RESIDUAL SCALING FOR LANGUAGE MODELS: AN EXACT LAYER SCORE AND AN ERROR BUDGET FOR COMPOSING IT

Anonymous authors
Paper under double-blind review

ABSTRACT

For scalar gains on the residual contributions of a frozen language model, the signed gradient–residual product is the exact loss derivative at identity. We locate where its predicted benefit is lost when composing multiple layers. The derivative ranks single-layer amplification at Spearman 0.83–0.99 on matched prompts and 0.45–0.98 on held-out prompts, whereas measured ablation ranks it at -0.19 to 0.02 on the same Qwen3.5-9B items. The gap between summed predictions and joint gain splits exactly into cross-item transfer, a single-coordinate finite step and cross-layer interaction. Across four checkpoints of three families, the finite step dominates on GSM8K at $\alpha = 1.3$; for three-layer profiles, interaction is at most 30% of it. Transfer grows linearly and the finite step locally quadratically with strength. Coefficients frozen at $\alpha = 1.1$ predict switches between these terms on ARC-Challenge: registered fp16 measurements on the original questions locate them at 1.31–1.35, and on disjoint questions at 1.22–1.27, within the registered 0.10 tolerance. Choosing triples by measured single-layer gains lowers joint-loss regret from 0.041 to 0.0014; measuring eight derivative-ranked candidate blocks recovers this choice at an estimated 45% of the recorded single-block stage time.

1 INTRODUCTION

Can a few scalar changes to a frozen language model be chosen from a cheap layer score? Activation scaling is a simple intervention family (Stoehr et al., 2024), and gradient–activation products are a common first-order proxy for the effect of an intervention (Kramár et al., 2024). Choosing a multi-layer intervention from such a score, however, chains three predictions: the score must predict the effect of each layer, that prediction must hold for the finite strength that is deployed and for the items it is deployed on, and the single-layer effects must compose.

We study scalars on native residual updates, $x_{l+1} = x_l + (1 + \delta_l) f_l(x_l)$. For this family the signed gradient–residual product is not an approximation of the score but the exact derivative of the loss at identity, so every gap between a predicted and a measured gain has a definite source. Ablation and importance scores usually decide which units to prune or which layer to edit (Molchanov et al., 2019; Meng et al., 2022), which makes ablation sensitivity a natural candidate for where to amplify. Ablation sets $\delta_l = -1$ and small amplification uses $0 < \delta_l \leq 0.3$; both have the same local derivative, so the family also separates what the derivative predicts from what a finite intervention does.

What is new. Work on scalar activation interventions (Stoehr et al., 2024), gradient attribution (Kramár et al., 2024), layer selection (Sun et al., 2025; Gadgil et al., 2026) and joint multi-layer learning (Rodríguez et al., 2025) evaluates an intervention or a score as a whole. This paper locates where the benefit predicted by an exact layer score is lost on the way to a joint loss gain: the gap splits exactly into cross-item transfer, the within-layer finite step and cross-layer interaction, each measured on the same items.

Contributions.

1. **Single-layer prediction holds, and ablation does not transfer to amplification.** The signed derivative ranks single-strength amplification at Spearman 0.83–0.99 on its own prompts in

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.

15 model–task units of three families and at 0.45–0.98 on held-out prompts in ten. Measured full ablation ranks the same amplification at only -0.19 to 0.02 on Qwen3.5-9B with all three quantities measured on the same items, where the derivative reaches 0.58–0.77, and at 0.19–0.64 on Qwen3-8B across runs, as Sections 4.2 and 4.3 show.

2. **An exact decomposition of the composition gap.** The difference between summed single-layer predictions and a measured joint gain is the sum of transfer, finite-step and interaction terms; single-layer gains certify a joint gain under a uniform mixed-derivative bound, and smooth counterexamples in Section 3.3 show that single-layer and pairwise checks alone cannot.
3. **A measured error budget on four checkpoints of three families.** Section 4.4 shows that on the same GSM8K items at $\alpha = 1.3$ the finite step is the largest term in every profile set of the four checkpoints, and for three-layer profiles interaction stays at most 30% of it. Transfer is exactly linear and the finite step locally quadratic in the strength; coefficients frozen at $\alpha = 1.1$ name the leading term at the already measured $\alpha = 1.3$ in all five checkpoint–task groups, and registered fp16 measurements on the same ARC-Challenge questions find the switch between the two terms at 1.35 and 1.31 against a predicted 1.32 and 1.33, within 0.08 for every single seed. Ranking candidates by measured rather than predicted single-layer gains cuts the joint-loss regret of the chosen triple from 0.041 to 0.0014, and measuring only the eight candidate blocks with the largest derivative recovers that choice at an estimated 45% of the recorded single-block stage time.

2 Related Work

Scalar and learned activation interventions. Activation Scaling (Stoehr et al., 2024) learns signed scalars on activations of a frozen model by gradient descent with effectiveness, faithfulness and sparsity objectives and can also take attribution-patching scores directly as scalars; LinEAS (Rodríguez et al., 2025) learns maps at many layers end-to-end because layer-local maps ignore downstream shifts. Scaling also acts on neurons (Akhtar et al., 2026) and fine-tuning residuals by depth (Wang et al., 2025), and targeted sub-network updates amplify circuits for mathematical reasoning (Prakash et al., 2025). Restricting the gains to native block residuals makes the layer score an exact derivative and each link from score to joint gain measurable on the same items.

Gradient attribution and its error. In this family the attribution-patching inner product becomes the exact derivative $\nabla_{o_l}\mathcal{L}^\top f_l$ of the loss at identity, the derivative of an identity-initialized scalar gate that first- and second-order pruning criteria use for structure removal (Molchanov et al., 2019); we use it signed, for positive gains of a frozen model. AtP* (Kramár et al., 2024) corrects its false negatives for patching; concurrent work (Zhang & Wang, 2026) traces its error mainly to downstream nonlinearity and adds a Hessian-vector correction. Gradient attributions also build steering vectors (Nguyen et al., 2026), and layer Jacobians support feedback steering (Skifstad et al., 2026). We measure where the prediction of this score loses fidelity for a finite positive gain and for a joint profile. On GSM8K at $\alpha = 1.3$ the single-coordinate finite step is the largest of three error terms on the four checkpoints we trace in Section 4.4, consistent with the patching diagnosis, and the composition condition of Eq. 5 states what a joint certificate additionally needs. Integrated Hessians (Janizek et al., 2021) attribute path interactions but give no such bound.

Choosing where to intervene. Steering layers are chosen by activation discriminability and consistency (Sun et al., 2025), logit-lens accessibility (Billa, 2026), a learned per-input selector (Gadgil et al., 2026), or, concurrently, single-layer marginal effects refined per instance (Nelwan & Wicaksono, 2026); other concurrent work asks which activations should supply the signal (Ye et al., 2026). These methods add vectors and are validated on behaviour targets. Causal tracing localizes factual recall (Meng et al., 2022), yet localization does not predict the best layer to edit (Hase et al., 2023), and patching conclusions depend on metric and corruption (Zhang & Nanda, 2024); our ablation-versus-amplification result is one instance of this gap. Our measurements cover each link between an exact layer score and a joint loss gain on the same items, with magnitude and depth priors as controls.

Reliability and evaluation of steering. Steering vectors (Turner et al., 2023; Zou et al., 2023; Li et al., 2023) vary across inputs and prompts (Tan et al., 2024), fail to help or degrade many of up to 36 tested models (Da Silva et al., 2025) and trail prompting on a large benchmark (Wu et al., 2025); evaluations should use task contexts, likelihoods and baselines (Pres et al., 2024), and

regex answer extraction is itself error-prone (Yu et al., 2025). MAT-Steer (Nguyen et al., 2025) and Activation Scaling's faithfulness term limit unwanted changes during learning. Layer removal with healing (Gromov et al., 2025) shows that many deeper layers can be removed with little loss, and structured massive activations (Sun et al., 2024) show that residual magnitude need not mark a useful direction, which is why magnitude and depth priors are controls here.

3 METHOD

3.1 RESIDUAL INTERVENTION AND CALIBRATION PROTOCOLS

Consider a pre-norm transformer with L layers whose residual stream evolves as $\mathbf{x}_{l+1} = \mathbf{x}_l + f_l(\mathbf{x}_l)$, where f_l is the full attention-plus-MLP block output added at layer l. The intervention studied throughout is a scalar gain on that contribution,

$$\mathbf{x}_{l+1} = \mathbf{x}_l + \alpha_l f_l(\mathbf{x}_l), \qquad \delta_l = \alpha_l - 1, \tag{1}$$

with $\alpha_l = 0$ being zero-ablation and $\alpha_l \in \{1.05, \ldots, 1.30\}$ being amplification. The backbone weights are frozen. The residual gains are calibration parameters: derivative and loss-based rules use gold-answer supervision, while norm, depth and random rules use no labels. The experiment tables distinguish the following four procedures.

1. **Atlas**, an exhaustive scan: for each layer independently, evaluate the calibration loss at $\alpha_l \in \{1.0, 1.05, 1.10, 1.15, 1.20, 1.30\}$ with all other layers at 1; keep the α_l^* that minimizes calibration loss, or set $\delta_l = 0$ if none improves. "Atlas top-k" composes the k layers with the largest calibration gain at $\alpha = 1.3$.
2. **GILS screen**: rank layers by the exact signed derivative of Section 3.2, from one forward–backward pass on the calibration split; "GILS top-k" applies one common α to the top k layers.
3. **Strength**: the ranking tables use $\alpha = 1.3$ for every layer; composed profiles apply one common $\alpha \in \{1.1, 1.3\}$ to a triple; only the atlas uses per-layer α_l^*.
4. **Shortlist profiles**, on Qwen3.5-9B only: a pool rule proposes six layers, all $\binom{6}{3} = 20$ triples are scored on a disjoint selection set at $\alpha = 1.3$, and the argmax is frozen as described in Appendix E; these frozen triples are the profiles whose loss chain Section 4.4 measures.

The profile objects differ in what chooses them. *Derivative top-k*, used in Tables 1 and 2 and in the Llama-3.1-8B chain, is chosen by calibration derivatives and judged by held-out loss; a Qwen3.5-9B *derivative triple* is chosen from the derivative's six-layer pool by selection-item first answers and judged on the chain of test-item losses; a *measured-sum* choice ranks candidates by summed calibration single-layer loss gains and is judged by the joint loss gain on disjoint items. All loss comparisons of rules use the same candidates, items and strength. *Ablation sensitivity* is the measured loss increase when layer l is zeroed, $S_l = \mathcal{L}(\alpha_l{=}0) - \mathcal{L}_{\text{base}}$; *amplifiability* is the measured loss decrease $A_l = \mathcal{L}_{\text{base}} - \mathcal{L}(\alpha_l{=}1.3)$, so positive values mean improvement for both. Every baseline called "ablation" in this paper is this measured S_l. Three magnitude rules appear as baselines and are named separately: the block-output norm $\sigma_l = \|f_l\|$ of Tables 1 and 2, the relative update $r_l = \|f_l\|/\|\mathbf{x}_l\|$ of Appendix Table 12 and the residual energy $e_l = \sum_t \|f_l(t)\|^2$, a Qwen3.5-9B pool rule; Appendix A.6 shows that they rank different layers.

3.2 THE SIGNED DERIVATIVE FOR A SINGLE-LAYER INTERVENTION

Treat α_l as a scalar input to the loss. Taylor expansion at $\alpha_l = 1$ gives

$$\mathcal{L}(1+\delta) - \mathcal{L}(1) = \delta\, \nabla\mathcal{L}^\top f_l + \tfrac{\delta^2}{2}\, f_l^\top \mathbf{H}_l f_l + O(\delta^3), \tag{2}$$

where $\nabla\mathcal{L} = \nabla_{\mathbf{o}_l}\mathcal{L}$ is the backward gradient at the block output and $\mathbf{H}_l$ the downstream Hessian along f_l; the remainder is $O(\delta^3)$ when $\alpha_l \mapsto \mathcal{L}$ is three times continuously differentiable on the segment between 1 and $1+\delta$, as it is for smooth activations and normalizations with nonzero variance. We define

$$\begin{aligned} \mathrm{GILS}_l &= -\delta\, \nabla\mathcal{L}^\top f_l \\ &= -\delta\, \|\mathbf{g}_l\|\, \sigma_l\, \varphi_l, \qquad \sigma_l = \|f_l\|,\ \varphi_l = \cos(\mathbf{g}_l, f_l), \end{aligned} \tag{3}$$

the first-order predicted gain. At fixed upstream gains, $f_l(x_l)$ does not depend on its own α_l, so this scalar derivative is exact; finite-step prediction is the approximation. Appendix H.1 factorizes

the score into gradient norm, residual norm and the gradient–residual cosine. This is signed Grad×Residual, the inner product of attribution patching (Kramár et al., 2024), here exact rather than approximate; its ranking is invariant to $\delta > 0$ by linearity, so δ only sets the scale. First, the derivative is identical for $\delta < 0$ and $\delta > 0$, so the first-order prediction of ablation is $\hat{S}_l^{(1)} = -\nabla\mathcal{L}^\top f_l$ and any dissociation between S_l and A_l must come from the finite-step terms of Eq. 2. Second, computing all L scores needs one forward and one backward pass, that is $O(LTd)$ inner products for sequence length T and width d. Unsigned magnitudes $|\nabla\mathcal{L}^\top f_l|$ discard whether strengthening helps or hurts, and Appendix Table 14 shows that the signed score is what separates constructive from harmful layers. In the ranking experiments, screening takes 7–22 seconds per model and task against 51–191 seconds for the single-strength sweep; Appendix G gives the timings.

3.3 When can single-layer scores justify a joint intervention?

Write $F(d) = \mathcal{L}(0) - \mathcal{L}(d)$ for the loss gain from a vector d of residual increments, with $F(0) = 0$ and $g = \nabla F(0)$. GILS estimates $d_l g_l$ one coordinate at a time. The gap between that quantity and a joint profile has two distinct parts:

$$F(d) - g^\top d = \sum_l \big[F(d_l e_l) - g_l d_l\big] + \underbrace{F(d) - \sum_l F(d_l e_l)}_{I(d)\text{: cross-layer interaction}}. \tag{4}$$

The first part is single-layer nonlinearity. The second can be nonzero even when every single-layer score is exact. Thus agreement with all single-layer sweeps does not validate composition.

A sufficient condition for composition. If F is twice continuously differentiable on the box between 0 and d and $|\partial_{ij} F(u)| \le M_{ij}$ throughout that box, then

$$|I(d)| \le \sum_{i<j} M_{ij} |d_i d_j|. \tag{5}$$

The sum of measured single-layer gains thus certifies a positive joint loss gain only if it exceeds this interaction bound. The proof in Appendix H.3 is a coordinate telescoping argument. The condition requires mixed-derivative bounds across the intervention region; neither a Hessian at the unmodified model nor finitely many pairwise endpoint probes supplies such a bound. The condition specifies what is missing from a single-layer certificate; it is a conditional certificate, not a measured bound.

For example,

$$F(d_1, d_2) = d_1 + d_2 - 10 d_1 d_2$$

has exact positive single-layer gains at $d_1 = d_2 = 0.3$, yet their joint gain is -0.3, and $F = -d_1 - d_2 + 10 d_1 d_2$ reverses both signs; such smooth counterexamples show why discarding low-ranked coordinates can exclude useful combinations, not which interactions occur in a particular model. Pairwise checks do not close the gap: for

$$F(d) = d_1 + d_2 + d_3 - 40 d_1 d_2 d_3$$

every single-coordinate gain at 0.3 is 0.3, every two-coordinate gain is 0.6 and every pairwise interaction with the third coordinate at zero vanishes, yet the three-coordinate gain is -0.18, as Appendix H.3 verifies. A certificate therefore needs the uniform bound; the chain below measures $I(d)$ directly.

The measured chain. In practice the derivative is computed on calibration items and the profile is applied to other items, so we measure four quantities per profile, all on the evaluation items unless stated: the *calibration prediction* $\sum_{l\in P} D_l^{\mathrm{cal}}$, the *same-item prediction* $\sum_{l\in P} D_l$, the sum of measured single-layer gains $\sum_{l\in P} A_l$, and the measured joint gain G. Their successive differences are *transfer*, same-item minus calibration prediction, the *finite step*, $\sum_l A_l$ minus the same-item prediction and thus the first sum of Eq. 4, and the *interaction* $I = G - \sum_l A_l$; the three add up exactly to the gap between the calibration prediction and G. The finite step is a single-coordinate error: it contains all downstream curvature along f_l, the Hessian of Eq. 2, and does not locate a nonlinearity inside block l; likewise I is the net pairwise and higher-order cross-layer non-additive effect, so a small I does not bound each pairwise term. Along a fixed profile direction with strength $t = \alpha - 1$,

transfer is exactly linear in t while the finite step and interaction are $O(t^2)$, so the relevant magnitudes are the coefficients $c_T = |T|/t$, $c_S = |S|/t^2$ and $c_I = |I|/t^2$: the finite step exceeds transfer once $t > t^* = c_T/c_S$. Since c_T is exact at every strength, coefficients estimated at one strength predict another as far as c_S is stable; Section 4.4 tests this with coefficients frozen at $\alpha = 1.1$.

4 EXPERIMENTS

4.1 SETUP AND EVIDENCE SCOPE

Models and tasks. The ranking experiments use pretrained Qwen3-8B (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024) and Mistral-7B-v0.1 (Jiang et al., 2023), with Qwen2.5-7B/3B/1.5B, Llama-3-8B and Mistral-7B-v0.3 for cross-scale cells; these checkpoints predate 2026 and are marked †. The same-condition loss chain of Section 4.4 adds the 2026-generation Qwen3.5-9B and registered chains on Llama-3.1-8B and Mistral-7B-v0.3 for GSM8K and ARC-Challenge. Tasks are Math on GSM8K (Cobbe et al., 2021), Science on ARC-Challenge (Clark et al., 2018), Code on the 257 sanitized MBPP test problems (Austin et al., 2021), Knowledge on MMLU subjects (Hendrycks et al., 2021) and Reasoning on BBH Boolean Expressions (Suzgun et al., 2023). For Code the gold target is a test assertion of the problem, so this unit measures the likelihood of assertion text given the task description, not of a function body. Prompt pools hold 200/200/257/300/135 items and are split 50/50 into calibration and held-out with a fixed seed; the exact split is retained for each experiment.

Metric and scope tags. Every quantity is a teacher-forced loss on gold targets: we report $\Delta\mathcal{L} = \mathcal{L}_{\text{base}} - \mathcal{L}_{\text{amp}}$, where positive is better, per layer or per profile, in nats per target token. On Qwen3.5-9B a 1,024-token teacher-forcing window shortens 72 of $4 \times 1{,}319$ test targets, as Appendix E details. Every table carries a scope tag: **fixed case** marks one checkpoint, one split and one seed with no inferential claim; **unit-level** marks summaries over n model–task units, reported as ranges and counts; **item-level** marks per-item measurements with question-level intervals. The layer-bootstrap and layer-permutation intervals of Appendix A.1 are descriptive only, because the 32–36 layers of one network are not exchangeable units.

