Sensitivity, Causality, and Repair Dissociate: A Layer-Wise Analysis of Perturbation Robustness and Its Scaling

Anonymous ACL submission

Abstract

Sensitivity (clean–perturbed cosine divergence), causality (patching recovery), and compensatory capacity (LoRA effectiveness) dissociate as per-layer properties of transformer language models under surface perturbation. Across a five-model panel we identify two propagation regimes—*spike-and-suppress* (Phi-3.5, Gemma-2-9B) and *late-accumulation* (Llama-3, Mistral, Qwen2.5-7B)—with sensitivity and patching anti-correlated on the two models meeting an 80% identity-patch gate (Phi-3.5, Llama-3; $\rho = -0.72$ to -0.88, $p < 0.001$, block bootstrap). Within-family scaling on Qwen2.5 (1.5B$\rightarrow$ 7B$\rightarrow$ 14B) yields a monotonically increasing late/early divergence ratio ($1.59 \rightarrow 2.90 \rightarrow 4.10$), with every per-perturbation entry monotonic; a two-point Llama scan (1B$\rightarrow$ 8B: $2.08 \rightarrow 5.29$) corroborates the direction on a second family. A cascade-disruption analysis explains why patching-implicated early layers are the wrong adapter target, and a fixed-harness layer sweep across four models (3.8–8B) confirms it: diagnostic-flagged sites are the most damaging windows on the adjudicable models. An earlier version of our pipeline reported gains from a representation-stability loss; those gains reverse under the fixed harness, and the reversal itself is independent evidence for cascade disruption.

1 Introduction

Language models degrade on surface perturbations such as typos, OCR noise, speech-like substitutions, and homophones (Pruthi et al., 2019; Rijhwani et al., 2020; Liu et al., 2023). When a failure happens, a natural question is which layers are responsible. Mechanistic analysis of transformer internals (Elhage et al., 2021) makes this question tractable. But "responsible" is ambiguous: it might mean the layers where the perturbation creates the most divergence, the layers where patching the representation back to its clean value recovers the prediction, or the layers where a lightweight adapter can correct the damage. The three notions look like they should converge on the same layers. They do not.

We introduce Layer-Wise Representation Divergence (LRD), a cosine-divergence metric between clean and perturbed hidden states—building on prior analyses of how perturbations propagate through model representations (Moradi and Samwald, 2021; Shi and Huang, 2020)—and use it alongside activation patching (Meng et al., 2022) and LoRA effectiveness to measure three distinct per-layer properties: **sensitivity**, **causality**, and **compensatory capacity**. All three dissociate, each produces a distinct layer profile, and the correlations between them depend on the model's propagation regime.

Our contributions are:

1. **A three-signal dissociation across five models.** LRD and patching recovery anti-correlate on the two models meeting the 80% identity-patch gate (Phi-3.5, Llama-3; $\rho = -0.72$ to -0.88, $p < 0.001$, block bootstrap), with Mistral as directionally consistent exploratory support. LRD $\times$ LoRA is directionally consistent with regime but does not reach per-model significance after autocorrelation correction.

2. **Two propagation regimes, with within-family scaling evidence.** Models cluster into *spike-and-suppress* (Phi-3.5, Gemma-2-9B) and *late-accumulation* (Llama-3, Mistral, Qwen2.5-7B). Holding family fixed, Qwen2.5 (1.5B$\rightarrow$ 7B$\rightarrow$ 14B) yields a monotonically increasing late/early LRD ratio ($1.59 \rightarrow 2.90 \rightarrow 4.10$), with monotonicity on every per-perturbation entry. A two-point Llama scan (1B$\rightarrow$ 8B: $2.08 \rightarrow 5.29$) corroborates the direction on a second family.

3. **Cascade disruption as the mechanism.** Early-layer LoRA breaks intact clean-input computation, with downstream damage peaking at layers 10–14 (downstream of the LoRA window itself). This explains why the patching-implicated early layers are the wrong adapter-placement target.

4. **A layer sweep confirms the mechanism, plus a negative result.** Under a fixed evaluation harness, patching-window and early-layer placements produce the largest regressions on the two adjudicable late-accumulation models (Qwen2.5-7B, Llama-3); Phi-3.5 is flat rather than positively repaired, and the only non-negative windows are the deepest available. An earlier version of our pipeline reported gains from a representation-stability loss; those gains do not survive the fixed harness, and the reversal is itself independent evidence for cascade disruption.

2 Related Work

Robustness to surface perturbations. Character-level and lexical perturbations degrade NLP systems across tasks (Wang et al., 2022; Belinkov and Bisk, 2018; Eger et al., 2019; Sun et al., 2020; Dong and Smith, 2018; Li et al., 2024). Rather than constructing worst-case inputs, we ask where in the network these failures localize.

Representational similarity. Probing (Belinkov, 2022), methods such as CKA (Kornblith et al., 2019) and SVCCA (Raghu et al., 2017), and model stitching (Bansal et al., 2021) measure what models encode at each layer. LRD differs: it tracks *perturbation-induced* change rather than absolute representational content, at every layer, giving a propagation trajectory rather than a snapshot.

Layer-wise function in transformers. Transformer layers are functionally heterogeneous; early layers handle surface structure and later layers encode more abstract semantic information (Tenney et al., 2019; Jawahar et al., 2019; Geva et al., 2022, 2023; Dar et al., 2023). The three-map dissociation adds a perturbation-robustness dimension: sensitivity, causality, and compensatory capacity each implicate different layer ranges, and which ranges depend on the propagation regime.

Activation patching and causal tracing. Replacing activations at specific layers with clean-run values is standard for causal localization (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Hanna et al., 2023; Conmy et al., 2023). We show that causal sensitivity is anti-correlated with compensatory capacity, so patching-guided placement produces poor intervention sites.

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) enables per-layer placement flexibility; we use this to measure compensatory capacity per layer.

3 Method

3.1 Layer-Wise Representation Divergence (LRD)

For mean-pooled hidden states $\bar{\mathbf{h}}^{(L)}_{\text{clean}}$ and $\bar{\mathbf{h}}^{(L)}_{\text{noisy}}$ at layer L on a (clean, perturbed) input pair,

$$\text{LRD}(L) = 1 - \cos\big(\bar{\mathbf{h}}^{(L)}_{\text{clean}}, \bar{\mathbf{h}}^{(L)}_{\text{noisy}}\big). \qquad (1)$$

Two summary statistics—*slope* (linear trend across layers) and *recovery rate* $\big(\text{LRD}(0) - \text{LRD}(L_{\max})\big)/\text{LRD}(0)$—distinguish spike-and-suppress (high recovery, non-positive slope) from late-accumulation (near-zero recovery, positive slope).

3.2 Perturbation Types

Six perturbation types: typos (5%), OCR (5%), whitespace (10%), case swap (10%), speech-like (10%), and homophones at three lexical (per-word) rates (20%, 40%, 50%; higher than the character-level rates so substitutions are not vanishingly rare). Unless noted, "Homophones" cells and the layer-sweep mean use the 20% severity.

3.3 Activation Patching

At each layer L we substitute the clean hidden state into the noisy forward pass and record the fraction of originally-failed examples (clean-correct, perturbed-incorrect) that now succeed; identity (all-layer) and random-noise patches bound the scale. Layers with substantial single-layer recovery form the **causal window**, with identity recovery $\geq 80\%$ gating results as *confirmed* (else *exploratory*).

3.4 Adapter Setup and Three-Map Signals

We attach LoRA adapters (r=4, α=8) to attention `q_proj`/`v_proj` (Phi-3.5 has fused attention, so we use `qkv_proj`/`o_proj` equivalently) with the base model frozen, training on perturbed GSM8K with clean supervision targets and cross-entropy on answer positions. The principal sweep is

CE-only; a representation-stability variant adding $L_{\text{stab}} = 1 - \cos(\bar{\mathbf{h}}^{\text{LoRA}}, \bar{\mathbf{h}}^{\text{frozen}})$ appears only for the ablation in §5.3. The three-map analysis compares three normalized per-layer signals: *sensitivity* (LRD(L)), *causality* (patching recovery at L), and *compensatory capacity* (perturbed-accuracy gain from a 5-layer LoRA window centered at L). Pairwise Spearman ρ is computed across layers per model; per-layer Cohen's d (Cohen, 1988) at L is the standardized mean difference in LRD(L) between failing and succeeding examples.

3.5 Cascade Disruption and Statistical Framework

Cascade disruption (§4.5) is the downstream LRD between frozen and adapted models on clean inputs. Two post-hoc intrinsic metrics, C3 (singular-value entropy of `q_proj`/`v_proj`) and C4 (Frobenius norm of the cross-entropy gradient on the same weights; full definitions in Supplemental), correlated with LoRA effectiveness on the original three models but failed prospective validation (§4.4); we report them as descriptive only. Because layer-wise signals autocorrelate (partial-autocorrelation decay to the 95% white-noise band by lags 5–8), we use the **moving-block bootstrap** (Künsch, 1989) (block size 5, 10k resamples) for correlation CIs and p-values; post-decorrelation effective N is typically 3–5.

4 Diagnostic Results

Setup. Diagnostic runs use n=500 per condition on Phi-3.5-mini-instruct (3.8B; Abdin et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3-8B-Instruct (Grattafiori et al., 2024), Gemma-2-9B base (Team et al., 2024), and Qwen2.5-7B-Instruct (Qwen et al., 2025) over GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021), and BBH (Suzgun et al., 2023). Clean GSM8K accuracies (model order above): 85.4%, 59.6%, 78.4%, 64.8%, 89.0%. A single fixed evaluation harness with sufficient `max_new_tokens` is used throughout (§5.3); full hyperparameters in Supplemental.

4.1 Two Propagation Regimes

Table 1 classifies the five models by recovery rate and slope (Qwen2.5-7B has an early transient spike at layers 0–3 but a dominant late-layer buildup, placing it in late-accumulation).

Per-perturbation recovery rates confirm the classification (e.g., typos: Phi-3.5 39.5% vs. Llama-3 0.5%; full breakdown in Supplemental).

Model	Regime	Clean	Recovery	Slope
Phi-3.5	S&S	85.4%	34–65%	≤ 0
Gemma-2-9B	S&S	64.8%	21–51%	≤ 0
Llama-3	Late	78.4%	0–5%	$+$
Mistral	Late	59.6%	0–5%	$+$
Qwen2.5-7B	Late	89.0%	0–5%[§]	$+$

Table 1: Five-model regime classification. S&S = spike-and-suppress. Recovery is the fraction of layer-0 divergence eliminated by the final layer; slope is the sign of the linear trend across layers. § Qwen2.5-7B homophones at 40–50% show recovery ≈40%; all other late-accumulation cells are near-zero.

