THE INTRUDER THRESHOLD: A SPECTRAL LAW FOR LORA FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

LoRA fine-tuning creates intruder dimensions, new leading singular vectors of the updated weight matrix $W + BA$ that are nearly orthogonal to every pretrained singular vector and that drive catastrophic forgetting, and no theory has predicted, layer by layer on measured spectra, when they appear. We derive a per-layer critical update strength $s^* = \bar{\theta}/(\gamma\sigma_1(BA))$ from the measured spectrum of W alone, through the rectangular spiked-deformation transform together with an exact secular-equation characterization of the updated spectrum, with no fitted parameters. In a pre-specified study of 18 adapters across seven base models, including a state-space model, a mixture of experts, and an encoder-decoder (9,840 layer scans), the law localizes the empirical threshold within a factor of two on 82% of layers, and a parameter-free combination of its two pre-specified edge evaluations reaches 98%, confirmed out of sample at 0.997 on six third-party adapters. The law separates intruder-bearing from intruder-free layers at deployment with a mean AUC of 0.89 and places the WikiText-2 forgetting knee at Spearman 0.84 over 16 adapters. Full fine-tuning disperses its update below the threshold of every layer, by a median of 100–$200\times$, which accounts for the asymmetry between LoRA and full fine-tuning. Norm-matched interventions show that threshold-crossing layers carry 1.5 to $270\times$ the forgetting of subcritical layers at equal injected norm, the larger figure where nearly every layer is supercritical, and a spike-budget rule derived from the thresholds, set from one SVD and one calibration run in place of a validation sweep, reduces forgetting by 66% on the most fragile model at no task cost and, applied stage by stage in a three-task continual-learning benchmark, removes over 70% of the accumulated WikiText-2 forgetting at higher task accuracy.

1 INTRODUCTION

Low-rank adaptation (LoRA) (Hu et al., 2022) updates a pretrained weight matrix $W \in \mathbb{R}^{n\times m}$ as $W + \gamma BA$ with $B \in \mathbb{R}^{n\times r}$, $A \in \mathbb{R}^{r\times m}$, $r \ll \min(n, m)$. Shuttleworth et al. (2025) showed that LoRA-tuned matrices contain *intruder dimensions*: top singular vectors y_j of the tuned matrix with $\max_i |\langle y_j, u_i\rangle| < \tau$ against *all* pretrained singular vectors u_i. Full fine-tuning produces none, and causal intervention on intruder singular values restores pretraining behavior, so intruders cause forgetting. The phenomenon is widely replicated, but its existing accounts are qualitative (a toy sketch in their Appendix A), live in representation space (Koubbi et al., 2026), or are span-containment arguments that cannot distinguish LoRA from full fine-tuning (Yao et al., 2026).

We ask at what update strength the intruder appears, and we answer with a per-layer critical strength computed from the pretrained spectrum alone, with no training curves and no fitted parameters. We test the answer on real LLMs of 1.7B to 16B parameters with a falsification protocol whose gates and bars were fixed in code before any scan ran. Our contributions are the following.

- *An exact instrument and a law with no fitted parameters* (§2): a core matrix no larger than W returns the singular system of $W + s\gamma BA$ and its overlaps with the pretrained basis exactly (Zha and Simon, 1999; Brand, 2006), and the spiked-deformation transform on the measured spectrum gives $s^* = \bar{\theta}/(\gamma\sigma_1(BA))$.
- *A pre-specified validation* (§3): across 18 adapters and seven base models the law localizes the threshold within a factor of two on 82% of layers, 98% with a parameter-free combination

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.

of its two edges that transfers to third-party adapters, it places the forgetting knee, and it holds for AdaLoRA and MiLoRA and under continual learning.

- *Prediction converted to control* (§4): predicted supercritical layers carry 1.5 to 270× the forgetting of norm-matched subcritical ones, and a spike budget set from one SVD and one calibration run cuts forgetting by 66% on the most fragile model at no task cost and removes over 70% of WikiText-2 forgetting under sequential fine-tuning.
- *The residual explained* (§5): real adapters produce intruders 1.2–1.5× earlier than predicted, an anti-alignment bias that two corrections computed from the trained update account for, and full fine-tuning stays a median 100–200× below threshold, which is why it creates almost no intruders.

2 THEORY: AN EXACTNESS LADDER

We characterize the spectrum of $W + s\gamma BA$ exactly and then take a controlled limit on the measured spectrum of each layer. The theory forms a ladder: the exact rung characterizes the updated spectrum with no assumptions and serves as the instrument; the universal rung assumes only that the update sits in generic position relative to the pretrained basis and yields a prediction with no fitted parameters, the hypothesis under test; the intermediate rung replaces parts of that assumption with measurements.