4.2 RANKING FIDELITY AND PROSPECTIVE SINGLE-LAYER SELECTION

Table 1 reports two things per model–task unit. The *matched* column is the Spearman correlation between the derivative and the measured single-strength brute-force sweep of all layers at $\alpha = 1.3$ on the same calibration prompts: it is a local-linearity check, expected to be high for an exact derivative, and it is 0.83–0.99 in all 15 units, with layer-permutation $p < 10^{-5}$ in every unit in Appendix Table 7. The *prospective* columns are the stricter test: the derivative is computed on the calibration split and evaluated against brute-force amplifiability on the held-out split of the 10 Qwen3-8B and Llama-3.1-8B units. Held-out Spearman is 0.45–0.98, and NDCG@3 is 0.76–0.99 on Qwen3-8B and 0.28–0.99 on Llama-3.1-8B. The same held-out arrays evaluate two label-free priors under the same top-3 rule. The block-output-norm rule $\sigma_l = \|f_l\|$ selects the same triple on all five tasks of a model, $(6, 35, 34)$ on Qwen3-8B and $(31, 1, 30)$ on Llama-3.1-8B. In Table 2, both priors select a harmful layer on all ten arrays and have a lower summed held-out gain than the derivative on all ten, on one split seed. The relative-update rule $\|f_l\|/\|\mathbf{x}_l\|$ of Appendix A.6 selects $(0, 6, 1)$ on Qwen3-8B and is again below the derivative on 5 of 5 tasks. On Mistral-7B-v0.1 the matched ranking replicates with the same protocol at Spearman 0.94–0.99, as Appendix A.2 shows. Table 13 shows where the signal lies: the gradient–residual cosine φ_l carries most of it, the block norm σ_l alone ranks amplifiability at $|\rho| \leq 0.31$, and the full product reaches 0.94–0.98; in Appendix Table 14, signed and unsigned versions of the same product differ by 0.73–1.13 in Spearman correlation.

4.3 MEASURED ABLATION IS NOT AN AMPLIFICATION PROXY

On Qwen3.5-9B the three quantities of Appendix Table 25 use the same 100 calibration items per seed: measured ablation S_l ranks the measured amplification A_l at ρ_s from -0.19 to 0.02 and its own first-order prediction at 0.17–0.36 with sign agreement in 10–15 of 32 layers, whereas the derivative ranks the same A_l at 0.58–0.77; by Eq. 2 the two predictions differ only in the step. Table 15 gives the Qwen3-8B view, a cross-run comparison with a separately constructed calibration split: S_l is not a norm ranking, with $\rho_s(S, \sigma) = 0.01$–0.34; nor a first-order ranking, since the derivative at

Table 1: **Scope: unit-level.** 15 matched units, 10 prospective units, seed 42. Matched: Spearman ρ between the derivative and the single-strength sweep on the same calibration prompts. Prospective: derivative on calibration, brute-force on held-out. NDCG@3 uses held-out gains clipped at zero; P@3 is the derivative's top-3 overlap with the held-out top-3; "Harmful G/N/D" counts selected layers with negative held-out amplifiability for the derivative, block-output-norm σ_l and last-three top-3 rules; regret is $1 - \Sigma_{\text{derivative}}/\Sigma_{\text{best}}$ of summed held-out gains. †2023–2025-generation checkpoints.

Model	Task	Matched ρ	Held-out ρ	NDCG@3	P@3	Harmful G/N/D	Regret
Qwen3-8B†, $L=36$	Math	0.87	0.45	0.96	3/3	0/1/1	0.00
	Science	0.95	0.90	0.86	2/3	0/1/1	0.15
	Code	0.97	0.92	0.76	1/3	0/1/1	0.20
	Knowledge	0.96	0.94	0.99	2/3	0/1/2	0.02
	Reasoning	0.98	0.92	0.98	3/3	0/1/1	0.00
Llama-3.1-8B†, $L=32$	Math	0.83	0.56	0.64	2/3	1/2/1	0.72
	Science	0.88	0.87	0.88	2/3	0/1/1	0.14
	Code	0.93	0.84	0.28	1/3	2/3/3	1.06
	Knowledge	0.97	0.93	0.95	2/3	0/3/3	0.06
	Reasoning	0.97	0.98	0.99	2/3	0/2/1	0.02
Mistral-7B-v0.1†, $L=32$	five tasks, App. A.2	0.94–0.99	—	—	—	—	—

Table 2: **Scope: unit-level.** Seed 42. Sum of held-out single-layer amplifiabilities of the three layers selected on the calibration split by each rule. "Best": the oracle held-out top-3. "Random": expectation over all triples, $3\bar{A}^{\text{held}}$. Measured-ablation ranking is available for Qwen3-8B only and comes from the separate run of Appendix Table 15.

Model	Task	Derivative	Best	Block-output norm σ_l	Last three	Random	Measured ablation
Qwen3-8B†	Math	+0.298	+0.298	+0.087	+0.081	+0.005	+0.228
Qwen3-8B†	Science	+0.348	+0.407	+0.027	+0.114	−0.023	+0.114
Qwen3-8B†	Code	+0.062	+0.078	+0.002	−0.004	−0.029	−0.055
Qwen3-8B†	Knowledge	+0.759	+0.772	−0.163	−0.486	−0.025	+0.759
Qwen3-8B†	Reasoning	+0.599	+0.599	−1.313	−1.072	−0.053	−0.019
Llama-3.1-8B†	Math	+0.022	+0.076	−0.048	−0.023	−0.027	—
Llama-3.1-8B†	Science	+0.495	+0.573	+0.357	+0.217	+0.030	—
Llama-3.1-8B†	Code	−0.001	+0.016	−0.097	−0.104	−0.042	—
Llama-3.1-8B†	Knowledge	+0.241	+0.255	−0.196	−0.169	−0.060	—
Llama-3.1-8B†	Reasoning	+0.381	+0.388	−0.709	−0.491	−0.091	—

$\delta = -1$ predicts it at $\rho_s = 0.18$–0.66 with sign agreement in 20–27 of 36 layers, against 0.87–0.98 for $\delta = +0.3$; and only a moderate guide to amplifiability, with $\rho_s(S, A) = 0.19$–0.64 here and 0.45–0.70 in the second measurement of Appendix Figure 1. The ablation top-3 contains a harmful amplification target on Science, Code and Reasoning. Ablation sensitivity is reproducible: two measurements in Appendix A.3 agree at Spearman 0.89–0.98. Across six checkpoints of three families the mean signed ablation–amplifiability correlation per model is 0.05–0.37 in Appendix Table 17.

4.4 Where the predicted joint gain is lost

Qwen3-8B, fixed profiles, fp32. We measured the chain of Section 3.3 for twenty fixed three-layer profiles of Qwen3-8B, ten deployed triples and ten fixed-seed random triples, together with all 36 single-layer interventions, on the same 100 held-out GSM8K items with the gold-token-weighted loss in fp32, as Appendices C and D detail. Table 3 gives the means. For the deployed triples at $\alpha = 1.3$ the prediction loses 12% of its value between calibration and evaluation items, a further 31% of the same-item prediction in the finite step, and 8% of $\sum_l A_l$ to interaction; all ten keep a positive joint gain, against four of ten random triples in Appendix Table 18, and interaction is small relative to it: mean $|I| = 0.0028$ against mean $|G| = 0.0232$, and the sign of G agrees with that of $\sum_l A_l$ in 18 of 20 profiles.

Qwen3.5-9B, 1,319 test items. The second chain was measured on Qwen3.5-9B under one fixed condition declared before launch, described in Appendix E with the registered endpoints in Appendix

Table 40: the plain three-shot GSM8K prompt of each of four exemplar seeds, fp16, 100 calibration and 100 selection items from the train split and all 1,319 test items per seed, and the 21 distinct three-layer profiles frozen from seven six-layer pools by selection-item first answers at $\alpha = 1.3$, so a "derivative triple" is not the derivative's top three. The whole-layer ranking carries over: on the selection items the derivative orders the 32 measured single-layer gains at Spearman 0.54–0.69 per seed, and at 0.43–0.72 from calibration items, as Appendix Table 19 lists. Table 4 shows that the level of the prediction is set by the finite step. On the test items, under the item-averaged loss that scored the pools, the derivative triples' prediction of 0.0072 nats/token becomes 0.0007 after the finite step and 0.0006 jointly: the finite step removes 91–103% of it under item-averaged or token-weighted loss in Appendix Table 20, with mean absolute transfer and interaction of 0.0021 and 0.0011, about a third and a sixth of the finite step's 0.0066. Over all 21 frozen triples the finite step is the largest term for 17, transfer for 3 and interaction for 1. On the 1,250 test questions whose targets fit the window in every seed the finite step removes 89–99% and the counts stay 17/3/1, as Appendix E reports. Interaction is minor here as well: over all 423 candidate triples of the seven pools in the four seeds, summed single-layer gains order the joint gains at Spearman 0.96 overall and 0.83–0.96 per seed, mean $|I|$ is 23% of mean $|G|$, and the signs disagree for 2.6% of triples. Prospectively, Appendix Table 21 shows that summed single-layer gains measured on the calibration items order the joint gains on the disjoint selection items at 0.90, and the top-ranked triple gains $+0.0021$ $[0.0011, 0.0031]$ jointly against a candidate mean of -0.014.

The hierarchy by strength. In Table 6, at $\alpha = 1.3$ the finite step has the largest group-average absolute value in every profile set, including the fp32 Qwen3-8B sets; on Qwen3.5-9B, the fp32 re-measurement of Appendix Table 24 moves single-layer gains and derivatives by at most 3.8% of the fp32 finite step. At $\alpha = 1.1$ the finite step shrinks on both checkpoints; it remains the largest term for 105 of 160 Qwen3.5-9B triples, while on Qwen3-8B transfer becomes the largest. Interaction is the smallest term for the Qwen3-8B sets and the derivative triples, and at most 30% of the finite step in every set at $\alpha = 1.3$. The coefficients of Section 3.3 account for the switch: for the deployed Qwen3-8B triples $c_T = 0.021$ and $c_S = 0.13$ at both strengths, so the finite step grows quadratically and $t^* \approx 0.16$ lies between the two strengths; for the Qwen3.5-9B derivative triples $t^* \approx 0.10$.

Two further families, registered. A cross-family chain registered before launch and described in Appendix F measures every term on Llama-3.1-8B and Mistral-7B-v0.3 for 200 GSM8K and 200 ARC-Challenge test questions, with four seeds each, all 32 single blocks, $\alpha \in \{1.05, 1.1, 1.3\}$ and three- and six-layer profiles fixed from calibration data only. On GSM8K the picture of the Qwen checkpoints recurs in both families in Table 5: at $\alpha = 1.3$ the finite step is the largest term over the three-layer profiles, with a margin over the next term of $[0.0136, 0.0147]$ for Llama and $[0.0064, 0.0070]$ for Mistral, interaction is 20% of it, and summed single-layer gains order the joint gains at 0.96 in both over all profiles. Its share falls as α goes to 1.05, where the derivative top-three keeps a positive joint test gain in both families. ARC-Challenge, with gold labels of one or two tokens, shows the other side of the same rule: calibration and test derivatives agree less, at Spearman 0.31–0.41, c_T is 14 to 17 times the GSM8K value of the same checkpoint, $t^* = 0.31$–0.34, and transfer is at least as large as the finite step up to $\alpha = 1.3$ in both families by point estimate, while the finite-step share rises with strength and composition stays tight. Six-layer profiles raise c_I two- to threefold and leave the GSM8K finite step the largest term; every registered outcome is listed in Appendix Table 41.

Predicting the leading term from one strength. Since c_T is exact, coefficients estimated at one strength predict another as far as c_S is stable. In a registered check that holds $\alpha = 1.3$ out of the estimate, c_T, c_S and c_I frozen at $\alpha = 1.1$ name the measured leading term at $\alpha = 1.3$ in all five checkpoint–task groups and predict the mean finite step within -6% to $+11\%$ of its measured value, as Appendix Table 28 shows. For ARC-Challenge the frozen coefficients place the switch at $\alpha \approx 1.32$ for Llama-3.1-8B and 1.33 for Mistral-7B-v0.3; new measurements at $\alpha \in \{1.2, 1.4, 1.5\}$, registered with these predictions, find the predicted leading term at every strength in both families and the switch at $\alpha = 1.35$ and 1.31 in Appendix Table 29. Over 19,999 question resamples the switch occurs by $\alpha = 1.5$ in 90% and 95% of them, single seeds switch within 0.08 of their own prediction, and six-layer profiles switch earlier, at 1.26 and 1.22 against a predicted 1.24, as Appendix Table 30 shows. On a 32-question subset, fp32 re-measurement preserves the leading term at $\alpha = 1.3$ and 1.4

in both families and every seed, as Appendix Tables 34 and 35 show; the switch locations above are fp16 measurements on the full question sets.

From the budget to a profile choice. The budget changes which rule should choose a profile when the target is the joint loss. On the same Qwen3.5-9B candidates, calibration items and strength $\alpha = 1.3$, ranking by summed *measured* calibration single-layer gains picks a top triple with joint selection-item gain $+0.0021$ $[0.0011, 0.0031]$ and regret 0.0014 against the bank maximum, whereas summed calibration *derivatives*, whose level the finite step removes, pick triples with the last block and regret 0.041, as Appendix Table 22 shows. Re-running the whole choice on resampled calibration and selection items gives $+0.0024$ $[0.0010, 0.0038]$. At $\alpha = 1.3$ the measured sums choose the best profile of each three- and six-layer GSM8K set of both families, and on ARC-Challenge a profile with a larger joint gain than the derivative sums in every set of Appendix Table 31. The correction need not measure every block: on the Qwen3.5-9B banks, measuring only the eight bank blocks with the largest derivative while keeping the derivative elsewhere chooses the same triple as the full measured sums in all four seeds at an estimated 45% of the recorded single-block stage time in Appendix Table 23; on the eight cross-family sets of Appendix Table 32, screen-8 has a lower regret than the derivative sums in all eight. A per-block second-order correction whose curvature comes from single-block probes at $\alpha = 1.1$ reaches the joint gain of the full measured sums at $\alpha = 1.3$ in seven of these eight sets, as Appendix Table 33 shows.

New evaluation questions. A registered validation keeps the original coefficients frozen and measures all eight ARC-Challenge units on 200 disjoint questions. The three-layer switch is 1.219 for Llama and 1.270 for Mistral, within the registered 0.10 tolerance of the frozen predictions 1.3165 and 1.3288, and the predicted mean finite step is within the registered 30% tolerance at all four new-question strengths in both families. The predicted leading term is the measured one at $\alpha = 1.2$, 1.4 and 1.5. The coefficients are not refit on the fresh questions, so both tolerances check the original $\alpha = 1.1$ estimates out of sample, on questions outside the original 200 on which they were estimated. Appendix F.1 reports the uncertainty estimates and per-seed switches, and Appendix Table 41 every registered decision.

5 Discussion

A derivative's ranking fidelity and predicted gain require separate assessment. The ranking can survive a finite intervention even when its predicted gain does not; measured single-layer gains correct the level for profile selection. The decomposition identifies which correction is needed. On the original questions, coefficients estimated at $\alpha = 1.1$ name the leading term at $\alpha = 1.3$ in all five groups and predict the measured switch within 0.04 for pooled profiles and 0.08 per seed. On disjoint ARC-Challenge questions the switch stays within the registered 0.10 tolerance and the finite-step magnitude within 30%, so the frozen coefficients locate the switch to within about 0.1 in strength, the resolution of the leading-term prediction near the switch.

Scope. Claims concern teacher-forced loss on 1.5B–9B Qwen, Llama and Mistral checkpoints; the registered first-answer endpoints of Appendix I are reported for completeness, and the paper makes no behavioural claim. Scalar scaling of native block outputs excludes trained modules such as adapters. The certificate in Eq. 5 assumes a uniform mixed-derivative bound, which pairwise tests alone do not certify. Profile selection is a loss-level diagnostic over fixed candidates, with the unmodified model as reference when no candidate has positive joint gain.

6 Conclusion

For native residual scaling the signed gradient–residual product is the exact derivative at identity and ranks small single-layer amplification in three model families, unlike measured ablation. The gap between summed single-layer predictions and a measured joint gain splits exactly into transfer, finite-step and interaction terms: on GSM8K the finite step is the largest at $\alpha = 1.3$ on four checkpoints, on the original evaluation questions, coefficients frozen at one strength name the leading term at another, and measured gains of a few derivative-ranked blocks choose lower-regret profiles than summed derivatives.

Table 3: **Scope: fixed case.** Qwen3-8B†, 100 held-out GSM8K items, fp32. Same-item loss chain in nats/token, mean over ten profiles per group. "Step" is the finite-step single-layer error $\sum_l A_l -$ same-item prediction, and $I = G - \sum_l A_l$ the interaction.

α	Profiles	n	Calib. pred.	Same-item pred.	$\sum_l A_l$	G	Step	I
1.1	Deployed	10	0.0140	0.0123	0.0110	0.0105	−0.0013	−0.0005
1.1	Random	10	0.0054	0.0031	0.0027	0.0027	−0.0004	+0.0000
1.3	Deployed	10	0.0419	0.0368	0.0253	0.0232	−0.0115	−0.0021
1.3	Random	10	0.0163	0.0094	0.0029	0.0035	−0.0066	+0.0006

Table 4: **Scope: item-level.** Qwen3.5-9B, 1,319 GSM8K test items per seed, $\alpha = 1.3$, nats/token, item-averaged loss. Test-item chain of the frozen derivative triples: predicted gain from calibration items and from the test items, sum of measured single-layer gains, joint gain G and the three error terms. Bottom rows: means over the four derivative triples, whose mean G has the question-level 95% interval $[0.0002, 0.0010]$, and over all 21 frozen triples.

Seed	Triple	Cal. pred.	Test pred.	$\sum_l A_l$	G	Transfer	Step	I
7	$(0, 1, 5)$	+0.0097	+0.0105	−0.0034	−0.0013	+0.0008	−0.0140	+0.0021
42	$(1, 2, 4)$	+0.0090	+0.0066	+0.0035	+0.0018	−0.0025	−0.0030	−0.0017
100	$(3, 6, 15)$	+0.0102	+0.0061	+0.0038	+0.0031	−0.0041	−0.0023	−0.0007
123	$(0, 4, 15)$	+0.0067	+0.0057	−0.0012	−0.0012	−0.0010	−0.0069	+0.0001
Mean	derivative	+0.0089	+0.0072	+0.0007	+0.0006	−0.0017	−0.0066	−0.0001
Mean	all 21 frozen	+0.0035	+0.0030	−0.0104	−0.0076	−0.0005	−0.0134	+0.0027

Table 5: **Scope: item-level.** Registered chains on Llama-3.1-8B† and Mistral-7B-v0.3†, 200 test questions, three-layer profiles of all seeds, item-averaged loss, nats/token. ρ_{same}: seed-mean Spearman over 32 blocks between same-item derivative and measured single gain; mean absolute transfer $|T|$, finite step $|S|$ and interaction $|I|$; step share $|S|/(|T| + |S| + |I|)$; $c_S = |S|/t^2$; $\rho(G, \sum A)$: seed-mean Spearman over the same three-layer profiles. L: Llama, M: Mistral. $c_T = |T|/t$: L, GSM8K 0.0058; L, ARC-C 0.0806; M, GSM8K 0.0032; M, ARC-C 0.0555. $\alpha = 1.05$, where fp16 rounding is of the size of the finite step, is in Appendix F.

Units	α	n	ρ_{same}	$\lvert T\rvert$	$\lvert S\rvert$	$\lvert I\rvert$	Step share	c_S	$\rho(G, \sum A)$
L, GSM8K	1.1	55	0.95	0.0006	0.0021	0.0004	0.68	0.21	0.97
L, GSM8K	1.3	55	0.79	0.0017	0.0178	0.0036	0.77	0.20	0.95
L, ARC-C	1.1	56	0.92	0.0081	0.0025	0.0008	0.22	0.25	0.99
L, ARC-C	1.3	56	0.67	0.0242	0.0214	0.0046	0.43	0.24	0.92
M, GSM8K	1.1	56	0.96	0.0003	0.0009	0.0002	0.65	0.09	0.98
M, GSM8K	1.3	56	0.86	0.0010	0.0084	0.0017	0.76	0.09	0.97
M, ARC-C	1.1	56	0.96	0.0056	0.0017	0.0004	0.22	0.17	0.98
M, ARC-C	1.3	56	0.82	0.0167	0.0162	0.0028	0.46	0.18	0.96

Table 6: **Scope: item-level summaries.** Error terms in nats/token by checkpoint, items, profile set and strength. Each cell: signed mean / mean absolute value; "L" counts profiles in which the term has the largest absolute value. Qwen3-8B: Table 3; Qwen3.5-9B: Appendix E, where "all candidates" is every candidate triple of the seven pools in four seeds and "deriv. + random" the 20 derivative-pool and 20 random-pool triples per seed, with derivatives rescaled to $\delta = 0.1$.