4.1.1 Within-Family Scaling of the Late-Accumulation Regime

Holding family fixed and varying only scale deconfounds regime from architecture. The primary scan is Qwen2.5 at 1.5/7/14B (a 9× range, all with $\geq$ 28 layers); a two-point Llama corroboration (1/8B) gives a second family with the same direction modulo a narrow-bin caveat at the 17-layer 1B point (Figure 2, Table 2).

Qwen is treated as primary evidence and Llama as direction-only corroboration; neither is a controlled multi-family sweep, and we make no universal-scaling claim. (Llama-3-8B uses n=200 due to a compute budget; the late/early ratio is stable to ∼0.2 under bootstrap resampling at that size.) A Mann–Whitney consistency check on final-layer LRD (wrong vs. right) for Llama-3.2-1B-Instruct (n=500, clean 68.4%) matches the per-perturbation pattern we observe on Llama-3-8B (4-of-6 significant; whitespace and case null on both scales).

4.2 Per-Layer Cohen's d: Where the Failure Signal Emerges

Final-layer LRD predicts failure weakly compared to per-layer d because aggregation across depth dilutes the signal; per-layer d shows when succeed/fail separation develops (figures in Supplemental). On Phi-3.5, d is near-zero through layers 0–14 for most perturbation types; succeed/fail separation develops at layers 20–27 and peaks at layers 28–32. The same deep-layer concentration appears on the four other models (per-model panels in Supplemental). The patching-implicated early layers and the mid-layer range that adapter sweeps interrogate both lie upstream of where failures crystallize.

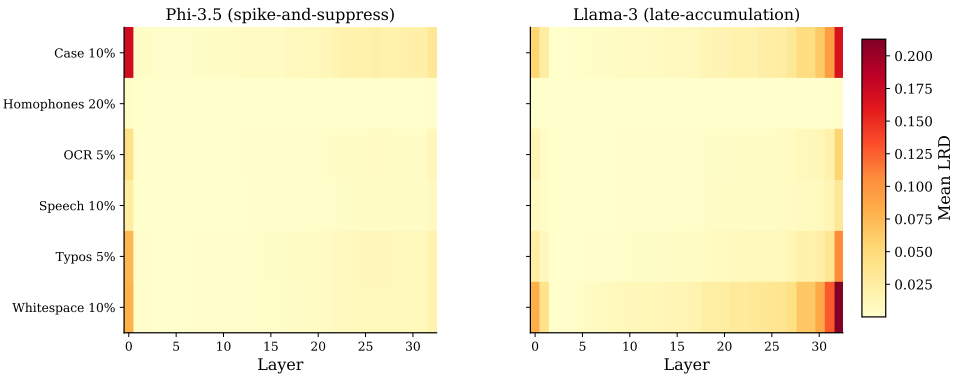


Figure 1: LRD heatmaps (GSM8K, typos 5%). Phi-3.5 shows spike-and-suppress; Llama-3 shows late-accumulation. Color encodes LRD magnitude on a shared scale; rows are layers, columns are examples. Gemma-2-9B, Qwen2.5-7B, and Mistral LRD profiles in Supplemental Materials.

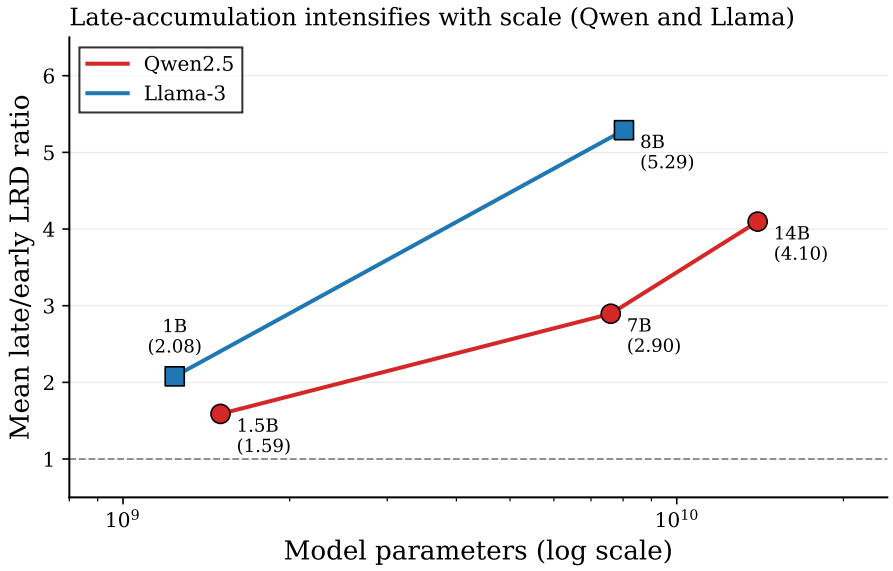


Figure 2: Within-family scaling of late accumulation. Mean late/early LRD ratio (deepest third over shallowest third, six-perturbation average) vs. parameter count (log x). Qwen2.5: $1.59 \rightarrow 2.90 \rightarrow 4.10$; Llama: $2.08 \rightarrow 5.29$. Every per-perturbation cell is monotonic within family (Table 2).

4.3 Activation Patching Locates the Causal Window

Single-layer clean patching (Figure 3) identifies the causal window on the two confirmed-gate models: Phi-3.5 (n=50 pairs, identity 100%) peaks at layers 1–8 and crosses baseline at $L18$; Llama-3-8B (n=100, identity 90%) peaks at $L1$ and stays above baseline through $L8$. Gemma-2-9B (identity 60%), Qwen2.5-7B (67%), and Mistral (identity recovery below the 80% gate) are exploratory.

Qwen at scale clears the patching gate. On Qwen2.5-14B (48-layer sweep, 84 clean-correct pairs of 100), identity recovery rises to 85.0%, clearing the gate the 7B variant did not (67%). The curve (Figure 4) is consistent with late-accumulation: early and mid thirds recover 73.1% and 71.1% of failed examples, while the late third ($L32$–47) falls to 42.3% (peak $L23$=82.1%). The random-patch control reached 20.5%, higher than expected—a caveat for effect-size attribution but not for the identity-ceiling promotion. The 14B curve is reported as a within-family consistency check, not a third entry in the pre-registered two-model Spearman panel.

4.4 Three-Map Dissociation

Overlaying the three per-layer signals (LRD, patching recovery, LoRA effectiveness; figures in Supplemental), the peaks do not coincide and the relationship between them differs by regime.

Family	Model	Typos	OCR	Whitespace	Case	Homophones	Speech	Mean
Qwen2.5	1.5B	1.65	1.53	1.68	1.42	1.34	1.92	**1.59**
	7B	3.19	3.40	2.66	2.40	2.24	3.49	**2.90**
	14B	4.20	4.88	3.55	3.17	3.84	4.93	**4.10**
Llama	3.2-1B	2.44	2.02	2.24	2.51	1.28	1.98	**2.08**
	3-8B[§]	5.90	5.86	4.50	5.24	3.25	6.99	**5.29**

Table 2: Late/early LRD ratio per perturbation for the two scaling scans; Mean averages the six entries. Every per-perturbation cell is monotonic within family. GSM8K, n=500, `max_new_tokens`=768; § Llama-3-8B used n=200.

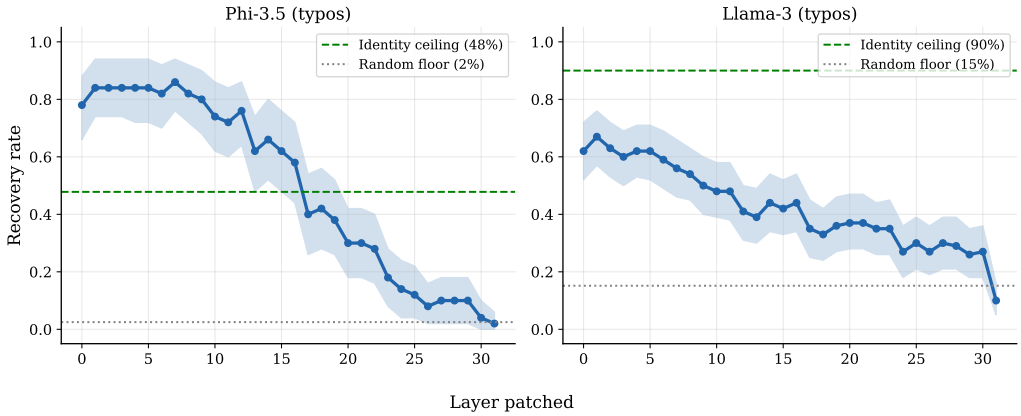


Figure 3: Activation patching recovery on the two confirmed-gate models (Phi-3.5, Llama-3; 3.8–8B). y-axes differ by identity ceiling (Phi-3.5 100%, Llama-3 90%). Early-layer patches recover the most failed examples.

LRD and patching are anti-correlated. On the two confirmed-patching models, $\rho = -0.72$ to -0.88, $p < 0.001$ (eff $N \approx 3$–5) on Phi-3.5 and Llama-3. Mistral matches the anti-correlation directionally but identity recovery falls below the 80% gate; we treat it as exploratory. Gemma-2-9B is an exception ($\rho = +0.389$); we conjecture this reflects the 60% identity ceiling, which we expect to corrupt the patching signal when perturbation effects are distributed across layers (the exception is not adjudicated by the panel-level analysis).

LRD × LoRA sign tracks regime. LoRA is most effective where LRD is suppressed (Phi-3.5, $\rho = -0.55$) or still accumulating (Llama-3/Mistral, $\rho = +0.63, +0.69$). After autocorrelation correction only Mistral reaches $p < 0.05$ individually ($p = 0.025$); we report the sign pattern across the panel rather than per-model significance. The regime evidence is correlational and mutually consistent across the anti-correlation, sign pattern, cross-task stability (§4.6), sweep asymmetry (§5), and Qwen scaling (§4.1.1); only the Qwen scan varies family independently of regime.

C3/C4 intrinsic metrics are post-hoc only. Frozen-model intrinsic metrics weight effective rank (C3) and gradient norm (C4) showed post-hoc correlation with LoRA effectiveness on the original three models ($|\rho| = 0.54$–0.75, regime-matched signs) but failed prospective validation: C3/C4 predicted L08–11 for Qwen and L24–29 for Gemma, but the fixed-harness sweep (§5) finds Qwen L08–11 to be the *worst* window (-7.8 pp) and Gemma flat at every completed window. We treat C3/C4 as descriptive only; full panel correlations are in Supplemental.

4.5 Cascade Disruption: Why Early Intervention Fails

For each trained LoRA window, we compute LRD between the frozen and adapted models on clean inputs at every downstream layer (Figure 5).

