vs. $+0.34$pp wrapped) (Zhu et al., 2026), in the *opposite* direction from generation. Generation magnitude is at best a coarse correlate of pair structure and is inverted on the format axis. We call the assumption that it carries the functional differences the **magnitude fallacy** (Figure 2a).

A cautionary result: naive bf16 estimates of G are mostly round-off. Recomputing G_ℓ on the *same* reference state with the block cast to float32 shrinks it to 14–23% of its bfloat16 value, with directions nearly uncorrelated across precisions (median cosine 0.15–0.24): 75–90% of the naively measured "local generation" is bfloat16 weight-rounding roughness, which is task-independent and therefore exactly why the band looks universal. Paired *contrasts* survive, since the rounding floor cancels in differences, and all causal results are computed in float32. Appendix A states the caution in full.

3.1 Transport dominates: an exact decomposition, plus the injection bounds

The accounting identity $I_{\ell+1} = G_\ell + P_\ell$ leaves P_ℓ mixing transport of the existing term with state–parameter coupling, so we decompose it *exactly*. Evaluating all four blocks (float32) both at the common reference $\bar{h}_\ell$ and at the four actual path states regroups the actual-state four-corner term, bitwise, as

$$I_{\ell+1} \;=\; G_\ell \;+\; \underbrace{F_\ell^{AB}(h_\ell^{AB}) - F_\ell^{AB}(\bar{h}_\ell)}_{T_\ell \text{ (transport of the existing } I_\ell)} + M_\ell, \tag{4}$$

where M_ℓ bundles the marginal-path state mismatches (the state–parameter coupling channel). Here $G_\ell + M_\ell$ is precisely the one-step regeneration a block produces when I_ℓ has just been erased. Over the 36 fp32 cells: transport carries $\sim$69% of the layerwise flux (regeneration share 0.31, 3/3 seeds), late-network transport exceeds regeneration by $2.3\times$, and *transport alone amplifies*, at $\|T_\ell\|/\|I_\ell\| = 1.08$ per late block, 3/3 seeds (Appendix X, Figure 4). A replication on Llama-3.2-1B reproduces both, at transport share 65%, late gain 1.14, 3/3 seeds. Transport dominates the measured flux at both tested family points; broader family invariance is not established. The coarser arithmetic exclusion agrees. If layers contributed independent increments of the measured $\|G_\ell\|$, cumulative growth is bounded by 0.088 (incoherent) and 0.369 (perfectly aligned), against an observed $\|I_{28}\| = 0.974$, which is $2.6\times$ even the coherent bound. Consistently, smooth local generation is small and flat, with mixed-derivative stencils at five centers of the composition rectangle level to within $1.4\times$, in agreement with CTL's approximate-linearity observation (Zhou et al., 2024), while relocating everything that observation leaves unexplained into propagation.

Refuted conjecture: we had predicted (55% prior) that mixed-derivative mass concentrates off-diagonal. Measured edge-to-diagonal ratios of 1.005–1.010 (max 1.37, zero blocks above the $2\times$ threshold) falsify it.

4 The residue affects a first-token non-additivity metric directionally

Section 3 shows that magnitude alone does not explain the measured pair and format contrasts. Here we test whether residue direction matters for the first-generated-token JSD interaction ratio R. This is an intervention on that metric, not a demonstration of task-performance repair. We intervene at five fixed depths with cross-term erasure and the four norm-matched structural controls of §2.3, with all three predictions fixed before implementation.

Erasure works; every norm-matched control fails or backfires. At $\lambda = 1$ on code prompts (code+safety, mean over five depths and three grid points), erasure removes roughly 19% of expressed interference, a depth-average the regrowth analysis below decomposes, and the specificity contrasts are positive against *every* control, 3/3 seeds (Table 1). The wrong-pair and coefficient-mismatch patches in fact have *negative* D: an equally large, equally structured displacement in the wrong direction makes interference worse. Conditioning on each control's overlap $\cos(\delta, -I_\ell)$, the effect is proportional to the displacement's projection on the erasure direction (e.g., at depth 14: correct $\cos = 1.00$, $D = +0.028$; wrong-pair $\cos = 0.25$, $D = +0.007$), so we claim *cross-term-direction specificity*, the direction being the mediator-facing variable, rather than task-identity specificity beyond it.

Dose–response dissociates direction from displacement. Along the correct direction, $D(\lambda)$ rises and saturates near $\lambda \approx 1$. Along the norm-matched cross-prompt direction it *worsens monotonically* (Figure 2b). No account in which perturbation per se helps or hurts survives both curves. In these

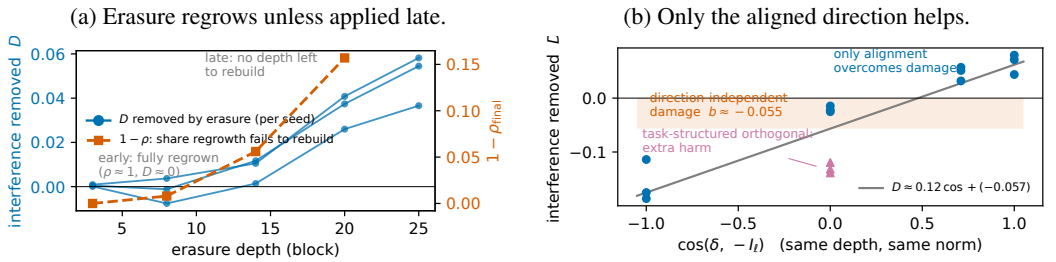


Figure 3: **Erasure dynamics for the first-token interaction ratio.** (a) Early erasure is followed by recovery of the measured residue coordinate; blue trajectories are individual seeds, orange is the unrecovered share. (b) At fixed depth and displacement norm, blue dots are individual seeds (three per direction), pink triangles mark the task-structured orthogonal control, and the grey line is a mean fit; it is not a pointwise law or a behavioural effect.

tested arms, direction and dose interact: the correct direction reaches its largest measured benefit near $\lambda \approx 1$, where the residue is fully erased.

Matched magnitudes, $14\times$–$337\times$ apart in removable interference. On the very code prompts where the two merges' local generation differs by at most $1.9\times$ in float32 (§3), erasure removes substantial interference from code+safety and almost nothing from code+math (Table 1, last row), giving pair ratios of $13.9/337/29.5$ (the middle value rides a near-zero denominator; in *unnormalized* JSD the reductions differ by 18–$52\times$). The pairs are matched on local generation but *not* on the carried term, where late-layer $\|I\|$ differs $2.4\times$ (0.65 vs. 0.27), disclosed because §4.3 says that is what matters. Scale still cannot produce the gap. At depth 25 erasure removes 56% of expressed interference from code+safety and -8% from code+math, a change of *sign*. The paper's central quantitative contrast: the pair difference lives in the orientation of the carried cross-term, far beyond what its size can explain.

Removal is not destruction. Erasure at $\lambda = 1$ leaves first-token behavior almost untouched, at top-1 agreement 0.94–0.99, patched-versus-unpatched JSD (0.003) two orders below the main effect, and 92–99% of continuation tokens preserved.

4.1 Erased early, re-created: the cross-term is not stored

Propagation undoes erasure except where depth runs out (cf. emergent self-repair, McGrath et al., 2023; Rushing & Nanda, 2024; distinguished in Appendix K). Self-repair is known to corrupt first-order ablation scores (Gong et al., 2026); our early-depth null is such a case, reported as regeneration rather than inertness. If interference were *stored* in the cross-term, erasing it should remove it for good. A follow-up measurement (window and priors fixed before data) tracked the cross-term downstream of a λ=1 erasure, expecting partial regrowth ($\rho = \|I^{\text{post}}\|/\|I^{\text{orig}}\| \in (0.3, 0.95)$ at the output). The window was *falsified from above*: erase at depth 3 or 8 and the network rebuilds the cross-term essentially exactly, at latest-tested-boundary $\rho = 1.00/0.99$ and direction cosine 0.99 to the original cross-term (the frozen direction-memory prediction, $\cos > 0.5$, held with room to spare). Regrowth only falls off when depth runs out, with $\rho = 0.94$ erasing at 14 and 0.84 at 20. This profile interlocks with the causal effect. Decomposing D by patch depth, early erasures achieve nothing ($D \approx 0$ at depths 3 and 8, 3/3 seeds) while late erasures carry the entire effect ($+0.026$ to $+0.058$ at depths 20–25). Removable interference is exactly what the remaining depth cannot rebuild (Figure 3a). The reading we draw (formed after these results): the carried cross-term is *continuously re-established from the untouched marginal paths* rather than stored. What returns is the cross-term coordinate, not the state, and a matched-displacement control finds no restoring force specific to this direction, and the perturbed trajectory never rejoins the unpatched one. This sharpens §3.1's exclusion into an interventional statement. Propagation does not merely amplify the carried term, it actively reconstructs it in norm and direction. That yields a design corollary. Erasure-style mitigation must act late, where reconstruction cannot catch up.

Table 1: **Cross-term erasure is direction-specific: the correct direction removes interference, every norm-matched control fails or backfires** (λ=1, code prompts, code+safety, mean over 5 fixed depths $\times$ 3 grid points; per-seed values). Brackets: 95% prompt-bootstrap CIs (5,000 paired draws/seed) from the bitwise-reproducing per-prompt re-run; all 15 exclude zero.

	seed 42	seed 123	seed 456
D(correct)	$+0.0301$ [.021, .042]	$+0.0289$ [.023, .037]	$+0.0160$ [.013, .020]
D(correct)$-D$(wrong-pair)	$+0.0556$ [.042, .074]	$+0.0587$ [.044, .077]	$+0.0487$ [.040, .059]
D(correct)$-D$(coeff-mismatch)	$+0.0699$ [.053, .091]	$+0.0719$ [.056, .091]	$+0.0583$ [.048, .069]
D(correct)$-D$(cross-prompt)	$+0.0460$ [.038, .055]	$+0.0391$ [.032, .047]	$+0.0257$ [.021, .031]
D(correct)$-D$(random)	$+0.0375$ [.027, .050]	$+0.0355$ [.029, .044]	$+0.0206$ [.017, .025]
D(correct), code+math	$+0.0022$	$+0.0001$	$+0.0005$

4.2 Orthogonal controls: a direction term over a damage intercept

Target and estimand share the additive counterfactual, so the reduction might reflect *any* displacement toward the additive neighbourhood rather than direction. Holding depth and norm fixed and varying *only* the angle to $-I_\ell$ excludes that reading: orthogonal displacements, one random and one task-structured, *increase* R ($D = -0.015$ to -0.139), while D varies monotonically with $\cos(\delta, -I_\ell)$ across five angles (3/3 seeds; Appendix T, Figure 3b). A norm-matched steepest-descent arm *backfires* (-0.17 to -0.65, 9/9) while $-I_\ell$, nearly orthogonal to it, helps, so what matters is the low-gain direction the quadratic model favours rather than loss alignment. The sweep decomposes the effect into $D \approx a\cos(\delta, -I_\ell) + b$ with slope $a = 0.09$–0.14 and *negative* intercept $b \approx -0.055$, an average negative intercept in the tested angle sweep, not a pointwise direction-independent law. The depth profile agrees: a construction-driven effect would appear at every depth, yet $D \approx 0$ at depths 3–8, in proportion to $1 - \rho$. The coupling's falsifiable consequence also fails. If moving the state onto $\bar{h}_\ell$ guaranteed moving the logits onto the additive prediction, erasure would help every pair at every depth; it does not, because $F(\bar{h}_\ell) \neq a$, and for code+math at depth 25 exact erasure *increases* expressed non-additivity in two of three seeds. A construction artifact cannot change sign. *Statistical basis, disclosed:* the frozen verdict rests on 3/3 seed-direction consistency; a per-prompt re-run reproducing every immutable scalar bitwise supplied the bootstrap, and 27 of 28 intervals exclude zero. A two-level seed$\times$prompt bootstrap agrees ($D_{\text{correct}} = +0.018$ $[+0.010, +0.031]$).

4.3 A local model of the angle response

A first-order expansion of the logit map at $\bar{h}_\ell$ (the displacement is 1.5–14% of the state norm) and a quadratic approximation to JSD give $D(\delta) = -2\langle u_0, J\delta\rangle_M - \|J\delta\|_M^2$. The quadratic contribution is non-positive for every displacement, but the sign of the full expression depends on the linear term. A cosine regression law follows only after an additional mean-zero assumption for residual directions; it is not a pointwise identity. The tested $-I_\ell$ direction is nearly orthogonal to steepest descent (cosine 0.03–0.12) and has low readout gain ($|b|/\|JI_\ell\|^2 = 1.15$–1.28). The model fits the frozen dose ladder with a two-parameter parabola ($R^2 = 0.99$–1.00); separate dose and angle fits estimate the same $\langle u_0, JI_\ell\rangle$ within 7–13%. This is an explanatory fit, not proof of a unique mechanism: the angle fit is imperfect and the measured $p_b = 1.75/1.63/1.71$ falls below the idealized value 2 (Appendix S).