Model	Items	Profiles	α	n	Transfer	L	Step	L	Interaction	L
Qwen3-8B	100 held-out	deployed	1.1	10	−0.0017 / 0.0021	7	−0.0013 / 0.0013	2	−0.0005 / 0.0006	1
Qwen3-8B	100 held-out	random	1.1	10	−0.0023 / 0.0023	9	−0.0004 / 0.0010	0	0.0000 / 0.0004	1
Qwen3-8B	100 held-out	deployed	1.3	10	−0.0051 / 0.0063	0	−0.0115 / 0.0115	10	−0.0021 / 0.0028	0
Qwen3-8B	100 held-out	random	1.3	10	−0.0069 / 0.0069	4	−0.0066 / 0.0072	6	+0.0006 / 0.0020	0
Qwen3.5-9B	selection	deriv. + random	1.1	160	−0.0005 / 0.0008	53	−0.0022 / 0.0022	105	+0.0001 / 0.0002	2
Qwen3.5-9B	selection	derivative	1.3	4	−0.0016 / 0.0016	1	−0.0073 / 0.0073	3	+0.0002 / 0.0014	0
Qwen3.5-9B	selection	all frozen	1.3	21	−0.0001 / 0.0017	4	−0.0137 / 0.0140	17	+0.0028 / 0.0035	0
Qwen3.5-9B	selection	all candidates	1.3	423	−0.0012 / 0.0027	44	−0.0212 / 0.0214	366	+0.0026 / 0.0035	13
Qwen3.5-9B	test	derivative	1.3	4	−0.0017 / 0.0021	1	−0.0066 / 0.0066	3	−0.0001 / 0.0011	0
Qwen3.5-9B	test	all frozen	1.3	21	−0.0005 / 0.0013	3	−0.0134 / 0.0135	17	+0.0027 / 0.0034	1

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to generate synthetic data sets, and support for qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and no personal data. All experiments use publicly released model checkpoints and the public benchmarks GSM8K, ARC-Challenge, MMLU, BBH and MBPP, and we release no new dataset; the supplementary records contain model losses and derivatives on these public items only. The intervention studied, a scalar gain on the residual updates of a frozen model, requires access to model weights and internals. Like other activation interventions it could be used to strengthen undesired behaviour; we evaluated no safety-relevant behaviour. Our claims concern teacher-forced loss and do not concern task accuracy, fairness or safety. We declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The intervention and the exact derivative are defined in Section 3, and the decomposition and the composition condition in Section 3.3, with complete proofs in Appendix H. Models, tasks, splits, prompts, precision, loss normalization and profiles are specified in Section 4.1 and Appendices G, C, D, E and F. Registered endpoints are listed in Appendix I, and Table 39 lists the records behind each experiment. The anonymous supplementary material contains the implementation of the intervention and the derivative, the measurement and analysis code, the registrations, and the per-layer, per-profile and per-item records from which the reported summaries are recomputed; the ranking sweeps are provided as layer-level measurements.

REFERENCES

Ryyan Akhtar, Payal Pahwa, and Monika Arora. Selective neuron amplification in transformer language models. *arXiv preprint arXiv:2604.07098*, 2026. URL https://arxiv.org/abs/2604.07098.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026. URL https://arxiv.org/abs/2604.15557.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19856–19882, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.974. URL https://aclanthology.org/2025.acl-long.974/.

Soham Gadgil, Chris Lin, and Su-In Lee. Where to steer: Input-dependent layer selection for steering improves LLM alignment. *arXiv preprint arXiv:2604.03867*, 2026. URL https://arxiv.org/abs/2604.03867.

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, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ngmEcEer8a.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 17643–17668, 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

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

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021. URL https://jmlr.org/papers/v22/20-1223.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024. URL https://arxiv.org/abs/2403.00745.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11264–11272, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Molchanov_Importance_Estimation_for_Neural_Network_Pruning_CVPR_2019_paper.html.

Muhammad Faishal Adly Nelwan and Alfan Farizki Wicaksono. Deployable per-instance multi-layer activation steering for large language models. *arXiv preprint arXiv:2608.08829*, 2026. URL https://arxiv.org/abs/2608.08829.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634, 2025. URL https://aclanthology.org/2025.acl-long.1007/.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. GrAInS: Gradient-based attribution for inference-time steering of LLMs and VLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46523–46543, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2159. URL https://aclanthology.org/2026.acl-long.2159/.

Nikhil Prakash, Donghao Ren, Dominik Moritz, and Yannick Assogba. Constructive circuit amplification: Improving math reasoning in LLMs via targeted sub-network updates. *arXiv preprint arXiv:2512.16914*, 2025. URL https://arxiv.org/abs/2512.16914.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. In *NeurIPS 2024 Workshop on Foundation Model Interventions*, 2024. URL https://openreview.net/forum?id=7xJcX2gbm9.

Pau Rodríguez, Michal Klein, Eleonora Gualdoni, Valentino Maiorca, Arno Blaas, Luca Zappella, Marco Cuturi, and Xavier Suau. LinEAS: End-to-end learning of activation steering with a distributional loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=EBONa3tT3K.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of LLMs enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*, 2026. URL https://arxiv.org/abs/2604.19018.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, 2024. doi: 10.18653/v1/2024.findings-emnlp.479. URL https://aclanthology.org/2024.findings-emnlp.479/.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. LayerNavigator: Finding promising intervention layers for efficient activation steering in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 101058–101080. Curran Associates, Inc., 2025. doi: 10.52202/085713-3383. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9280bfb61f159c768dfdeb729a41ccd0-Abstract-Conference.html.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2402.17762.

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. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 139179–139212, 2024. doi: 10.52202/079017-4417. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.17146.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Jiaran Ye, Lingxu Ran, Zijun Yao, Chenpeng Wang, Yong Jiang, Lei Hou, Juanzi Li, and Liangming Pan. Where steering signals come from: Activation source selection in activation steering. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. URL https://arxiv.org/abs/2607.25270. arXiv:2607.25270.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xFinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=7UqQJUKaLM.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Luyang Zhang and Jialu Wang. When attribution patching lies: Diagnosis and a second-order correction. *arXiv preprint arXiv:2606.09899*, 2026. URL https://arxiv.org/abs/2606.09899.

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

A Additional Layer-Score Analyses

A.1 Exact Significance of the Matched-Prompt Rank Correlations

Table 1 reports Spearman correlations without intervals, because the layers of one network are not exchangeable units. Table 7 adds, for every model–task pair, the asymptotic two-sided p-value and a layer-permutation test with 10^5 permutations of the amplifiability vector. All 15 pairs, the ten primary pairs plus the five Mistral-7B replication pairs of Appendix A.2, are individually significant: the largest asymptotic p is 4.8×10^{-9}, for Llama-3.1-8B Math, and in no pair did any permutation reach the observed $|\rho|$, so every permutation $p < 10^{-5}$. Because each test uses $n = 32$–36 layers of one network, it concerns the within-configuration layer-ranking association only; it does not replace the held-out transfer check of Table 1, and the main text does not use these p-values as inference.

Table 7: Spearman ρ between GILS and brute-force amplifiability on matched calibration prompts, with asymptotic two-sided p-values. Layer-permutation $p < 10^{-5}$ for every entry, with zero exceedances in 10^5 permutations.

	Qwen3-8B, $n=36$		Llama-3.1-8B, $n=32$		Mistral-7B, $n=32$	
Task	ρ	p	ρ	p	ρ	p
Math	0.87	6.3×10^{-12}	0.83	4.8×10^{-9}	0.98	9.1×10^{-22}
Science	0.95	3.1×10^{-19}	0.88	4.1×10^{-11}	0.96	8.0×10^{-19}
Code	0.97	8.1×10^{-23}	0.93	5.1×10^{-15}	0.94	1.4×10^{-15}
Knowledge	0.96	2.4×10^{-20}	0.97	9.8×10^{-20}	0.99	2.8×10^{-25}
Reasoning	0.98	3.4×10^{-24}	0.97	1.5×10^{-20}	0.98	3.9×10^{-24}

A.2 Third-Architecture Replication: Mistral-7B

The main text establishes the norm–cosine dissociation and GILS ranking fidelity on Qwen3-8B and Llama-3.1-8B. As a replication on a third architecture family, the identical matched-prompt σ–φ protocol, with the same task files, identical calibration splits, the same $\alpha = 1.3$ and seed 42, was run on Mistral-7B-v0.1 with $L = 32$ layers. Table 8 reports the same predictor comparison as Table 13. The qualitative picture replicates: the full GILS product is the strongest predictor on every task, with Pearson 0.94–0.99 and Spearman 0.94–0.99 at asymptotic $p \leq 1.4 \times 10^{-15}$ throughout, the cosine factor dominates the block-output norm, and σ_l and $|\varphi_l|$ remain weakly correlated. The decorrelation is $|\rho(\sigma, |\varphi|)| \leq 0.38$, largest on Science at -0.371, and σ_l alone reaches $|\rho| = 0.43$ on Math. Table 11 displays all 15 values, including Llama. This replication covers the rank-correlation and dissociation claims.

Table 8: Mistral-7B-v0.1 replication of the predictor comparison: Pearson ρ with brute-force amplifiability A_l across all 32 layers on matched calibration prompts. Last column: matched-prompt Spearman correlation of the full GILS score.

Task	$\rho(\sigma, A)$	$\rho(\varphi, A)$	$\rho(\sigma \cdot \varphi, A)$	ρ(GILS, A)	$\rho(\sigma, \lvert\varphi\rvert)$	GILS Spearman ρ
Math	-0.43	-0.68	-0.87	$+0.95$	$+0.01$	0.98
Science	$+0.01$	-0.93	-0.56	$+0.98$	-0.37	0.96
Code	-0.22	-0.90	-0.67	$+0.94$	-0.30	0.94
Knowledge	-0.17	-0.94	-0.79	$+0.99$	-0.17	0.99
Reasoning	-0.22	-0.72	-0.84	$+0.99$	-0.25	0.98

A.3 Direct Test of the Magnitude Hypothesis and of the Ablation–Amplifiability Dissociation

A direct magnitude test asks whether measured ablation sensitivity follows residual norm; their relationship need not be assumed from either definition. Table 9 reports the direct measurement on Qwen3-8B: the zero-ablation loss increase S_l and $\sigma_l = \|f_l\|$ were measured together on one fixed calibration split per task, and compared with the amplifiability A_l and GILS scores of Table 13, which come from a separate run with a separately constructed calibration split; the two splits are

not item-identical, and Knowledge has 150 against 149 items, so these are cross-run comparisons; the matched comparison is the Qwen3.5-9B one of Section 4.3. Three facts follow. First, the magnitude hypothesis is false as stated: Spearman $\rho(S, \sigma)$ is 0.01–0.34 across the five tasks and 0.08–0.43 after removing layer 0, so ablation rankings are not norm rankings. Second, ablation is not a first-order quantity: the first-order prediction of S_l, the same signed inner product that defines GILS evaluated at $\delta = -1$, has Spearman only 0.18–0.66 with the measured S_l, and its sign agrees with the measured ablation sign in only 20–27 of 36 layers per task, whereas the same term predicts small positive amplification at Spearman 0.87–0.98 in Table 1. Third, the ablation–amplifiability dissociation is moderate rather than total: Spearman $\rho(S, A)$ is 0.19–0.64 here and 0.45–0.70 in the second, single-run measurement of Table 10; the three most ablation-sensitive layers contain a layer whose amplification is harmful, with $A_l < 0$, on three of the five tasks, namely layer 35 on Science, layers 0 and 35 on Code and layer 0 on Reasoning, and on Knowledge the ablation top-3 and amplifiability top-3 coincide. We therefore describe the finding as "ablation-sensitive layers are unreliable amplification targets," not as a categorical dissociation. Two independent measurements exist for both quantities. Ablation sensitivity from the cross-scale measurement, with 512-token truncation and one split construction, and from the direct measurement, with 1,024 tokens and a different split construction, agree at Spearman 0.89–0.98 per task, and their top-3 sets coincide on four of five tasks, all but Science.

Table 9: Direct measurement on the 36 layers of Qwen3-8B. ρ_s: Spearman; ρ_p: Pearson; $\hat{S}_l^{(1)} = -\nabla\mathcal{L}^\top f_l$ is the first-order prediction of ablation sensitivity. "Sign" counts layers where $\mathrm{sign}(S_l) = \mathrm{sign}(\hat{S}_l^{(1)})$. "Harmful" counts layers among the ablation top-3 with $A_l < 0$.

Task	n_{cal}	$\rho_s(S, \sigma)$	$\rho_s(S, \sigma)$ w/o L0	$\rho_s(S, A)$	$\rho_p(S, A)$	$\rho_s(S, \hat{S}^{(1)})$	Sign	Harmful
Math	100	$+0.34$	$+0.43$	$+0.47$	$+0.52$	$+0.54$	21/36	0/3
Science	100	$+0.17$	$+0.18$	$+0.37$	$+0.11$	$+0.34$	23/36	1/3
Code	128	$+0.28$	$+0.38$	$+0.42$	-0.03	$+0.50$	20/36	2/3
Knowledge	150	$+0.01$	$+0.08$	$+0.64$	$+0.41$	$+0.66$	27/36	0/3
Reasoning	67	$+0.28$	$+0.38$	$+0.19$	$+0.10$	$+0.18$	21/36	1/3

A.4 Per-Cell Cross-Scale Results

Figure 1 shows the Qwen3-8B cells and Table 17 summarises six models. Table 10 gives every model–task cell: the signed Pearson and Spearman correlation between the calibration ablation and amplification profiles, and how many of the 20 random 3-layer profiles the atlas profile beats on the held-out split. The atlas beats all 20 random profiles in 16 of 26 cells and more than half in 24; these counts compare the atlas with chance-level selection, and the label-free priors of Table 1 are the harder comparison. Spearman correlations between ablation and amplification profiles range from -0.14 to $+0.84$ and are highest on Reasoning for five of the six models and on Knowledge for Qwen3-8B, so the dissociation is task-dependent rather than uniform.

A.5 Prospective Single-Layer Selection: Summed Held-Out Gains

Table 1 summarises prospective selection by NDCG@3, P@3 and harmful counts. Table 2 in Section 4.2 gives the underlying quantity: the sum of held-out single-layer amplifiability values A_l^{held} over the three layers each rule selects on the calibration split. This is an *additive proxy* that sums three single-layer effects, not an evaluation of the composed three-layer profile; Section 4.4 measures how far composed profiles depart from such sums. Regret in Table 1 is $1 - \Sigma_{\mathrm{derivative}}/\Sigma_{\mathrm{best}}$, where "best" is the held-out top-3; it exceeds one on Llama Code, where the derivative's summed held-out gain is -0.001 in Table 2. On 10 of 10 task arrays the derivative rule has a larger sum than the block-output-norm rule σ_l, which selects the same triple on all five tasks of a model as Appendix A.6 shows, than the last-three rule and than the random expectation; the first two are fixed model-level priors re-scored on ten arrays under one split seed. Against measured ablation, available for Qwen3-8B only, it is larger on four tasks and identical on Knowledge, where the two rules select the same set. All of this uses one split seed, 42.

Table 10: Per-cell cross-scale results. ρ_p/ρ_s: Pearson/Spearman correlation between calibration ablation sensitivity and single-layer amplifiability across layers. "Beats": number of the 20 random 3-layer profiles whose held-out $\Delta\mathcal{L}$ the atlas profile exceeds. Cells with $n_{\text{held}} = 5$ are pilot-only.

Model	Task	n_{held}	$\rho_p(S, A)$	$\rho_s(S, A)$	Beats/20
Qwen3-8B	Math	100	$+0.59$	$+0.56$	20
Qwen3-8B	Science	100	$+0.46$	$+0.52$	20
Qwen3-8B	Code	129	-0.04	$+0.45$	20
Qwen3-8B	Knowledge	150	$+0.40$	$+0.70$	20
Qwen3-8B	Reasoning	68	$+0.17$	$+0.51$	20
Llama-3-8B	Math	100	$+0.22$	-0.08	20
Llama-3-8B	Science	100	$+0.30$	$+0.21$	20
Llama-3-8B	Code	129	$+0.05$	$+0.18$	20
Llama-3-8B	Knowledge	150	-0.32	$+0.01$	19
Llama-3-8B	Reasoning	68	$+0.27$	$+0.29$	20
Mistral-7B	Math	100	$+0.17$	$+0.21$	14
Mistral-7B	Science	100	-0.05	$+0.36$	20
Mistral-7B	Code	5	-0.51	-0.09	9
Mistral-7B	Reasoning	68	$+0.59$	$+0.67$	20
Qwen2.5-7B	Math	100	$+0.16$	-0.14	13
Qwen2.5-7B	Science	100	$+0.56$	$+0.19$	20
Qwen2.5-7B	Code	5	-0.11	$+0.23$	15
Qwen2.5-7B	Reasoning	68	$+0.81$	$+0.79$	20
Qwen2.5-3B	Math	100	$+0.22$	-0.01	19
Qwen2.5-3B	Science	100	$+0.17$	$+0.36$	19
Qwen2.5-3B	Code	5	$+0.35$	$+0.20$	1
Qwen2.5-3B	Reasoning	68	$+0.75$	$+0.80$	20
Qwen2.5-1.5B	Math	100	$+0.17$	-0.06	12
Qwen2.5-1.5B	Science	100	$+0.59$	$+0.29$	19
Qwen2.5-1.5B	Code	5	-0.33	$+0.04$	20
Qwen2.5-1.5B	Reasoning	68	$+0.65$	$+0.84$	20

Table 11: All matched-calibration Pearson correlations $\rho(\sigma, |\varphi|)$. Values are recomputed from the stored layer arrays; they are descriptive correlations, not independent layer trials.

Task	Qwen3-8B	Llama-3.1-8B	Mistral-7B-v0.1
Math	0.055	-0.092	0.014
Science	0.026	0.008	-0.371
Code	-0.226	-0.336	-0.299
Knowledge	-0.092	-0.148	-0.174
Reasoning	0.026	-0.101	-0.253

A.6 Three Magnitude Rules Called "Residual Norm"

Three different magnitude statistics were used as "norm" baselines, and they rank different layers. The block-output norm $\sigma_l = \|f_l\|$, the Frobenius norm of the block output over the whole prompt averaged over prompts, is the quantity of Table 13, the "N" rule of Table 1 and the norm column of Table 2; its top-3 is the same triple on all five tasks of a model: $(6, 35, 34)$ on Qwen3-8B, $(31, 1, 30)$ on Llama-3.1-8B and on Mistral-7B-v0.1. The relative update $r_l = \|f_l\|/\|\mathbf{x}_l\|$, the token-mean ratio of block output to residual-stream input averaged over prompts, is a second label-free prior; it selects $(0, 6, 1)$ on Math, Science and Knowledge and $(6, 0, 1)$ on Code and Reasoning, because the residual stream is small at the first blocks. The residual energy $e_l = \sum_t \|f_l(t)\|^2$ is one of the Qwen3.5-9B pool rules of Appendix E. Table 12 re-scores the r_l triple on the held-out arrays of Table 2: the derivative's sum is larger on 5 of 5 Qwen3-8B tasks under either norm definition; the r_l triple selects a harmful layer on 3 of 5 tasks, the σ_l triple on 5 of 5. r_l was measured on Qwen3-8B.

A.7 Main-Text Ranking Tables

Tables 13–15 and Figure 1 give the fixed-case factor, signedness and ablation measurements summarized in Sections 3.2, 4.2 and 4.3.

Table 12: Held-out single-layer sums on Qwen3-8B† with seed 42 for the triples selected by the two norm definitions, the last-three rule and the derivative; "harm" is the number of selected layers with $A_l^{\text{held}} < 0$, given in parentheses for the last-three rule.