Early-window LoRA (L00–04, L05–09) induces clean-input disruption peaking at layers 10–14 on the three models with full coverage (Phi-3.5, Llama-3, Mistral); later windows disrupt much less, and partial Gemma/Qwen-7B coverage (Supplemental) is directionally consistent. Total clean disruption rank-orders inversely with perturbed accuracy gain (scatter in Supplemental): LoRA placed where patching points breaks clean-input processing, so the anti-correlation between patching win-

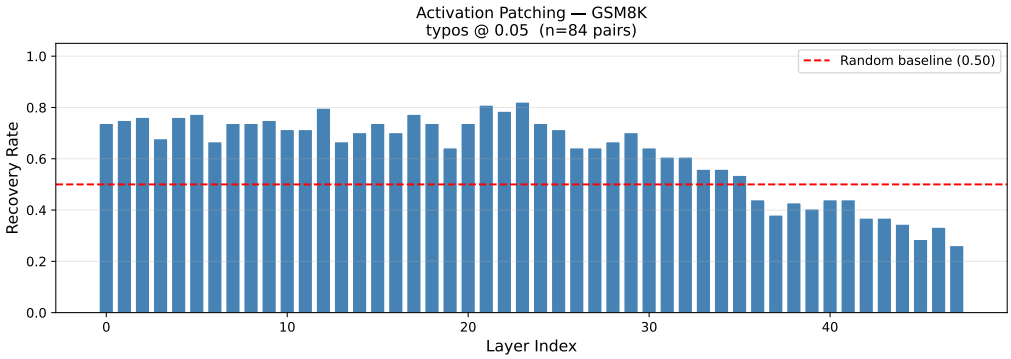


Figure 4: Activation patching recovery on Qwen2.5-14B (48 layers, identity 85.0%). Peak at $L23$=82.1%; the late third ($L32$–47) collapses to 42.3%, consistent with late-accumulation.

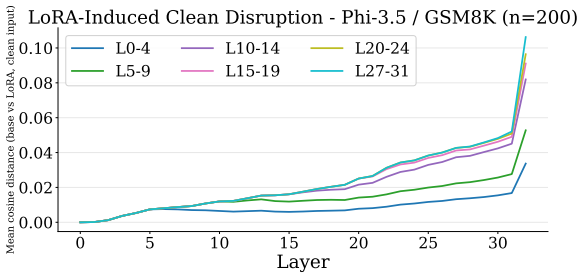


Figure 5: Clean-input disruption for Phi-3.5 (one line per LoRA window). Early windows peak at L10–14; later windows disrupt less. Other models: Supplemental.

dow and best LoRA window is mechanistic, not coincidental. Independent re-runs reproduce the disruption measurements within the bf16 numerical floor (max abs. diff. 0.0025 on Mistral, $\leq$0.0002 on Phi-3.5 / Llama-3).

4.6 Cross-Task Stability of Propagation Regimes

LRD profiles on Phi-3.5 and Mistral (the two models for which we collected cross-task data) are near-identical across GSM8K, MMLU, and BBH ($\rho > 0.95$ in both models): task modulates LRD magnitude but not profile shape (figure in Supplemental).

4.7 Held-Out Perturbation Types

To test whether the regime signature is a property of the model rather than of the perturbation distribution, we ran LRD on three held-out perturbations (`char_insert`, `char_swap`, `qwerty`; absent from training and the main analysis), one model per regime. Final-layer LRD on GSM8K: Phi-3.5 (S&S) 0.016, 0.023, 0.020; Llama-3-8B (Late) 0.066, 0.105, 0.082—a $\sim$5$\times$ regime gap. Severity and error rate also dissociate: Llama-3-8B has higher LRD but higher raw perturbed accuracy (87–90% vs. Phi-3.5 at 60–72%), consistent with LRD measuring cascade severity rather than error rate.

5 Layer Sweep: Diagnostic-Flagged Sites Resist Repair

The cascade-disruption result (§4.5) predicts that the patching-implicated early layers should be *the worst*, not the best, sites for perturbation-correction adapters. We test that prediction with a non-overlapping layer sweep on four small-to-mid models (Phi-3.5 at 3.8B; Qwen2.5-7B, Mistral, Llama-3 at 7–8B) plus the 9B Gemma base, under the fixed harness used throughout.

Setup. We train CE-only LoRA (r=4, α=8, attention `q_proj`/`v_proj`; λ_{stab}=0) on perturbed GSM8K with clean supervision, sweeping non-overlapping 5-layer windows across the depth of each model with three seeds per window. The perturbed Δ is the mean change in perturbed accuracy over the six perturbation types relative to the no-adapter baseline, under the fixed harness (§5.3).

5.1 Per-Model Layer Sweeps

Phi-3.5 (S&S; clean 85.4%, 32 layers). Every mid-layer window is strongly negative; the only non-negative window is L27–31 (+0.2 pp, effectively null). Seed std ranges 0.07–0.92 pp, no window crosses zero under seed variation, and the earlier-harness L15–19 "optimum" falls to -3.5 under the fixed harness.

Other panel models. Table 3 gives the full grid. Qwen2.5-7B's worst window is the C3/C4-predicted L08–11 (-7.8 pp); its least-regressive is the deepest, L24–27 (+0.3). Llama-3-8B has

Model (clean)	L00–04	L05–09	Early-mid	Mid	Mid-late	Late
Phi-3.5 (85.4%)	-1.4 ± 0.7	-0.9 ± 0.1	-5.3 ± 0.9	-3.5 ± 0.8	-0.1 ± 0.2	$+0.2 \pm 0.3$
Qwen2.5-7B (89.0%)	-0.6 ± 0.2	-4.0 ± 0.3	-7.8 ± 1.7	-3.4 ± 0.2	-0.3 ± 0.1	$+0.3 \pm 0.0$
Llama-3-8B (78.4%)	-2.5 ± 0.2	-6.4 ± 0.5	—	-2.7 ± 0.8	$+0.4 \pm 0.2$	$+1.1 \pm 0.1$
Mistral-7B-v0.3 (59.6%)	-7.8 ± 2.5	-17.7 ± 1.1	—	-19.4 ± 0.9	-8.5 ± 0.4	-4.3 ± 0.3
Gemma-2-9B (64.8%)[s]	$+0.2$	-0.2	—	$+0.5$	$+0.2$	0.0

Table 3: Mean perturbed Δ (with-adapter minus no-adapter, six perturbations on 500 GSM8K items) per cell; mean $\pm$ std across three seeds. Nominal 5-layer windows vary slightly by partition: Early-mid L10–14 (Phi-3.5) / L08–11 (Qwen); Mid L15–19 (L15–20 for Gemma); Mid-late L20–24 (L20–23 for Qwen; L25–30 for Gemma); Late L27–31 (L24–27 for Qwen; L35–40 for Gemma). Dashes (—) mark windows absent by design. [s] Gemma single-seed per window (seed-42 for L00–04/L05–09/L15–20; seed-43 for L25–30 and L35–40); L00–04 also replicates at seed-44 (+0.23). *Mistral row shown for completeness but excluded from adjudication (lr artifact; §5.2).* Qwen L08–11 is the C3/C4 prediction and Qwen's worst window—a pre-registered falsifier of C3/C4 as a placement rule.

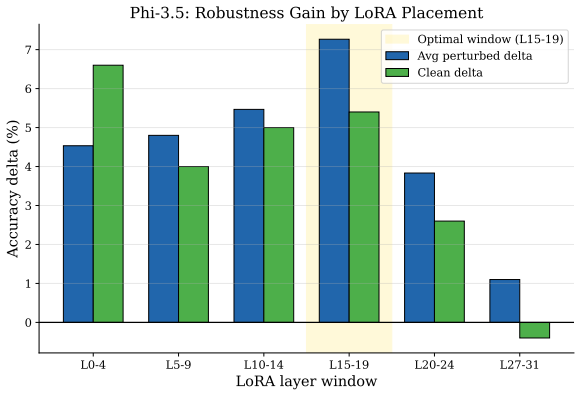


Figure 6: Fixed-harness CE-only LoRA layer sweep. Mean perturbed Δ (six perturbations, three seeds) on Phi-3.5, Qwen2.5-7B, Llama-3, and Mistral. On the three adjudicable models the mid-layer L08–L20 range is uniformly most damaging; non-negative windows are the deepest available. Mistral's row reflects an lr artifact (§5.2). Gemma-2-9B is adapter-insensitive ($\pm$0.7 pp) and reported separately.

L20–24 and L27–31 non-negative (+0.4, +1.1); the latter is the cleanest late-layer positive signal in the panel. Mistral-7B-v0.3 is uniformly negative at the shared lr; a single-cell audit at lr$=1\mathrm{e}{-5}$ recovers to +0.13 pp, consistent with an lr artifact (§5.2). Gemma-2-9B is adapter-insensitive (every window within $\pm$0.7 pp of zero) and we treat it as "no measurable signal."

5.2 The Mid-Layer Wall and the Pre-Registered Placement Tally

Two structural observations follow from Table 3. First, **mid-layer windows uniformly backfire**: across the three adjudicable models (Phi-3.5, Qwen-7B, Llama-3), every window in the L05–L20 range is strongly negative—the range cascade disruption (§4.5) flags as most harmful when broken—and the patching-implicated early windows (L00–04) are also negative, though less catastrophically. Second, **least-regressive windows are the deepest ones**: the only non-negative windows on the three models are the deepest available, confirming the diagnostic prediction that placement should be *away from* where divergence and patching causality concentrate.

Adjudicated tally. Excluding Mistral (lr artifact) and Gemma (non-adjudicable, flat at every window), the diagnostic predicted that windows away from the patching-implicated early layers would be less damaging than the flagged layers; the (post-hoc) C3/C4-implicated mid layers were additionally pre-registered as Qwen-7B's worst. Qwen ✓ (L24–27 beats L08–11 by 8.1 pp); Llama-3 ✓ (L27–31 beats L05–09 by 7.5 pp); Phi-3.5 × (no window meaningfully above zero). Two holds and one fail. We claim the mid-layer wall, not a positive placement recipe; Phi-3.5's flatness is part of why placement is demoted to a confirmation of cascade disruption.