Let $W = U\Sigma V^\top$ be the reduced SVD of a pretrained weight matrix, with $p = \min(n, m)$ singular values $\sigma_1 \geq \cdots \geq \sigma_p$ and aspect ratio $c = p/\max(n, m)$, and let the deployed update be $\Delta(s) = s\gamma BA$ with LoRA scaling $\gamma = \alpha/r$ (or $\alpha/\sqrt{r}$ for rsLoRA (Kalajdzievski, 2023)).

2.1 THE EXACT RUNG: CORE MATRIX AND SECULAR EQUATION

Write $P_B = U^\top B$ and $P_A = V^\top A^\top$, and let QR be the QR factorization of the component of the update outside the span of the incomplete side of the basis.

Proposition 1. *The nonzero singular values of $W + s\gamma BA$ equal those of the $(p{+}r_b) \times (p{+}r_a)$ core matrix $C(s) = \begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix} + s\gamma \begin{bmatrix}P_B\\ R_B\end{bmatrix}\begin{bmatrix}P_A\\ R_A\end{bmatrix}^\top$, where $r_b = r$ if $n > m$ and 0 otherwise, and $r_a = r$ if $m > n$ and 0 otherwise; for square W both vanish and $C(s) = \Sigma + s\gamma P_B P_A^\top$. Moreover, if $C = U_c\Sigma_c V_c^\top$, then for $i \leq p$ the entry (i, j) of U_c (resp. V_c) is exactly the overlap of the j-th new left (right) singular vector with the pretrained u_i (v_i).*

The proof is an orthonormal completion argument (Appendix B). The entire intruder diagnostic (top singular values, the max-cosine criterion against all pretrained directions, and the subspace overlaps $m_K(s)$) therefore comes exactly from one SVD of the core, a matrix no larger than W, per (layer, s) pair.

Embed W in the symmetric matrix $H_0 = \left(\begin{smallmatrix}0 & W\\ W^\top & 0\end{smallmatrix}\right)$, whose eigenpairs are $\pm\sigma_i$ with eigenvectors $(u_i, \pm v_i)/\sqrt{2}$, together with $|n-m|$ zero modes. A rank-r update of W becomes a symmetric update of rank at most $2r$, and a standard determinant identity states that a value $\lambda \notin \text{spec}(H_0)$ belongs to the updated spectrum exactly when $\det\left(J - K^\top(\lambda - H_0)^{-1}K\right) = 0$, where K collects the update factors and J swaps its two blocks (Bunch et al., 1978). Expanding the resolvent in the eigenbasis of H_0 reduces every entry of this determinant to the weights $\lambda/(\lambda^2 - \sigma_i^2)$, $\sigma_i/(\lambda^2 - \sigma_i^2)$, and $1/\lambda$. For a rank-one update $\theta uv^\top$ the determinant is two by two and the condition reads

$$\left(X(\lambda) - \frac{1}{\theta}\right)^2 = \Phi_u(\lambda)\,\Phi_v(\lambda), \tag{1}$$

where

$$\begin{aligned}\Phi_u(\lambda) &= \sum_i \langle u, u_i\rangle^2 \frac{\lambda}{\lambda^2 - \sigma_i^2}, \qquad \Phi_v(\lambda) = \sum_i \langle v, v_i\rangle^2 \frac{\lambda}{\lambda^2 - \sigma_i^2} + \frac{\|v_\perp\|^2}{\lambda},\\ X(\lambda) &= \sum_i \langle u, u_i\rangle\langle v, v_i\rangle \frac{\sigma_i}{\lambda^2 - \sigma_i^2},\end{aligned} \tag{2}$$

and $v_\perp$ is the component of v in the null space of W (convention $n \leq m$, transpose otherwise, so the null-space term appears on the right only). Equation (1) is exact: given the overlaps of (u, v)

with the pretrained basis it determines every outlier singular value of $W + \theta uv^\top$; the later rungs are controlled limits of it, the outliers of the fp32 scans lie within a relative 1.1×10^{-3} of its float64 roots in all but 5 of 5,133 checked cases (Appendix A), and the bias dissection of §5 operates on its terms.