Task	σ_l triple	ΣA	harm	r_l triple	ΣA	harm	Last three ΣA (harm)	Derivative ΣA
Math	$(6, 35, 34)$	$+0.087$	1	$(0, 6, 1)$	$+0.129$	1	$+0.081$ (1)	$+0.298$
Science	$(6, 35, 34)$	$+0.027$	1	$(0, 6, 1)$	$+0.124$	0	$+0.114$ (1)	$+0.348$
Code	$(6, 35, 34)$	$+0.002$	1	$(6, 0, 1)$	-0.021	2	-0.004 (1)	$+0.062$
Knowledge	$(6, 35, 34)$	-0.163	1	$(0, 6, 1)$	$+0.526$	0	-0.486 (2)	$+0.759$
Reasoning	$(6, 35, 34)$	-1.313	1	$(6, 0, 1)$	-0.116	1	-1.072 (1)	$+0.599$

Table 13: **Scope: fixed case.** Qwen3-8B, one fixed calibration split per task, 36 layers. Signed Pearson ρ of each factor of Eq. 3 with measured amplifiability A_l on matched calibration prompts; last column is the σ–$|\varphi|$ decorrelation. Comparisons in the text use magnitudes.

Task	$\rho(\sigma, A)$	$\rho(\varphi, A)$	$\rho(\sigma \cdot \varphi, A)$	$\rho(\text{GILS}, A)$	$\rho(\sigma, \lvert\varphi\rvert)$
Math	$+0.31$	-0.87	-0.66	$+0.94$	$+0.06$
Science	$+0.03$	-0.93	-0.76	$+0.97$	$+0.03$
Code	$+0.14$	-0.81	-0.67	$+0.97$	-0.23
Knowledge	$+0.07$	-0.86	-0.62	$+0.98$	-0.09
Reasoning	-0.31	-0.60	-0.93	$+0.97$	$+0.03$

B Additional Results

B.1 Per-Layer Amplification Profiles

Figure 2 shows the raw per-layer amplification effect at $\alpha = 1.3$ on Qwen3-8B from a separate fixed-prompt diagnostic. Profiles are sparse, task-specific and signed: Math benefits from layers 0/11/35, Knowledge from 25/6/12, Code from 10/31/33, and on Reasoning the adjacent late layers 34 and 35 flip sign. This diagnostic is used for visualization only; $k = 3$ is a fixed design choice.

B.2 $\sigma \cdot \varphi$ Verification

We verify the norm–cosine factorization across all five tasks on Qwen3-8B. For each task, we compute the block-output norm $\sigma_l = \|f_l\|$ and gradient–residual cosine φ_l at every layer, then compare σ_l alone vs. $\sigma_l \cdot \varphi_l$ as predictors of amplifiability. Table 13 reports these correlations. The decorrelation $|\rho(\sigma, |\varphi|)| \leq 0.23$ holds uniformly on Qwen3-8B, and remains at most 0.35 across both 8B-class architectures.

Predictor ablation. Table 16 disaggregates the factors in the first-order gain $D_l = -\delta \cdot \|\mathbf{g}_l\| \cdot \sigma_l \cdot \varphi_l$, which equals GILS_l. The gradient–residual cosine φ_l alone is a strong predictor at $|\rho| = 0.60$–0.93, confirming it carries most of the task-specific signal. The block-output norm σ_l alone is weak at $|\rho| \leq 0.31$, and the gradient norm $\|\mathbf{g}_l\|$ alone is inconsistent, also at $|\rho| \leq 0.31$. The full GILS product, which combines all three factors, achieves the highest correlations, $|\rho| = 0.94$–0.98, consistent with its use as the screening score. Table 14, discussed in Section 4.2, further isolates the importance of sign: replacing signed GILS with unsigned $|\nabla \mathcal{L}^\top f_l|$ collapses the Spearman rank correlation on every task, because large harmful alignments become indistinguishable from large helpful ones.

B.3 Cross-Scale Summary

Table 17 summarises a six-checkpoint loss-level measurement: per model, the number of tasks on which the exhaustive atlas top-3 exceeds more than half of 20 random three-layer profiles in held-out $\Delta\mathcal{L}$, the fraction of all random profiles it exceeds, and the mean signed correlation $\bar{\rho}_{\text{SA}}$ between ablation sensitivity and amplifiability. Per-cell values are in Table 10.

Table 14: **Scope: fixed case.** Matched calibration prompts, seed 42. Signedness ablation for gradient–residual attribution. Entries are Spearman ρ with brute-force amplifiability on matched calibration prompts. Signed Grad×Residual is GILS; the unsigned variant is a magnitude-only ablation, not the AtP* algorithm, and loses the help-vs.-hurt distinction.

Task	Qwen signed	Qwen unsigned	Llama signed	Llama unsigned
Math	0.87	0.10	0.83	0.03
Science	0.95	−0.01	0.88	0.15
Code	0.97	0.05	0.93	−0.20
Knowledge	0.96	−0.04	0.97	0.02
Reasoning	0.98	0.02	0.97	−0.15

Table 15: **Scope: fixed case, cross-run.** Qwen3-8B†, seed 42, 36 layers; S, σ and A, GILS come from two runs with separately constructed calibration splits. Spearman correlations of measured ablation sensitivity S with residual norm σ, single-layer amplifiability A and its own first-order prediction $\hat{S}^{(1)} = -\nabla\mathcal{L}^\top f_l$; "Sign" counts layers where $\operatorname{sign} S_l = \operatorname{sign} \hat{S}_l^{(1)}$; "Harmful" counts ablation-top-3 layers with $A_l < 0$; last column: the derivative's matched Spearman with A from Table 1.

Task	$\rho_s(S,\sigma)$	$\rho_s(S,A)$	$\rho_s(S,\hat{S}^{(1)})$	Sign	Harmful	$\rho_s(\text{GILS},A)$
Math	0.34	0.47	0.54	21/36	0/3	0.87
Science	0.17	0.37	0.34	23/36	1/3	0.95
Code	0.28	0.42	0.50	20/36	2/3	0.97
Knowledge	0.01	0.64	0.66	27/36	0/3	0.96
Reasoning	0.28	0.19	0.18	21/36	1/3	0.98

C QWEN3-8B INTERACTION MEASUREMENTS

The diagnostic evaluates twenty three-layer corners of Qwen3-8B at $\alpha = 1.3$, ten deployed profiles and ten fixed-seed random triples, together with all thirty-six single-layer interventions on the same 100 held-out GSM8K examples, which contain 11,622 gold-answer tokens and have identity NLL 0.8231 nats per token. The profiles were fixed before this diagnostic; the deployed profiles are not newly selected. Evaluation uses fp32 arithmetic and batch size two.

Definitions. Let $\mathcal{L}(\alpha)$ be the gold-token-weighted teacher-forced NLL with gains $\alpha \in \mathbb{R}^{36}$, and $\mathbf{1}$ the identity. For a triple $P = \{i, j, k\}$ with corner α^P, where $\alpha_l = 1.3$ for $l \in P$ and $\alpha_l = 1$ otherwise, the joint gain, the measured single gains and the interaction are

$$G = \mathcal{L}(\mathbf{1}) - \mathcal{L}(\alpha^P),$$
$$A_l = \mathcal{L}(\mathbf{1}) - \mathcal{L}(\alpha^{\{l\}}),$$
$$I = G - \textstyle\sum_{l\in P} A_l.$$

Findings. Table 18 lists all twenty profiles. All ten deployed profiles have a positive joint gain, against four of ten random triples. Interaction is small relative to the joint gain: mean $|I| = 0.0028$ against mean $|G| = 0.0232$ for the deployed profiles and 0.0020 against 0.0121 for the random ones, and the sign of G agrees with the sign of $\sum_l A_l$ in eighteen of twenty profiles. Once the single-layer gains are measured rather than predicted, their sum tracks the joint gain at this strength.

D SAME-ITEM LOSS CHAIN FOR FIXED QWEN3-8B PROFILES

The chain of Section 3.3 was measured on the same 100 held-out GSM8K items for the twenty fixed Qwen3-8B profiles of Appendix C at $\alpha \in \{1.1, 1.3\}$, with the teacher-forced gold-token NLL in nats per token and fp32 arithmetic. For each profile, the *calibration* prediction is the sum of the three derivatives computed on the calibration items, the *same-item* prediction uses derivatives on

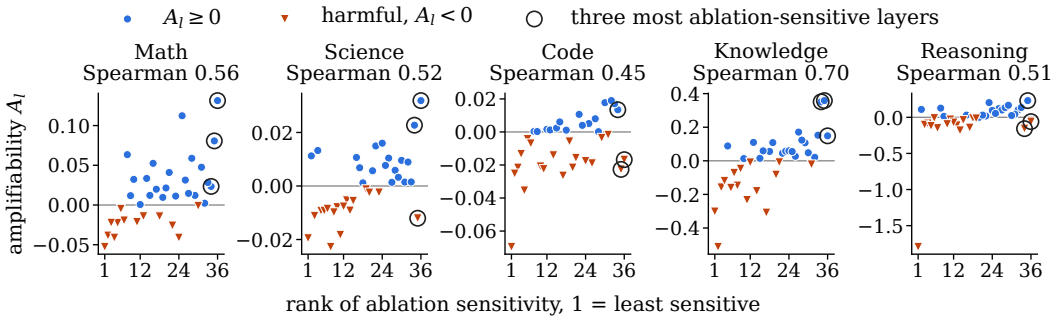


Figure 1: **Scope: fixed case.** Qwen3-8B† cross-scale measurement, calibration split. Rank of ablation sensitivity on the x-axis against single-layer amplifiability on the y-axis per layer; circles mark the three most ablation-sensitive layers. Spearman 0.45–0.70; on Science, Code and Reasoning the three most ablation-sensitive layers include a harmful amplification target. This measurement differs from Table 15 in input truncation and split construction.

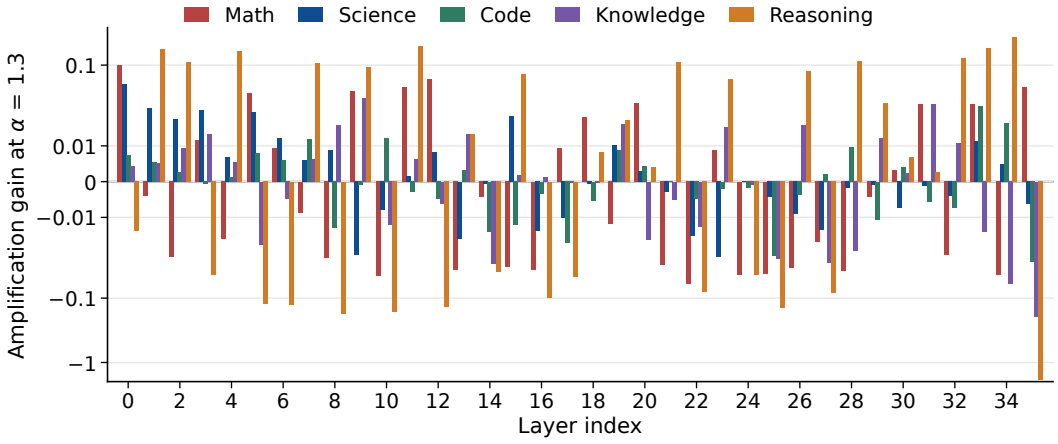


Figure 2: Per-layer amplification effect, the Δ log-probability from $\alpha = 1.3$, across five task families on Qwen3-8B† from a separate fixed-prompt diagnostic used for visualization only; the y-axis is symmetric-log with linear region $|\Delta| \leq 0.02$.

the evaluation items, $\sum_l A_l$ and G are measured, the finite-step error is $\sum_l A_l$ minus the same-item prediction, and $I = G - \sum_l A_l$; derivative and finite differences use the same token-weighted normalization. Table 3 in Section 4.4 gives the group means and Table 6 the signed and absolute terms with largest-term counts; the per-profile values are in the Qwen3-8B loss-chain table of the supplement's data folder.

For the deployed triples at $\alpha = 1.3$ the prediction loses 12% of its value between calibration and evaluation items, a further 31% of the same-item prediction in the finite step, and 8% of $\sum_l A_l$ to interaction; the finite step is the largest term for all ten. At $\alpha = 1.1$ the order changes: transfer between item sets is the largest term for 7 of 10 deployed and 9 of 10 random profiles, while the finite step falls from 0.0115 to 0.0013 for the deployed triples.

E SAME-CONDITION LOSS CHAIN ON QWEN3.5-9B

Condition and measurements. The chain of Eq. 4 was measured on Qwen3.5-9B under one fixed condition: fp16, the plain three-shot GSM8K prompt of each of the four exemplar seeds 7, 42, 100 and 123, and that seed's 100 calibration and 100 selection items, drawn disjointly from the train split,

Table 16: Predictor ablation on Qwen3-8B: Pearson ρ between each factor or combination and brute-force amplifiability A_l, measured across all 36 layers on matched calibration prompts.

Predictor	Math	Science	Code	Know.	Reason.
σ_l alone	$+0.31$	$+0.03$	$+0.14$	$+0.07$	-0.31
$\|\mathbf{g}_l\|$ alone	$+0.05$	-0.01	$+0.01$	$+0.31$	$+0.11$
φ_l alone	-0.87	-0.93	-0.81	-0.86	-0.60
$\sigma_l \cdot \varphi_l$	-0.66	-0.76	-0.67	-0.62	-0.93
Full GILS	$+0.94$	$+0.97$	$+0.97$	$+0.98$	$+0.97$

Table 17: Cross-architecture summary; all six checkpoints are of the 2023–2025 generation†. $\bar{\rho}_{\text{SA}}$: mean signed ablation–amplifiability correlation across tasks. Rows with four tasks omit Knowledge due to prompt-format differences; Code splits for Mistral-7B and Qwen2.5 models contain only $n = 5$ held-out examples.

Model	Layers	Tasks won	Atlas > random	$\bar{\rho}_{\text{SA}}$
Qwen3-8B	36	5/5	100%	0.32
Llama-3-8B	32	5/5	99%	0.11
Mistral-7B	32	3/4	79%	0.05
Qwen2.5-7B	28	4/4	85%	0.35
Qwen2.5-3B	36	3/4	74%	0.37
Qwen2.5-1.5B	28	4/4	89%	0.27

and all 1,319 test items. The profiles are the ones frozen by a shortlist protocol before this experiment: seven pool rules, namely derivative, unsigned attribution, residual energy, measured ablation, last six, seeded random and a contiguous window by derivative mass, each propose six layers per seed, and the triple among the 20 with the most correct first answers on the selection items is frozen, which gives 21 distinct seed–triple instances; depth and residual energy froze the same triple in every seed. The design and primary endpoints were registered before launch, as Appendix I lists, and nothing was re-tuned. Per item, the teacher-forced loss is the mean cross-entropy over the gold solution tokens, the loss that scored the pools; per-item token counts are kept so that the token-weighted loss is also available. Prompt and gold solution are teacher-forced in one window of at most 1,024 tokens, so the loss of a longer target covers its first tokens, as the paragraph *Teacher-forcing window* below quantifies. The derivative is

$$D_l = -(\alpha - 1)\, \nabla_{\mathbf{o}_l} \mathcal{L}^\top f_l$$

computed at batch 1 and rescaled linearly for $\alpha = 1.1$. Measured single-layer gains A_l cover all 32 blocks at $\alpha = 1.3$ on the calibration and selection items, all 32 blocks at $\alpha = 1.1$ on the selection items, and, on the test items, the layers of the frozen triples and of the derivative pool. Joint gains G cover the 21 frozen instances on the calibration, selection and test items, and every candidate triple of the seven pools on the selection items. As a reproduction control, regenerating each seed's frozen derivative triple on its 100 selection items reproduced all 100 stored outputs token for token in every seed. Every per-item unit and the analysis scripts are in the supplement.

Ranking and level. The same-item derivative orders the measured single-layer gains at 0.54–0.69 per seed and matches their sign in 19–23 of 32 layers; derivatives from the calibration items give 0.43–0.72 and 20–25 signs, so transfer between these two item sets costs little. At $\alpha = 1.1$ the same-item Spearman is 0.63, 0.81, 0.78 and 0.99 for seeds 7, 42, 100 and 123. The size of the finite step is visible in the levels: the summed absolute gap between measured and predicted single-layer gains is 1.3–3.0 times the summed absolute prediction at $\alpha = 1.3$ and 0.47–1.26 times at $\alpha = 1.1$. For the four derivative triples the same-item prediction of 0.0073 becomes 0.0000 after the finite step and $+0.0001$ jointly on the selection items in Table 19; the test items of Table 4 give the same picture, and the question-level 95% interval of the derivative triples' mean joint test gain is $[0.0002, 0.0010]$ from 19,999 resamples of test questions with the four frozen triples held fixed, so the interval is conditional on the frozen profiles. Table 6 gives every term as a signed mean and a mean absolute value.

Table 18: **Scope: fixed case.** Qwen3-8B†, 100 held-out GSM8K items, fp32, $\alpha = 1.3$, nats per gold token. Summed measured single-layer gains, joint gain G and interaction I for all twenty fixed profiles.

Group	Layers	$\sum_l A_l$	G	I	Group	Layers	$\sum_l A_l$	G	I
Deployed	(0,13,33)	$+0.0222$	$+0.0176$	-0.0046	Random	(13,16,35)	-0.0004	$+0.0031$	$+0.0035$
Deployed	(33,34,35)	$+0.0321$	$+0.0325$	$+0.0003$	Random	(0,13,23)	$+0.0010$	-0.0023	-0.0033
Deployed	(32,34,35)	$+0.0270$	$+0.0218$	-0.0052	Random	(13,31,34)	$+0.0332$	$+0.0313$	-0.0018
Deployed	(30,34,35)	$+0.0202$	$+0.0197$	-0.0006	Random	(6,8,35)	-0.0160	-0.0113	$+0.0047$
Deployed	(32,33,34)	$+0.0443$	$+0.0385$	-0.0058	Random	(2,7,34)	$+0.0274$	$+0.0276$	$+0.0002$
Deployed	(30,31,32)	$+0.0128$	$+0.0118$	-0.0011	Random	(31,33,35)	$+0.0126$	$+0.0162$	$+0.0036$
Deployed	(30,31,33)	$+0.0180$	$+0.0142$	-0.0037	Random	(12,21,31)	-0.0028	-0.0024	$+0.0004$
Deployed	(30,33,34)	$+0.0375$	$+0.0346$	-0.0029	Random	(1,12,23)	-0.0155	-0.0155	-0.0001
Deployed	(32,33,35)	$+0.0154$	$+0.0184$	$+0.0030$	Random	(5,25,29)	-0.0044	-0.0062	-0.0018
Deployed	(13,34,35)	$+0.0238$	$+0.0232$	-0.0007	Random	(8,22,32)	-0.0065	-0.0055	$+0.0010$

Table 19: **Scope: item-level.** Same-condition chain on Qwen3.5-9B with 100 selection items per seed, $\alpha = 1.3$, nats/token and item-averaged loss. Left: Spearman over 32 layers between the derivative and the measured single-layer gain on the selection items, with the derivative from the same items, from the calibration items, and with both sides token-weighted. Right: the frozen derivative triple; predicted gain from calibration and from the same items, sum of measured single-layer gains, joint gain G and interaction $I = G - \sum_l A_l$.