5.3 An Attempted Intervention and a Harness Artifact

During this project's development we explored a representation-stability loss $L_{\text{stab}} = 1 - \cos(\bar{\mathbf{h}}^{\text{LoRA}}, \bar{\mathbf{h}}^{\text{frozen}})$ added to CE LoRA. An earlier (pre-fix) harness with `max_new_tokens=100` reported +7.3 pp on Phi-3.5 at L15–19 and a regime-dependent split with smaller (or negative) reported gains on Qwen2.5-7B. Under the fixed harness (`max_new_tokens=768`) those gains reverse (Figure 7): Phi-3.5 L15–19 records −3.5 pp, and the Qwen2.5-7B L24–27 vanilla CE baseline reads −7.5 pp (was reported +11.6 pp on the same checkpoint). The short-decode harness scored

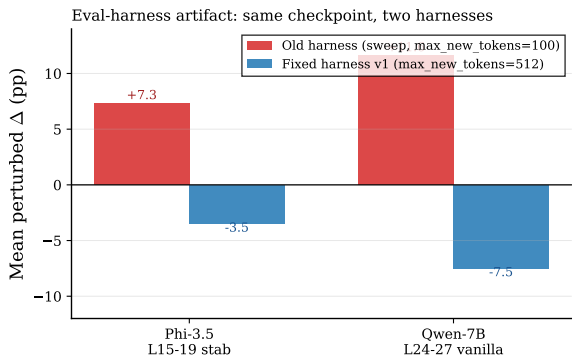


Figure 7: Same checkpoints, two harnesses. Mean perturbed Δ for Phi-3.5 L15–19 (earlier-harness $+7.3$ pp) and the Qwen2.5-7B L24–27 vanilla CE baseline (earlier-harness $+11.6$ pp) under `max_new_tokens`$=100$ (blue) vs. 768 (orange); n=500. The earlier gains were truncated chain-of-thought scored as empty.

truncated chain-of-thought generations as empty answers, so conditions producing shorter (often less accurate) output were credited with higher accuracy. Holding the fixed harness and toggling only λ_{stab}, adding the stability term ($\lambda_{\text{stab}} = 1.0$) *hurts* perturbed accuracy by 1.2–3.9 pp across the three seeds on Phi-3.5 L15–19 (the same window/condition where the earlier harness reported $+7.3$ pp). The diagnostic, mechanism, and scaling results in §4 are independent of this intervention claim, and the reversal rules out the regime-dependent stabilizer-constraint hypothesis we earlier entertained in favor of cascade disruption.

6 Discussion

Restoration vs. adaptation. The causal window (layers 1–8) identifies where *restoring* representations helps, not where adaptation is effective. Cascade disruption explains the shared anti-early-layer rule across regimes, and the layer sweep instantiates it—the diagnostic-flagged sites are the worst, not the best. Within our panel, regime initially tracked family; the Qwen2.5 and Llama within-family scans weaken that confound but do not eliminate it, and we make no universal-scaling claim.

Believing intervention claims, and open questions. The §5.3 reversal shows intervention gains on chain-of-thought tasks can be dominated by generation-length artifacts: for any future representation-engineering intervention, the relevant null is the same intervention under a generation budget long enough for the model to emit its answer. Two follow-ups would tighten the picture: a controlled multi-family scan varying architecture and scale independently, and extending the layer sweep to non–chain-of-thought tasks where harness-length artifacts cannot dominate.

7 Conclusion

Sensitivity, causality, and compensatory capacity dissociate per layer, and within-family scaling on Qwen2.5 and Llama weakens the family/regime confound that a single five-model panel cannot. Cascade disruption predicts that patching-implicated layers resist repair, and a fixed-harness layer sweep confirms this on three adjudicable models (3.8–8B). An earlier representation-stability intervention does not survive fixed-harness re-evaluation; we report the reversal alongside the diagnostic that anticipates it.

Limitations

Scope of intervention claims. We provide no positive intervention recipe. All placement-related claims are relative-ranking statements under a fixed harness ("window A is less regressive than window B"); absolute gains over no-adapter baselines are not claimed. The earlier-pipeline representation-stability gains were a harness artifact (§5.3); the underlying cascade-disruption mechanism is independently supported by the diagnostic.

Panel size. Our regime taxonomy is observed on a five-model panel. The primary de-confounding evidence is the three-point Qwen2.5 scan (§4.1.1); the two-point Llama scan corroborates the direction on a second family but has a narrow-bin caveat at the 1B end (17 hidden layers). Two families is not a controlled multi-family sweep, and confirming regime as a general transformer property still requires within-family scans on additional families and, ideally, controlled pretraining recipes.

Task and language scope. The layer-sweep adjudication uses GSM8K only. Cross-task LRD stability is measured on two models (Phi-3.5 and Mistral; §4.6) across GSM8K/MMLU/BBH, but we do not run a fixed-harness layer sweep on the other tasks; results may not transfer to non–chain-of-thought settings where harness-length artifacts cannot dominate. All experiments are English-only.

Patching ceiling effects. Gemma-2-9B and Qwen2.5-7B identity ceilings (60%, 67%) limit

causal interpretation when perturbation effects distribute across layers. Qwen-14B clears the 80% gate (85.0%) but with a 20.5% random-patch warning that weakens effect-size attribution on the 14B curve.

Statistical power. Three seeds per layer-sweep window characterizes each cell's direction and magnitude but limits the precision of fine-grained between-window comparisons. The block-bootstrap correlation analyses use block size 5 from the partial-autocorrelation decay; we did not run a block-size sensitivity sweep, and the post-correction effective $N \approx 3$–5 is already small.

Mistral learning-rate sensitivity. At the shared-grid lr ($5\mathrm{e}{-5}$) every Mistral window is strongly negative; a single-cell re-run at $1\mathrm{e}{-5}$ recovers to $+0.13$ pp, consistent with an lr artifact rather than a capacity issue, though we did not re-run every window. Mistral is excluded from the placement-prediction adjudication.

Code and data. Code, configurations, and per-model evaluation logs will be released at camera-ready.

Use of AI assistants. AI coding assistants were used during this project for code scaffolding, experiment orchestration (job-launch scripts, log parsing, plot generation), and writing assistance (copy-editing and reorganization of paper text). All scientific claims, experimental designs, statistical analyses, and result interpretations are the authors' own; AI-generated text and code were reviewed and verified by the authors before inclusion.

Artifact use and licensing. All models and benchmarks used here are publicly released research artifacts, cited at first use in §4 and listed in references. Licenses, to the best of our knowledge at the time of writing: Phi-3.5-mini-instruct (MIT), Mistral-7B-Instruct-v0.3 (Apache 2.0), Llama-3-8B-Instruct and Llama-3.2-1B-Instruct (Llama-3 Community Licenses), Gemma-2-9B (Gemma Terms of Use), Qwen2.5-{1.5B, 7B, 14B}-Instruct (Apache 2.0), GSM8K (MIT), MMLU (MIT), BBH (Apache 2.0). Our usage—measuring representational properties and training diagnostic LoRA adapters for academic research—is consistent with the intended use stated by each artifact's release. The benchmarks (GSM8K, MMLU, BBH) consist of grade-school math word problems, multiple-choice exam questions, and reasoning challenges respectively, and to our knowledge contain no personally identifying information or offensive content; each is documented in its release paper (cited).

Potential risks. This work is diagnostic and does not release new models, training data, or adversarial tools. The perturbation types we study (typos, OCR noise, whitespace, case, speech-like substitutions, homophones) are commonplace and not novel attack vectors. The diagnostic itself (LRD, patching recovery, layer sweeps) could in principle inform adversarial perturbation design or targeted intervention against deployed models; we judge this risk low because the techniques and perturbations are already public. The more practical risk is *misuse of the regime taxonomy*: extrapolating the spike-and-suppress / late-accumulation labels to models outside our five-model panel without re-running the diagnostic, or treating the post-hoc C3/C4 metrics as a placement rule despite the falsification in §4.4. The Limitations above are written to make both kinds of over-reach explicit. Finally, the layer-sweep and diagnostic experiments are GPU-intensive (multi-model LoRA sweeps with three seeds per window, plus activation patching at every layer); we report exact configurations in Supplemental so practitioners can scope re-runs to the cells they need.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias

Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.

Rui Dong and David Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, Melbourne, Australia. Association for Computational Linguistics.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Hans R. Künsch. 1989. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2024. Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics*, 150:104598.

Hanyu Liu, Chengyuan Cai, and Yanjun Qi. 2023. Expanding scope: Adapting English adversarial attacks to Chinese. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 276–286, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2020. Robustness to modification with shared words in paraphrase identification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 164–171, Online. Association for Computational Linguistics.

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *Preprint*, arXiv:2003.04985.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

This supplement provides extended results referenced in the main paper. §A contains LRD heatmaps for the three additional models (Gemma-2-9B, Qwen2.5-7B, Mistral) not displayed in main paper Figure 1, and §B adds Qwen-1.5B and Qwen-14B heatmaps that complement the within-family scaling figure (main paper Figure 2, §4.1.1). §C gives the full 48-layer patching curve underlying main paper Figure 4. §D–§F give per-perturbation recovery rates, the full panel of pairwise Spearman correlations across all five models, and the intrinsic-predictor (C1–C4) definitions and C3/C4 falsification table. §G–§J contain per-model figure panels for the three-map overlay (all five models), Cohen's d (all five models), intrinsic capacity metrics (Llama-3/Mistral), and cascade disruption (Llama-3/Mistral). §K–§L provide the disruption-vs-gain scatter (Phi-3.5, Llama-3, Mistral) and the cross-task LRD profiles (Phi-3.5, Mistral). §M contains the per-model layer-sweep panels underlying main paper Table 3. §N briefly catalogs investigations performed before the evaluation harness was fixed—the cosine vs. MSE stability-loss matched-rate controls, the λ_{stab} scaling sanity-check, the LoRA capacity experiments, and a residual-adapter stabilizer tested on Mistral and Qwen2.5-7B—and explains why those numbers are excluded from this revision. §O lists reproducibility settings for the fixed-harness runs.

A LRD Profiles: Gemma-2-9B, Qwen2.5-7B, and Mistral

Figure 8 shows LRD heatmaps for the three models not displayed in main paper Figure 1. Gemma-2-9B exhibits the spike-and-suppress profile (large early divergence, suppressed through the middle layers, with a mild secondary hump near the final layers). Qwen2.5-7B exhibits the late-accumulation profile, with an early transient spike at layers 0–3 and dominant late buildup through layers 20–28. Mistral shows a gradual late-accumulation buildup through depth.

B Within-Family Scaling: Qwen-1.5B and Qwen-14B Heatmaps

Figure 9 shows LRD heatmaps for Qwen-1.5B and Qwen-14B; the Qwen-7B heatmap is in Figure 8b. Together with the within-family scaling figure in the main paper (Figure 2), these visualize the same late-accumulation regime at three Qwen scales: an early transient spike near layers 0–3 that fades, followed by a depth-dominant buildup whose late/early ratio intensifies monotonically with scale ($1.59 \rightarrow 2.90 \rightarrow 4.10$).

C Qwen2.5-14B Per-Layer Patching Recovery

Figure 10 gives the full 48-layer activation-patching recovery curve for Qwen2.5-14B on GSM8K typos 5% (84 clean-correct pairs of 100 attempted). The main paper summarises the curve into per-region recovery and the per-layer peak ($L23$, 82.1%) and worst ($L47$, 26.2%); the figure shows that the curve is broadly consistent with late-accumulation, with the highest recovery concentrated in the early-to-mid third and a clear collapse in the deepest layers.

D Per-Perturbation Recovery Rates

Table 4 gives per-perturbation recovery rates for all five models (GSM8K). Cascade slope summaries are in main paper Table 1.