5 Complete residue erasure is insufficient in the tested collapse cells

Behavioural interventions reveal a partial-dose exception. For these k-task merges, the erased quantity is the aggregate residue relative to the base plus all k single-task paths, not just the two-task factorial term used above. At 1.5B (k=6, α=1.5), one-time complete erasure at block input 25, the latest tested boundary of 28, leaves the 12-item code score unchanged in all three seeds (Table 2). At 7B (k=3, α=2), one-time erasure yields no positive recovery above 4% of the expert gap; the latest tested boundary worsens one seed by 7.4% and leaves two unchanged (Appendix O). Persistent complete clamping at three tested boundaries, during prefill and at every generated position, also fails to consistently repair the 1.5B code scores. This does not imply that the residue is behaviourally

irrelevant: persistent *partial* erasure (λ=0.25) improves code scores in all three seeds while worsening expert-reference NLL in all three. Complete erasure is therefore insufficient as a repair intervention in these cells, not evidence for a unique cause of collapse. A same-coefficient single-task control is still needed to distinguish merge-specific failure from over-scaling of an individual task.

Table 2: **Complete versus partial residue erasure at the 1.5B collapse cell.** Code scores average the fraction of unit tests passed within each of 12 coding items (not binary pass@1); seed order is 42/123/456 throughout this paper. Single erasure changes the last prompt-token state once at block input 25; persistent erasure clamps prefill and every generated token at one or three tested inputs. Expert-reference NLL change is relative to no patch; positive is worse. A dash means this comparison is not reported here.

Intervention	Code 42	Code 123	Code 456	ΔNLL (42/123/456)
No patch	.403	.194	.667	0/0/0
Single λ=1, input 25	.403	.194	.667	—
Persistent λ=1, input 25	.250	.389	.333	—
Persistent λ=1, inputs 14/20/25	.250	.389	.333	—
Persistent λ=0.25, inputs 14/20/25	.500	.639	.681	+.202/+.127/+.009

The quarter-dose code changes computed from Table 2 are approximately $+0.097/+0.445/+0.014$ (seed 42/123/456; differences of displayed rounded scores); this exploratory three-seed pattern has no equivalence or significance claim. The behavioural instrument is small (12 fractional-test code items; 8 binary math items), and the unpatched code scores vary substantially across seeds. The same-coefficient single-task baseline and token-level output-difference audit for the one-time collapse-cell patch have not been run; the $+I$ arm demonstrates behavioural sensitivity of the shared hook, but does not substitute for those checks. In the original coefficient-one endpoint audit, one seed's base code score (0.958) already exceeded its single-task expert (0.917); these instruments detect degradation relative to that expert but do not certify expert advantage over the base (Appendix P).

6 What existing metrics see, and what they miss

Can the metrics the field already uses *predict* interference for unseen pairs and *explain* the causal structure of §4–5? Across 15 pairs of six tasks (6 calibration, 9 held-out) we compare CTL linearity and residual, parameter cosine, direction features and cumulative magnitude against measured interference, fitting only feature *signs* in calibration.

Neither linearity nor magnitude separates what matters. CTL's central observation replicates *more strongly than reported* (cosines 0.93–0.98, every family), but precisely because it is uniform it cannot explain why one pair interferes at 2× another or a wrapper changes expression 20×; parameter cosine fares little better ($\rho = +0.25$ held out). The magnitude family sees coarse pair structure and nothing else: SurgeryV2's bias orders the pair contrast (2.9×, 3/3) yet moves between -1% and $+34\%$ on the format bridge while expression collapses ∼95%, and cumulative $\|I\|$ predicts held-out pair ordering at $\rho = +0.63$–0.65 (directional, not significant). Both stay blind on formats, where every magnitude points the *wrong way*; *our own estimand belongs on this list*, since it too fails to rank merging methods (§5).

The family splits in state space. The metric-family disagreement, in the audited cells: in the collapse cell, SurgeryV2-style bias and a CTL-style state residual, the *same* relative non-additivity as R read before the readout, rank the methods *correctly* (-0.8 to -1.0, 3/3 seeds) while the per-method R convention ranks them backwards; under a common reference its ranking is uninformative. The per-method inversion is absent under a common reference, so it cannot by itself localize a cause to the readout map (§4.3). A 7B gate (GSM8K-test, executed HumanEval) confirms the state-space ranking (-1.0, both endpoints) and flattens R's range to 0.06 in the deepest cell, the interventional counterpart of the correlational finding of Cao et al. (2026). Our own direction features failed, the best ($\rho = +0.62$) being magnitude in disguise, so direction's causal role rests on intervention.

6.1 THE READOUT INVERSION, ITS SCALE BOUNDARY, AND HOW TO STEP AROUND IT

The method audit has three distinct output-side outcomes. At 1.5B, per-method counterfactuals produce an outright inversion: R ranks the best method last. Holding the counterfactual common in the same collapse cell removes the inversion but also the information (range 0.006–0.015, Spearman $-0.4/0/+0.4$). At 7B the inversion does not reproduce and R ranks methods correctly in seven of nine cells, but in the deepest-degradation cell its range collapses to 0.06 while the state-space measures remain perfectly ordered. We therefore use **readout inversion** for the sharp 1.5B instance, not as a scale-invariant law; the broader failure is regime-dependent loss or reversal of method signal. Read the same relative non-additivity before the readout map, using SurgeryV2-style bias or the CTL-style state residual at a late boundary, and it ranks the four methods correctly in both scale audits and on both 7B real benchmarks. The serialization gate is separate in scope and strengthens at 7B, where it closes completely. These scope conditions from §6 and §5 carry over unchanged.

6.2 FORMAT DEPENDENCE OF THE OUTPUT RATIO

On the same 60 public code prompts, Alpaca wrapping suppresses the first-token output interaction contrast reported by Zhu et al. (2026) from $+6.91$pp to $+0.34$pp, while local generation increases (3/3 seeds). Two additional plain-text instruction frames unseen during adapter training also suppress the ratio; ChatML and a length-matched non-instruction prefix do not. Thus training exposure is not the identified cause, and the decisive serialization feature remains unknown.

In the audited two-task case, wrapping inflates the main-effect JSD denominator about 13-fold (0.025 to 0.32), without a corresponding shrinkage of absolute interaction JSD. The reduction in the ratio after identical residue erasure is likewise 13-fold smaller under wrapping (3/3 seeds); this comparison shares the format-sensitive denominator and is not independent evidence for an internal gate. At six tasks the format effect reverses (Appendix U). These are first-token measurement results, not claims about generation quality or behavioural repair.

Positioning. Prior diagnostics presuppose an account of where interference lives; this paper tests that account. Appendix B states the distance to CTL, SurgeryV2, TSV, APL, RAIN-Merging and the concurrent intervention-validity work.

7 DISCUSSION, LIMITATIONS, AND CONCLUSION

Limitations and scope. The evidence spans three model points and two families but only rank-16 LoRA, supporting structural replication rather than universality across dense fine-tuning, architectures, or adaptation ranks. Prompt-bootstrap intervals are within-seed; three seeds give coarse exact tests, and one 3/3 gate has $p{=}1/8$. Controlled prompts diagnose output non-additivity, while GSM8K and HumanEval anchor the target failure behaviourally; neither makes non-additivity reduction a general repair objective. Finally, the 7B criterion fires in one of three seeds (Appendices U, Q), precluding a seed-invariant null. Our claim is therefore about the audited collapse cells, not every possible merging failure.

Conclusion. The non-additive residue is exactly defined and its generation, transport, and measured regrowth have structure. Complete erasure leaves five of six audited one-time code scores unchanged; the sixth worsens, and persistent complete erasure at 1.5B does not consistently repair the tested failure. Yet persistent quarter-dose erasure improves code scores in three of three seeds while worsening expert-reference NLL in all three. It is a causal determinant of expressed non-additivity, but in over-scaled collapse its removal is not sufficient to repair: the failure persists under this specified additive-state intervention, without identifying which component causes it. Because both cells are reached by over-scaling, the single-task control at the collapse coefficient is needed to separate the target task's own displacement from cross-task leakage, and we have not run it. What a merge carries, what its readout expresses, and what controls its failure are three different questions.

REPRODUCIBILITY STATEMENT

All hypotheses, thresholds, layer positions, dose ladders, and control constructions were frozen before the corresponding data existed, in a versioned preregistration (six freeze versions, each superseding

change logged); all fourteen predictions and their outcomes, three verdict documents, 1,296+90+54 immutable per-cell results plus the 54 per-prompt intervention records and their bootstrap-CI report, the second-family replication preregistration, verdict, and its 54+54 cells, the frozen regrowth follow-up and its 18 cells, the frozen SurgeryV2-bias comparison and its 36 cells, the frozen scale point and its 54+54 cells, the frozen transport-decomposition/fp32-grid experiment and its 36 cells, the intervention-divergence measurement and its 6 cells, the frozen merging-method audit with its metric-family, common-counterfactual, and 7B real-benchmark replications (36+36+12+51 cells and four verdicts), the frozen collapse-regime behavioural loop with its small-dose ladder and 7B second-parameterization replication (78+54+63 cells, three verdicts), the multi-position readout check (3×60 prompts), the frozen steepest-descent control (27 cells) and three-layer reanalysis with their verdicts, the numerical-validity audits (determinism nulls, ULP floors, bitwise equivalence audits), the exact software stack, and the pinned base-model commit are released with the code. Every figure is generated by a script whose numbers come from the immutable cell records or the frozen verdict tables, never retyped.

Statement on AI usage

Generative AI tools assisted with drafting and editing the paper, literature discovery, dataset reformatting, experiment-code implementation, data-analysis scripting, and interpreting results. The authors reviewed AI-assisted text and tested generated code. Hypotheses, predictions, and decision rules were specified or approved by the authors before the corresponding data existed. Reported measurements derive from immutable experiment records, and released scripts regenerate the figures from those records. The authors reviewed the theoretical arguments and mathematical claims and take responsibility for the final text, code, results, and other artifacts.

References

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Mei-Yen Chen, Thi Thu Uyen Hoang, Michael Hahn, and M. Saquib Sarfraz. Position: Pause recycling loras and prioritize mechanisms to uncover limits and effectiveness. In *arXiv preprint arXiv:2506.13479*, 2025a. URL https://arxiv.org/abs/2506.13479.

Yan-Lun Chen, Yi-Ru Wu, Chia-Yi Yang, et al. Layer-aware task arithmetic: Disentangling task-specific and instruction-following knowledge. *arXiv preprint arXiv:2502.20186*, 2025b.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.08099.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Zitao Fang, Guodong Du, Shuyang Yu, Yifei Guo, Yiwei Zhang, Jing Cao, Jie Ma, Sim Kuan Goh, and Ho-Kin Tang. Disentangling task interference within neurons: Model merging in alignment with neuronal mechanisms. *arXiv preprint arXiv:2503.05320*, 2025.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.00081.

Ting Gong and Shitan Xu. Twistedmerge: Certified higher-order diagnostics and abstention for model merging. *arXiv preprint arXiv:2607.20887*, 2026.

Zhiren Gong, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026. URL https://arxiv.org/abs/2607.01940.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

Yifeng Guo, Jin-Hong Du, and Xiang Chen. Beyond importance: Interchange-sobol sensitivity reveals task-specific content channels in transformer components. *arXiv preprint arXiv:2606.20678*, 2026. URL https://arxiv.org/abs/2606.20678.

Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. Localize-and-stitch: Efficient model merging via sparse task arithmetic. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=9CWU8Oi86d.

Stefan Horoi, Guy Wolf, Eugene Belilovsky, and Gintare Karolina Dziugaite. From memorization to parameter interference: How overtraining experts harms model merging. *arXiv preprint arXiv:2506.14126*, 2025.

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

Zhehao Huang, Yuhang Liu, Baijiong Lin, Yixin Lou, Zhengbao He, Hanling Tian, Tao Li, and Xiaolin Huang. RAIN-merging: A gradient-free method to enhance instruction following in large reasoning models with preserved thinking format. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.22538.

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

Jiaming Ji, Donghai Chen, et al. PKU-SafeRLHF: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*, 2024.

Abdallah Khemais. A theory of conditional collapse under low-rank weight-space ablations: I. the single-block theory and synthetic validation. *arXiv preprint arXiv:2608.03620*, 2026a. URL https://arxiv.org/abs/2608.03620.

Abdallah Khemais. Cross-layer interaction under weight-space ablation: A closed-form attention jacobian bound and a test on a real pretrained model. *arXiv preprint arXiv:2608.03629*, 2026b. URL https://arxiv.org/abs/2608.03629.

Fanshuang Kong, Richong Zhang, Zhijie Nie, Ziqiao Wang, and Qiang Sun. Rethink the evaluation protocol of model merging on classification task. *arXiv preprint arXiv:2412.13526*, 2024a.

Fanshuang Kong, Richong Zhang, Ziqiao Wang, and Yongyi Mao. Activated parameter locating via causal intervention for model merging. *arXiv preprint arXiv:2408.09485*, 2024b.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.04959.

Michael S. Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2111.09832.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, 2018.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.12827.

Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Defeating the training-inference mismatch via FP16. *arXiv preprint arXiv:2510.26788*, 2025.

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

Pratik Ramesh, George Stoica, Arun Iyer, Leshem Choshen, and Judy Hoffman. Resolving interference (RI): Disentangling models for improved model merging. *arXiv preprint arXiv:2603.13467*, 2026.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *International Conference on Machine Learning*, 2024.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the KnOTS. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.19735.