Seed	ρ same	ρ cal	ρ token	Triple	Cal. pred.	Same pred.	$\sum_l A_l$	G	I
7	0.54	0.43	0.53	$(0, 1, 5)$	$+0.0097$	$+0.0085$	-0.0061	-0.0041	$+0.0020$
42	0.69	0.64	0.68	$(1, 2, 4)$	$+0.0090$	$+0.0049$	$+0.0021$	$+0.0010$	-0.0010
100	0.64	0.72	0.62	$(3, 6, 15)$	$+0.0102$	$+0.0096$	$+0.0061$	$+0.0046$	-0.0015
123	0.68	0.59	0.71	$(0, 4, 15)$	$+0.0067$	$+0.0064$	-0.0022	-0.0010	$+0.0012$
Mean				derivative	$+0.0089$	$+0.0073$	0.0000	$+0.0001$	$+0.0002$
Mean				all 21 frozen	$+0.0035$	$+0.0034$	-0.0103	-0.0075	$+0.0028$

Teacher-forcing window. Teacher forcing truncates prompt plus gold solution at 1,024 tokens, so for long targets the measured loss is the loss of a solution prefix; within a seed all intervention arms see the same tokens, so the decomposition is unaffected. Comparing each item's stored target-token count with a re-tokenization of the untruncated target, with the identical tokenizer and prompt and checked item by item against the stored counts, shows that seeds 7 and 42 have no shortened target; seed 100 has 3 shortened test targets with 30 gold tokens omitted; and seed 123 has 69 shortened test targets with 3,319 tokens omitted, 7 calibration targets with 146 and 2 selection targets with 244. The shortened test targets belong to 69 distinct questions, which include the three of seed 100. Table 20 repeats the test-item analysis on the 1,250 questions whose targets are complete in every seed, with per-question flags and statistics in the supplement; without the two shortened selection items, the registered composition endpoint E2 of Table 40 is 0.960, with mean $|I|$/mean $|G| = 0.233$ and 10 of 423 sign disagreements.

Table 20: Test-item chain of the four frozen derivative triples and largest-term counts over the 21 frozen triples on Qwen3.5-9B at $\alpha = 1.3$ in nats/token, on all 1,319 test questions and on the 1,250 questions whose prompt plus gold solution fits the 1,024-token window in every seed. Each row uses one normalization throughout: Cal. pred. averages the calibration-item derivatives over items, or weights them by their gold-token counts in the token-weighted rows. Removed: finite step as a fraction of the same-item prediction. Interval: question-level bootstrap of the mean joint gain with 19,999 resamples of the common questions, shared across seeds, with the frozen triples fixed.

Questions	Loss	Cal. pred.	Test pred.	$\sum_l A_l$	G	Removed	Largest: step/transfer/interaction
1,319	item-averaged	0.0089	0.0072	0.0007	0.0006	91%	17/3/1
1,319	token-weighted	0.0080	0.0060	-0.0002	-0.0001	103%	16/3/2
1,250	item-averaged	0.0089	0.0074	0.0008	0.0007 [0.0003, 0.0011]	89%	17/3/1
1,250	token-weighted	0.0080	0.0063	0.0000	0.0001	99%	17/3/1

Prospective profile ranking from measured single-layer gains. The registered layer-rank endpoint E1 of Table 40 concerns the order of the 32 layers; choosing a profile also needs the level of each layer's gain, which the finite step changes. Measured single-layer gains at $\alpha = 1.3$ on the 100 calibration items give that level. Table 21 ranks every candidate triple of each seed's bank, the 423 candidates of E2, by the sum of its three calibration single-layer gains and evaluates the ranking against the joint gain measured on the disjoint selection items. The top-ranked candidate is within 0.0025 of the best of the bank in every seed, against a bank mean of -0.014, and 17 of the 20 top-five candidates have $G \geq 0$. Intervals come from 4,000 resamples of the selection items with calibration fixed, and the pooled row is the mean of the four seeds; per-seed and pooled predictor statistics and the per-candidate losses are in the supplement. The derivative-pool triples of Table 4 are a different object: they were frozen by first answers on the selection items from the derivative's six-layer pool.

Table 21: Prospective ranking of the candidate triples by summed calibration single-layer gains on Qwen3.5-9B at $\alpha = 1.3$, with item-averaged loss in nats/token and joint gains G on the 100 selection items. ρ: Spearman between the predicted sum and G over the candidates; top-1: top-ranked candidate and its G; top-3: mean G of the three top-ranked; regret: best G in the bank minus top-1 G; harmful: top-five candidates with $G < 0$; bank: mean and maximum G over all candidates.

Seed	n	ρ	Top-1 triple	Top-1 G	Top-3 G	Regret	Harmful	Bank mean	Bank max
7	99	0.899	$(1, 4, 5)$	-0.0021	-0.0014	0.0025	3/5	-0.0166	0.0004
42	115	0.915	$(1, 2, 4)$	$+0.0010$	$+0.0022$	0.0024	0/5	-0.0139	0.0035
100	117	0.948	$(1, 3, 6)$	$+0.0074$	$+0.0069$	0.0006	0/5	-0.0113	0.0080
123	92	0.823	$(3, 4, 18)$	$+0.0021$	$+0.0011$	0.0000	0/5	-0.0151	0.0021
Pooled	423	0.896		$+0.0021$	$+0.0022$	0.0014	3/20	-0.0142	
95% interval		$[0.871, 0.898]$		$[0.0011, 0.0031]$	$[0.0013, 0.0031]$	$[0.0010, 0.0023]$		$[-0.0151, -0.0133]$	

Rules compared on one target. Table 22 compares four ranking rules on the same banks, calibration items, strength and target, the joint loss gain on the selection items. The calibration derivative orders the 32 layers well, as endpoint E1 shows, but its summed level is shifted by the finite step at $\alpha = 1.3$: its top-ranked triple contains the last block in every seed, for which the derivative predicts the largest gain of any block, 0.005–0.017 nats/token, while the measured single-layer gain at $\alpha = 1.3$ is -0.026 to -0.043, a finite step larger than the prediction itself. The last row of the table repeats the whole measured-sum procedure on 2,000 joint resamples of the calibration items, re-estimating every single-layer gain and re-selecting, and of the selection items, so its interval includes the uncertainty of the choice itself, not only of the evaluation of a fixed triple. Across seed–resample pairs the original triple is re-selected in 45, 34, 80 and 54% for seeds 7, 42, 100 and 123, and the re-selected triple has $G \geq 0$ in 76% overall: in 94%, 99.8% and 82% for seeds 42, 100 and 123, and in 26% for seed 7, whose bank maximum is 0.0004.

Table 22: Ranking rules on the same 423-triple Qwen3.5-9B candidate banks at $\alpha = 1.3$ with item-averaged loss; means over four seeds, with 95% intervals from 4,000 selection-item resamples with calibration fixed except in the last row. Harmful: top-five candidates with $G < 0$, summed over seeds. Last row: 2,000 full-procedure resamples of calibration and selection items, regret against the resampled bank maximum.

Rule on calibration items	ρ with G	Top-1 G	Regret	Harmful
Summed measured single-layer gains	0.896 $[0.871, 0.898]$	$+0.0021$ $[0.0011, 0.0031]$	0.0014	3/20
Summed derivatives	-0.40 $[-0.44, -0.37]$	-0.0372 $[-0.0398, -0.0346]$	0.0407	20/20
Summed residual energy	-0.82 $[-0.83, -0.78]$	-0.0262 $[-0.0283, -0.0242]$	0.0297	20/20
Depth, sum of indices	-0.55 $[-0.58, -0.50]$	-0.0218 $[-0.0237, -0.0200]$	0.0253	20/20
Measured sums, full procedure	—	$+0.0024$ $[0.0010, 0.0038]$	0.0013 $[0.0003, 0.0025]$	—

Cheaper corrections of the derivative level. Table 23 adds, on the same banks, items, strength and target, two rules registered before they were computed that measure only part of what the measured sums need: *screen-m* measures the single-block gain of the m bank blocks with the largest calibration derivative and keeps the derivative for every other block, and *sub-n* measures every block on the first n of the 100 calibration items. Screen-8 chooses the same triple as the full measured sums in all four seeds at an estimated 45% of their stored single-block stage time, a derivative pass plus a quarter of the single-block pass against the full single-block pass; screen-4 and sub-25 keep the regret at 0.0017

at a quarter to a third of it, and sub-50 has regret 0.0013 with no harmful choice. The identity profile, with $G = 0$, is the reference for harm; the bank maximum is 0.0004–0.0080 per seed.

Table 23: Loss-only choice by rules of increasing measurement cost on the Qwen3.5-9B banks at $\alpha = 1.3$, from calibration items to the joint gain on the selection items; means over four seeds with 95% intervals from 4,000 selection-item resamples with calibration fixed. Harmful: seeds whose chosen triple has $G < 0$. Passes: nominal forward-pass equivalents per calibration item, counting the derivative as three; stage time: stored seconds per seed for the passes the rule needs.

Rule	Blocks measured	Passes	Stage time in s	Top-1 G	Regret	Harmful
Summed derivatives	0	3	149–198	-0.0372 $[-0.0399, -0.0346]$	0.0407	4/4
Screen-4	4	7	243–328	$+0.0018$ $[0.0008, 0.0028]$	0.0017	1/4
Screen-8	8	11	338–458	$+0.0021$ $[0.0011, 0.0031]$	0.0014	1/4
Sub-25, 25 items	32	8	189–260	$+0.0017$ $[0.0007, 0.0028]$	0.0017	1/4
Sub-50, 50 items	32	16	377–520	$+0.0022$ $[0.0011, 0.0033]$	0.0013	0/4
Measured sums, all	32	32	755–1,040	$+0.0021$ $[0.0011, 0.0031]$	0.0014	1/4

Numerical implementation. The derivative is computed at batch 1 and the measured losses in batches of 32; identity and intervened losses use the same batched routine, so the measured gains are differences of equally batched quantities. The registered control of Table 24 re-measured all 32 single-layer gains at $\alpha = 1.3$ in fp32 on 32 selection items of two seeds and computed the exact derivative by fp32 automatic differentiation: the fp16 and fp32 gains differ by 0.8% and 1.6% of the fp32 finite-step term against a registered threshold of 20%, rank the layers at Spearman 0.997 and 0.998, and the batch-1 fp16 derivative ranks the layers as the exact fp32 derivative does, at Spearman 0.995 and 0.997. Separately, precision and the residual write each change the single-layer gains by under 2% of the fp32 finite-step term, precision by 0.8% and 1.6% and the write by 0.8% and 1.9%, and batch size and $\alpha = 1$ interventions change per-item losses by at most 1.2×10^{-3} in Table 24. A second registered analysis of the same records checks the magnitude, not only the order, of the derivative: the fp16 derivative differs from the fp32 derivative by 1.2–3.8% of the fp32 finite step summed over blocks; for each seed's frozen derivative triple, every combination of fp16 and fp32 gains and derivatives keeps the removed fraction within an 11-point range; and the $\alpha = 1$ interventions change the mean loss by $+2.6 \times 10^{-5}$ and -0.1×10^{-5}, with intervals inside $\pm 1.5 \times 10^{-4}$, below the mean joint gains discussed in the main text. The fp32 Qwen3-8B chain of Table 3 gives the same ordering of terms at $\alpha = 1.3$.

Matched ablation. Section 4.3 joins three quantities measured on the same 100 calibration items of each seed: the measured ablation S_l, which is the item-averaged cross-entropy increase when block l's output is replaced by its input and comes from the pool-scoring run with per-item rows in the supplement, and the measured gain A_l at $\alpha = 1.3$ and the derivative prediction D_l, both item-averaged and from this experiment. The pool-scoring and selection measurements use item-identical calibration sets in all four seeds, checked before measuring; Table 25 gives the correlations.

Loss normalization. The pools were scored with the item-averaged loss above, whereas the Qwen3-8B chain of Appendix D uses the gold-token-weighted loss, with gradient sums divided by the total number of gold tokens, for the derivative and for the finite differences alike; the supplementary code contains both normalizations for this reason. Where both were measured on the same items, the ordering of terms does not depend on the choice: with both sides token-weighted the per-seed Spearman is 0.53–0.71 in Table 19; pairing the item-averaged derivative with token-weighted measured gains changes it by at most 0.06; and the chains from the same-item prediction to $\sum_l A_l$ to G are

21 frozen triples, selection items:	$0.0034 \to -0.0101 \to -0.0074$	token-weighted,
	$0.0034 \to -0.0103 \to -0.0075$	item-averaged;
derivative triples, test items:	$0.0060 \to -0.0002 \to -0.0001$	token-weighted,
	$0.0072 \to \ 0.0007 \to \ 0.0006$	item-averaged.

Table 24: Registered numerical control of the fp16 implementation on Qwen3.5-9B, using the first 32 selection items of seeds 7 and 42 and all 32 blocks; the registration is in the supplement. A_l: measured single-layer gain; $D_l^{\mathrm{fp32}} = -0.3\,\partial\mathcal{L}/\partial\alpha_l$ from exact fp32 automatic differentiation. The first row is the registered precision ratio R at $\alpha = 1.3$, displayed above the rows. Cells with two values list them in the order of the row label.

$$R = \frac{\sum_l |A_l^{\mathrm{fp16}} - A_l^{\mathrm{fp32}}|}{\sum_l |A_l^{\mathrm{fp32}} - D_l^{\mathrm{fp32}}|}$$

Registered quantity	Seed 7	Seed 42
Precision ratio R, decision threshold 0.2	0.008	0.016
Spearman A^{fp16} vs. A^{fp32} at $\alpha = 1.3$; alternative residual write, mean $\lvert\Delta A_l\rvert$	0.997; 2.6×10^{-5}	0.998; 4.9×10^{-5}
Spearman of the fp16, batch-1 bank prediction D_l vs. exact fp32 D_l^{fp32}	0.995	0.997
Spearman automatic differentiation vs. central difference at $\alpha = 1 \pm 0.01$, fp32; fp16	0.996; 0.867	0.998; 0.883
Layers with |automatic differentiation−difference| > 0.25 |difference|, fp32; fp16	5/32 (0, 3, 4, 7, 25); 20/32	2/32 (6, 7); 21/32
Derivative magnitude $R_D = \sum_l \lvert D_l - D_l^{\mathrm{fp32}}\rvert / \sum_l \lvert A_l^{\mathrm{fp32}} - D_l^{\mathrm{fp32}}\rvert$: bank D_l; fp16 automatic differentiation	0.038; 0.017	0.012; 0.014
Frozen derivative triples $(0, 1, 5)$ and $(1, 2, 4)$: summed D_l from the bank and from fp32; removed fraction over all fp16/fp32 choices of A and D	0.0104; 0.0124; 158–169%	0.0049; 0.0049; 88–89%
Mean paired per-item $\Delta\mathcal{L}$ with 95% interval, $\alpha = 1$ interventions on all blocks	$+2.6\times10^{-5}$; $[-8.2, 14.7]\times10^{-5}$	-0.1×10^{-5}; $[-9.5, 8.9]\times10^{-5}$
Max per-item $\lvert\mathcal{L}\rvert$ change: $\alpha = 1$ interventions on all blocks; batch 1 vs. 32	1.2×10^{-3}; 8.2×10^{-4}	5.1×10^{-4}; 9.7×10^{-4}
No-op generation at $\alpha = 1$ on the frozen triple: identical token sequences; correct, with the identity count in parentheses	97/100; 89 (89)	94/100; 88 (89)

Table 25: Matched ablation on Qwen3.5-9B with 100 calibration items per seed and 32 blocks; Spearman correlations over layers. $D_l/0.3$ is the first-order prediction of S_l; "signs" counts layers with $\operatorname{sign} S_l = \operatorname{sign} D_l$.

Seed	$\rho(S, A)$	$\rho(S, D/0.3)$	Signs	$\rho(D, A)$
7	-0.188	$+0.283$	14/32	$+0.581$
42	-0.089	$+0.167$	13/32	$+0.754$
100	$+0.025$	$+0.359$	15/32	$+0.760$
123	-0.003	$+0.251$	10/32	$+0.772$

F Registered cross-family chain on Llama-3.1-8B and Mistral-7B-v0.3

Design. The design was registered before any unit was launched, with the registration in the supplement, and nothing was tuned afterwards. Checkpoints: the base Llama-3.1-8B and Mistral-7B-v0.3 with 32 blocks each, fp16 weights and cross-entropy in fp32; Mistral-7B-v0.3 was named as the confirmatory family of the pre-stated predictions. Tasks: GSM8K with a three-shot "Question/ Answer" prompt and the gold solution as target, and ARC-Challenge with a three-shot multiple-choice prompt and the gold label as target, one token for 194 of the 200 evaluation questions and two for six, and one token for 96–99 of the 100 calibration items. Per seed, the train split is shuffled with the seed; three items are the exemplars and the next 100 the calibration items; the evaluation items are 200 test questions shared by all seeds. Prompts plus targets fit the 2,048-token window for every item. Intervention and derivative are those of Section 3, with the per-item derivative at batch 1 and predicted gain $-(\alpha - 1)\,\partial\mathcal{L}_i/\partial\alpha_l$, at $\alpha \in \{1.05, 1.1, 1.3\}$. For every unit we measure the derivatives and all 32 single-block losses on calibration and evaluation items, then fix the profiles from calibration data

only before any joint loss is measured: for $k \in \{3, 6\}$ the top-k blocks by calibration derivative, by residual energy, by depth, that is the last k, and by measured calibration single-block gain at $\alpha = 1.3$, and ten seeded random profiles, with the three-block ones nested in the six-block ones; the joint losses of every unique profile are then measured at every strength on both item sets. An fp32 control repeats identity, all single blocks at $\alpha = 1.05$ and 1.3 and the derivatives on the first 16 evaluation items. Intervals are question-level bootstraps with 19,999 resamples of the 200 shared questions, conditional on each seed's exemplar and calibration draw. The design has 16 units, from two checkpoints, two tasks and four seeds; all 16 were completed with the registered settings and are reported in Table 41.

Chain by profile set. Table 26 lists the chain for the derivative, measured-sum and random three-block profiles. On GSM8K the three terms order as on the Qwen checkpoints in both families: at $\alpha = 1.3$ the finite step is the largest term for every chosen profile and for 39 of 40 Llama-3.1-8B and 38 of 40 Mistral-7B-v0.3 random profiles. At $\alpha \leq 1.1$ transfer leads for most chosen Llama-3.1-8B profiles, while on Mistral-7B-v0.3, whose derivatives transfer better between item sets with $c_T = 0.0032$, the finite step leads for the chosen profiles from $\alpha = 1.1$ on. The Mistral-7B-v0.3 measured-sum profiles keep a positive joint test gain at every strength, $+0.0014$ $[0.0006, 0.0021]$ at $\alpha = 1.3$. On ARC-Challenge, where calibration and test derivatives agree less, the calibration prediction of a chosen profile is several times its test-item prediction, a transfer effect; ARC-Challenge differs from GSM8K in both task and target length, with one or two label tokens, and the comparison covers the two jointly; the finite step of random profiles grows with strength and is the largest term for 20 of 40 Llama-3.1-8B and 17 of 40 Mistral-7B-v0.3 random profiles at $\alpha = 1.3$.

Coefficients by strength and profile size. Table 27 divides the mean absolute transfer over all unique profiles of a size by t and the mean absolute finite step and interaction by t^2. Transfer is exactly linear in t. Between $\alpha = 1.1$ and 1.3 the three-block finite-step coefficient changes by -5% on GSM8K and -7% on ARC-Challenge for Llama-3.1-8B and by $+1\%$ and $+7\%$ for Mistral-7B-v0.3, so the finite step is close to quadratic over this range; the interaction coefficient roughly doubles from three to six blocks. For profiles chosen on the calibration items, transfer also contains the selection on those items: for the Llama-3.1-8B GSM8K derivative top-three, $c_T = 0.010$ and $c_S = 0.10$ at both $\alpha = 1.1$ and 1.3 in Table 26, so $t^* \approx 0.10$, consistent with transfer and finite step being level for these profiles at $\alpha = 1.1$, at 0.0010 each, and the finite step leading at $\alpha = 1.3$.