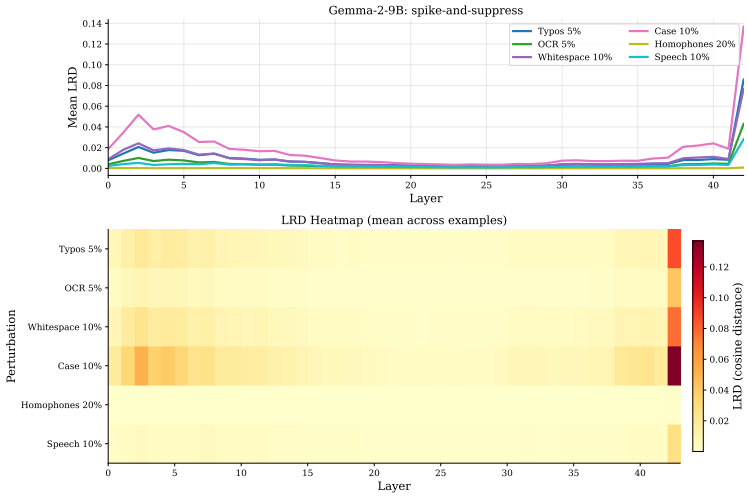
E Pairwise Spearman Correlations: All Five Models

F Intrinsic Predictor Details

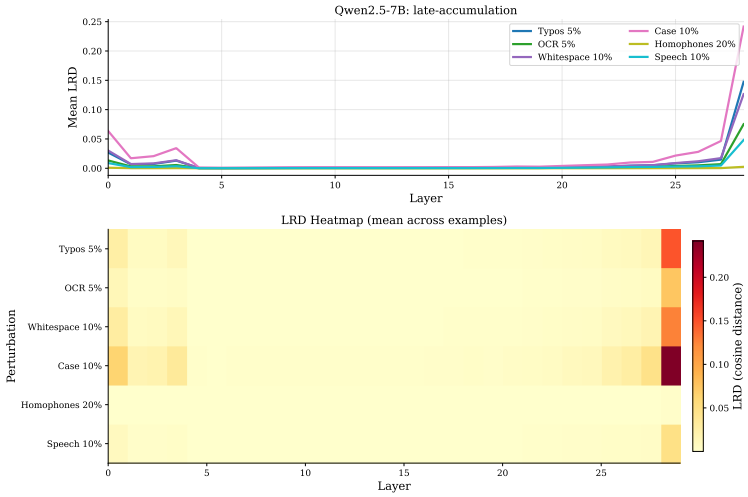
- **C1: Update ratio.** $C_1(L) = \|\mathbf{h}^{(L)} - \mathbf{h}^{(L-1)}\|_2 \,/\, \|\mathbf{h}^{(L-1)}\|_2$, where $\mathbf{h}^{(L)}$ is the hidden state at layer L, mean-pooled over valid (non-padded) token positions per example. Computed on $n = 200$ clean GSM8K examples and averaged across examples.

- **C2: Attention entropy.** $C_2(L) = -\sum_k p_k \log p_k$, where p_k are the attention probabilities at layer L (extracted via `output_attentions=True`). Entropy is computed per-head per-query-position, then averaged across (batch, heads, query positions) on $n = 200$ clean GSM8K examples.

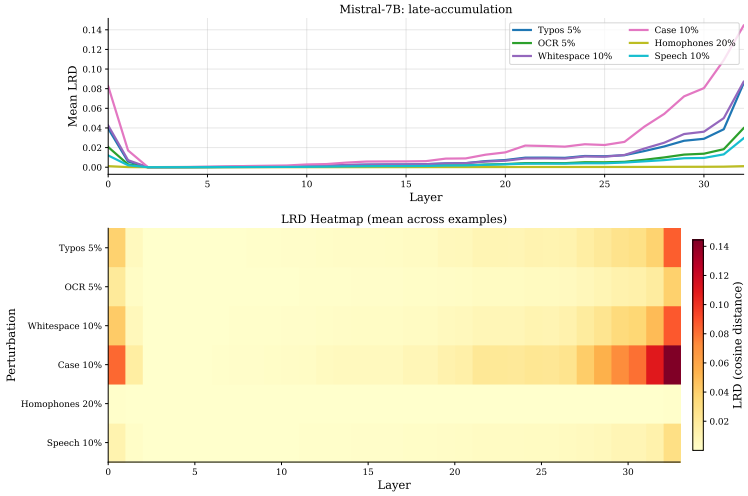
- **C3: Weight effective rank (stable rank).** $C_3(L) = \exp(-\sum_i \tilde{\sigma}_i \log \tilde{\sigma}_i)$, with $\tilde{\sigma}_i = \sigma_i / \sum_j \sigma_j$ and σ_i the singular values of the layer's weight matrix. This is a smooth, differentiable rank proxy (not a threshold-based count). Computed statically on `q_proj` and `v_proj`; for Phi-3.5 fused `qkv_proj`, the Q and V slices are extracted and the per-layer score is $(r_Q + r_V)/2$. No forward pass required.



(a) Gemma-2-9B. Spike-and-suppress profile with a mild secondary hump at layers 30–41 and a sharp jump at the final layer.

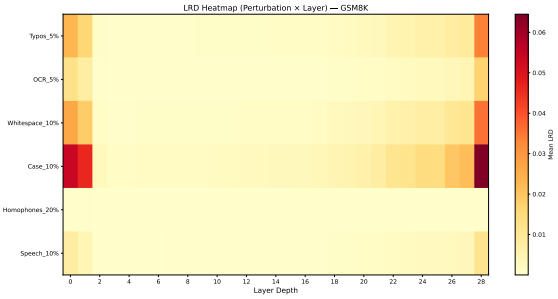


(b) Qwen2.5-7B. Late-accumulation profile with an early transient spike at layers 0–3 and dominant late buildup through layers 20–28.

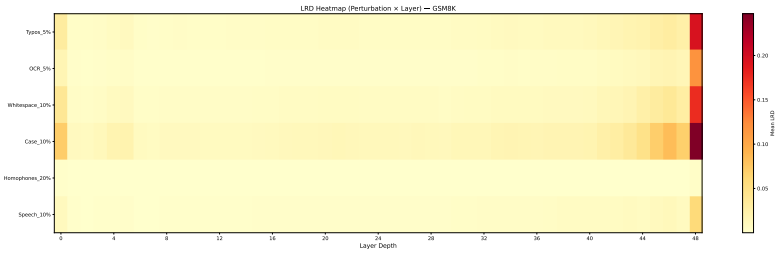


(c) Mistral. Late-accumulation profile; gradual buildup through depth.

Figure 8: LRD heatmaps (GSM8K, typos 5%) for the three additional models. Compare Phi-3.5 and Llama-3 in main paper Figure 1.



(a) Qwen2.5-1.5B (late/early ratio 1.59). Late-accumulation profile is present but mild; depth-driven divergence is visible after roughly the midpoint of layers.



(b) Qwen2.5-14B (late/early ratio 4.10). Late-accumulation profile is sharper: deep-layer divergence dominates earlier and on every perturbation type.

Figure 9: LRD heatmaps (GSM8K, typos 5%) for Qwen2.5-1.5B and Qwen2.5-14B. Compare with Qwen2.5-7B in Figure 8b. The late-accumulation regime persists at all three scales and intensifies monotonically with size.

Perturbation	Phi-3.5	Gemma-2-9B	Llama-3	Mistral	Qwen2.5-7B
Typos 5%	39.5%	34.5%	0.5%	1.5%	1.8%
OCR 5%	60.0%	35.5%	3.5%	4.5%	5.2%
Whitespace 10%	43.5%	38.5%	0.0%	0.0%	0.8%
Case 10%	47.5%	36.5%	0.0%	0.0%	0.0%
Homophones 20%	56.5%	51.5%	34.5%	38.0%	39.8%
Speech 10%	34.5%	21.0%	1.0%	0.5%	1.6%

Table 4: Per-perturbation recovery rates for all five models (GSM8K). Recovery is the fraction of layer-0 divergence eliminated by the final layer. Spike-and-suppress models (Phi-3.5, Gemma-2-9B) show partial recovery on all types; late-accumulation models (Llama-3, Mistral, Qwen2.5-7B) show near-zero recovery, except homophones at 20% severity (the only condition producing meaningful late-model recovery). Sample sizes: Phi-3.5, Mistral, Qwen2.5-7B $n = 500$; Llama-3, Gemma-2-9B $n = 200$.

Model	Pair	ρ	95% CI	p
Phi-3.5	LRD × Patch	-0.881	$[-0.991, -0.787]$	$<0.001^{***}$
	LRD × LoRA	-0.550	$[-0.884, +0.624]$	0.33
	Patch × LoRA	$+0.482$	$[-0.840, +0.909]$	0.44
Llama-3	LRD × Patch	-0.723	$[-0.982, -0.450]$	$<0.001^{***}$
	LRD × LoRA	$+0.631$	$[-0.018, +0.908]$	0.054
	Patch × LoRA	-0.642	$[-0.760, +0.091]$	0.072
Mistral	LRD × Patch	-0.752	$[-0.969, -0.550]$	$<0.001^{***}$
	LRD × LoRA	$+0.685$	$[+0.132, +0.854]$	0.025^{*}
	Patch × LoRA	-0.731	$[-0.824, -0.080]$	0.033^{*}
Gemma-2-9B	LRD × Patch	$+0.389$	$[-0.43, +0.83]$	0.264
	LRD × LoRA	$+0.091$	$[-0.38, +0.69]$	0.418
	Patch × LoRA	$+0.735$	$[+0.26, +0.92]$	$0.006^{**\dagger}$
Qwen2.5-7B	LRD × Patch	$+0.025$	$[-0.575, +0.728]$	0.866
	LRD × LoRA	$+0.413$	$[-0.763, +0.726]$	0.802
	Patch × LoRA	-0.167	$[-0.529, +0.356]$	0.602

Table 5: Pairwise Spearman correlations (block bootstrap, block size 5, 10k resamples). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. † Gemma-2-9B patching is exploratory (60% identity ceiling). LRD × LoRA is directionally consistent with regime but significant only on Mistral after autocorrelation correction.

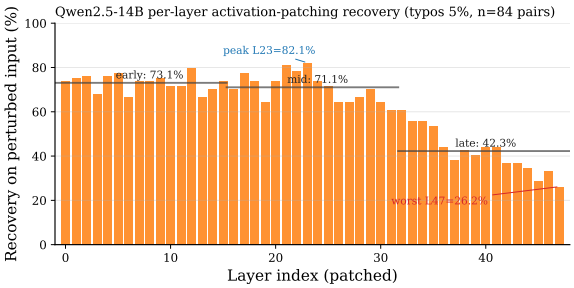


Figure 10: Qwen2.5-14B activation-patching recovery across all 48 layers on GSM8K typos 5%. Identity recovery 85.0% (above the 80% gate). The peak is at $L23$ (82.1%); the worst single layer is $L47$ (26.2%). The random-patch control on this model recovered 20.5% (logged warning, see main paper).