2.2 THE UNIVERSAL RUNG: THE FREE-POSITION LIMIT

Suppose the spike directions are generic: for a uniformly random unit vector u the weighted sums of the squared overlaps $\langle u, u_i\rangle^2$ (mean $1/n$ each) concentrate at their means, likewise for v, and the cross products carry independent signs and sum to about zero, so the sums in Eq. (2) lose all dependence on the directions, $\Phi_u \to \varphi$, $\Phi_v \to c\varphi + (1-c)/\lambda$, and $X \to 0$, with

$$\varphi(z) = \frac{1}{p}\sum_i \frac{z}{z^2 - \sigma_i^2}, \qquad D(z) = \varphi(z)\Big[c\,\varphi(z) + \frac{1-c}{z}\Big]. \tag{3}$$

Equation (1) then collapses to $D(\lambda) = 1/\theta^2$, which is the rectangular spiked-deformation condition of Benaych-Georges and Nadakuditi (2012), evaluated here on the empirical measure of the actual layer rather than on a limiting distribution. Since D decreases on (σ_1, ∞) to zero, an outlier beyond an evaluation point $z_0 > \sigma_1$ exists exactly when $\theta > D(z_0)^{-1/2}$, the classical sharp transition on a limiting measure and, on a discrete spectrum, escape from the edge by the margin $z_0 - \sigma_1$ (§2.3). For a LoRA update with leading spike $s\gamma\sigma_1(BA)$ this gives a per-layer critical scale:

$$\boxed{s^* = \bar\theta \,/\, \big(\gamma\,\sigma_1(BA)\big), \qquad \bar\theta = D(z_0)^{-1/2}} \tag{4}$$

where z_0 is an evaluation point just above the spectrum, discussed next. Above threshold the outlier location solves $D(\sigma_{\mathrm{out}}(s)) = (s\gamma\sigma_1(BA))^{-2}$, and the limiting overlap of the new singular vector with the spike direction is $-2\varphi(\sigma_{\mathrm{out}})/(\theta^2 D'(\sigma_{\mathrm{out}}))$ on the smaller side. Every quantity derives from one SVD of W.

2.3 THE EVALUATION POINT AND ITS SENSITIVITY

The empirical measure is discrete, so φ has a pole at every σ_i: the $i = 1$ term behaves as $\sigma_1/\big((z-\sigma_1)(z+\sigma_1)\big)$ and diverges as $z \to \sigma_1^+$. Evaluating D at the top of the spectrum would return $\bar\theta = 0$ for any model, so the evaluation point sits slightly outside, $z_0 = \sigma_1(1+\varepsilon)$, where $\varepsilon > 0$ is a choice rather than a derived quantity; all reported predictions use $\varepsilon = 10^{-3}$, fixed in the scan code before any adapter was scanned. The prediction is meaningful only if it does not depend materially on this choice, so we treat $\bar\theta$ as a function of ε and report, for every layer, the sensitivity ratio

$$R = \max_{\varepsilon\in E} \bar\theta(\varepsilon) \Big/ \min_{\varepsilon\in E} \bar\theta(\varepsilon), \qquad E = [10^{-4},\, 3\times 10^{-2}]. \tag{5}$$

A small R means that $\bar\theta(\varepsilon)$ has a plateau and the threshold is a collective property of the whole spectrum; a large R means that the few singular values closest to the edge dominate: a synthetic heavy-tailed spectrum with a sparse top ($p = 4096$) has a plateau ($R = 1.9$), one with half of its mass within 0.5% of the top has none ($R = 22.7$). Gate G2 declares the prediction degenerate when the median R across layers exceeds 10; the observed medians are 2.2–2.4 on the five dense and state-space base models and 3.4–4.3 on the encoder-decoder and mixture-of-experts models. A second choice is which edge to stand above: the full edge uses σ_1 itself, while the bulk edge first trims isolated top outliers by a gap test (the rule is stated in Appendix A). Both are pre-specified and stored per layer; the primary results use the full edge, and §3 analyzes the difference. The law therefore has no fitted parameters rather than no free parameters: ε and the edge are declared choices, constrained by the sensitivity analysis and never tuned to data.

2.4 THE INTERMEDIATE RUNG: ALIGNMENT CORRECTION

Replacing $\sigma_1(BA)$ by $\sigma_1(P_K^\perp\, BA\, P_K^\perp)$, the spike strength in the orthogonal complement of the top-K pretrained subspace, gives an alignment-corrected prediction ($K \in \{16, 64, 256\}$ reported). Where a real adapter falls on this ladder is itself a measurement of how far trained updates are from free position.