Predicting the leading term from one strength. Because c_T is exact, a prediction from one strength rests on the stability of c_S. The registered check A4 of Table 28 estimates the coefficients at $\alpha = 1.1$ only and predicts the mean absolute terms at $\alpha = 1.3$: the predicted leading term is the measured one in all five groups, the predicted mean finite step is within -6% to $+11\%$ of the measured value, and all five checks meet the registered criterion. For ARC-Challenge the coefficients frozen at $\alpha = 1.1$ put the switch at $\alpha^*_{1.1} = 1.32$ for Llama-3.1-8B and 1.33 for Mistral-7B-v0.3, just above $\alpha = 1.3$, where transfer still leads narrowly, 0.0242 against 0.0214 and 0.0167 against 0.0162. The registered measurement A6 tests this switch prospectively. Before any new strength was measured, the $\alpha = 1.1$ coefficients were frozen and their predictions for $\alpha \in \{1.2, 1.4, 1.5\}$ recorded; the unchanged measurement code then measured all single blocks and all profiles of the ARC-Challenge units at these strengths on the same items; Table 29 reports the three-block profiles of all seeds. In both families the predicted leading term is the measured one at all three strengths; the measured switch, obtained by linear interpolation of $\log(|\overline{S}|/|\overline{T}|)$ between adjacent strengths, lies at $\alpha = 1.35$ against the predicted 1.32 on Llama-3.1-8B and at 1.31 against 1.33 on Mistral-7B-v0.3, and the predicted mean finite step is within $+3\%$ to $+14\%$ on Llama-3.1-8B and -6% to -5% on Mistral-7B-v0.3 of the measured value. c_S decreases slowly with strength on Llama-3.1-8B, from 0.255 at $\alpha = 1.1$ to 0.224 at $\alpha = 1.5$, so its frozen prediction slightly exceeds the measured finite step at the largest strengths, and stays within 0.169–0.181 on Mistral-7B-v0.3. The identity losses re-measured in these runs equal the stored ones in every unit, and all three registered hypotheses hold in both groups of Table 41.

Uncertainty of the switch, A10. Table 30 resamples the 200 evaluation questions 19,999 times, with the resamples shared by all seeds, strengths and profiles and with exemplars, calibration items and profiles fixed, and recomputes in every resample both the switch $\alpha^*_{1.1}$ predicted from the $\alpha = 1.1$ coefficients and the measured switch α^* over $\alpha \in \{1.1, 1.2, 1.3, 1.4, 1.5\}$, interpolated as in Table 29. With three-block profiles the switch falls inside the measured range in 89.8% of resamples for

Table 26: Test-item chain of the three-block profiles with item-averaged loss in nats/token, averaged over the profiles of all seeds. Columns: calibration prediction, same-item prediction, sum of measured single-block gains, joint gain G with question-level 95% interval, signed transfer, finite step and interaction, and the number of profiles in which transfer/step/interaction is the largest term.

α	Profiles	n	Cal. pred.	Same pred.	$\sum A$	G	Transfer	Step	I	Largest
Llama-3.1-8B, GSM8K, seeds 7, 42, 100 and 123										
1.05	derivative	4	$+0.0017$	$+0.0012$	$+0.0009$	$+0.0009$ $[0.0006, 0.0012]$	-0.0005	-0.0003	0.0000	3/1/0
1.05	measured sum	4	$+0.0014$	$+0.0009$	$+0.0007$	$+0.0008$ $[0.0006, 0.0010]$	-0.0004	-0.0002	0.0000	4/0/0
1.05	random	40	-0.0002	-0.0002	-0.0006	-0.0006 $[-0.0007, -0.0005]$	0.0000	-0.0004	0.0000	15/25/0
1.1	derivative	4	$+0.0033$	$+0.0023$	$+0.0014$	$+0.0013$ $[0.0008, 0.0019]$	-0.0010	-0.0010	0.0000	2/2/0
1.1	measured sum	4	$+0.0027$	$+0.0018$	$+0.0012$	$+0.0012$ $[0.0008, 0.0016]$	-0.0009	-0.0006	0.0000	3/1/0
1.1	random	40	-0.0005	-0.0004	-0.0019	-0.0019 $[-0.0021, -0.0017]$	$+0.0001$	-0.0015	0.0000	6/34/0
1.3	derivative	4	$+0.0100$	$+0.0070$	-0.0017	-0.0015 $[-0.0029, -0.0000]$	-0.0030	-0.0088	$+0.0003$	0/4/0
1.3	measured sum	4	$+0.0081$	$+0.0055$	$+0.0001$	-0.0001 $[-0.0012, 0.0010]$	-0.0026	-0.0054	-0.0002	0/4/0
1.3	random	40	-0.0015	-0.0013	-0.0142	-0.0139 $[-0.0147, -0.0131]$	$+0.0002$	-0.0129	$+0.0003$	1/39/0
Llama-3.1-8B, ARC-C, seeds 7, 42, 100 and 123										
1.05	derivative	4	$+0.0119$	$+0.0017$	$+0.0008$	$+0.0006$ $[-0.0025, 0.0036]$	-0.0102	-0.0010	-0.0001	4/0/0
1.05	measured sum	4	$+0.0099$	$+0.0030$	$+0.0027$	$+0.0025$ $[-0.0002, 0.0051]$	-0.0069	-0.0003	-0.0002	4/0/0
1.05	random	40	$+0.0005$	$+0.0004$	-0.0001	-0.0001 $[-0.0011, 0.0009]$	0.0000	-0.0005	0.0000	36/4/0
1.1	derivative	4	$+0.0238$	$+0.0035$	-0.0001	-0.0009 $[-0.0071, 0.0052]$	-0.0203	-0.0035	-0.0008	4/0/0
1.1	measured sum	4	$+0.0198$	$+0.0059$	$+0.0046$	$+0.0035$ $[-0.0016, 0.0084]$	-0.0138	-0.0013	-0.0012	4/0/0
1.1	random	40	$+0.0010$	$+0.0009$	-0.0015	-0.0013 $[-0.0034, 0.0009]$	-0.0001	-0.0024	$+0.0002$	34/6/0
1.3	derivative	4	$+0.0713$	$+0.0104$	-0.0170	-0.0178 $[-0.0368, 0.0012]$	-0.0610	-0.0274	-0.0007	4/0/0
1.3	measured sum	4	$+0.0593$	$+0.0178$	$+0.0062$	$+0.0016$ $[-0.0129, 0.0160]$	-0.0415	-0.0117	-0.0046	4/0/0
1.3	random	40	$+0.0029$	$+0.0026$	-0.0177	-0.0169 $[-0.0248, -0.0087]$	-0.0003	-0.0202	$+0.0008$	20/20/0
Mistral-7B-v0.3, GSM8K, seeds 7, 42, 100 and 123										
1.05	derivative	4	$+0.0010$	$+0.0008$	$+0.0007$	$+0.0007$ $[0.0005, 0.0008]$	-0.0002	-0.0001	0.0000	3/1/0
1.05	measured sum	4	$+0.0008$	$+0.0007$	$+0.0007$	$+0.0006$ $[0.0005, 0.0008]$	-0.0001	0.0000	-0.0001	2/0/2
1.05	random	40	-0.0003	-0.0002	-0.0003	-0.0003 $[-0.0004, -0.0003]$	$+0.0001$	-0.0002	0.0000	23/16/1
1.1	derivative	4	$+0.0020$	$+0.0016$	$+0.0012$	$+0.0011$ $[0.0008, 0.0014]$	-0.0003	-0.0005	-0.0001	0/4/0
1.1	measured sum	4	$+0.0015$	$+0.0014$	$+0.0011$	$+0.0011$ $[0.0008, 0.0013]$	-0.0001	-0.0003	-0.0001	0/4/0
1.1	random	40	-0.0005	-0.0004	-0.0011	-0.0011 $[-0.0012, -0.0009]$	$+0.0002$	-0.0007	0.0000	11/29/0
1.3	derivative	4	$+0.0059$	$+0.0049$	$+0.0006$	$+0.0001$ $[-0.0007, 0.0009]$	-0.0010	-0.0043	-0.0005	0/4/0
1.3	measured sum	4	$+0.0045$	$+0.0042$	$+0.0018$	$+0.0014$ $[0.0006, 0.0021]$	-0.0003	-0.0024	-0.0004	0/4/0
1.3	random	40	-0.0016	-0.0011	-0.0075	-0.0073 $[-0.0077, -0.0068]$	$+0.0005$	-0.0064	$+0.0003$	2/38/0
Mistral-7B-v0.3, ARC-C, seeds 7, 42, 100 and 123										
1.05	derivative	4	$+0.0095$	$+0.0023$	$+0.0019$	$+0.0019$ $[-0.0006, 0.0042]$	-0.0072	-0.0004	-0.0001	4/0/0
1.05	measured sum	4	$+0.0086$	$+0.0025$	$+0.0024$	$+0.0021$ $[-0.0010, 0.0049]$	-0.0061	-0.0001	-0.0003	4/0/0
1.05	random	40	-0.0003	-0.0005	-0.0009	-0.0010 $[-0.0016, -0.0003]$	-0.0002	-0.0004	-0.0001	35/2/3
1.1	derivative	4	$+0.0190$	$+0.0046$	$+0.0023$	$+0.0030$ $[-0.0015, 0.0074]$	-0.0144	-0.0023	$+0.0007$	4/0/0
1.1	measured sum	4	$+0.0172$	$+0.0051$	$+0.0034$	$+0.0032$ $[-0.0027, 0.0086]$	-0.0121	-0.0017	-0.0002	3/1/0
1.1	random	40	-0.0006	-0.0010	-0.0025	-0.0027 $[-0.0040, -0.0013]$	-0.0004	-0.0015	-0.0002	31/8/1
1.3	derivative	4	$+0.0571$	$+0.0139$	-0.0077	-0.0035 $[-0.0151, 0.0081]$	-0.0433	-0.0216	$+0.0042$	3/1/0
1.3	measured sum	4	$+0.0517$	$+0.0152$	$+0.0007$	-0.0006 $[-0.0160, 0.0133]$	-0.0364	-0.0145	-0.0013	3/1/0
1.3	random	40	-0.0018	-0.0029	-0.0175	-0.0174 $[-0.0229, -0.0118]$	-0.0012	-0.0146	$+0.0001$	23/17/0

Llama-3.1-8B and 94.6% for Mistral-7B-v0.3, otherwise above $\alpha = 1.5$; it lies within 0.10 of the resampled prediction in 86.5% and 91.6% of resamples, and each seed alone switches within 0.08 of its own prediction. Six-block profiles, for which the $\alpha = 1.1$ prediction was computed after the new strengths had been measured, switch earlier, at 1.26 and 1.22 against a predicted 1.24, inside the measured range in 97.7% and 99.5% of resamples. This analysis was registered as descriptive, without a decision rule.

Loss-only choice within each profile set. Table 31 reports, per rule, the test joint gain of the profile it ranks first among the unit's candidates, with the same candidates, strength and items for every rule, the regret against the best candidate and the number of seeds in which the chosen profile is harmful. At $\alpha = 1.3$ the measured sums have a higher seed-mean joint gain and a lower regret than the derivative sums in all eight three- and six-block sets of the two checkpoints, and they choose the best profile of every GSM8K set; when even the best profile of a set has $G < 0$, as for Llama-3.1-8B GSM8K at $\alpha = 1.3$, the unmodified model is the reference choice. Table 32 adds the cheaper rules of Table 23: screen-8 has a lower regret than the derivative sums in all eight sets and the seed-mean joint gain of the full measured sums in four of them.

Second-order correction from a small-strength probe. The registered analysis A9 corrects the derivative block by block with the quadratic model of Section 3.3. With a_l the exact derivative per

Table 27: Error-term coefficients on the test items: $c_T = \overline{|T|}/t$, $c_S = \overline{|S|}/t^2$, $c_I = \overline{|I|}/t^2$ over all unique k-block profiles of all seeds, and the finite-step share $\overline{|S|}/(\overline{|T|} + \overline{|S|} + \overline{|I|})$. L: Llama-3.1-8B; M: Mistral-7B-v0.3.

Units	k	α	n	c_T	c_S	c_I	Step share
L, GSM8K	3	1.05	55	0.0058	0.210	0.054	0.55
L, GSM8K	3	1.1	55	0.0058	0.208	0.042	0.68
L, GSM8K	3	1.3	55	0.0058	0.198	0.040	0.77
L, GSM8K	6	1.05	56	0.0073	0.313	0.103	0.56
L, GSM8K	6	1.1	56	0.0073	0.310	0.091	0.65
L, GSM8K	6	1.3	56	0.0073	0.298	0.088	0.73
L, ARC-C	3	1.05	56	0.0806	0.279	0.148	0.14
L, ARC-C	3	1.1	56	0.0806	0.255	0.082	0.22
L, ARC-C	3	1.3	56	0.0806	0.238	0.051	0.43
L, ARC-C	6	1.05	56	0.1052	0.458	0.255	0.16
L, ARC-C	6	1.1	56	0.1052	0.439	0.163	0.27
L, ARC-C	6	1.3	56	0.1052	0.408	0.124	0.46
M, GSM8K	3	1.05	56	0.0032	0.086	0.018	0.51
M, GSM8K	3	1.1	56	0.0032	0.093	0.017	0.65
M, GSM8K	3	1.3	56	0.0032	0.094	0.019	0.76
M, GSM8K	6	1.05	56	0.0041	0.132	0.041	0.52
M, GSM8K	6	1.1	56	0.0041	0.146	0.038	0.65
M, GSM8K	6	1.3	56	0.0041	0.147	0.044	0.72
M, ARC-C	3	1.05	56	0.0555	0.190	0.102	0.14
M, ARC-C	3	1.1	56	0.0555	0.169	0.040	0.22
M, ARC-C	3	1.3	56	0.0555	0.181	0.031	0.46
M, ARC-C	6	1.05	55	0.0674	0.322	0.172	0.18
M, ARC-C	6	1.1	55	0.0674	0.284	0.086	0.27
M, ARC-C	6	1.3	55	0.0674	0.318	0.082	0.51

Table 28: Registered frozen-strength prediction A4: coefficients estimated at $\alpha = 1.1$ only predict the mean absolute terms at $\alpha = 1.3$ for the three-block profiles of all seeds and, on Qwen3-8B, for the twenty fixed profiles of Table 3. $\alpha^*_{1.1} = 1 + c_T/c_S$ with the $\alpha = 1.1$ coefficients. For all groups except Mistral-7B-v0.3 GSM8K the $\alpha = 1.3$ records had been analysed before; the check holds the strength out of the estimate.

Units	n	c_T	$c_S(1.1)$	$c_S(1.3)$	$\alpha^*_{1.1}$	$\lvert S\rvert$ pred. / meas.	Error	Leader pred. / meas.	Decision
Q3, GSM8K, 20 fixed	20	0.0220	0.115	0.104	1.19	0.0104 / 0.0093	+0.11	finite step / finite step	met
L, GSM8K	55	0.0058	0.208	0.198	1.03	0.0187 / 0.0178	+0.05	finite step / finite step	met
L, ARC-C	56	0.0806	0.255	0.238	1.32	0.0229 / 0.0214	+0.07	transfer / transfer	met
M, GSM8K	56	0.0032	0.093	0.094	1.03	0.0084 / 0.0084	−0.01	finite step / finite step	met
M, ARC-C	56	0.0555	0.169	0.181	1.33	0.0152 / 0.0162	−0.06	transfer / transfer	met

unit strength and $A_l(t_0)$ the measured calibration single-block gain at the probe $t_0 = 0.1$, that is $\alpha = 1.1$, the curvature is $b_l = (A_l(t_0) - a_l t_0)/t_0^2$ and

$$Q_l(t) = a_l\, t + b_l\, t^2$$

predicts the single-block gain at $t = 0.3$, that is $\alpha = 1.3$. Scoring profiles by $\sum_l Q_l(0.3)$ reaches the seed-mean joint gain of the full measured sums at $\alpha = 1.3$ in seven of the eight sets and has a regret no larger than the derivative sums in all eight sets of Table 33; it raises the block-level Spearman correlation with the measured gains at $\alpha = 1.3$ from 0.79–0.90 to 0.90–0.98 and removes 52–89% of the calibration finite step. The probe needs as many forward passes as the full measurement, at a strength three times smaller, and the quadratic model then serves every strength. The $\alpha = 1.05$ probe, at which fp16 rounding is of the size of the finite step as the next paragraph shows, is listed for completeness.

Precision. The fp32 control compares fp16 and fp32 single-block gains with the fp32 finite step, $\sum_l |A_l^{\mathrm{fp16}} - A_l^{\mathrm{fp32}}| / \sum_l |A_l^{\mathrm{fp32}} - D_l^{\mathrm{fp32}}|$, on 16 evaluation items per unit: at $\alpha = 1.3$ it is 0.011–0.028 on GSM8K and 0.075–0.235 on ARC-Challenge over both checkpoints, with Spearman 0.993–1.000 between fp16 and fp32 gains; at $\alpha = 1.05$ it is 0.43–0.90 and 1.9–7.9, with Spearman 0.91–0.99, so at $\alpha = 1.05$ fp16 rounding is of the size of the finite step and the $\alpha = 1.05$ finite-step values are indicative only. A registered fp32 re-measurement, A11, covers the $\alpha = 1.1$ anchor and $\alpha = 1.3$ and

Table 29: Registered new-strength measurement A6 on the ARC-Challenge units: measured mean absolute transfer, finite step and interaction over the three-block profiles of all seeds at each strength, with $\alpha = 1.2$, 1.4 and 1.5 newly measured, $c_S = \overline{|S|}/t^2$, the number of profiles in which transfer/step/interaction is the largest term, the measured leading term, and the mean finite step and leading term predicted from the coefficients frozen at $\alpha = 1.1$ before the measurement.

Units	α	n	$\lvert T\rvert$	$\lvert S\rvert$	$\lvert I\rvert$	c_S	Largest	Leader	$\lvert S\rvert$ pred.	Leader pred.
L, ARC-C	1.1	56	0.0081	0.0025	0.0008	0.255	48/8/0	transfer	anchor	anchor
L, ARC-C	1.2	56	0.0161	0.0099	0.0022	0.247	40/16/0	transfer	0.0102	transfer
L, ARC-C	1.3	56	0.0242	0.0214	0.0046	0.238	29/27/0	transfer	0.0229	transfer
L, ARC-C	1.4	56	0.0322	0.0368	0.0075	0.230	23/33/0	finite step	0.0407	finite step
L, ARC-C	1.5	56	0.0403	0.0560	0.0118	0.224	19/37/0	finite step	0.0636	finite step
M, ARC-C	1.1	56	0.0056	0.0017	0.0004	0.169	43/12/1	transfer	anchor	anchor
M, ARC-C	1.2	56	0.0111	0.0071	0.0012	0.178	35/21/0	transfer	0.0068	transfer
M, ARC-C	1.3	56	0.0167	0.0162	0.0028	0.181	29/27/0	transfer	0.0152	transfer
M, ARC-C	1.4	56	0.0222	0.0288	0.0050	0.180	23/33/0	finite step	0.0270	finite step
M, ARC-C	1.5	56	0.0278	0.0444	0.0079	0.178	15/41/0	finite step	0.0422	finite step

Table 30: Uncertainty of the ARC-Challenge switch, A10. $\alpha^*_{1.1}$: switch predicted from the $\alpha = 1.1$ coefficients; α^*: measured switch. Intervals are 95% percentile intervals over 19,999 shared question resamples with exemplars, calibration items and profiles fixed. In range: share of resamples with the switch in $[1.1, 1.5]$, over which the α^* and difference intervals are taken; within 0.10: share of all resamples in range with $|\alpha^* - \alpha^*_{1.1}| \le 0.10$; per-seed α^*: seeds 7/42/100/123 alone, with the largest per-seed $|\alpha^* - \alpha^*_{1.1}|$; last column: $\overline{|S|} - \overline{|T|}$ at $\alpha = 1.5$. The six-block rows, marked by an asterisk, are post hoc.