Wenju Sun, Qingyong Li, Wen Wang, Yang Liu, Yangliao Geng, and Boyang Li. Towards minimizing feature drift in model merging: Layer-wise task vector fusion for adaptive knowledge integration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Derek Tam, Yash Kant, Brian Lester, Igor Gilitschenski, and Colin Raffel. Realistic evaluation of model merging for compositional generalization. In *International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2409.18314.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *LREC*, 2012.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *International Conference on Machine Learning*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2203.05482.

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

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xingwei Wang, Xiaocun Cao, Jie Zhang, and Dacheng Tao. Surgeryv2: Bridging the gap between model merging and multi-task learning with deep representation surgery. *arXiv preprint arXiv:2410.14389*, 2024a.

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

Kotaro Yoshida, Yuji Naraki, Takafumi Horie, Ryotaro Shimizu, and Hiroki Naganuma. DisTaC: Conditioning task vectors via distillation for robust model merging. *arXiv preprint arXiv:2508.01148*, 2026. URL https://arxiv.org/abs/2508.01148.

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

Han-Chen Zhang, Zi-Hao Zhou, Mao-Lin Luo, Shimin Di, Min-Ling Zhang, and Tong Wei. DC-Merge: Improving model merging with directional consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.06242.

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

Zhanpeng Zhou, Zijun Chen, Yilan Chen, Bo Zhang, and Junchi Yan. On the emergence of cross-task linearity in the pretraining-finetuning paradigm. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.03660.

Chencheng Zhu, Xiaoyang Li, and Taotao Cai. When do task vectors interfere? mapping the validity boundaries of weight-space composition. *arXiv preprint arXiv:2608.09490*, 2026.

A THE BF16 CAUTION IN FULL

Activation-difference analyses run at training precision can be dominated by quantization roughness while passing casual sanity checks (cf. the BF16 training–inference mismatch in RL fine-tuning, cured by FP16; Qi et al., 2025). Because the rounding floor is condition-independent it cancels in paired contrasts, which is why the pair and format contrasts of §3 survive the precision audit while the absolute band does not.

B RELATED WORK

Relation to prior measurement work. We reuse three things from Zhu et al. (2026) and should be read against that work. The training and measurement protocol is theirs, used here without modification: the adapters, seeds, prompt strata, and the 7×7 composition grid with its aggregation order. The output-side estimand R is theirs, including the ratio-of-means convention. And the format result our expression-gating section builds on is theirs: on these same 60 public code prompts, wrapping in an Alpaca template collapses expressed interference from $+6.91$pp to $+0.34$pp. That study establishes the phenomenon at the output; it does not open the model.

What is new here is everything that requires the hidden states. The four-state factorial ledger isolates the layerwise residue I_ℓ exactly, which no output-side measurement can do, and the identity $I_{\ell+1} = G_\ell + T_\ell + M_\ell$ splits it into local generation, transport, and marginal mismatch. That split yields the numerical-validity result—naive bfloat16 estimates of G are 75–90% round-off—which applies to activation-difference analyses generally and has no counterpart in that work. It also makes the quantity *interveneable*: the erasure experiments, their norm-matched controls, dose ladders and angle sweep, and the collapse-cell verdict all act on I_ℓ directly and are new. Finally, the mechanism-level reading of the format gate is new even though the gate is not: we show the suppression is a denominator effect and that the same relative non-additivity read one step upstream does not suffer it. The honest summary is that Zhu et al. (2026) measure expressed interference and map where that measurement is valid, while this paper asks what the underlying quantity causally does, and finds that complete residue erasure is insufficient to repair the tested over-scaled failures.

Method landscape. Merging is improved by weight averaging and by resolving parameter-space conflicts, and diagnosed by predictors that vary with method, scale and evaluation protocol; Appendix C surveys both lines. All of it presupposes an account of where interference lives, which our results test and revise. "Direction" here is three claims: *parameter-space* (TSV, DC-Merge), *activation-space* (the orientation of I_ℓ, measured) and *causal* (what intervening does, tested). The third is load-bearing.

Linearity of features under weight interpolation. Cross-Task Linearity (CTL; Zhou et al., 2024) is the closest prior work, observing that layerwise features of merged models are approximately linear combinations of the endpoint models' features. Our I_ℓ *is* the residual of that approximation, refined further in float32 (cosines 0.93–0.98). The difference is in the intervention: CTL establishes an approximate state-space linearity; we test how removing its residual changes an output non-additivity metric and the separate behavioural endpoint. CTL contains no intervention, no format factor, and no propagation analysis.

Format sensitivity of merged models. RAIN-Merging (Huang et al., 2026) documents, at the behavioral level, that merging disrupts instruction-format handling and repairs it in the null space. Expression gating is the mechanism-level counterpart: the wrapper closes an output gate on internally amplified interference. They repair behaviour; we measure a format-dependent dissociation between hidden and output interaction summaries (§5).

Validity of activation patching. Activation-level counterfactuals carry two known caveats: patched states may interact with unpatched components (Vaidyanathan et al., 2026), and interventions can push states off the natural distribution (Grant et al., 2026). Every effect we report is the response to a specified, norm-matched intervention with its overlap disclosed, and at $\lambda{=}1$ the patched state is exactly $\bar{h}_\ell$; Appendix M reports the displacement magnitudes and the context-sensitivity checks Grant et al. (2026) ask for.

C The merging method and diagnostic landscape

Task arithmetic and model merging. Task vectors compose skills by weight addition (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023). A large literature improves merging by weight averaging (Wortsman et al., 2022; Matena & Raffel, 2022; Yang et al., 2024b), by resolving parameter-space conflicts (Yadav et al., 2023; Chen et al., 2025b; He et al., 2024; Gargiulo et al., 2025; Stoica et al., 2025; Marczak et al., 2025; Fang et al., 2025), by repairing representation bias post hoc (Yang et al., 2024a), or by minimizing input-subspace interference, feature drift and singular-space inconsistency (Cheng et al., 2025; Sun et al., 2025; Ramesh et al., 2026; Zhang et al., 2026). Task-vector norm disparities are themselves a failure mode (Yoshida et al., 2026), which is why our endpoints are norm-matched. Diagnostically, collapse is predicted correlationally by representational incompatibility rather than parameter conflict (Cao et al., 2026), task information survives in merged weights (Wang et al., 2024), and mergeability predictors vary by method (Zhou et al., 2026), degrade with overtraining (Horoi et al., 2025; Gong & Xu, 2026) and can be forecast from early training (Tang et al., 2026). Evaluation itself is contested: gains shrink under compositional testing (Tam et al., 2025), classification-time protocols conflate representations with classifier mismatch (Kong et al., 2024a), and a position paper argues for mechanisms over new recycling algorithms (Chen et al., 2025a). We take that position literally. All of this *presupposes* an account of where interference lives, which our results test and revise. "Direction" here is three claims: *parameter-space* (TSV, DC-Merge), *activation-space* (the orientation of I_ℓ, measured) and *causal* (what intervening does, tested). The third is load-bearing. APL (Kong et al., 2024b) intervenes on parameters; ours on the cross-term itself. Exact interaction decompositions under weight-space ablation exist in other settings (Khemais, 2026a;b), but not for merging, and not paired with an audit of what the decomposed quantity controls. The conflation we attack is recognized elsewhere. Guo et al. (2026) show that a single importance score merges two roles, transporting content and being necessary for the computation, and separate them with paired interventions on Qwen2.5-1.5B; their finding that replacement and deletion are not interchangeable is one reason we replace the state with the additive reconstruction rather than zeroing it. Our question differs in object: not which role a component plays, but whether a derived quantity carries a specific failure at all.

D Comparison against prior interference quantities

E Full estimator and aggregation definitions

All definitions below are transcribed from the frozen design document (§F); none was altered after data existed.

Table 3: **No prior quantity is causal, directional, and format-aware; none explains the 2×/20× structure** (empirical evaluation in §6). Rows: CTL (Zhou et al., 2024), SurgeryV2 (Yang et al., 2024a), TSV (Gargiulo et al., 2025), NeuroMerging (Fang et al., 2025), APL (Kong et al., 2024b), RAIN-Merging (Huang et al., 2026). Mag./Dir.: magnitude vs. direction quantity; Interv.: causal intervention on the quantity itself. Only CTL, parameter cosine, and SurgeryV2's representation bias are *measured* in our conditions (§6, Appendix L); their empirical columns report what we found. "n/e" marks quantities we did *not* evaluate, so those blanks are not negative results, and the remaining methods were never designed for the format axis. The held-out correlation for our own quantity is directional and not significant, and the table compares only how each quantity behaves on the 2×/20× structure: it is not a claim that ours diagnoses collapse, which §5 shows it does not.

	Mag./Dir.	Interv.	Format	2× pairs	20× formats	Held-out
CTL	mag.	—	—	uniform	no	—
SurgeryV2	mag.	—	—	2.9×	no	—
Parameter cosine	dir. (param.)	—	—	weak	no	ρ=.25
TSV	dir. (param.)	—	—	n/e	n/e	n/e
NeuroMerging	mag.	—	—	n/e	n/e	n/e
APL	mag.	param.	—	n/e	n/e	n/e
RAIN-Merging	(behav.)	repair	✓	n/e	behav.	n/e
Ours	dir. (activ.)	✓ (4 ctrl.)	✓	14–337×	13× gate	.63–.65 (n.s.)

Four weight-state paths. For task vectors Δ_A, Δ_B and coefficients α, β, we instantiate four weight states M_{00}, $M_{\alpha 0}$, $M_{0\beta}$, and $M_{\alpha\beta}$ and, for a prompt x, record the residual state at the input of block ℓ under each: h_ℓ^{00}, $h_\ell^{\alpha 0}$, $h_\ell^{0\beta}$, and $h_\ell^{\alpha\beta}$. All hidden analyses use the *last prompt token* only, aligning with the first-generated-token logit proxy of Zhu et al. (2026) and avoiding position-averaging bias between raw and wrapped sequences of different lengths.

Cumulative cross-term (descriptive only). $I_\ell = h_\ell^{\alpha\beta} - h_\ell^{\alpha 0} - h_\ell^{0\beta} + h_\ell^{00}$. Because I_ℓ contains every nonlinearity produced and propagated by blocks $1..\ell{-}1$, depth curves of I_ℓ are frozen as descriptive: they cannot by themselves locate where cross-terms are *generated*.

Common-state local generation (KT1 primary object). With F_ℓ^{ab} the full block-ℓ map (residual update included) of weight state M_{ab}, and the additive reference $\bar{h}_\ell = h_\ell^{\alpha 0} + h_\ell^{0\beta} - h_\ell^{00}$,

$$G_\ell(\bar{h}_\ell) = F_\ell^{\alpha\beta}(\bar{h}_\ell) - F_\ell^{\alpha 0}(\bar{h}_\ell) - F_\ell^{0\beta}(\bar{h}_\ell) + F_\ell^{00}(\bar{h}_\ell).$$

Because all four block paths receive the identical input, G_ℓ isolates the four-corner term contributed by the combined weights at the current reference state. The exact ledger $I_{\ell+1} = G_\ell + P_\ell$ defines the *state-conditioned remainder* P_ℓ, which is frozen as *not* interpretable as pure propagation (it mixes propagation of existing interaction, reference-state mismatch, state–parameter coupling, and higher-order block nonlinearity).

Profiles, AUCs, and companions. The primary layer profile is $g_\ell^{\mathrm{RMS}} = \mathrm{RMS}(G_\ell)$ with full-depth AUC as the headline summary (early-depth AUC over normalized depth $[0, 0.5]$ is a frozen secondary). The functionally weighted companion applies $\phi(z) = W_U\,\mathrm{FinalNorm}(z)$ to each corner before differencing; untuned logit-lens is confirmatory only at normalized depth $d \geq 0.5$, descriptive earlier. The two frozen contrasts are $C_{\text{pair}} = \mathrm{AUC}(G; \text{code+safety}) - \mathrm{AUC}(G; \text{code+math})$ and $C_{\text{format}} = \mathrm{AUC}(G; \text{raw}) - \mathrm{AUC}(G; \text{wrapped})$.

Output estimand and intervention effect. R is the ratio-of-means convention of Zhu et al. (2026) (never mean-of-ratios):

$$R = \frac{\mathrm{mean}_x\ \mathrm{JSD}\left(p_x^{\alpha\beta}, p_x^{\mathrm{add}}\right)}{\mathrm{mean}_x\ \mathrm{JSD}\left(p_x^{\alpha\beta}, p_x^{00}\right)}, \qquad D_\ell^{(q)} = R_{\mathrm{original}} - R_{\mathrm{patched},\ell}^{(q)},$$

with the denominator of R_{patched} *frozen to the unpatched main effect*—a preregistered choice: the patched model's own main effect also moves, and a self-normalized ratio would confound two

movements in one number. Unnormalized-JSD shrinkage is a clearly-separated descriptive secondary. R/D are formed once from the concatenated prompt set (micro-batch averaging of D is forbidden by the freeze).

Aggregation order and statistics. Track A computes every hidden estimand on the same norm-matched 7×7 composition grid's 36 interior points, prompts, seeds, and aggregation order as Zhu et al. (2026); KT2 uses the frozen subset $\{(1,1),(0.6,1),(1,0.6)\}$ and five frozen boundaries $\{3,8,14,20,25\}$. Seeds are the preregistered outer unit (3/3 direction agreement required); within seed, paired prompt bootstrap with 5,000 draws (frozen RNG seed 20260712) yields the CIs reported for KT1 contrasts.