- **C4: Gradient norm.** $C_4(L) = \frac{1}{2}\left(\left\|\partial\mathcal{L}_{\text{CE}}/\partial W_Q^{(L)}\right\|_F + \left\|\partial\mathcal{L}_{\text{CE}}/\partial W_V^{(L)}\right\|_F\right)$, the mean of the Frobenius gradient norms of the Q and V weight matrices under teacher-forced cross-entropy loss. Computed on $n = 50$ clean GSM8K examples (subset of the 200) with batch size 2; requires forward+backward pass.

C3 and C4 show the strongest *post-hoc* correlations with LoRA effectiveness on the original three models (Table 6); C1 and C2 are reported in the same table for completeness. As reported in the main paper, the prospective validation on Qwen-7B (C3/C4 predicted L08–11) is actively falsified by the fixed-harness layer sweep: L08–11 is the worst window for Qwen-7B (-7.8 pp mean perturbed Δ, three seeds), not the best. We do not recommend C3/C4 as a placement rule.

Metric	Phi-3.5 ρ	Llama-3 ρ	Mistral ρ
C1: Update ratio	-0.517^{**}	$+0.394^{*}$	$+0.430^{*}$
C2: Attn entropy	$+0.211$	-0.712^{***}	-0.296
C3: Weight eff. rank	-0.627^{***}	$+0.751^{***}$	$+0.670^{***}$
C4: Gradient norm	$+0.537^{**}$	-0.714^{***}	-0.594^{***}

Table 6: Intrinsic capacity metric correlations with LoRA effectiveness (original three models; i.i.d. bootstrap, anti-conservative given eff $N \approx 3$–5; significance markers are indicative). C3 and C4 flip sign between Phi-3.5 and the late-accumulation models.

Table 7 summarises C3/C4 predictions versus the least-regressive window under the fixed-harness layer sweep (§5 of the main paper). The original-three-model fits stay post-hoc; on the two pre-registered models, the prediction is falsified on Qwen-7B and is moot on Gemma (flat sweep).

Model	C3/C4 predicted	Least-regr. window	Outcome
Phi-3.5	L15–19	L27–31	Post-hoc only; sign flipped
Llama-3	L20–24	L27–31	Post-hoc only; close, not best
Mistral	L20–24	L27–31	Post-hoc only; close, not best
Qwen2.5-7B	L08–11	L24–27	**Falsified** (predicted = worst)
Gemma-2-9B	L24–29	(flat)	Moot ($\rho \approx 0$)

Table 7: C3/C4 predicted window versus the fixed-harness least-regressive window. The Phi-3.5/Llama-3/Mistral entries are post-hoc fits to the broken-harness sweep; under the fixed harness no window is non-negative on Phi-3.5 or Mistral, and the least-regressive window is L27–31 on all three. On Qwen-7B, the C3/C4 prediction is the *worst* window in the sweep (-7.8 pp; pre-registered falsifier). Gemma's sweep is flat.

G Three-Map Overlays: All Five Models

Figure 11 shows the per-layer overlay of the three signals (LRD sensitivity, patching recovery, LoRA effectiveness) for each of the five models in the panel. Numerical correlations across the panels are in Table 5.

H Per-Layer Cohen's d: All Five Models

Figure 12 shows the per-layer Cohen's d on failing vs. succeeding examples for each model. Effect sizes peak in the deep layers, downstream of every adapter window tested in the fixed-harness sweep (§5 of the main paper).

I Intrinsic Capacity Metrics: Llama-3 and Mistral

Figure 13 shows C3 (weight effective rank) and C4 (gradient norm) for Llama-3 and Mistral. Both models show the opposite sign pattern from Phi-3.5 (numerical correlations in Table 6). These curves are descriptive: C3/C4 was demoted to post-hoc only after the prospective validation on Qwen-7B was falsified by the fixed-harness layer sweep (main paper §4.4, Table 7).

J Cascade Disruption: Llama-3 and Mistral

Figure 14 replicates the Phi-3.5 cascade disruption finding (main paper Figure 5) on Llama-3 and Mistral.

K Clean Disruption vs. Perturbed Gain: Scatter

Figure 15 shows, for each of the three models with full window-coverage cascade data (Phi-3.5, Llama-3, Mistral), the rank-order relationship between total clean-input disruption (cf. main paper Figure 5 and Figure 14) and perturbed accuracy gain. Higher disruption tracks lower gain in all three.

L Cross-Task LRD Profiles: Phi-3.5 and Mistral

Figure 16 shows the cross-task LRD comparison for the two models on which we collected GSM8K/MMLU/BBH data (main paper §4.6). Both models give $\rho > 0.95$ between tasks, indicating that task modulates LRD magnitude but not profile shape.

M Per-Model Layer Sweep Panels

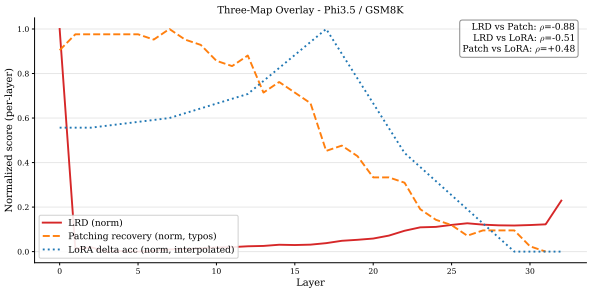
Figure 17 shows, for each model in the panel, the fixed-harness CE-only LoRA layer sweep broken out by window with three-seed dispersion visible per cell. These panels underlie the per-model layer-sweep table in the main paper. The 7–8B models share a clear mid-layer wall (every window in the L05–L20 range is strongly negative); the deepest available window is the least-regressive. Mistral is uniformly negative; Gemma-2-9B is adapter-insensitive across every window we tested.

N Earlier Investigations (Pre-Fixed-Harness)

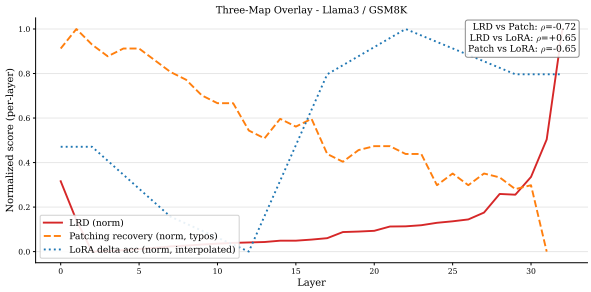
This section catalogs investigations carried out before the evaluation harness bug was identified (main paper §5.3). All numerical results in those experiments were generated by an evaluation harness with `max_new_tokens=100`, which truncated chain-of-thought generations before the `#### NUMBER` extraction line and scored truncated answers as empty. The same checkpoints read between $+10$ and $+20$ pp differently under the fixed harness, depending on condition (see main paper Figure 7 and §5.3). The qualitative conclusions of these investigations are preserved below; full numerical tables are omitted from this revision because they are not trustworthy on absolute scale.

Cosine versus MSE stability loss, with matched-rate control. We ran cosine and MSE variants of the representation-stability loss at the original variable per-perturbation rates and at a matched 10% rate. The original rates produced a per-condition split (cosine winning on typos/OCR/speech/homophones, MSE on whitespace/case); the matched-rate control collapsed that split. Cohen's d values for the so-called "directional" and "uniform" groups overlapped completely (0.40–0.48 vs. 0.15–0.48). Pooled interaction CI: $[+2.5\%, +8.7\%]$ at variable rates, $[-2.5\%, +3.5\%]$ at matched rates. The taxonomy claim from this loss-geometry work was disproved by the matched-rate control; the underlying intervention has now also been disproved by the harness fix.

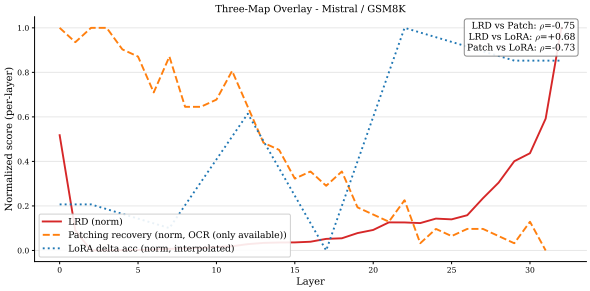
λ_{stab} scaling sanity check. At $\lambda_{\text{stab}} = 0.3$, L_{stab} contributed $\leq 0.6\%$ of the total loss in every run. Setting $\lambda_{\text{stab}} = 3.0$ raised the contribution to 6–8%. This established that earlier cosine-vs-MSE comparisons at $\lambda_{\text{stab}} = 0.3$ were essentially noise.



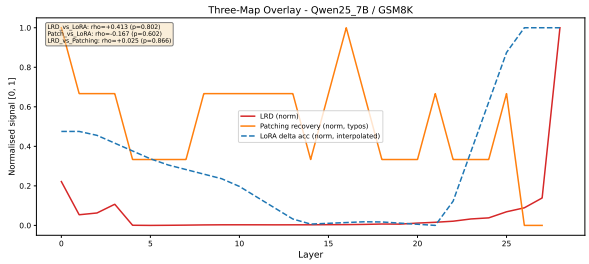
(a) Phi-3.5 (S&S): LoRA peaks where LRD is suppressed.



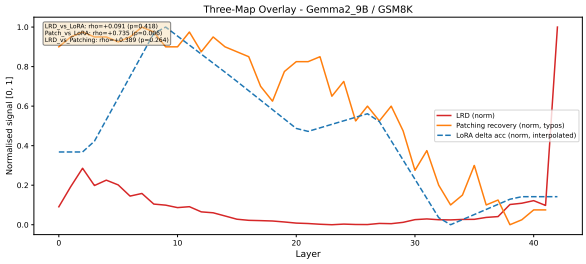
(b) Llama-3 (Late): LoRA peaks where LRD is building.



(c) Mistral (Late): LoRA anti-correlated with patching window.

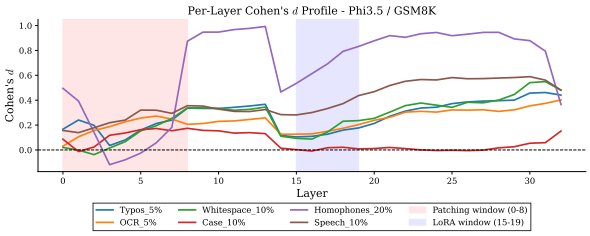


(d) Qwen2.5-7B (Late): LRD$\times$Patch unusable (n=3).