Units	k	$\alpha^*_{1.1}$	α^*	$\alpha^*_{1.1}$ interval	In range	α^* interval	$\alpha^* - \alpha^*_{1.1}$	Within 0.10	Per-seed α^*	Max diff.	$\overline{\lvert S\rvert} - \overline{\lvert T\rvert}$ $\alpha = 1.5$
L, ARC-C	3	1.32	1.35	$[1.22, 1.50]$	89.8%	$[1.24, 1.49]$	$[-0.01, +0.11]$	86.5%	1.43/1.36/1.28/1.34	0.08	$[-0.0071, +0.0315]$
L, ARC-C	6*	1.24	1.26	$[1.16, 1.47]$	97.7%	$[1.17, 1.46]$	$[-0.04, +0.08]$	96.6%	1.26/1.26/1.27/1.25	0.05	$[+0.0007, +0.0744]$
M, ARC-C	3	1.33	1.31	$[1.24, 1.58]$	94.6%	$[1.26, 1.47]$	$[-0.10, +0.07]$	91.6%	1.43/1.28/1.20/1.34	0.04	$[-0.0038, +0.0248]$
M, ARC-C	6*	1.24	1.22	$[1.17, 1.47]$	99.5%	$[1.18, 1.40]$	$[-0.12, +0.04]$	95.4%	1.21/1.26/1.19/1.21	0.05	$[+0.0111, +0.0602]$

1.4 on the first 32 evaluation and 32 calibration questions of all eight ARC-Challenge units: identity, derivatives, all single blocks and all profiles, compared with the fp16 records of the same questions; Table 34 pools the four seeds of each family and Table 35 gives every seed. Block by block, fp16 and fp32 gains at $\alpha = 1.1$ differ by 0.47–1.68 for Llama-3.1-8B and 0.52–1.31 for Mistral-7B-v0.3 times the fp32 finite step, against 0.05–0.18 and 0.08–0.16 at $\alpha = 1.3$, with Spearman at least 0.984 throughout, so at the anchor fp16 rounding is of the size of a single block's finite step and the fp32 rows of Table 34 are the reference there. The leading term at $\alpha = 1.3$ and 1.4 is the same in both precisions in both families and in every seed, and pooled c_T differs by about 0.1% or less between precisions, and by at most 1.3% in single seeds. For Llama-3.1-8B the block-level differences cancel in the four-seed pooled profile means: c_S at $\alpha = 1.1$ differs by 0.5% and the predicted switch by 0.002, while single seeds differ by -40% to $+50\%$ in c_S, so this agreement is a pooled one. For Mistral-7B-v0.3 the pooled c_S at $\alpha = 1.1$ is 0.208 in fp16 and 0.161 in fp32, a 29% difference that moves the predicted subset switch from 1.72 to 1.92, while the leader at $\alpha = 1.3$ and 1.4 is the same in both precisions. Mean absolute interaction at $\alpha = 1.1$ is larger in fp16, 0.0017 and 0.0009 against 0.0010 and 0.0007 in fp32. In fp32 the second-order probe of A9 removes 71% of the calibration finite step on these questions for Llama-3.1-8B and 81% for Mistral-7B-v0.3; the fp16 values of the $\alpha = 1.1$ probe on 32 questions carry the rounding above. The fp32 identity losses equal the stored fp32 control on the 16 shared questions in every unit. On these 32 questions transfer is about twice its size on all 200, $c_T = 0.17$ against 0.08 and 0.15 against 0.06, so the switch predicted for this subset lies beyond the largest measured strength 1.4 in both precisions, at 1.61 for Llama-3.1-8B and at 1.72 and 1.92 for Mistral-7B-v0.3; the single-seed subset switches of Table 35 are likewise extrapolations, not measured crossovers.

Table 31: Profile chosen by summed calibration derivatives, summed measured calibration single-block gains, summed residual energy or depth, within each unit's fixed candidate set. Cells: seed-mean test joint gain G, followed in parentheses by the mean regret and the number of harmful seeds. L: Llama-3.1-8B; M: Mistral-7B-v0.3.

Units	α	k	Derivative	Measured sum	Energy	Depth
L, GSM8K	1.05	3	$+0.0009$ (0.0000; 0)	$+0.0009$ (0.0000; 0)	-0.0024 (0.0034; 4)	-0.0022 (0.0031; 4)
L, GSM8K	1.3	3	-0.0015 (0.0014; 3)	-0.0001 (0.0000; 2)	-0.0425 (0.0424; 4)	-0.0352 (0.0351; 4)
L, GSM8K	1.3	6	-0.0056 (0.0035; 4)	-0.0021 (0.0000; 4)	-0.0287 (0.0266; 4)	-0.0275 (0.0254; 4)
L, ARC-C	1.05	3	$+0.0006$ (0.0039; 2)	$+0.0006$ (0.0039; 2)	-0.0016 (0.0061; 4)	-0.0020 (0.0065; 4)
L, ARC-C	1.3	3	-0.0178 (0.0285; 3)	$+0.0016$ (0.0091; 1)	-0.0279 (0.0387; 4)	-0.0254 (0.0362; 4)
L, ARC-C	1.3	6	-0.0235 (0.0390; 4)	$+0.0080$ (0.0075; 1)	-0.0177 (0.0332; 4)	-0.0215 (0.0370; 4)
M, GSM8K	1.05	3	$+0.0007$ (0.0000; 0)	$+0.0007$ (0.0000; 0)	-0.0015 (0.0022; 4)	-0.0016 (0.0023; 4)
M, GSM8K	1.3	3	$+0.0001$ (0.0012; 2)	$+0.0014$ (0.0000; 0)	-0.0221 (0.0234; 4)	-0.0200 (0.0213; 4)
M, GSM8K	1.3	6	-0.0001 (0.0006; 2)	$+0.0005$ (0.0000; 2)	-0.0143 (0.0148; 4)	-0.0152 (0.0157; 4)
M, ARC-C	1.05	3	$+0.0019$ (0.0011; 0)	$+0.0012$ (0.0018; 0)	-0.0040 (0.0070; 4)	-0.0041 (0.0071; 4)
M, ARC-C	1.3	3	-0.0035 (0.0114; 3)	-0.0006 (0.0085; 3)	-0.0379 (0.0458; 4)	-0.0354 (0.0434; 4)
M, ARC-C	1.3	6	-0.0096 (0.0152; 4)	$+0.0022$ (0.0033; 2)	-0.0381 (0.0437; 4)	-0.0344 (0.0400; 4)

Table 32: Registered cheaper choice rules A5 on the cross-family profile sets at $\alpha = 1.3$: derivative sums; screen-m, the measured gain for the m candidate blocks with the largest calibration derivative and the derivative elsewhere; sub-n, measured gains on the first n of 100 calibration items; full measured sums. Nominal forward-pass equivalents per calibration item: 3, 7, 11, 8, 16, 32. Cells: seed-mean test joint gain G, followed in parentheses by the mean regret and the number of harmful seeds.

Units	k	Derivative	Screen-4	Screen-8	Sub-25	Sub-50	Measured sum
L, GSM8K	3	-0.0015 (0.0014; 3)	-0.0004 (0.0003; 2)	-0.0006 (0.0005; 2)	-0.0001 (0.0000; 2)	-0.0002 (0.0001; 2)	-0.0001 (0.0000; 2)
L, GSM8K	6	-0.0056 (0.0035; 4)	-0.0021 (0.0000; 4)	-0.0030 (0.0010; 4)	-0.0023 (0.0002; 4)	-0.0023 (0.0002; 4)	-0.0021 (0.0000; 4)
L, ARC-C	3	-0.0178 (0.0285; 3)	$+0.0049$ (0.0058; 1)	$+0.0016$ (0.0091; 1)	$+0.0007$ (0.0101; 1)	-0.0083 (0.0191; 3)	$+0.0016$ (0.0091; 1)
L, ARC-C	6	-0.0235 (0.0390; 4)	$+0.0024$ (0.0132; 2)	$+0.0080$ (0.0075; 1)	$+0.0064$ (0.0091; 1)	$+0.0064$ (0.0091; 1)	$+0.0080$ (0.0075; 1)
M, GSM8K	3	$+0.0001$ (0.0012; 2)	$+0.0006$ (0.0008; 1)	$+0.0011$ (0.0003; 0)	$+0.0011$ (0.0003; 0)	$+0.0007$ (0.0007; 2)	$+0.0014$ (0.0000; 0)
M, GSM8K	6	-0.0001 (0.0006; 2)	$+0.0004$ (0.0002; 2)	$+0.0002$ (0.0003; 2)	$+0.0003$ (0.0002; 2)	$+0.0005$ (0.0000; 2)	$+0.0005$ (0.0000; 2)
M, ARC-C	3	-0.0035 (0.0114; 3)	-0.0006 (0.0085; 3)	-0.0006 (0.0085; 3)	$+0.0033$ (0.0046; 2)	-0.0053 (0.0132; 4)	-0.0006 (0.0085; 3)
M, ARC-C	6	-0.0096 (0.0152; 4)	-0.0005 (0.0061; 2)	$+0.0022$ (0.0033; 2)	-0.0071 (0.0127; 3)	-0.0044 (0.0100; 3)	$+0.0022$ (0.0033; 2)

F.1 Frozen predictions on disjoint evaluation questions

The registered A12 validation uses test positions 200–399 after the original fixed ARC-Challenge shuffle, disjoint from the original 200 evaluation questions. All eight units are complete: Llama-3.1-8B and Mistral-7B-v0.3, seeds 7, 42, 100 and 123. Checkpoints, calibration data, exemplars, profiles, fp16 precision and batching remain fixed. Every unit contains identity, derivative, 32 single-block and all unique profile measurements at $\alpha \in \{1.1, 1.2, 1.3, 1.4, 1.5\}$; each has 200 questions, with no target truncation and no prefix mismatch.

The three-block predictions are frozen from the original $\alpha = 1.1$ coefficients; the fresh questions are not used to refit these confirmatory predictions. The registration states three criteria: F1 requires the predicted leading term at the four strengths $1.2, 1.3, 1.4, 1.5$, F2 an observed switch within 0.10 of the frozen prediction, and F3 a relative error of at most 30% for mean $|S|$ at all four strengths. Table 36 reports the switch estimates, and Table 41 lists all three decisions. The observed switches are within 0.10 of the frozen predictions in both families, and the predicted mean finite step is within the 30% tolerance at every strength. The predicted leading term is the measured one at $\alpha = 1.2$, 1.4 and 1.5 in both families. At $\alpha = 1.3$, which lies between the observed and the frozen switch, mean $|S|$ minus mean $|T|$ is $+0.0066$ for Llama-3.1-8B and $+0.0020$ for Mistral-7B-v0.3, with 95% question-level intervals $[-0.0035, 0.0102]$ and $[-0.0100, 0.0065]$. Table 41 lists the measured leader at every strength with the F1 decision.

At $\alpha = 1.2$, 1.3, 1.4 and 1.5 the relative finite-step errors, predicted minus measured over measured, are -18.2%, -13.9%, -11.0% and -7.9% for Llama and -27.0%, -23.1%, -20.0% and -17.6% for Mistral. The per-seed observed three-block switches, in seed order 7/42/100/123, are $1.246/1.204/1.206/1.221$ and $1.418/1.240/1.188/1.261$. Resampling the 200 shared questions jointly across seeds with 19,999 draws, 81.7% and 82.9% of all draws have an interior switch within

Table 33: Registered second-order rule A9 on the cross-family profile sets at $\alpha = 1.3$. Cells: seed-mean test joint gain G, with the mean regret in parentheses, of the profile chosen by derivative sums (Der.), by second-order sums $\sum_l Q_l(0.3)$ with the $\alpha = 1.1$ probe (Quad) and by full measured sums (Meas.); seed-mean block-level Spearman correlation of the derivative / second-order prediction with the measured calibration gains at $\alpha = 1.3$; removed fraction $1 - \sum_l |Q_l - A_l| / \sum_l |D_l - A_l|$ of the calibration finite step; and the same two quantities with the $\alpha = 1.05$ probe. L: Llama-3.1-8B; M: Mistral-7B-v0.3.

Units	Der.	Quad	Meas.
$k = 3$			
L, GSM8K	-0.0015 (0.0014)	-0.0001 (0.0000)	-0.0001 (0.0000)
L, ARC-C	-0.0178 (0.0285)	$+0.0016$ (0.0091)	$+0.0016$ (0.0091)
M, GSM8K	$+0.0001$ (0.0012)	$+0.0014$ (0.0000)	$+0.0014$ (0.0000)
M, ARC-C	-0.0035 (0.0114)	-0.0006 (0.0085)	-0.0006 (0.0085)
$k = 6$			
L, GSM8K	-0.0056 (0.0035)	-0.0021 (0.0000)	-0.0021 (0.0000)
L, ARC-C	-0.0235 (0.0390)	$+0.0080$ (0.0075)	$+0.0080$ (0.0075)
M, GSM8K	-0.0001 (0.0006)	$+0.0003$ (0.0002)	$+0.0005$ (0.0000)
M, ARC-C	-0.0096 (0.0152)	$+0.0022$ (0.0033)	$+0.0022$ (0.0033)

	Probe $\alpha = 1.1$		Probe $\alpha = 1.05$
Units	ρ Der. / Quad	Removed	ρ Quad; removed
L, GSM8K	0.79 / 0.98	0.89	0.87; 0.66
L, ARC-C	0.80 / 0.90	0.52	0.64; -0.58
M, GSM8K	0.86 / 0.98	0.89	0.88; 0.64
M, ARC-C	0.90 / 0.93	0.54	0.74; -0.75

Table 34: Registered fp32 control at the anchor, A11, on the ARC-Challenge units of Llama-3.1-8B, L, and Mistral-7B-v0.3, M: first 32 evaluation and 32 calibration questions of each seed, three-block profiles pooled over the four seeds of each family, item-averaged loss. Coefficients at $\alpha = 1.1$ and the switch $\alpha^*_{1.1} = 1 + c_T/c_S$ they predict; mean $|T|$ / $|S|$ / $|I|$ and the leading term at $\alpha = 1.3$ and 1.4; last column: removed fraction of the calibration finite step at $\alpha = 1.3$ by the second-order rule of Table 33 with the $\alpha = 1.1$ probe, seed mean.

Units	Precision	Seeds	c_T	c_S	c_I	$\alpha^*_{1.1}$	$\lvert T\rvert$ / $\lvert S\rvert$ / $\lvert I\rvert$, $\alpha = 1.3$	Leader	$\lvert T\rvert$ / $\lvert S\rvert$ / $\lvert I\rvert$, $\alpha = 1.4$	Leader	Removed
L, ARC-C	fp16	4	0.170	0.280	0.174	1.61	0.0509 / 0.0220 / 0.0077	transfer	0.0678 / 0.0371 / 0.0134	transfer	0.34
L, ARC-C	fp32	4	0.169	0.279	0.099	1.61	0.0508 / 0.0219 / 0.0076	transfer	0.0678 / 0.0371 / 0.0134	transfer	0.71
M, ARC-C	fp16	4	0.149	0.208	0.092	1.72	0.0446 / 0.0142 / 0.0049	transfer	0.0595 / 0.0241 / 0.0076	transfer	-0.07
M, ARC-C	fp32	4	0.149	0.161	0.068	1.92	0.0446 / 0.0138 / 0.0049	transfer	0.0595 / 0.0238 / 0.0075	transfer	0.81

0.10 of the frozen prediction. These are empirical resampling summaries, not success probabilities on an unobserved population.

For six-block profiles, a registered descriptive analysis compares frozen switches 1.2399 and 1.2369 with observed 1.1452 and 1.2187; conditional intervals are $[1.121, 1.246]$ and $[1.169, 1.393]$, with 99.8% and 99.7% of resamples in range. No six-block confirmatory decision is inferred. All per-strength and per-seed quantities are in the supplement. This new-question validation uses fp16 like the original measurements; the fp32 comparison is the 32-question control of Table 34.

G Experimental Protocol and Implementation

Notation. Table 37 collects the symbols used throughout.

Unified protocol summary. Table 38 summarizes the protocol for each reported result. The ranking and ablation experiments use fixed 50/50 calibration/held-out partitions of each prompt pool; the loss chains use 100 held-out GSM8K items on Qwen3-8B and the train/test protocol of Appendix E on Qwen3.5-9B. Every quantity is a teacher-forced loss; Table 1 reports both the matched-prompt ranking fidelity and the calibration-to-held-out transfer of the derivative.

Table 35: Per-seed breakdown of the fp32 control, A11s, registered as descriptive, with the questions, profiles and definitions of Table 34 and each seed's own three-block profiles. Entries are fp16 / fp32. $\Delta c_S = (c_S^{\text{fp16}} - c_S^{\text{fp32}})/c_S^{\text{fp32}}$. $\alpha^*_{1.1}$ is the switch predicted by the seed's $\alpha = 1.1$ coefficients on the 32-question subset, an extrapolation when it exceeds 1.4; "200 q." is the same prediction in fp16 on all 200 evaluation questions of Table 30. Leader at $\alpha = 1.3$ and 1.4: T transfer, S finite step. Removed: fraction of the calibration finite step at $\alpha = 1.3$ removed by the $\alpha = 1.1$ probe of Table 33.

	Seed	c_T	c_S	Δc_S	$\alpha^*_{1.1}$	200 q.	Leader 1.3	Leader 1.4	Removed
L	7	0.170 / 0.170	0.251 / 0.167	+50%	1.68 / 2.02	1.35	T/T	T/T	0.34 / 0.73
L	42	0.239 / 0.240	0.161 / 0.267	−40%	2.48 / 1.90	1.32	T/T	T/T	0.20 / 0.64
L	100	0.175 / 0.173	0.328 / 0.342	−4%	1.53 / 1.51	1.23	T/T	T/T	0.43 / 0.78
L	123	0.094 / 0.095	0.380 / 0.338	+12%	1.25 / 1.28	1.41	S/S	S/S	0.38 / 0.69
M	7	0.131 / 0.131	0.129 / 0.145	−11%	2.02 / 1.91	1.47	T/T	T/T	−0.59 / 0.75
M	42	0.123 / 0.123	0.193 / 0.164	+18%	1.64 / 1.75	1.32	T/T	T/T	−0.16 / 0.78
M	100	0.173 / 0.173	0.336 / 0.199	+69%	1.52 / 1.87	1.19	T/T	T/T	0.13 / 0.88
M	123	0.167 / 0.167	0.172 / 0.136	+27%	1.97 / 2.23	1.37	T/T	T/T	0.35 / 0.81

Table 36: A12, four completed seeds per model, 200 disjoint questions, three-block profiles. Switch intervals condition on a switch within the measured range; the next column gives the unconditional fraction of resamples in that range. F2 and F3 are the registered point-estimate decisions on the switch and on the finite-step magnitude.

Model	Frozen switch	Observed	95% interval	In range	F2	F3
Llama-3.1-8B	1.3165	1.2193	[1.190,1.354]	100.0%	Yes	Yes
Mistral-7B-v0.3	1.3288	1.2702	[1.205,1.464]	96.9%	Yes	Yes

Amplification protocol. The α grid is $\{1.0, 1.05, 1.10, 1.15, 1.20, 1.30\}$. For each layer, the calibration split selects the α that minimizes calibration loss; if no $\alpha > 1$ improves over baseline, $\delta_l = 0$. The composed profile applies all selected α_l^* simultaneously.

Timing. Timings were measured on a single GPU. In the matched-prompt measurements used by Table 1, GILS scoring takes 7.2–18.8 seconds and the single-strength 32/36-layer sweep takes 51–182 seconds. The stricter held-out Qwen runs take up to 22.1 seconds for GILS and 191 seconds for the layer sweep; timings are hardware- and prompt-length-specific. For the Qwen3.5-9B profile choice, with 100 calibration items of the three-shot GSM8K prompt and four seeds, the stored stage times are 149–198 seconds for the calibration derivatives and 755–1,040 seconds for all 32 single-block losses at $\alpha = 1.3$; the cheaper rules of Table 22 scale these by the fraction of blocks or items they measure.

GILS computation. A single forward pass computes all f_l; a single backward pass computes all $\nabla_{\mathbf{o}_l}\mathcal{L}$. The first-order loss-change predictor is $\delta \cdot \nabla\mathcal{L}^\top f_l$. The reported GILS gain uses the negated convention of Eq. 3,

$$\text{GILS}_l = -\delta \cdot \nabla\mathcal{L}^\top f_l,$$

so larger scores indicate larger predicted loss reduction. For Table 1, we use $\delta = 0.3$, that is $\alpha = 1.3$, as the reference amplification strength; because GILS_l is linear in δ, the ranking is identical for every $\delta > 0$ and δ only sets the scale of the predicted gain.