F PREREGISTRATION TIMELINE AND THE FOURTEEN PREDICTIONS

Freeze chain. The design document was versioned v1.0–v1.6; every version's full text is retained verbatim, and any change to hypotheses, gates, units, doses, statistics, or the decision tree produced a new version rather than a silent overwrite. Table 4 lists the chain. Versions v1.0–v1.4 all predate any hidden-state result. v1.5 is explicitly a *post-KT1* revision—it does not rewrite the KT1 verdict (archived under v1.4 rules) but freezes the hypotheses and pass criteria for the *next* experiments (KT2, direction features, CTL comparison, adversarial arm) before any of their data existed, and adds the "stable reversal" rule the original decision tree lacked. v1.6 adds the bf16 dtype audit with its reading rules frozen before the audit ran. A separate point-prediction memo (P1–P6) was frozen after two pilot cells but before the 1,296-cell sweep launched.

Version	Content of the change
v1.0	Initial freeze (2026-08-02): hypotheses, estimators, units, KT1–KT3 gates, decision tree.
v1.1	D denominator fixed to unpatched main effect; determinism null split from ULP floor; patch-quantization audit added. Pre-data.
v1.2	Hardware lock (single A100, `sdpa`, deterministic, TF32 off); precision lock (bf16 KT1/KT3, fp32 KT2). Pre-data.
v1.3	Batching lock (micro-batch 8, global pad width); aggregation order; equivalence-audit requirement. Pre-data.
v1.4	Per-stage execution manifests (restores the frozen staged order; relaxes no gate). Pre-data.
v1.5	Post-KT1: PK2-1/2/3 frozen before KT2 implementation; F1–F3 and CTL-table analysis preregistration; adversarial arm; stable-reversal rule.
v1.6	bf16 G dtype audit with frozen reading rules and prior (55% "G real").
P1–P6 memo	Point predictions with priors, frozen before sweep launch.

Table 4: The preregistration freeze chain. Full documents and sidecars are released with the code.

The fourteen predictions. Table 5 lists the original study's fourteen frozen predictions with their prior confidences and outcomes: 7 held, 4 failed, 3 split. Counting every preregistered prediction across this document and the follow-up preregistrations (E12–E26) gives 81 gates, each frozen with its own hash sidecar and judged in a released verdict document. The additionally frozen dtype-audit prior (55% that bf16 G was real signal) also fell on the wrong side (§3); it is listed separately as an audit belief rather than an experimental prediction.

G NUMERICAL VALIDITY AUDITS

The four-corner difference is a catastrophic cancellation: the four paths differ at $O(\alpha)$ while I_ℓ is $O(\alpha\beta)$. Every estimator therefore carries its own validity chain, all items frozen before confirmatory execution.

Determinism null (must be exactly zero). Pointing all four weight states at the same model and running four independent forwards must produce a *bitwise-zero* four-corner difference under deterministic kernels—this validates determinism only and is explicitly *not* a noise floor (under

#	Prediction (frozen)	Statement	Prior	Outcome
1	P1-direction	$C_{\text{pair}} > 0$, both strata, 3/3 seeds	70%	**held**
2	P1-magnitude	AUC ratio in 1.3–2.0	70%	failed (1.02–1.15)
3	P2	math-stratum contrast $< \frac{1}{2}$ of code, still positive	high	**held**
4	P3	raw $G >$ wrapped G, ratio > 1.5	60%	failed (*reversed*, 3/3)
5	P4	generation share 4–8%	high	split (carried-dominance holds; window wrong: mean 12.6%)
6	P5	$\cos(G, B)$ depth profile	high	split (shape yes; variance far larger)
7	P6	all numerical gates green	high	**held** (1,296/1,296)
8	PK2-1	erasure direction-specific vs. all controls	65%	**held** (3/3)
9	PK2-2	expression gating: $D_{\text{raw}} \gg D_{\text{wrapped}}$	55%	**held** (3/3, 13×)
10	PK2-3	pair contrast carried by direction, ratio > 1.3	50%	**held** (14×–337×)
11	CTL separation	no CTL/TSV metric separates pair or format ordering	70%	split (cumulative $\|\|I\|\|$ *does* order pairs; format still invisible)
12	F1–F3	$\geq$1 direction feature passes held-out gate	45%	failed (0/3; F1 magnitude-contaminated)
13	Adversarial band	extreme-geometry pairs stay in G band	60%	**held**
14	Off-diagonal mass	edge stencil $\|B\| > 2\times$ diagonal	55%	failed (max 1.37)

Table 5: All frozen predictions and outcomes. "Split": qualitative claim held, quantitative window failed. No outcome was rewritten after data; reversals are reported as falsifications with the reversed reading tested by *new* frozen predictions (rows 8–10).

a deterministic backend it is identically zero and can prove nothing about resolution). All 1,296 confirmatory cells passed with exact zeros.

Propagated-ULP cancellation floor (the real noise floor). The resolution limit is set by storage precision: we perturb each of the four stored hidden states by $\pm\frac{1}{2}$ ULP representation uncertainty and re-form the four-corner difference, reporting per-layer floor RMS (mean/p99/max) under frozen draws and RNG seed. One implementation subtlety is instructive: propagating the jitter in float32 *silently returns zero*—sub-ULP perturbations round back to the original value 100% of the time—so the propagation is done in float64 and validated by the exact 2^{15} ratio between fp32 (23 mantissa bits) and bf16 (8 mantissa bits) floors. At the main composition point $(1, 1)$ the signal-to-floor ratio is min 10.2 / median 38.2 / max 92.7 across confirmatory layers; layers below ratio 1 would be flagged *below numerical resolution* and barred from mechanistic interpretation (none were). Boundary 0 (embedding) is excluded as a definitional zero, not a resolution failure.

First-block identities. With LoRA leaving the embedding untouched, $I_1 = 0$ exactly and $I_2 = G_1$ to within the ULP floor. Both identities are asserted per cell; they jointly validate the hook boundary, the reference-state construction, and the block-map definition.

Bitwise equivalence audit of performance refactors. Single-GPU feasibility required three optimizations—capture-and-replay direct block invocation, prompt micro-batching, and (initially) `logits_to_keep=1`. Each had to prove semantic invariance *on the locked A100* (kernel behavior is part of the estimand; equivalence does not transfer across hardware). The audit caught two real issues before any confirmatory data existed. (i) `logits_to_keep=1` changes the lm-head GEMM shape and hence accumulation order, producing a 1-ULP deviation (max 0.0625 at $|\text{logit}| \in [16, 32)$) against the full-sequence path; the optimization was removed rather than tolerated. (ii) `output_hidden_states` returns the *final-norm-transformed* state at the last boundary, not the last block's raw output—mixing spaces in the ledger (observed direct-block discrepancy 67.94). Fixed with a forward hook capturing the raw output, a uniform `residual_state()` accessor, and a regression test locking the contract. After both fixes, all equivalence checks are exactly 0.0 at every audited depth.

The bf16 dtype audit (frozen reading rules). Recomputing G_ℓ on the identical $\bar{h}_\ell$ with block weights and arithmetic cast to float32: fp32/bf16 RMS ratio median 0.230 (code+safety) and 0.136 (code+math), minima 0.073/0.040; direction cosines median 0.239/0.152, minimum -0.031. Both frozen reading rules place this in the "G is predominantly weight-rounding artifact" branch (§3). A related earlier catch: an all-bf16 mixed-derivative stencil—which extracts an $O(\epsilon^2)$ signal from $O(\epsilon)$ differences, a worse cancellation than G itself—produced pure noise (≈50× the true value), exposed by its own frozen gates (probe convergence ≈ 1.0 everywhere; $|B| > 50|G|$ is physically impossible). All stencils were moved to fp32 block evaluation with their own floors. The general lesson we would offer practitioners: *every new activation-difference estimator needs its own propagated noise floor and convergence check, or it will discover noise*.

H CONTROL CONSTRUCTIONS AND THE CROSS-PROMPT DERANGEMENT

At each frozen boundary $\ell \in \{3, 8, 14, 20, 25\}$ the intervention replaces the merged model's residual state h_ℓ^{AB} with $h_\ell^{AB} + d_\ell^{(q)}$ and continues the forward pass in float32. The correct erasure has displacement $d^{(\text{correct})} = -\lambda I_\ell$ per prompt. Every control is *row-wise norm-matched*: its raw displacement is rescaled per prompt to $\|I_\ell(x)\|$, so all conditions apply identically-sized perturbations and differ only in structure.

- **Identity/no-op** ($\lambda = 0$): validates that the hook mechanism itself leaves the output bitwise unchanged.
- **Wrong-pair**: the additive reconstruction built with the *other* task pair's second axis, $d = (h^{A0} + h^{0B'} - h^{00}) - h^{AB}$, where B' is a third adapter (for code+safety the math axis and vice versa). Preserves the additive construction; destroys task-pair identity.
- **Coefficient-mismatch**: the frozen wrong-coefficient template $(+1, -1, +1)$ in the same affine subspace, $d = (h^{A0} - h^{0B} + h^{00}) - h^{AB}$, norm-matched. Same subspace, same prompts, wrong combination.
- **Cross-prompt**: the correct additive displacement computed on a *donor* prompt x' and applied to x. Donors come from a deterministic length-binned derangement frozen before implementation (master seed 20260802): prompts are binned by token length, and within each bin a derangement guarantees no prompt donates to itself while keeping donor and recipient lengths comparable. This is the direct null for prompt-conditioned structure and the second arm of the dose ladder.
- **Distance-matched random**: per-prompt Gaussian direction scaled to $\|I_\ell\|$; frozen as a secondary sensitivity control, never a substitute for the structured nulls.

Overlap accounting. For every condition we report $\cos(d^{(q)}, -I_\ell)$ and interpret specificity contrasts conditional on it (frozen rule §8.3/§8.8 of the design). The controls span overlap cleanly: correct $= 1.0$ by construction; wrong-pair ≈ 0.25 (its additive scaffold shares the A-axis); coefficient-mismatch, cross-prompt, and random near zero. The observed effect is approximately proportional to the projection on $-I_\ell$ (at boundary 14: correct $\cos = 1.00$, $D = +0.028$; wrong-pair $\cos = 0.25$, $D = +0.007$; its depth-averaged D is nonetheless negative)—which is exactly why the paper claims *cross-term-direction specificity* rather than task-identity specificity beyond the direction (§7).

Patch-quantization audit. Writing patched states back to bf16 could silently cancel low-dose rungs ($\|\text{bf16}(h - \lambda I) - (h - \lambda I)\| / \|\lambda I\|$ exceeding a frozen 0.3 threshold would mark a rung *not actually applied*). The v1.2 freeze resolved this preemptively by running all of KT2 in float32 (with TF32 off, since its 10-bit mantissa would defeat the purpose); the audit is retained as residual-risk monitoring and never triggered.

I AMPLIFICATION PROFILES (EXPLORATORY)

Everything in this appendix is computed from existing immutable cell records after the confirmatory verdicts; it motivated no gate and is labeled exploratory throughout.

The exclusion arithmetic (confirmatory input to §3.1). At the $(1, 1)$ code cell, if layers contributed independent increments of the measured sizes $\|G_\ell\|$, cumulative growth would be bounded by $\sqrt{\sum_\ell \|G_\ell\|^2} = 0.088$ (incoherent) and by $\sum_\ell \|G_\ell\| = 0.369$ even under perfect alignment of every increment. Observed: $\|I_{28}\| = 0.974$—11× and 2.6× the respective bounds, implying an average multiplicative gain of ≈ 1.19 per block on the carried term. Since the bounds are arithmetic in the measured quantities, no per-layer injection account survives regardless of the increments' provenance (including the rounding-artifact component of bf16 G, which only strengthens the exclusion by shrinking the true injections).

Layer-resolved gain. The per-block gain $\|I_{\ell+1}\|/\|I_\ell\|$ is above 1 through most of the depth, with pair- and format-dependent profile: gains are similar across the two task pairs early and diverge in the second half of the network—consistent with §6, where *cumulative* magnitude (post-amplification) coarsely orders pairs even though local generation does not. Across the 15-pair capture grid, cumulative $\|I\|$ spans a 3× range (0.084–0.256) on a generation floor that is nearly pair-independent.

Super-coherent amplification by format. Defining the super-coherence ratio as observed $\|I\|$ over the coherent accumulation bound: raw prompts 1.0–1.7×, instruction-wrapped prompts 3.8–4.1×—amplification is *strongest* exactly where expressed interference is smallest.

Depth relocation (3/3 seeds). Splitting mean gain at normalized depth 0.5: wrapped early-gain 1.55 vs. raw 1.36, while the raw−wrapped *late*-gain contrast is $+0.019/+0.012/+0.031$ across seeds—positive 3/3. The wrapper front-loads amplification and suppresses it late; raw amplification persists to the output. This profile was recorded (as an exploratory expectation, not a gate) *before* KT2 ran, where it correctly anticipated the causal gating result: erasure finds interference to remove on raw prompts and almost none under the wrapper (§5).