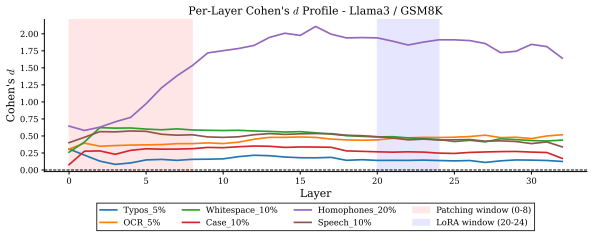


(e) Gemma-2-9B (S&S): exception, ρ=+0.389 (60% ceiling).

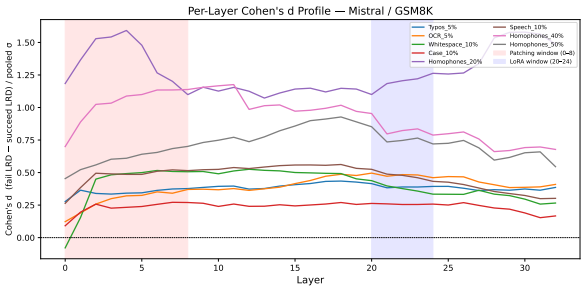
Figure 11: **Three-map overlays, all five models.** LRD sensitivity, patching recovery, and LoRA effectiveness per layer.



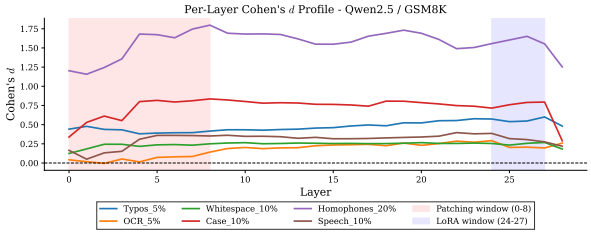
(a) Phi-3.5 (S&S): peak at L28–32.



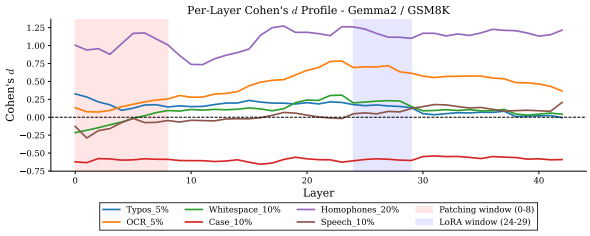
(b) Llama-3 (Late).



(c) Mistral (Late).

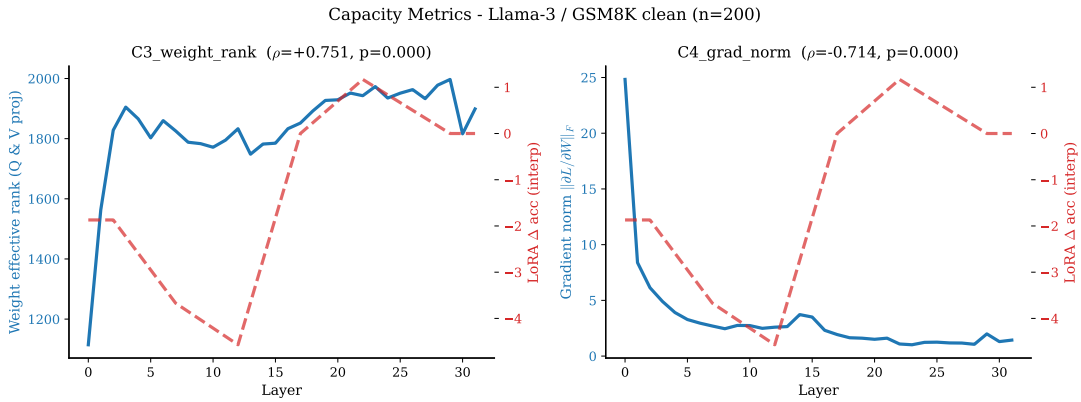


(d) Qwen2.5-7B (Late): peak $d \approx 1.80$ at L8 (Homophones 20%).

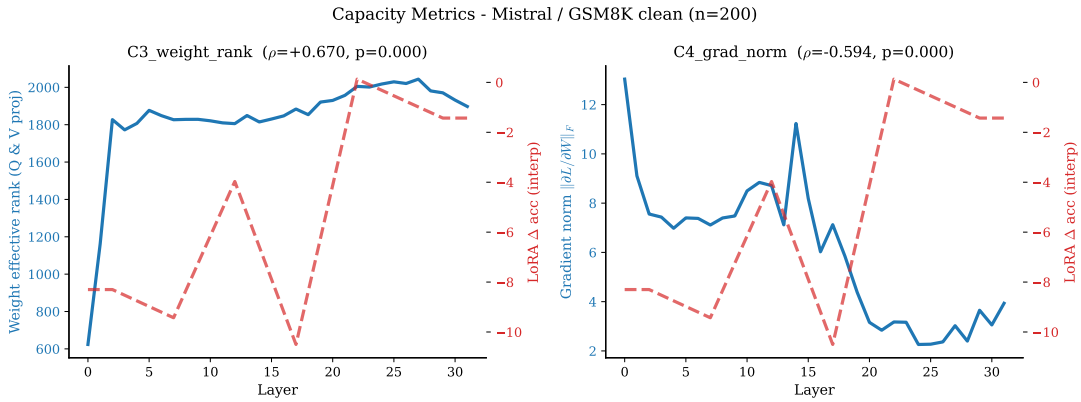


(e) Gemma-2-9B (S&S): peak $d \approx 1.27$ at L18 (Homophones 20%).

Figure 12: **Per-layer Cohen's d, all five models.** Shaded band: activation-patching causal window. Effect sizes peak in the deep layers, downstream of every adapter window tested under the fixed harness.

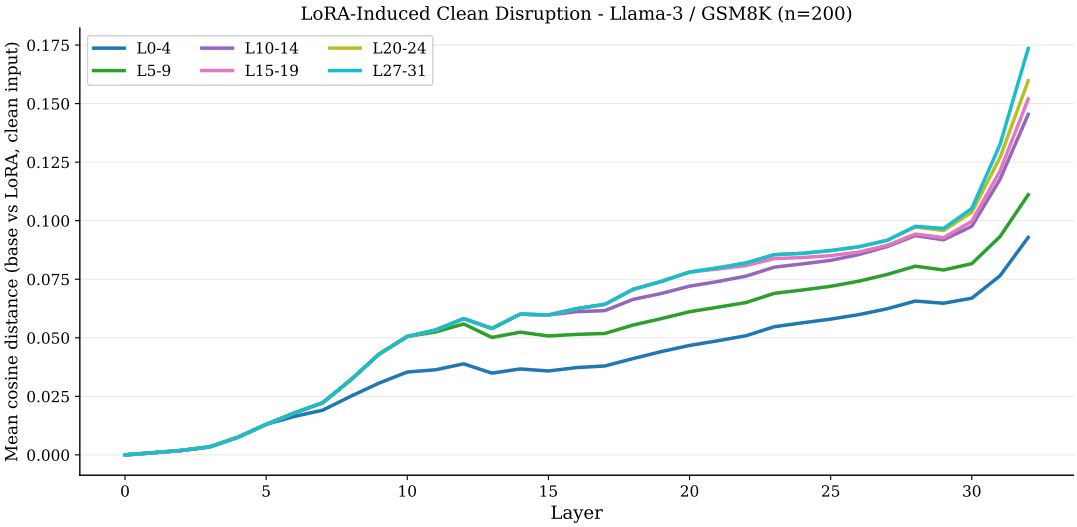


(a) Llama-3.

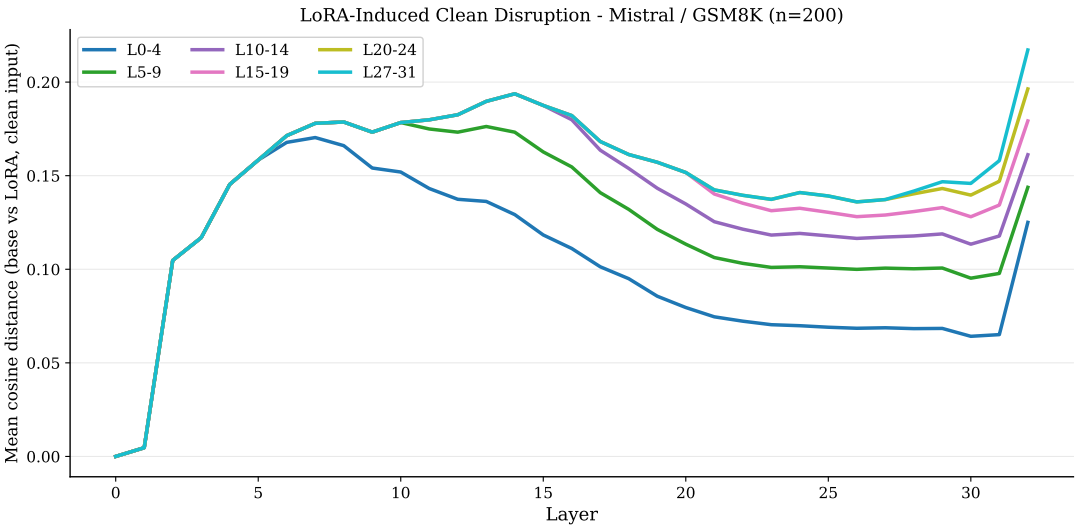


(b) Mistral.

Figure 13: **Intrinsic Capacity Metrics: Llama-3 and Mistral.** C3 (weight effective rank, left) and C4 (gradient norm, right) for each model. Both models show the opposite sign pattern from Phi-3.5 (correlations in Table 6). C1 and C2 correlations are reported in Table 6.

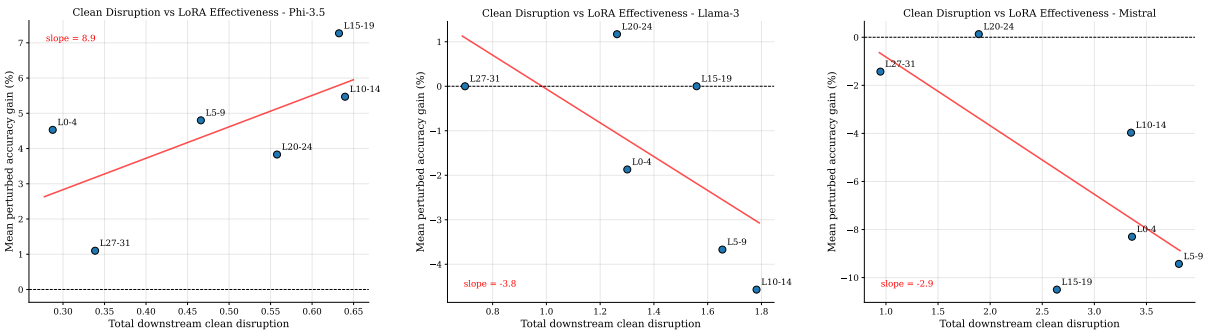


(a) Llama-3.