G.1 Records Behind Each Experiment

Table 39 lists, for each experiment, the records that exist and are included in the supplementary archive, with their folder in the last column.

Table 37: Notation; all gains are in nats per target token, and positive is better.

Symbol	Meaning
f_l, $\mathbf{x}_l$	output of block l added to the residual stream; residual stream entering block l
α_l, $\delta_l = \alpha_l - 1$, t	gain on f_l; its deviation from identity; common strength $t = \alpha - 1$ of a profile
D_l, GILS$_l$	first-order predicted gain $-\delta \nabla_{\mathbf{o}_l} \mathcal{L}^\top f_l$; D_l^{cal} on calibration items
A_l	measured single-layer gain $\mathcal{L}_{\text{base}} - \mathcal{L}(\alpha_l)$
S_l	measured ablation sensitivity $\mathcal{L}(\alpha_l{=}0) - \mathcal{L}_{\text{base}}$
σ_l, φ_l, r_l, e_l	block-output norm $\|f_l\|$; cosine of $\nabla_{\mathbf{o}_l}\mathcal{L}$ and f_l; relative update $\|f_l\|/\|\mathbf{x}_l\|$; residual energy $\sum_t \|f_l(t)\|^2$
P, k, G	profile, that is the set of amplified blocks; its size; measured joint gain
T, S, I	transfer $\sum_P D_l - \sum_P D_l^{\text{cal}}$; finite step $\sum_P A_l - \sum_P D_l$; interaction $G - \sum_P A_l$
c_T, c_S, c_I, t^*	$\overline{\lvert T\rvert}/t$, $\overline{\lvert S\rvert}/t^2$, $\overline{\lvert I\rvert}/t^2$; switch point $t^* = c_T/c_S$
ρ, ρ_s	Spearman rank correlation

Table 38: Protocol for each reported result; the selection objective uses calibration data and the metric uses held-out or evaluation items.

Table	Models	Selection	Metric	Notes
1: ranking	3 families	GILS	ρ, NDCG@3	matched; held-out
15: ablation	Qwen3	S, σ, atlas	ρ_s	layer-level, cross-run
25: ablation	Qwen3.5	S, A, D	ρ_s	layer-level, same items
3, 18: chain	Qwen3	fixed profiles	$\Delta\mathcal{L}$ terms	fp32, 100 items
4, 19: chain	Qwen3.5	frozen triples	$\Delta\mathcal{L}$ terms	fp16; test in 4, selection in 19
17: cross-scale	6 models	Loss	Win rate	vs. random

H Derivations and Proofs

H.1 Norm–cosine factorization of the derivative

The first-order predicted gain is $D_l = -\delta \cdot \nabla\mathcal{L}^\top f_l$. Write $\mathbf{g}_l = \nabla_{\mathbf{o}_l}\mathcal{L}$. By the definition of the inner product and of the cosine,

$$\nabla\mathcal{L}^\top f_l = \|\mathbf{g}_l\| \cdot \|f_l\| \cdot \cos(\mathbf{g}_l, f_l) \tag{6}$$

$$= \|\mathbf{g}_l\| \cdot \sigma_l \cdot \varphi_l. \tag{7}$$

Substituting into D_l gives

$$D_l = -\delta \cdot \|\mathbf{g}_l\| \cdot \sigma_l \cdot \varphi_l. \qquad \square$$

H.2 Covariance decomposition of the ablation–amplifiability correlation

This identity splits the Pearson correlation between ablation sensitivity S and amplifiability A over layers into a first-order part and a finite-step part, with $R = A - D$. It holds whenever the standard deviations over layers satisfy $\sigma_S, \sigma_A, \sigma_D, \sigma_R > 0$; if $\sigma_D = 0$ or $\sigma_R = 0$ the corresponding covariance vanishes and that term is omitted, and the correlation is undefined if $\sigma_S = 0$ or $\sigma_A = 0$. It concerns Pearson, not Spearman, correlation.

Step 1: split of the amplifiability. Write the positive-signed amplifiability as

$$A_l = \mathcal{L}_{\text{base}} - \mathcal{L}_{\text{amp}} = D_l + R_l,$$

where $D_l = -\delta \cdot \nabla\mathcal{L}^\top f_l$ is the first-order predicted gain and $R_l = A_l - D_l$ is the finite-intervention residual.

Step 2: linearity of covariance.

$$\text{Cov}(S, A) = \text{Cov}(S, D + R) \tag{8}$$

$$= \text{Cov}(S, D) + \text{Cov}(S, R). \tag{9}$$

Table 39: Records behind each experiment. "Supplement" names the folder of the supplementary archive that holds the records, inside its data folder unless stated; folder names are written as plain words.

Experiment	Used in	Records	Supplement
Matched-prompt sweep, Qwen3-8B / Llama-3.1-8B	Tables 13, 1	per-layer arrays	ranking sweeps
Matched-prompt sweep, Mistral-7B-v0.1	Table 1	per-layer arrays	ranking sweeps
Held-out sweep, Qwen3-8B / Llama-3.1-8B	Table 1	per-layer arrays, both splits	ranking sweeps
Direct S, σ measurement, Qwen3-8B	Table 15	per-layer arrays	ranking sweeps
Cross-scale cells, six checkpoints	Fig. 1, Table 10	per-layer summaries	ranking sweeps
Qwen3-8B loss chain and interactions	Tables 3, 18	profile-level values	CSV files in the data folder
Same-condition chain, Qwen3.5-9B	Tables 4, 19, 22, 40	every per-item unit; pool scores; selection records	same condition
Numerical control, Qwen3.5-9B	Table 24	per-item, per-layer values	numerical control
Cross-family chain, Llama-3.1-8B and Mistral-7B-v0.3	Tables 5, 26, 27, 31, 41	every per-item unit; profiles	cross family
fp32 control at the anchor, A11 and A11s	Tables 34, 35, 41	per-item fp32 losses and derivatives	fp32 anchor control
New strengths on ARC-Challenge, A6	Tables 29, 41	per-item single-block and profile losses; frozen predictions	strength extension
Fresh ARC-Challenge questions, A12	Table 36	per-item losses and derivatives; frozen predictions	fresh question validation
Additional analyses A1, A2, A4, A5, A9 and A10	Tables 24, 22, 28, 23, 32, 33, 30, 42	analysis outputs from the records above	additional analyses

Step 3: normalization. Dividing both sides by $\sigma_S\sigma_A$ and multiplying each term by σ_D/σ_D or σ_R/σ_R,

$$\rho(S,A) = \frac{\mathrm{Cov}(S,D)}{\sigma_S\sigma_A} + \frac{\mathrm{Cov}(S,R)}{\sigma_S\sigma_A} \tag{10}$$

$$= \frac{\sigma_D}{\sigma_A}\cdot\frac{\mathrm{Cov}(S,D)}{\sigma_S\sigma_D} + \frac{\sigma_R}{\sigma_A}\cdot\frac{\mathrm{Cov}(S,R)}{\sigma_S\sigma_R} \tag{11}$$

$$= \frac{\sigma_D}{\sigma_A}\cdot\rho(S,D) + \frac{\sigma_R}{\sigma_A}\cdot\rho(S,R). \qquad \square \tag{12}$$

H.3 Proof of the composition bound

Let $d^{(j)} = (d_1,\dots,d_j,0,\dots,0)$ and $d^{(0)} = 0$.

Step 1: coordinate telescoping of the joint gain. By the fundamental theorem of calculus, adding one coordinate at a time,

$$F(d) = \sum_j \int_0^{d_j} \partial_j F\big(d^{(j-1)} + te_j\big)\,dt.$$

Step 2: subtracting the single-layer gains. Each single-layer gain has the same form with the other coordinates at zero,

$$\sum_j F(d_j e_j) = \sum_j \int_0^{d_j} \partial_j F(te_j)\,dt.$$

Subtracting, the interaction is

$$I(d) = \sum_j \int_0^{d_j} \Big[\partial_j F\big(d^{(j-1)} + te_j\big) - \partial_j F(te_j)\Big]dt.$$

Step 3: telescoping each derivative difference. The difference in brackets moves the base point from te_j to $d^{(j-1)} + te_j$ one coordinate $i < j$ at a time, so it is an integral of $\partial_{ij}F$ along a path inside the same box:

$$\partial_j F\big(d^{(j-1)} + te_j\big) - \partial_j F(te_j) = \sum_{i<j} \int_0^{d_i} \partial_{ij} F\big(d^{(i-1)} + se_i + te_j\big)\,ds.$$

Hence $I(d)$ is a sum over pairs $i < j$ of iterated integrals of $\partial_{ij}F$ over paths inside the box.

Step 4: bounding. Inserting Step 3 into Step 2, taking absolute values and using $|\partial_{ij}F| \le M_{ij}$ on the box,

$$\begin{aligned} |I(d)| &\le \sum_{i<j} \left| \int_0^{d_j} \int_0^{d_i} \partial_{ij} F\big(d^{(i-1)} + s e_i + t e_j\big)\, ds\, dt \right| \\ &\le \sum_{i<j} M_{ij} \left| \int_0^{d_j} \int_0^{d_i} ds\, dt \right| \\ &= \sum_{i<j} M_{ij}\, |d_i d_j|, \end{aligned}$$

which proves Eq. 5. Negative coordinates are covered by reversing the integration limits. Subtracting $g^\top d$ then gives Eq. 4. The bound is a conditional certificate for teacher-forced loss.

Why endpoint pairs are not a certificate. A pairwise finite difference at other coordinates zero is an integrated interaction on one face of the box, not a uniform bound over all faces. For example, let

$$F(d) = d_1 + d_2 + d_3 - 40 d_1 d_2 d_3.$$

Every single-coordinate gain at 0.3 is 0.3, every two-coordinate gain is 0.6, and every pairwise interaction evaluated with the third coordinate zero is zero. The three-coordinate gain is nevertheless

$$F(0.3, 0.3, 0.3) = 0.9 - 40(0.3)^3 = -0.18.$$

This elementary example is a logical caution, not a measured model result. An empirical interaction probe is useful diagnostically; to certify Eq. 5 it would additionally need justified uniform bounds across the region. An endpoint-average interaction cannot be used as M_{ij}.

I REGISTERED ENDPOINTS

Table 40 lists the primary endpoints of the registration of Appendix E as specified before launch, with all their results; every registered statistic, per seed, is in the summaries folder of the supplement. The first-answer rows are associations within this fixed bank and procedure, from which the paper draws no conclusion. Table 41 lists the pre-stated predictions of the cross-family chain of Appendix F, the status of every unit of its design, the strength predictions A4 and A6 registered before they were computed or measured, the second-order rule A9, the descriptive switch-uncertainty analysis A10, the fp32 anchor control A11 and the new-question validation A12, whose eight units are all complete; Table 42 lists the additional registered analyses of the Qwen3.5-9B records. All registered decisions are listed.

Table 40: Registered primary endpoints of the same-condition experiment, whose registration is in the supplement. E2 pools the candidates of the four seeds, E3 is seed-stratified; all units completed in all four seeds. First-answer counts use the registered endpoint, the first complete answer line under greedy decoding with a cap of 1,024.

Endpoint	Registered quantity	Result
E1	Spearman over 32 layers, D_{sel} vs. A_{sel} and D_{cal} vs. A_{sel}, per seed at $\alpha = 1.3$	$0.54/0.69/0.64/0.68$; $0.43/0.64/0.72/0.59$ for seeds 7/42/100/123
E2	All candidate triples on selection items: Spearman$(G, \sum A)$; mean $\lvert I\rvert$/mean $\lvert G\rvert$; fraction with $\operatorname{sign} G \neq \operatorname{sign} \sum A$	423 triples, 99/115/117/92 per seed; pooled: 0.96, per seed 0.83–0.96; 0.23; 0.026
E3	Spearman$(G_{\mathrm{sel}}$, selection first-answer count), within-seed permutation with 10,000 draws, triple bootstrap	423 triples, association within the fixed bank on selection items: 0.48; 0/10,000 exceedances, $p = 1/10{,}001$; $[0.39, 0.55]$
E4	Test-item chain of the frozen derivative triples, then first-answer change vs. identity	$0.0089 \to 0.0072 \to 0.0007 \to 0.0006$; -1.42 pp $[-2.16, -0.70]$
E5	Policies with identity fallback, P13-FB, and at $\alpha = 1.1$ or $k = 1$, P11 and K1: mean paired first-answer change vs. identity, four seeds	P13-FB: derivative -1.23 $[-1.82, -0.64]$, unsigned -1.55 $[-2.24, -0.87]$, energy and last six -0.66 $[-1.23, -0.09]$, ablation -1.35 $[-2.01, -0.68]$, window -0.97 $[-1.55, -0.38]$, random -0.04 $[-0.72, 0.64]$; identity chosen in 1 of 4 seeds, for random in 0. P11 at $\alpha = 1.1$: derivative $+0.17$ $[-0.34, 0.68]$, random -0.06 $[-0.51, 0.40]$, identical with fallback, identity chosen in 0 of 4 seeds. K1: derivative -0.47 $[-1.00, 0.06]$, with fallback 0.00 $[-0.42, 0.44]$, identity chosen in 1 of 4 seeds

Table 41: Pre-stated predictions of the cross-family chain and additional registered analyses. L: Llama-3.1-8B; M: Mistral-7B-v0.3. Decisions use the point estimate as registered and are not significance tests; brackets give the question-level 95% interval. H1–H3 use all three-block profiles; "met" states whether the registered inequality holds.

Registered quantity	L, GSM8K	L, ARC-C	M, GSM8K	M, ARC-C
Units analysed; design: seeds 7, 42, 100, 123	seeds 7, 42, 100, 123	seeds 7, 42, 100, 123	seeds 7, 42, 100, 123	seeds 7, 42, 100, 123
H1: at $\alpha = 1.3$, mean $\lvert S\rvert$ minus the larger of mean $\lvert T\rvert$, mean $\lvert I\rvert$; criterion > 0	0.0142 $[0.0136, 0.0147]$; met	-0.0028 $[-0.0145, 0.0044]$; not met	0.0067 $[0.0064, 0.0070]$; met	-0.0004 $[-0.0110, 0.0029]$; not met
H2: step share at $\alpha = 1.05$ minus at $\alpha = 1.3$; criterion < 0	-0.215 $[-0.271, -0.205]$; met	-0.290 $[-0.327, -0.187]$; met	-0.246 $[-0.323, -0.226]$; met	-0.320 $[-0.338, -0.176]$; met
H3: mean $\lvert I\rvert$/mean $\lvert S\rvert$ at $\alpha = 1.3$; criterion ≤ 0.3	0.204 $[0.198, 0.212]$; met	0.213 $[0.187, 0.314]$; met	0.204 $[0.196, 0.214]$; met	0.170 $[0.162, 0.256]$; met
H4: seed-mean Spearman$(G, \sum A)$ at $\alpha = 1.3$, all profiles; criterion > 0.8	0.960 $[0.950, 0.962]$; met	0.919 $[0.835, 0.951]$; met	0.958 $[0.942, 0.962]$; met	0.963 $[0.917, 0.976]$; met
H5: measured-sum minus derivative-sum choice, G at $\alpha = 1.3$, $k = 3$; criterion > 0	0.0014 $[0.0004, 0.0023]$; met	0.0194 $[-0.0002, 0.0390]$; met	0.0012 $[0.0005, 0.0019]$; met	0.0029 $[-0.0097, 0.0146]$; met
H6: same-item Spearman at $\alpha = 1.05$ minus at $\alpha = 1.3$; criterion ≥ 0	0.191 $[0.132, 0.237]$; met	0.284 $[0.094, 0.310]$; met	0.117 $[0.082, 0.163]$; met	0.150 $[0.041, 0.212]$; met
A4: leader and mean $\lvert S\rvert$ at $\alpha = 1.3$ from coefficients frozen at $\alpha = 1.1$; criterion: leader equal and $\lvert$error$\rvert \leq 0.30$	finite step predicted and measured; $\lvert S\rvert$ error $+0.05$; met	transfer predicted and measured; $\lvert S\rvert$ error $+0.07$; met	finite step predicted and measured; $\lvert S\rvert$ error -0.01; met	transfer predicted and measured; $\lvert S\rvert$ error -0.06; met
A6 HS1: leader at $\alpha = 1.2, 1.4, 1.5$ as predicted from $\alpha = 1.1$	—	transfer, finite step, finite step; met	—	transfer, finite step, finite step; met
A6 HS2: measured crossover within 0.10 of the predicted $1 + t^*$	—	predicted 1.32, measured 1.35; met	—	predicted 1.33, measured 1.31; met
A6 HS3: $\lvert S\rvert$ error ≤ 0.30 at $\alpha = 1.2, 1.4, 1.5$	—	$+0.03, +0.11, +0.14$; met	—	$-0.05, -0.06, -0.05$; met
A9: second-order rule, regret $\leq$ derivative sums for $k = 3$ and 6, and removed fraction ≥ 0.5	regret 0.0000 vs 0.0014, 0.0000 vs 0.0035; removed 0.89; met	regret 0.0091 vs 0.0285, 0.0075 vs 0.0390; removed 0.52; met	regret 0.0000 vs 0.0012, 0.0002 vs 0.0006; removed 0.89; met	regret 0.0085 vs 0.0114, 0.0033 vs 0.0152; removed 0.54; met
A10: switch under question resampling, $k = 3$, descriptive without a decision rule	—	in range 89.8%; within 0.10 of prediction 86.5%	—	in range 94.6%; within 0.10 of prediction 91.6%
A11: fp32 control at the anchor on 32 evaluation and 32 calibration questions per unit; PA: c_S within 25% and switch within 0.05; PB: leader at $\alpha = 1.4$; PC: probe removes ≥ 0.5 in fp32 and within 0.15 of fp16, given as fp32 / fp16, per seed in Table 35	—	PA: 0.5%, 0.002; met. PB: transfer in both; met. PC: 0.71 / 0.34; not met	—	PA: 29.0%, 0.208; not met. PB: transfer in both; met. PC: 0.81 / -0.07; not met
A12 F1: leader at $\alpha = 1.2, 1.3, 1.4, 1.5$ on 200 disjoint questions as predicted from the frozen $\alpha = 1.1$ coefficients	—	predicted transfer, transfer, finite step, finite step; measured transfer, finite step, finite step, finite step; not met	—	predicted transfer, transfer, finite step, finite step; measured transfer, finite step, finite step, finite step; not met
A12 F2: observed switch within 0.10 of the frozen prediction	—	frozen 1.317, observed 1.219; met	—	frozen 1.329, observed 1.270; met
A12 F3: $\lvert S\rvert$ error ≤ 0.30 at $\alpha = 1.2, 1.3, 1.4, 1.5$	—	$-0.18, -0.14, -0.11, -0.08$; met	—	$-0.27, -0.23, -0.20, -0.18$; met

Table 42: Additional registered analyses of the Qwen3.5-9B records, whose registrations are in the supplement.

Registered quantity	Seed 7	Seed 42
A1, 32 items: R_D of the bank derivative; criterion < 0.2	0.038; met	0.012; met
A1: removed fraction of the derivative triple over fp16/fp32 choices	158–169%	88–89%
A1: mean paired no-op $\Delta\mathcal{L}$ [95%]	$+2.6\times10^{-5}$ $[-8.2, 14.7]\times10^{-5}$	-0.1×10^{-5} $[-9.5, 8.9]\times10^{-5}$

A2, full-procedure resampling of the candidate banks with 2,000 resamples: measured-sum choice $G = +0.0024$ $[0.0010, 0.0038]$, regret 0.0013, re-selected triple harmful in 24% of seed–resample pairs, original triple re-selected in 45/34/80/54% for seeds 7/42/100/123; derivative-sum choice $G = -0.0351$ $[-0.0402, -0.0244]$, regret 0.0388, harmful in 99.9%.
A3 on the cross-family coefficients and A7 on composition over three-block profiles are descriptive and appear in Tables 27 and 5; A5 on cheaper choice rules is descriptive and appears in Tables 23 and 32; A8 on stage times is descriptive and appears in Appendix G.