J SECOND-FAMILY REPLICATION (LLAMA-3.2-1B)

A reduced replication with the base family as the *only* varied factor: `meta-llama/Llama-3.2-1B` (commit-pinned; 16 blocks, GQA, 128k tokenizer), nine LoRA adapters trained with the identical recipe (datasets, samples, epochs, LoRA shape, seeds), intervention boundaries $\{2, 5, 8, 11, 14\}$ mapped from the frozen Qwen set at equal normalized depths, and a new 9-delta core-median norm radius recorded before any kill-test cell ran. Four predictions (RP1–RP4, with priors 70/75/60/high%) were frozen before any adapter on this family existed. Execution: 54 reduced KT1 cells (KT2 subgrid × {controlled code, raw, wrapped} × 3 seeds × 2 pairs), the amplification-bound analysis, and the full 54-run KT2 with per-prompt persistence, under the same numerical-validity chain—which passed unmodified on the new architecture (RP4: bitwise determinism nulls, first-block identities, zero equivalence audits).

RP1 (band + output mismatch): held, with reduced margins. The bf16 G-AUC band ratio across all 18 $(1, 1)$ cells is 1.201 (gate <1.3; absolute level ≈0.0013, an order below Qwen's—hidden width and norm scale differ, and only within-band uniformity is the cross-family claim). Output interference again exceeds what the band allows: pair ratio 1.51 (gate >1.5; per-seed 1.59/1.82/1.06), format ratio 2.03 (gate >2; per-seed 2.82/1.71/1.93). Both gates pass at the frozen mean-level reading but are seed-variable and far smaller than Qwen's 2×/20×; we grade this held-with-reduced-margins. Not lost in the reduction: $G_{\text{wrapped}} > G_{\text{raw}}$ replicates 3/3 seeds—the falsified-then-reversed KT1 prediction is not a Qwen artifact.

RP2 (injection exclusion): held, more strongly than Qwen. Cumulative-to-bound ratios per condition: 6.7–24.9× incoherent (gate >3), 2.19–7.85× coherent (gate >1.2). The depth-relocation profile of Appendix I also transfers: super-coherent amplification is again largest under the wrapper (5.4–7.9× vs. 2.2–3.2× raw), and the raw−wrapped late-gain contrast is positive 3/3 seeds ($+0.042/+0.053/+0.070$).

RP3 (causal triple): held; the gating leg by its frozen letter only. With the frozen bootstrap template: (a) $D_{\text{correct}} = +0.0233/+0.0333/+0.0192$, all 15 intervals (effect and four specificity contrasts

per seed) exclude zero. (c) Implied pair ratios $D(\text{code+safety})/D(\text{code+math}) = 2.84/5.46/2.56$ (gate $>$1.3, 3/3), contrast CIs excluding zero—and for the seed whose *output* pair ratio is only 1.06, erasure still separates the pairs 2.6×: the direction-carried contrast is more stable than the output ratio itself. (b) $D_{\text{raw}} - D_{\text{wrapped}} > 0$ holds 3/3 by sign, satisfying the frozen criterion, but margins are weak ($+0.0052^*/+0.0019/+0.0039^*$; the middle CI crosses zero) and the removable-interference ratio is only ≈2×—consistent with this family's weak output-side format collapse. The gate's *direction* transfers; its *strength* is a family property.

K REGROWTH AFTER ERASURE (PREREGISTERED FOLLOW-UP)

Three predictions were frozen before any regrowth data existed (priors 60/60/65%). Setup: at each frozen boundary $b \in \{3, 8, 14, 20\}$, apply the λ=1 correct erasure and continue the forward pass; at every later frozen boundary (plus the raw final-block output) form the post-patch cross-term against the *unpatched* marginals, $I_\ell^{\text{post}} = \tilde{h}_\ell^{AB} - h_\ell^{A0} - h_\ell^{0B} + h_\ell^{00}$, and report the per-prompt recovery ratio ρ_ℓ and direction cosine to I_ℓ^{orig}. Scope: 18 cells (2 pairs × 3 strata × 3 seeds, $(1, 1)$, float32).

Outcomes. (PE1-1, *falsified*): latest-tested-boundary ρ after erasing at depth 8 was predicted in $(0.3, 0.95)$; observed $0.987/0.998/0.992$ across seeds—regrowth is essentially complete, outside the window on the "transient" side. (PE1-2, *held*): direction memory $\cos(I^{\text{post}}, I^{\text{orig}}) > 0.5$ predicted; observed $0.987/0.987/0.986$. (PE1-3, fail by letter): ρ decreasing in patch depth holds in 2/3 seeds; the third inverts between depths 3 and 8 by $+0.0003$, a noise-level violation we report as the frozen criterion requires. Full profile (seed means): $\rho = 1.000/0.992/0.944/0.843$ erasing at depths $3/8/14/20$.

Interlock with the causal effect. Decomposing the KT2 effect by patch depth from the per-prompt records (code+safety, code stratum, mean over grid): $D \approx 0$ at depths 3–8 ($+0.0005/-0.0012$, $+0.0009/+0.0037$, $+0.0001/-0.0076$ per seed), rising to $+0.0374/+0.0408/+0.0260$ at depth 20 and $+0.0545/+0.0582/+0.0366$ at depth 25. The unrecovered share $1 - \rho$ and the durable output effect D rise together: what erasure removes is what the remaining depth cannot rebuild. The regeneration reading was subsequently subjected to a preregistered basin test (gate frozen first: latest-tested-boundary $\cos > 0.8$ from every start): patching at depths 8 and 14 with correct erasure at $\lambda \in \{1, 2\}$ and all four norm-matched structural controls, the downstream cross-term reconverges to the original direction from *every* start—per-condition means 0.90–0.99, worst single cell 0.83, 18/18 (both pairs, 3/3 seeds). Convergence of direction is compatible with the divergent *output* effects of §4: wrong-direction patches commit their damage to the expression pathway before late-network regrowth completes, exactly as the depth decomposition shows.

Relation to emergent self-repair. The Hydra effect (McGrath et al., 2023) and its systematic follow-up (Rushing & Nanda, 2024) show that ablating a transformer component is partially compensated downstream: an output-side, incomplete, prompt-noisy recovery of the component's direct effect, with final-LayerNorm rescaling explaining at most 30%. The regeneration here is a different object: what returns is a four-path *differential* coordinate, rebuilt to 99% of its norm at cosine 0.99 from three marginal paths the intervention never touched, while the state itself never rejoins its trajectory and no direction-specific restoring force exists (Appendix R). LayerNorm rescaling acts within the patched trajectory and could in principle contribute to the recovered *magnitude*; it does not by itself supply the direction memory the basin test measures from every start, nor the marginal-path provenance the ledger decomposition establishes. A quantitative accounting of how much of the regrowth known self-repair mechanisms explain is open.

L SURGERYV2 REPRESENTATION BIAS, HEAD-TO-HEAD

Two predictions frozen before any bias data existed (priors 70/50%). Operationalization: in the factorial setup the task experts are the marginal paths, so layerwise bias is $\|h_\ell^{AB} - h_\ell^{A0}\|$ (and against the right expert, $\|h_\ell^{AB} - h_\ell^{0B}\|$), last prompt token, bfloat16 measurement lock, full-depth AUC over the frozen boundaries; 36 cells (2 pairs × 6 strata × 3 seeds, $(1, 1)$).

Format axis (frozen prediction: wrapped $\geq$ raw, all 6 pair$\times$seed conditions). Held in 5/6; one condition (code+safety, seed 42) inverts by 0.8% (16.36 raw vs. 16.23 wrapped)—reported as a falsification of the letter. The substantive comparison is one-sided everywhere: bias ratios wrapped/raw span 0.99–1.34 while the expressed-interference ratio on the same prompts is $\sim$0.05 (a $20\times$ collapse). Representation bias does not track expression in any condition.

Pair axis (frozen two-sided). Bias orders the pairs as the output does: bias(code+safety)/bias(code+math) $= 2.83/3.05/2.96$ across seeds on code prompts. This places representation bias in the same family as cumulative $\|I\|$ (§6): a coarse pair-level magnitude signal—expected, since it contains the amplified cumulative term—with no visibility into expression gating.

M Intervention validity: introduced divergence and context sensitivity

Causal interventions can displace states off the model's natural distribution, where they recruit pathways that never fire naturally; Grant et al. (2026) accordingly ask intervention papers to (i) report the representational divergence they introduce and (ii) test interventions for context sensitivity. We address both.

The displacement is the estimand's own counterfactual, not an arbitrary perturbation. Our patch sends $h_\ell^{AB} \mapsto h_\ell^{AB} - \lambda I_\ell$, so at λ=1 the patched state is exactly $\bar{h}_\ell = h_\ell^{A0} + h_\ell^{0B} - h_\ell^{00}$: the additive reconstruction against which expressed interference is defined, and the state the network would occupy if composition were additive at that depth. This is the minimal displacement that answers the causal question, in contrast to the multiplicative feature scalings (up to $15\times$) that motivate the concern.

Introduced divergence, measured. We report the displacement itself, not only its behavioral consequences. Over the same six cells (2 pairs $\times$ 3 seeds, code prompts, float32), the λ=1 displacement is $\|I_\ell\|/\|h_\ell^{AB}\| = 1.5$–$3.6\%$ of the state norm at depth 3, rising to 5.3–14.3% at depth 25; measured against the natural prompt-to-prompt spread of merged states at the same depth, it is 0.06–0.13 of that spread early and at most 0.60 late. The patched state therefore stays inside the envelope of variation the model already exhibits across prompts—a different regime from the multiplicative feature scalings that motivate the concern. Output-side the divergence is correspondingly small: patched-versus-unpatched first-token JSD 0.003 against a main effect of 0.236, top-1 agreement 0.94–0.99, and 92–99% of continuation tokens preserved (§4). Behavioral invariance alone cannot exclude dormant divergences—precisely Grant et al.'s point—which is why we also report the geometry above, and treat $\lambda > 1$ as an explicitly off-manifold regime where only ordered contrasts, never absolute effects, are interpreted.

The displacement's own depth profile does not explain the effect. Relative displacement grows with depth, so one might ask whether the depth profile of D merely tracks how much we move the state. It does not, quantitatively: from depth 3 to 25 the displacement grows 3–4$\times$ while D grows roughly $100\times$ ($+0.0005 \to +0.055$). The erasure is complete at every depth by construction (λ=1 removes the entire cross-term present there), and what differs is whether propagation rebuilds it—measured directly as ρ (Appendix K), not inferred. The angle sweep isolates the same point with displacement norm held fixed.

Context sensitivity. A hidden-pathway artifact would show up as an effect that exists in one configuration and vanishes or reverses arbitrarily elsewhere. The same intervention is applied at five depths, on three prompt distributions, for two task pairs, at three model points, and—in the angle sweep (Appendix T)—at five orientations with depth and norm held fixed. The outcome is not configuration-idiosyncratic but systematic in the intervention's geometry: ≈ 0 at early depths and rising with $1-\rho$; monotone in $\cos(\delta, -I_\ell)$; sign-determined by that angle rather than by displacement size. A dormant-pathway account would have to reproduce this two-parameter structure across all four axes, which is a far stronger commitment than the effect itself.

N AUDITING MERGING METHODS WITH THE LEDGER

Four weight-space merging methods, applied post hoc to the same LoRA task vectors with no training: **task arithmetic** (the composition used throughout this paper), **trim-only** (keep the top 20% of $|\Delta|$ per task), **TIES** (Yadav et al., 2023) (trim, sign-elect, disjoint mean) and **DARE** (Yu et al., 2024) (p=0.9 random drop with $1/(1-p)$ rescaling). For method M the merged weights are $W_0 + \alpha \sum_t \Delta_t^M$ and the additive counterfactual is built from that method's *own* marginals $W_0 + \alpha\Delta_t^M$, so I_ℓ always measures the non-additivity of the composition step and is comparable across methods; deltas are written into the base weights with the adapter path disabled and restored exactly afterwards. Grid: 3 seeds $\times$ 4 methods $\times$ $\{k$=2, α=1; k=6, α=1; k=6, α=1.5$\}$, scored in the templated regime. Pipeline check (not a gate): task arithmetic at k=2 gives $R = 0.0031/0.0044/0.0025$ against $0.0031/0.0044/0.0025$ measured independently in Appendix P—the weight-writing and adapter-scaling paths agree to three or four significant figures.

Result. In the collapse cell (k=6, α=1.5), seed-mean R and code score: TIES $0.93/0.84$; trim-only $0.77/0.73$; DARE $0.67/0.18$; task arithmetic $0.64/0.42$. The frozen prediction that some method would cut R by $\geq$30% failed 0/3. The frozen pooled-correlation prediction ($\rho \leq -0.4$) passed at -0.42, but the within-cell correlations are $+0.20$ (k=2), $+0.07$ (k=6) and $+0.93$ (k=6, α=1.5), so the pooled value reflects cell composition, not method ranking. We report both.