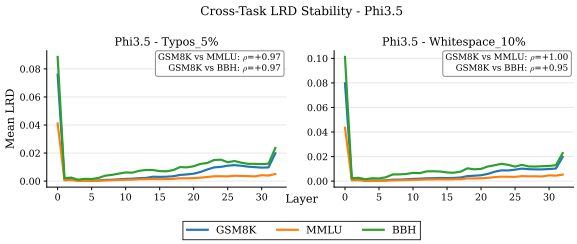
(b) Mistral.

Figure 14: **Cascade Disruption: Llama-3 and Mistral.** Clean-input disruption (one line per LoRA window). Early-window LoRA induces disruption peaking at layers 10–14 on both models, consistent with Phi-3.5 (main paper Figure 5).

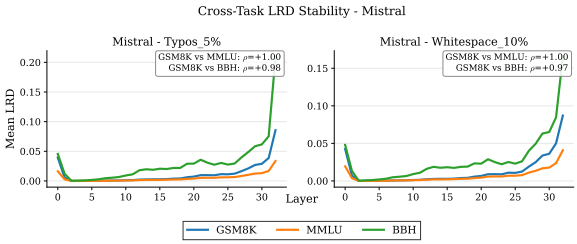


(a) Phi-3.5. (b) Llama-3. (c) Mistral.

Figure 15: Total clean-input disruption vs. perturbed accuracy gain on the three models with full window-coverage cascade sweeps.

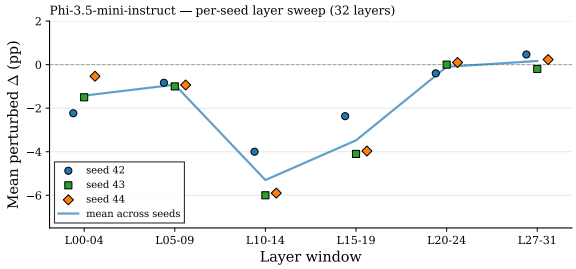


(a) Phi-3.5 (spike-and-suppress).

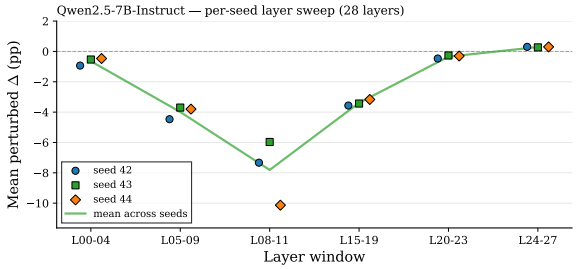


(b) Mistral (late-accumulation).

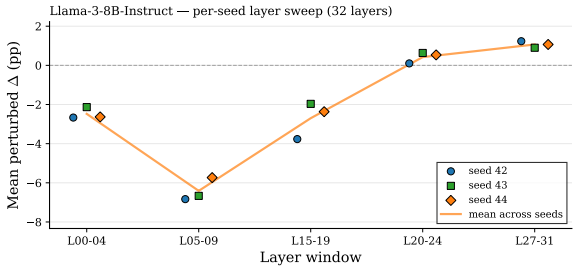
Figure 16: Cross-task LRD profiles (GSM8K, MMLU, BBH; typos 5%) for Phi-3.5 and Mistral. Both produce near-identical shapes across tasks ($\rho > 0.95$).



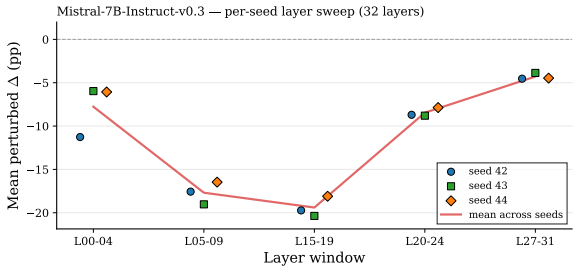
(a) Phi-3.5 (spike-and-suppress; clean 85.4%).



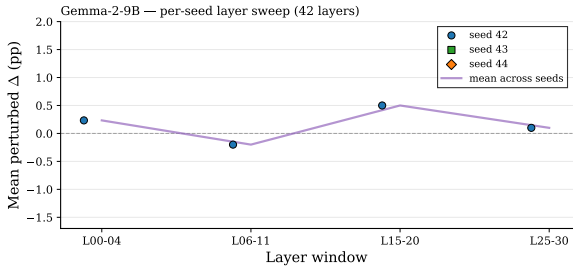
(b) Qwen2.5-7B (late-accumulation; clean 89.0%).



(c) Llama-3-8B (late-accumulation; clean 78.4%).



(d) Mistral-7B-v0.3 (late-accumulation; clean 59.6%; panel outlier).



(e) Gemma-2-9B base (spike-and-suppress; clean 64.8%; single-seed per window with one cross-seed replication at L00–04).

Figure 17: **Per-Model Layer Sweep Panels.** Fixed-harness CE-only LoRA, mean perturbed Δ per non-overlapping 5-layer window, three seeds per window (seeds 42, 43, 44). The numerical summary is in the per-model layer-sweep table in the main paper.

Under the fixed-harness ablation reported in the main paper, however, even the properly weighted $\lambda_{\text{stab}} = 1.0$ regime *hurts* CE LoRA by 1.2–3.9 pp on the conditions and seeds we tested. The stability loss is not a useful component once the evaluation harness is correct.

LoRA capacity experiments. We swept LoRA ranks $\{16, 32, 64\}$ at $\{3{,}000, 5{,}000\}$ training steps on Llama-3 and Mistral and observed monotonic degradation with rank and steps. The qualitative conclusion—that the prior stabilizer's failure on late-accumulation models is not a capacity issue—survives the harness fix. A separate "capacity concentration" result claimed that a concentrated 5-layer LoRA at L15–L19 outperforms an all-layer LoRA by $+5.1$ pp perturbed; that $+5.1$ number is broken-harness and is excluded from this revision. The cascade-disruption explanation in the main paper does not depend on it.

Transfer evaluation. GSM8K-trained adapters with the prior stability loss degraded MMLU by 1.5–4.5 pp and were near-zero on BBH. Task-specific MMLU training produced near-zero average gains. Diagnostic-side cross-task stability (LRD profile $\rho > 0.95$ across GSM8K/MMLU/BBH on Phi-3.5 and Mistral) is harness-independent and is retained in the main paper.

Alternative residual-adapter stabilizer. Independently of the LoRA-based stabilizer, we tested a lightweight two-layer bottleneck residual adapter $x \mapsto x + g \cdot W_2\sigma(W_1 x)$ inserted after the attention block at four placement configurations on Qwen-7B (layers 2/6/10/14; 8/12/16/20; 12/16/20/24; 4/10/18/26) and at the first eight layers on Mistral. Training mixed clean and OCR-perturbed inputs with clean targets. Every configuration regressed: Qwen-7B baseline outperformed all four stabilized variants by approximately 9–11 pp on GSM8K OCR-5%; Mistral baseline exceeded the stabilized variant by approximately 4.5 pp. The qualitative result (multiple stabilizer designs, including a non-LoRA one, fail to improve robustness on these models) is consistent with the main-paper negative result. As with the rest of this section, the absolute numbers are pre-fixed-harness.

Full per-condition numbers from each of these investigations are available on request and will be released with the public code drop.

O Reproducibility Settings

Component	Configuration
Eval harness	`max_new_tokens`=768, CoT, #### extraction
Diagnostics sample size	$n = 500$ per condition
Scale-variation sample size	$n = 500$ test; $n_{\text{clean-correct}}$=200
Patching pairs	Phi-3.5 n=50; Llama-3 n=100; Qwen-14B n=84/100
LoRA rank / alpha	4 / 8
LoRA targets	`q_proj`, `v_proj`
LoRA loss	cross-entropy only (λ_{stab}=0)
Layer-sweep windows	non-overlapping 5-layer windows across depth
Layer-sweep seeds	$\{42, 43, 44\}$ (three per window)
Training steps	300 (Phi-3.5); 1,000–3,000 (Llama-3, Mistral, Qwen, Gemma)
Bootstrap	block size 5, 10,000 resamples
Hardware	NVIDIA A40 (46GB), A100 (80GB), V100 (32GB); single-GPU jobs
Cluster	SLURM
Total compute	≈900 GPU-hours (reported experiments; excludes exploratory and failed runs)

Table 8: Reproducibility settings for the fixed-harness experiments in the main paper. The λ_{stab}=1.0 ablation referenced in main paper §5.3 uses the same settings with the stability-loss term re-enabled. The pre-fixed-harness investigations in §N used different settings (`max_new_tokens`=100, λ_{stab}=3.0, fixed optimal windows).

Software stack. Single-GPU PyTorch `2.10.0+cu128` with the following libraries: `transformers 4.44.2`, `peft 0.11.0`, `datasets 4.5.0`, `accelerate 1.12.0`, `numpy 2.2.5`, `scipy 1.15.3`. Exact pinned versions and the full `requirements.txt` accompany the code release.

Model checkpoints and loading. Weights are loaded from the official HuggingFace releases: `microsoft/Phi-3.5-mini-instruct`, `mistralai/Mistral-7B-Instruct-v0.3`, `meta-llama/Meta-Llama-3-8B-Instruct`, `google/gemma-2-9b`, `Qwen/Qwen2.5-7B-Instruct`, `Qwen/Qwen2.5-1.5B-Instruct`, `Qwen/Qwen2.5-14B-Instruct`, and `meta-llama/Llama-3.2-1B-Instruct`. All models load in `torch_dtype=bfloat16` with `attn_implementation="eager"` where required by the activation-patching hooks.

LoRA configuration (`peft`). `r=4`, `lora_alpha=8`, `target_modules=["q_proj","v_proj"]` for Mistral, Llama-3, Gemma-2, and all Qwen scales; `target_modules=["qkv_proj","o_proj"]` for Phi-3.5-mini (fused attention). Inference-mode flag toggled per call; otherwise default `peft` arguments.

Datasets (`datasets`). `openai/gsm8k` (main config, test split) drawn with shuffle seed 42 and `n_samples` $\in \{200, 500\}$ per condition as reported in Table 8; MMLU and BBH loaded from their respective official HuggingFace releases (cited in the

main paper Setup).

Optimizer (`torch`). `torch.optim.AdamW` with default $(\beta_1, \beta_2, \epsilon)$ and weight decay; learning rates as specified in the main paper (shared-grid sweep at $5\mathrm{e}{-5}$, Mistral audit at $1\mathrm{e}{-5}$). Training in `bfloat16` except where noted.

Statistics (`scipy`, `numpy`). `scipy.stats.pointbiserialr`, `scipy.stats.mannwhitneyu`, `scipy.stats.levene`, and `scipy.stats.spearmanr` are used at default arguments for the per-perturbation significance checks and the pairwise correlations reported above. The moving-block bootstrap is an in-house NumPy implementation (block size 5, $10\mathrm{k}$ resamples).