A common counterfactual (preregistered follow-up). The per-method convention above leaves TIES' marginal semantically different from a standalone expert (its limit below), so a frozen follow-up recomputed R for all four methods against ONE shared reference: the additive prediction from the task-arithmetic marginals ($W_0 + \alpha\Delta_t$, the actual standalone experts) and the same base logits. The pipeline check is exact: R^{common}(task arithmetic) reproduces the frozen per-method value to 10 decimal places, 3/3 seeds. The frozen expectation (55%: still no correct ranking) held, in a stronger form than predicted: under the common reference the output-side ratio carries almost no method information at all—four-method range $0.006/0.015/0.006$ against behavioural spreads of 0.6–0.9, Spearman $0.0/-0.4/+0.4$. The between-method differences of the per-method convention (0.64–0.93) are therefore largely a property of the counterfactual choice, not of the merged models: under either convention the output-side ratio cannot rank methods, while the state-space measures rank the same models correctly (Appendix O).

Reading, and its limits. The two methods that preserve behaviour both *trim*; the two that collapse both keep full-support displacements (DARE drops 90% of entries but rescales to preserve norm). What these methods control is the support and magnitude of the displacement; R measures non-additivity, and in the collapse regime the two can move in opposite directions under the per-method reference convention. This reading is post hoc and untested. Four limits belong with it: TIES and DARE were designed for dense full-model deltas, and their statistical assumptions may not transfer to rank-16 LoRA deltas—the audit's method axis is scoped accordingly; for TIES the per-task marginal is not a standalone expert (it is that task's share of the elected merge, divided by the agreement count), which changes what its additive counterfactual means; four methods give a coarse rank correlation, so the claim rests on 3/3 seed agreement in direction; and the behavioural instruments see collapse but not competence (Appendix P), which is why the audit is stated for the collapse cell only.

O THE METRIC FAMILY ON THE SAME MODELS, AND THE AUDIT AT 7B

Two preregistered follow-ups scope the merging audit (Appendix N); all gates below were frozen before any of their data existed.

The family, same models, same scores. On the identical merged models and prompts, with behavioural scores read from Appendix N's immutable records (nothing regenerated), we computed four family members: our output-side R; SurgeryV2-style representation bias ($\text{mean}_\ell \, \text{mean}_x \|h_\ell^{\text{merged}} - h_\ell^{\text{expert}}\|$); a CTL-style state residual ($\text{mean}_\ell \, \|h_\ell^{AB} - \bar{h}_\ell\| / \|h_\ell^{AB} - h_\ell^{00}\|$—the same relative non-additivity as R, read before the readout); and mean pairwise parameter cosine. Two frozen expectations both *failed*, in the direction the preregistration marked as the more valuable outcome: some

members do rank correctly (65% prior against), and R opposes rather than tracks bias (55% prior for agreement). Collapse-cell seed means:

	code score	R	bias	CTL residual
TIES	**0.84**	0.93	**26.7**	**2.00**
trim-only	0.73	0.77	31.2	2.22
task arithmetic	0.42	0.64	38.1	2.33
DARE	0.18	0.68	44.7	2.37

Bias and the residual align with score rank-for-rank (Spearman -0.8 to -1.0, 3/3 seeds); R is their mirror image ($+0.8$ to $+1.0$); parameter cosine is weakly backwards ($+0.4$ to $+0.8$).

The audit at 7B with real endpoints. Qwen2.5-7B carries three adapters (math, code, safety), so collapse is approached through the coefficient: 4 methods $\times$ $\alpha \in \{1, 1.5, 2\}$ $\times$ 3 seeds, templated regime, endpoints fixed before any data: GSM8K *test* split, first 200 problems, exact final-number match (Cobbe et al., 2021) (training used the train split), and HumanEval, 164 problems with executed unit tests (Chen et al., 2021); templated mode is primary with completion mode as instrument control, the code endpoint taking the better of the two. The instrument check passed *before* measurement for the first time in four behavioural attempts: base scores 0.315 (GSM8K) and 0.701 (HumanEval), and merged models at α=1 *exceed* base on GSM8K (up to 0.60)—these instruments see competence, not only collapse.

Collapse by the letter, and the gates. The frozen collapse criterion (mean code-endpoint drop $\geq$25% against each method's own expert, at the highest such α) fires in one seed of three (α=2: 30.2%, against 24.7% and 23.2% in the others—a binary gate missing by 0.3 points, reported as such). In that cell the state-space measures rank the four methods correctly against *both* real endpoints (Spearman -1.0; GSM8K and HumanEval give identical method orderings: TIES 0.67/0.41, trim 0.61/0.31, task arithmetic 0.52/0.13, DARE 0.31/0.06), and trimming wins as at 1.5B (frozen prediction held). Two frozen expectations *failed*: the output-side reversal did not persist (Spearman(R, score) $= -0.2$, gate required ≥ 0), and R no longer opposes bias ($+0.2$, gate required ≤ 0). Across the full nine-cell grid the 1.5B dissociation is absent everywhere—$\rho(R, \text{bias}) = +0.2$ to $+1.0$—and R ranks the methods correctly in seven of nine cells, but its four-method range collapses to 0.06 in the deepest-degradation cell while the state-space measures stay at -1.0.

The three-layer decomposition (preregistered reanalysis). A further question is whether "magnitude points the wrong way" is partly a normalization artifact. On the frozen payloads only (statistics never previously computed; gates frozen before computing them), we correlated the *absolute* interaction JSD, the normalized R, and the main-effect denominator with behavioural degradation, per axis. On the *coefficient* axis (E14's k=6 column and the 7B per-method sweeps; 15 instances) the absolute numerator ranks degradation no worse than R in 11/15—the normalization critique partly holds there. On the *method* axis it does not: under the common counterfactual the absolute interaction is as method-blind as the ratio (relative range under 10% of the behavioural range, 3/3 seeds). And on the *format* axis the failure lives in the denominator (absolute 0.4–$1.0\times$, ratio $13\times$, behaviour $20\times$). Three axes, three distinct failure modes; the blanket claim is retired in favour of this table, whose per-axis records are released with the artifact.

Reading, and limits. The reversal of Appendix N is thereby bounded to the small-scale collapse regime, and the exploratory diagnostic suggestion—compare state-space deviation alongside output-side ratios—is supported by the four audited methods at both scales, but is not a general method-selection rule. Consistent with §4.3, the failure lives in the state-to-output readout, whose gain is regime- and scale-dependent: backwards where the 1.5B collapse readout compresses the cross-term direction, signal-free where deep degradation saturates numerator and denominator together. Limits: four methods give coarse rank correlations; the 7B collapse cell exists in one seed by the letter; collapse at 7B is approached through α rather than task count, so the two collapse cells are not the same parameterization; and the HumanEval mode comparison drifts across cells (the endpoint takes the better mode, per the frozen instrument clause).

P THE DEGRADATION MAP: WHERE MERGING ACTUALLY BREAKS

Behavioural validation failed three times (Appendix U). Before designing a fourth attempt we asked the prior question—*does a large, seed-stable degradation exist anywhere accessible?*—by measuring, with no intervention, a grid over the number of merged tasks $k \in \{1, 2, 4, 6\}$ and a uniform coefficient $\alpha \in \{0.5, 1, 1.5\}$, scored in the templated regime that matches the training format, with R recorded in the same cells. Two predictions were frozen first; both failed, and the reasons are the content.

The frozen normaliser was degenerate: the experts do not beat the base model on these endpoints. We had specified degradation as a fraction of the expert's advantage over base. That advantage is non-positive here: on code, base scores 0.958 against the expert's 0.917; on math both score 0.875—while the same adapters move the output distribution enormously (main-effect JSD ≈ 0.56). The criterion was therefore unsatisfiable and we record it as failed by its letter. This is the third instance of one disease (Appendix U): **our behavioural instruments can see collapse but not competence.**

Degradation exists—but not where our causal experiments are. In absolute terms the merge does collapse at the edge of the grid: at k=6, α=1.5 the code score falls from $0.917/0.917/0.889$ (single-task) to $0.403/0.194/0.667$ (seed 42/123/456) (3/3 seeds down, mean -0.49), and math falls to 0.250 in one seed. Monotonicity in k nonetheless failed (3/3): at α=1 the mean absolute drop is 0.056 (k=2), 0.032 (k=4), 0.120 (k=6), with one seed *improving* at k=4—interference from different tasks can partially cancel, as the preregistration's stated counter-argument anticipated.

The gate is task-count dependent, which bounds §5. In the same templated regime where two-task merging gives $R \approx 0.003$, expressed interference rises steeply with k at α=1: 0.0033 (k=2) $\rightarrow 0.198$ (k=4) $\rightarrow 0.821$ (k=6), a factor of $\sim$250. At k=6 the interaction JSD is of the same order as the main effect. The instruction template pins the output distribution when two task vectors are added; it does not when six are. **Our gating result is measured at k=2 and should be read with that scope.**

R is not a sufficient statistic for degradation. Across all 27 grid cells the correlation between R and absolute drop is $+0.39$; *within* k=6 it is -0.32, because R saturates near 1 while degradation keeps growing with α. Expressed interference tracks the number of merged tasks; degradation tracks both k and α. At the seed level the association is visible but the sample is three: at k=4, α=1 the seed with $8\times$ lower R is the only one without degradation.

Why this closes the behavioural question honestly. Our causal experiments all sit at k=2, α=1, where the merge is nearly lossless on these endpoints; the degradation lives at $k \geq 6$, α=1.5, where interference is also strongly expressed. The two regimes are disjoint. That is a structural reason for the failure of behaviour-level validation—not a matter of instrument precision—and it names the coordinates a future causal experiment should use.

Q THE BEHAVIOURAL LOOP IN THE COLLAPSE REGIME

One question dominates the reading of the anatomy: the causal experiments sit at k=2, α=1, where the merge is near-lossless—does erasing the cross-term repair behaviour *where merging actually breaks*? A preregistered fourth behavioural attempt (both outcomes' obligations frozen in advance, prior 35% for repair) ran the intervention in E14's collapse cell itself.

Design. k=6 adapters at α=1.5, templated regime, float32, 3 seeds. The k-way cross-term $I_b = h_b^{\text{merged}} - \left(h_b^{00} + \sum_t (h_b^t - h_b^{00})\right)$ is patched at the last prompt position at boundary $b \in \{8, 14, 20, 25\}$ during prefill (the KT2 intervention point; the KV cache carries it into greedy generation). Four norm-matched arms: none, correct ($-I$), $+I$, random-orthogonal (the E10/E13 construction and seed). Endpoints: the templated code and math scores—*independent of the additive counterfactual*, with collapse-cell sensitivity certified by E14's frozen payload before any data existed. Pipeline check: the no-patch baselines reproduce E14's frozen collapsed scores bitwise ($0.403/0.194/0.667$), 3/3 seeds.

Outcome: no repair (frozen gate failed 0/3), and an asymmetry. No late boundary reaches the preregistered $\geq$25% recovery of the collapse gap (per-boundary means $-0.12/-0.87/0.00$ at b=14/20/25). At block input 25, the latest tested boundary of 28, the result is exact: erasing the measured residue—a displacement of 1.9–2.1$\times$ the state norm here—leaves the collapsed code score *bitwise unchanged* in all three seeds, while the equal-norm $+I$ arm is destructive ($0.40 \rightarrow 0.00$; $0.67 \rightarrow 0.22$). The residue can do damage in the $+$ direction; complete one-time removal did not repair these code scores. Partial persistent removal has a different behavioural effect, reported below. The direction-specificity gate was conditional on repair and is recorded as not triggered.

Reported both ways (unregistered directions, no claims). At b=8—excluded from the gate because regrowth was expected to undo early erasure—two of three seeds show substantial recovery ($+0.32/+0.31/-0.50$), the reverse of the first-token depth profile; the math endpoint (8 items, 0.125 granularity) shows no 3/3 pattern anywhere; at b=20 the correct arm is harmful in 2/3 seeds—mid-depth erasure of a displacement this large does damage of its own. In this cell $\|I_b\|/\|h_b\| = 0.83$–2.36: the intervention is far outside the small-displacement regime of Appendix M, so only ordered contrasts between equal-norm arms are interpreted, per the paper's standing policy.

The small-dose ladder (specified follow-up). The λ=1 null leaves a cancellation reading: removal of a harmful component and off-manifold damage of the same size. The first-order logic discriminates: a carried component gives an $O(\lambda)$ benefit, off-manifold damage is $O(\lambda^2)$, so the small-dose slope decides. We ran $\lambda \in \{0.1, 0.25, 0.5, 1.0\}$ (at λ=0.1 the displacement is 17–21% of the state norm, comparable to Appendix M's regime) at $b \in \{14, 25\}$ with a continuous endpoint fixed in advance: teacher-forced NLL of frozen expert reference continuations, whose dynamic range was certified *before* any arm data (collapse gap = 34–50$\times$ the expert's cross-seed spread); the λ=1 discrete scores reproduce the main experiment bitwise. The frozen repair gate (30% prior) *failed*: at the latest tested boundary the correct arm's dose curve is flat—ΔNLL ≈ -0.002 regardless of λ, a tenfold dose change with no effect—while the orthogonal arm degrades quadratically (-0.073 to -0.087 at λ=1), the observed NLL pattern in that one-time intervention; at mid-depth a $+0.003/+0.004$ small-dose signal appears in 2/3 seeds, is absent in the third, and even taken at face value bounds any carried share at a few percent of the collapse gap. These measurements disfavor a simple cancellation account on that endpoint, but do not rule out all endpoint-dependent or nonlinear effects; the persistent quarter-dose result below supplies such a counterexample for code accuracy.

A second collapse parameterization (preregistered follow-up). The loop was repeated where collapse is reached through the coefficient instead of the task count: Qwen2.5-7B, three adapters at α=2.0 (the deepest degradation E18 measured), real-benchmark endpoints, arms and construction unchanged, $b \in \{14, 25\}$. Pipeline check: the no-patch baselines reproduce E18's weight-written α=2 scores *bitwise* on the code endpoint (3/3)—two independent merge implementations agreeing exactly. The repair gate failed again (25% prior): correct-arm recovery is at most 4% of the collapse gap at any boundary, and at the latest tested boundary the code score is bitwise unchanged in 2/3 seeds—the null repair transfers. The destructive-doubling gate (55% prior) also failed: $+I$ costs only 0.006–0.049 on code (gate: ≥ 0.10), against $0.40 \rightarrow 0.00$ at 1.5B. The geometry explains the difference: here $\|I_b\|/\|h_b\| = 0.41$–1.02, less than half the 1.5B collapse cell's 1.9–2.1 at the latest tested boundary—at this displacement scale both directions are behaviourally inert (GSM shows a partial $+I$ cost in 2/3 seeds, below the gate; reported, not claimed). The failure of complete one-time erasure to produce substantial positive recovery appears in both tested parameterizations; destructive doubling is scale-dependent.

Persistent every-position erasure (preregistered follow-up). A review located the sharpest alternative reading: the erasure above acts once, at the last prompt position, while every generated position can re-form its cross-term—so the null repair might reflect temporal under-intervention. We therefore clamped the merged state toward the *step-wise* additive reference, $h \leftarrow (1-\lambda)h + \lambda \bar{h}_b^{(t)}$, at every position (full prefill and every generation step) and up to three late boundaries at once, with $\bar{h}$ built online from eight parallel weight states (base, six marginals, and the merged model), advanced on the common token history alongside the merged state being clamped. Pipeline check: the unpatched arm reproduces the frozen collapsed scores bitwise (3/3); the per-position cross-term removed is large and measured (norms 33–228). The repair gate (25% prior) *failed* again: full clamping recovers $+0.27/-0.30/-1.50$ of the gap—in two of three seeds it hurts. At the clamped

boundaries, λ=1 sets the tested state to its step-wise additive reference, yet code scores remain below the experts (0.25–0.39 against 0.89–0.92). This establishes insufficiency of this intervention, not a unique cause of collapse. Direction remains special under persistence (clamping deviates the NLL endpoint 3–4$\times$ less than the norm-matched orthogonal arm, which destroys both endpoints; frozen gate held 3/3). Reported both ways: quarter-dose clamping (λ=0.25) raises the code score in 3/3 seeds (approximately $+0.097/+0.445/+0.014$ in seed order 42/123/456) while worsening the expert-NLL in 3/3 and leaving math mixed—an endpoint-dependent, non-monotone response in the tested arms. No jointly beneficial intermediate dose is established. Single-boundary clamping equals three-boundary clamping at λ=1, consistent with the measured recovery of the residue coordinate after early erasure. The claim is scoped accordingly: complete erasure does not consistently repair the tested failures, while partial persistent erasure trades off endpoints.

Arithmetic correction to the archived E26 verdict. The frozen verdict lists the quarter-dose code scores correctly but prints their changes as $+0.24/+0.10/+0.01$. Reordering to the paper's seed order 42/123/456, subtraction of the displayed rounded scores gives $0.500-0.403=+0.097$, $0.639-0.194=+0.445$, and $0.681-0.667=+0.014$, as reported in Table 2. The immutable scores and the 3/3 direction are unchanged; this corrects derived numbers, not the underlying experiment.

Reading. The coefficient sweep supplies a distinct behavioural association: scaling the displacement down restores behaviour monotonically (the 7B code drop rises with α within every seed, reaching $30.2/24.7/23.2\%$ by seed at α=2, Appendix O; the 1.5B grid of Appendix P agrees), while complete erasure at fixed α does not reliably restore the tested endpoint. These observations do not isolate a unique mediator: coefficient changes also change the residue, and partial persistent erasure changes code accuracy. The present interventions do not identify which part of the over-scaled weight displacement drives the behavioural loss: the matching single-task control has not been run, and partial erasure is beneficial on one endpoint. The supported conclusion is that complete residue removal is not a reliable repair for the audited cells, even though the residue causally affects output non-additivity.

Convergent external evidence. Independently, Cao et al. (2026) characterize task-level merging collapse *correlationally*: representational incompatibility between the expert models predicts collapse across merging methods while parameter-conflict metrics do not, with a rate-distortion bound on mergeability. Their correlational results motivate testing marginal incompatibility, but our intervention results do not isolate it as the cause; persistent erasure was tested at 1.5B only, and the partial-dose benefit further limits that interpretation. The recommendation of §6 should be read with that priority: the direction was anticipated correlationally; the specified erasure responses are what this paper adds.

R Displacement-specific recovery: is the cross-term direction privileged?

The regeneration result is computed against three untouched marginal paths, so a network that simply washes out *any* perturbation would reproduce it. We therefore put every direction on one ruler, in state space: for a displacement δ applied at depth b, $r(\ell)=\|\tilde{h}_\ell - h_\ell^{AB}\|/\|\delta\|$, the surviving fraction of the perturbation. Four equal-norm directions ($-I_b$, random-orthogonal, task-structured-orthogonal, $+I_b$; the same construction and random seed as Appendix T) at $b\in\{8,14,20\}$, 2 pairs $\times$ 3 seeds, code prompts, float32. Two predictions were frozen first.

Result. Deviations at the output are 1.1–3.1$\times$ the displacement in absolute terms, but the residual stream itself grows 2.5–4.7$\times$ over the same span; dividing it out, the surviving relative deviation is 0.26–0.53 (patch at 8), 0.51–0.71 (14) and 0.79–1.02 (20). Perturbations are therefore *diluted* relative to the stream, the more so the more depth remains—but not uniformly: a random orthogonal displacement is diluted about twice as strongly as a structured one at every depth and in every seed (e.g. 0.26 vs. 0.47–0.53 at depth 8). A direction-independent contraction predicts equal dilution of all four arms and is excluded by this margin.

The frozen prediction of a privileged direction failed. We had predicted (50% prior) that the cross-term direction would be *corrected more strongly* than the orthogonal ones. It is not. At every

depth, $-I_b$ survives at the same level as $+I_b$ and the task-structured orthogonal arm, and more than the random arm (0/3 seeds pass). We report this as falsified and, accordingly, make no claim that the cross-term direction carries a restoring force. The second prediction—that recovery is direction-dependent at all (spread > 0.10)—passed 3/3 (spreads 0.45–1.79).

What this leaves. Two facts hold together: the perturbed trajectory never rejoins the unpatched one at any patch depth, yet after erasure the *cross-term coordinate* returns to 99–100% of its norm at cosine 0.99 (Appendix K). What propagation restores is the cross-term, re-created from the marginal paths the intervention did not touch—not the state. We therefore say "regenerated" rather than "attractor" throughout: we measured the restoration of a coordinate, not a restoring force in state space. (Relative deviations divide by the last measured stream norm, at depth 25, for a probe at the latest tested boundary, so the dilution reported is conservative.)

S THE MINIMAL MODEL: DERIVATION AND THE NORM-SCALING TEST

Derivation. Fix a depth b and displace only the merged path, $h_b^{AB} \mapsto h_b^{AB} + \delta$; the three marginal paths are untouched, so the additive logit prediction $a = \ell^{A0} + \ell^{0B} - \ell^{00}$ is frozen under the intervention and the expressed interaction is $u(\delta) = F(h_b^{AB} + \delta) - a$ exactly. Expanding F about $\bar{h}_b = h_b^{AB} - I_b$ under (A1) gives $u(\delta) \approx g + J(I_b + \delta)$ with $g = F(\bar{h}_b) - a$, and (A2) then yields the expression in §4.3. Writing $\delta = -s\cos\cdot I_b + \delta_\perp$ with $\|\delta\| = s\|I_b\|$ and taking $\delta_\perp$ generic ($\mathbb{E}\langle u_0, J\delta_\perp\rangle = 0$),

$$\mathbb{E}[D] = \underbrace{2s\langle u_0, JI_b\rangle}_{a(s)}\cos(\delta, -I_b)\ \underbrace{-\,s^2\bar{\sigma}^2\|I_b\|^2}_{b(s)} + O\big(s^2\cos^2(\|JI_b\|^2 - \bar{\sigma}^2\|I_b\|^2)\big)\,,$$

where $\bar{\sigma}^2 = \mathbb{E}_{\hat{d}}\|J\hat{d}\|_M^2$ is the mean squared readout gain over generic unit directions. The residual term vanishes when the cross-term direction has typical gain; empirically $|b|/\|JI_b\|^2 = 1.15$–1.28, i.e. **the cross-term direction is a *low*-gain direction of the readout**. It is worth removing not because it moves the output most, but because the output residual is aligned with it.

Corollaries. (i) For generic residual directions whose linear term averages to zero under the stated sampling assumptions, the expected quadratic contribution is non-positive. This does not determine the sign for every individual direction; the tested structured controls are empirical observations. (ii) D is governed by JI_b, not $\|I_b\|$; where propagation reconstructs the cross-term, the measured output effect can be small even when $\|I_b\|$ is large. (iii) The dose curve is a downward parabola through the origin peaking at $\lambda^* > 1$ exactly when regeneration is positively aligned with transport.

Norm-scaling test (preregistered, PT-1/PT-2). Design: the E10 angle sweep repeated at $s \in \{0.5, 1.5\}$ (2 pairs $\times$ 3 seeds, $b \in \{20, 25\}$, code prompts, float32), reusing E10's frozen s=1 slice and its displacement construction and random seed unchanged; one s=1 audit cell was re-run months later and reproduces the frozen E10 payload *exactly* (max $|\Delta D| = 0$ over all ten condition$\times$depth values). Per-scale angle fits and the resulting exponents:

	$s = 0.5$	$s = 1.0$ (E10)	$s = 1.5$
slope a	0.066 / 0.067 / 0.042	0.135 / 0.135 / 0.086	0.205 / 0.203 / 0.133
intercept b	-0.018 / -0.019 / -0.015	-0.061 / -0.058 / -0.051	-0.121 / -0.111 / -0.099
R^2 of $a\cos+b$	0.89 / 0.84 / 0.72	0.86 / 0.81 / 0.60	0.86 / 0.82 / 0.59

$p_a = 1.026/1.007/1.047$ (frozen window $[0.7, 1.3]$, predicted 1) and $p_b = 1.748/1.628/1.708$ (frozen window $[1.6, 2.4]$, predicted 2); both gates pass 3/3. The joint surface fit $D = 2As\cos - Cs^2$ over all 30 measurements per seed gives $2A = 0.135/0.135/0.088$ and $C = 0.056/0.051/0.045$ at $R^2 = 0.87/0.83/0.64$, against $2A_{\text{dose}} = 0.120/0.126/0.079$ and $B_{\text{dose}} = 0.049/0.046/0.035$ obtained from the independently frozen dose ladder.

Where it is honest to stop. p_b passes its window but sits consistently below 2 in all three seeds; we read this as the bounded range of the JSD compressing the arms with the largest $\|u\|$ (plus_I,

orth_task), which is a post hoc reading and is not itself tested. The linear-in-$\cos$ fit leaves $R^2 = 0.60$–0.89, so a $\cos^2$ component of the size the derivation predicts is not excluded. The model is a local, first-order account under (A1)–(A2): it is not claimed to hold at displacement scales beyond those measured, and it says nothing about which weight configurations produce a large I_ℓ in the first place.

T ORTHOGONAL CONTROLS: EXCLUDING THE CONSTRUCTION ARTIFACT

Design: at depths 20 and 25, five per-prompt displacements all rescaled to $\|I_\ell\|$, differing only in angle to $-I_\ell$ (measured cosines confirm the construction to 10^{-10}): $-I_\ell$ ($\cos = +1$); a random direction with its I_ℓ component projected out (0); the wrong-pair displacement with its I_ℓ component projected out (0, but task-structured); their equal-weight combination ($+0.71$); and $+I_\ell$ (-1). Two pairs $\times$ 3 seeds, code prompts, $(1, 1)$, float32; predictions frozen first.

Result. D decreases monotonically with the angle in all three seeds: $+0.044$ to $+0.080$ at $\cos = +1$; $+0.032$ to $+0.057$ at $+0.71$; -0.015 to -0.025 (random) and -0.120 to -0.139 (task-structured) at $\cos = 0$; -0.114 to -0.186 at -1. The orthogonal arms therefore *worsen* expressed non-additivity, contradicting the prediction of a generic-contraction account, under which displacement into the additive neighbourhood suffices.

Two-component reading. Regressing D on $\cos(\delta, -I_\ell)$ gives slope $0.135/0.135/0.086$ and intercept $-0.061/-0.058/-0.051$ ($R^2 = 0.90/0.85/0.64$): a direction-independent damage term plus a direction term. Net removal requires alignment with $-I_\ell$; the reported $D \approx +0.03$ at $\lambda{=}1$ is thus a *net* figure that understates the direction term. One frozen criterion failed by its letter: we had required $|D_{\text{orth}}| < D_{\text{correct}}/3$, i.e. that orthogonal displacements be near-inert; they are instead strongly harmful. The two directional clauses of that criterion passed 3/3, and we report the amplitude clause as falsified.

The steepest-descent control (preregistered follow-up). Sharing the additive counterfactual between target and estimand leaves one alternative the angle sweep does not touch: $-I_\ell$ might simply be a loss-reducing direction aligned with $-\nabla_h \mathrm{JSD}(p(h), p_{\mathrm{add}})$. We computed that gradient per prompt (autograd through the remaining depth and readout) and ran three norm-matched arms at $\|I_b\|$, $b \in \{14, 20, 25\}$, 3 seeds. Both frozen gates *failed*, in the direction that excludes the alternative: the negative-gradient arm was predicted (65%) to be at least as effective per unit norm—instead it backfires catastrophically ($D = -0.17$ to -0.65, against $+0.012$ to $+0.098$ for $-I_\ell$, 9/9 cells), and $-I_\ell$ is nearly orthogonal to it ($\cos = 0.03$–0.12, 9/9). The quadratic model explains both: the gradient direction has high readout gain, so an $\|I\|$-norm step is deep in its $\|J\delta\|^2$ regime, while the cross-term direction's low gain ($|b|/\|JI\|^2 \approx 1.2$, Appendix S) lets a full-norm step remain in the linear-benefit regime and terminate exactly at the additive reference. The third arm ($-I_\ell$ with its gradient-aligned component removed, renormalized) is null-to-mildly-harmful (-0.004 to -0.036; the frozen halving criterion held in 2/3 seeds, recorded as failed by the letter): the benefit travels through a small aligned component that only the low-gain carrier can deliver at full norm. Among every direction tested in this paper—random, cross-prompt, wrong-pair, coefficient-mismatched, task-structured-orthogonal, $+I$, steepest-descent—$-I_\ell$ remains the only one that helps at full norm.

U BEHAVIORAL ENDPOINTS: THREE ATTEMPTS, REPORTED IN FULL

Attempt 1–2 (discrete scoring, null instruments). Unit-test pass rate on frozen code tasks and reference-answer containment on math, scored on prefill-patched greedy generations at depths 20 and 25 (2 pairs $\times$ 3 seeds). At 1.5B every code condition scored 0.0 (the code expert itself: 0.03/0.08/0.00) and math saturated at ceiling with the merge *above* its experts. At 7B the pattern replicated: code expert 0.000 against merge 0.083—the instrument does not separate models, so no verdict about the mechanism is possible from it. We report both as null instruments, not as evidence.

Attempt 3 (continuous NLL, valid instrument, negative result). Endpoint: the merged model's mean teacher-forced NLL of a frozen 48-token continuation generated once by the task expert—continuous, threshold-free, discriminative on raw prompts, and independent of the additive counterfactual. A preregistered validity gate required the instrument to see merge degradation at all; it passed 3/3

(merge−expert $+0.0256/+0.0218/+0.0114$; base further at $+0.09$). The frozen predictions then failed: D-style erasure at depths 20/25 changed the endpoint by $-0.00003/+0.00012/+0.00212$ (recovery share $+0.1\%/-0.6\%/-18.5\%$), and 11 of 12 control contrasts went the wrong way, with the wrong-pair control giving the *lowest* NLL in one seed. Post hoc readings we flag as untested: the effect being sought is the behavioral projection of a $\sim$19% reduction in expressed non-additivity; expert-likeness is not task correctness; and merge degradation need not lie along the expert direction. A final power-maximized test persisted per-prompt NLL and per-prompt interaction JSD for four angle-arm conditions at depth 25: the paired regression of behavioral gain on interference strength has *negative* slope in all seeds ($-0.074/-0.034/-0.037$), and the per-prompt coupling between removing interaction JSD and improving reference NLL is $-0.076/+0.011/-0.345$ (240 paired points per seed)—the two quantities are empirically decoupled at this scale, which also removes the rationale for stronger (persistent) interventions. The bounded conclusion is final for this setting: erasure controls expressed non-additivity, not behavioral expert-likeness.

The multi-position readout check (preregistered follow-up). The instruments above, and the estimand itself, read a single position; a review asked whether the decoupling is an artifact of that choice. The merged model greedily continued each prompt for 48 tokens; the same sequence was teacher-forced through all four weight states, giving the interaction and main-effect JSDs at *every* continuation position. Three readouts—first-token, the 48-position aggregate $\Sigma\mathrm{inter}/\Sigma\mathrm{main}$, and the mean per-position ratio—were correlated per prompt with the reference-NLL instrument. The frozen gate (35% prior that accumulation restores coupling) failed: no readout reaches $|\rho| \geq 0.3$ in any seed and signs are inconsistent (multi-position $+0.11/-0.00/+0.12$; first-token $+0.09/-0.09/+0.01$). The per-position profile explains why accumulation cannot help: the interaction JSD is largest at the first token and decays 5–35$\times$ over the continuation (the main effect 6–18$\times$)—expression fades with generation depth, so the single-token readout was the choice *most* favourable to finding coupling, not least. The decoupling stands against four instruments.

V MULTI-WRAPPER GENERALIZATION OF THE GATE

Frozen predictions before any data: three semantically equivalent instruction templates (Alpaca plus two new phrasings) and a token-length-matched non-instruction prefix control, on both pairs $\times$ 3 seeds $\times$ 60 prompts, $(1, 1)$. Result: every instruction template inflates the main-effect denominator 11–35$\times$ over raw and reduces R (ratios 0.09–0.60) in 18/18 conditions—the frozen "$R \leq$ raw/2" letter held in 17/18, missing once at 0.60, reported as the criterion requires. The length control *falsified* our frozen expectation that its denominator would stay within 2$\times$ (observed 3.5–6.7$\times$): long prefixes do shift the distribution. The discriminating quantity the data identify is R itself: the non-instruction prefix inflates numerator and denominator together (numerator 3.4–8.3$\times$), leaving R directionally unchanged (0.59–2.4), whereas instruction templates pin the distribution faster than they excite the interaction. Across all wrappers, including the control, the *absolute* interaction numerator grows in 23 of 24 cells (0.98–11$\times$; the one sub-unity cell is code+safety seed 456 under Alpaca)—the fifth independent replication of internal amplification under wrapping. An out-of-sample probe agrees: a follow-up in Zhu et al. (2026) re-serializes the same prompts in the model's native ChatML template—untrained for these adapters—and finds numerator and denominator moving together with expressed interference intact ($+12.5$pp), the signature of the non-instruction class (Zhu et al., 2026).

W SCALE POINT (QWEN2.5-7B)

Same family, larger scale, as the third grid point; 28 blocks, so the frozen boundaries carry over unchanged. Nine adapters with the identical recipe; reduced KT1 (54 cells), amplification bounds, and the *full* 54-run KT2 in float32 with per-prompt persistence (run on a single A800-80GB, the same GA100 die as the A100; the card-type amendment was logged before any KT2 data existed). Four predictions frozen before any 7B data.

Band and output mismatch (frozen gate missed by 0.1%). The G-band ratio across the 18 $(1, 1)$ cells is 1.301 against the frozen $<$1.3 gate—a falsification by the letter, reported as such. The substantive contrast sharpens with scale: output pair ratios $6.34/12.20/7.19$ (vs. 2$\times$ at 1.5B)

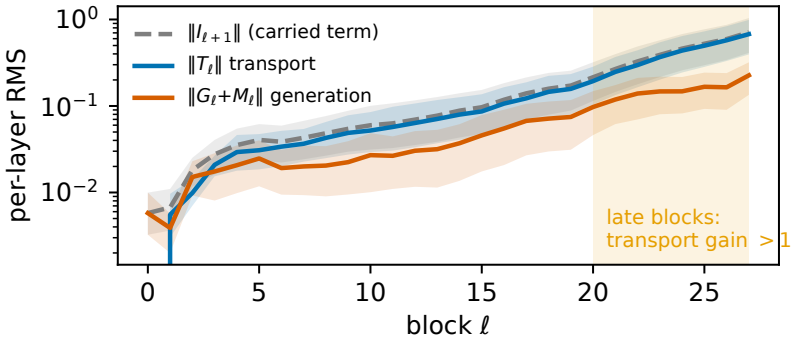


Figure 4: **Transport dominates local generation.** Per-layer RMS of the carried cross-term, its transported component, and local generation across two task pairs and three seeds; shaded bands span cells. Late-block transport gain exceeds one.

and format ratios $6.45/7.54/6.33$, against a band spread of $\leq$1.30; and $G_{\text{wrapped}} > G_{\text{raw}}$ replicates 3/3—the third model point with the reversal.

Amplification. $\|I\|$ exceeds the coherent bound by 1.80–8.31$\times$ across conditions (frozen gate $>$1.2, 3/3). Super-coherent amplification is again largest under the wrapper in both pairs, and the raw$-$wrapped late-gain contrast is $+0.091/+0.091/+0.073$—positive 3/3, the largest of the three model points. The pair-axis late-gain contrast is not seed-consistent (exploratory, reported).

Causal triple (all three legs, all bootstrap intervals excluding zero). $D_{\text{correct}} = +0.0274/+0.0549/+0.0270$ with all 15 specificity intervals excluding zero. Expression gating is *total* at this scale: $D_{\text{wrapped}} = +0.0040/-0.0038/+0.0007$, each CI containing zero, while $D_{\text{raw}} - D_{\text{wrapped}} = +0.0658/+0.0596/+0.0475$, all excluding zero—under the wrapper there is nothing left to erase. Pair contrasts $+0.0241/+0.0525/+0.0231$ all exclude zero, implying removable-interference ratios $8.3/22.9/6.9$ between pairs whose generation magnitudes differ by $\leq$1.3$\times$.

X Exact transport decomposition and the fp32 generation grid

Both experiments were preregistered together (five predictions with priors, frozen before any decomposition or fp32-grid data existed) as the data-level adjudication of two review-critical questions: whether the ledger's remainder can be attributed to transport, and whether the bf16 uniformity band survives at float32. Scope: 36 cells (2 pairs $\times$ 6 strata $\times$ 3 seeds, $(1, 1)$, all prompts, all 28 layers); the model is loaded in bf16 (the measurement state of the main sweep) and each block is evaluated with float32 weights and captured call arguments. The regrouping identity $I_{\ell+1} = G_\ell + T_\ell + M_\ell$ is asserted numerically at every layer of every cell.

E4 (transport): both frozen predictions held; the two-sided share resolved to transport-dominant. Late-network RMS(T_ℓ) exceeds RMS$(G_\ell + M_\ell)$ by 2.3$\times$ (0.32 vs. 0.14, 3/3 seeds); the transport gain $\|T_\ell\|/\|I_\ell\|$ averages 1.078–1.079 over the late half (3/3); the regeneration share of the total flux is 0.307–0.311 (3/3), so transport carries $\sim$69% under the norm accounting. Because the split is non-orthogonal in principle, we also computed the components' alignment and an energy (symmetric-projection) accounting on the code cells: late-network $\cos(T_\ell, G_\ell{+}M_\ell) = -0.014$ to -0.031 (6/6 near-orthogonal), and the energy-accounting transport share is 0.83–0.89—the attribution is essentially accounting-independent. Combined with the E1 regrowth result—regeneration alone suffices to rebuild the full term, in norm and direction, given $\geq$14 blocks—the consistent picture is a transport-dominated flux whose fixed direction the regeneration channel actively restores.

E5 (fp32 grid): both frozen predictions falsified, and reported as the criteria require. The fp32 AUC(G) band across the 36 cells is 3.36$\times$ (0.0012–0.0040; frozen gate $<$2), and the pair ratio is 1.81–1.90$\times$ (frozen gate $<$1.5): float32 local generation is *not* uniform, and it coarsely tracks the

pair ordering—joining cumulative norm, parameter cosine, and representation bias in the coarse-magnitude family. The format axis preserves the dissociation: fp32 generation is 1.66–1.84× *larger* under the wrapper (raw/wrapped ratios 0.54–0.60, 3/3 seeds) while expressed interference collapses $\sim$20×. The bf16 $\pm$15% band of §3 is therefore confirmed to be floor-compression—the cautionary tale stands in its strongest form, and the magnitude-fallacy claim rests on the causal contrast and the format inversion, not on band uniformity.

Llama replication (frozen before its data): transport invariant, magnitude behavior not. On Llama-3.2-1B (36 cells, 16 layers) both transport predictions held again—late-network RMS(T) exceeds RMS$(G + M)$ by $\sim$2×, transport gain 1.14 per late block, regeneration share 0.35 (all 3/3). The fp32 magnitude gates, by contrast, resolved *opposite* to Qwen—band ratio 1.65 (within the $<$2 gate), pair ratio 1.06 (no pair separation at all, against an output ratio of $\sim$1.5×)—falsifying our frozen expectation that the Qwen pattern would replicate isomorphically. The cross-family summary is sharper than either result alone: transport dominance and the format contrast ($G_{\text{wrapped}} > G_{\text{raw}}$; raw/wrapped 0.54–0.60 on Qwen, 0.78–0.84 on Llama, each 3/3) recur at the two tested family points, while whether local-generation magnitude tracks pairs at all is itself a family property—magnitude is not merely a coarse signal, it is an *inconsistent* one.