3 PRE-SPECIFIED VALIDATION

Protocol. All gates and bars were fixed in the analysis code before any scan ran (pre-specified in this internal sense; no external registry was used). *Census*: full-layer SVD census of Qwen2.5-7B, Llama-3.1-8B, Mistral-7B-v0.3, and OLMo-2-7B; gate G2 terminates the study if $\bar{\theta}$ is degenerate. *Falsifier*: 8 LoRA adapters (2 models $\times$ {MetaMathQA (Yu et al., 2024), alpaca-cleaned (Taori et al., 2023)} $\times$ $r \in \{16, 256\}$, $\alpha = 2r$). For each adapted matrix, called a layer throughout, we scan $W + s\gamma BA$ over two decades of s and record the exact spectrum and overlap diagnostics of Proposition 1; the empirical s^*_{emp} is the first grid point with an intruder under the original criterion ($\tau = 0.5$, top-10; Appendix A). Bars: the universal law passes if $\geq 70\%$ of evaluable layers satisfy $|\log_2(s^*_{\text{pred}}/s^*_{\text{emp}})| \leq 1$ and Spearman $\rho \geq 0.6$. Throughout, hit@$\times 2$ is the fraction of evaluable layers whose prediction lands within a factor of two of the measurement in this sense, ρ the Spearman correlation between predicted and empirical thresholds across an adapter's layers, pooled values average per-adapter results, and every scanned layer is evaluable. The same protocol was then applied unchanged to adapters on Mistral-7B-v0.3, OLMo-2-7B, Mamba-2.8b, OLMoE-1B-7B, and FLAN-T5-XL (math task, $r \in \{16, 256\}$), for 18 adapters in total.

Census: gate G2 passed. On all seven base models, 100% of layers have a finite $\bar{\theta}$, the median sensitivity ratio stays far below the degeneracy bound of 10 (§2.3), and the threshold varies across layers (coefficient of variation 0.44–0.91 on the four dense models): the per-layer threshold is well defined and discriminative on real, heavy-tailed spectra (Martin and Mahoney, 2021).

Falsifier: the law holds. Averaged over the 18 adapters (9,840 layer scans), hit@$\times 2$ is 82.2% with a bootstrap 95% confidence interval of $[0.785, 0.862]$ over adapters, entirely above the pre-specified bar, and $\rho = 0.661$ with interval $[0.583, 0.736]$, whose point estimate clears the bar while the lower bound sits marginally below it. The transition of every layer was captured within the scan grid, so censoring cannot manufacture hits (Appendices A and I; per-adapter results in Table 1, per-layer examples in Figure 1). At s=1, the fraction of layers carrying intruders in the r=256 adapters spans 0% (OLMo-2), 8–17% (Qwen, Llama), 45% (Mamba), and 91% (Mistral), an order-of-magnitude spread that mirrors the per-layer thresholds and reproduces Shuttleworth et al. (2025) qualitatively.

What the law adds over references. The law supplies an absolute scale available before any scan, a per-layer ranking ($\rho = 0.66$ under the full edge, 0.90 out of sample under the geometric-mean edge), and layer-level classification of deployment risk, separating intruder-bearing from intruder-free layers at s=1 with a mean ROC-AUC of 0.89 across the 14 adapters that contain both classes. Of the two references in Table 1, an oracle constant that assigns every layer its adapter's median empirical threshold reaches hit@$\times 2 = 0.83$, since thresholds cluster within an adapter, but it requires the scans it claims to predict, carries no ranking, and scores AUC 0.5; the toy condition of Shuttleworth et al. (2025), $\sigma_{\max}$ in place of $\bar{\theta}$, reaches 0.71 with $\rho = 0.72$, a ranking it shares with the law through the common factor $1/(\gamma\sigma_1(BA))$; Appendix C separates that factor from the spectral part.

Edge variants: a parameter-free combination is nearly unbiased. The geometric mean of the two pre-specified edges reaches mean hit@$\times 2 = 0.98$ (10 of 18 adapters at 1.00), mean $\rho = 0.84$, and median emp/pred ratio 1.07. The full edge (primary, all results above) is systematically late on module types whose spectra carry large isolated outliers, worst for `o_proj` in OLMo-2 (Appendix I), which depresses the pooled ρ of the low-ρ families by misordering module types (and, within `o_proj` of OLMo-2, layers of one type). The bulk-edge variant repairs the ordering on every family ($\rho = 0.72$–0.96, mean 0.90; OLMo-2 $r16$ $0.33 \rightarrow 0.82$) at the cost of an early prediction (median ratio 1.5). The two variants bracket the effective edge, and their geometric mean adds no parameter. Chosen on these 18 adapters and pre-specified as a secondary evaluation of the second batch, it reaches 0.995 on its 53 trained adapters (Appendix K) and out of sample 0.997 with mean $\rho = 0.90$ on six third-party adapters from the HuggingFace Hub, on which the unmodified full edge performs as on our own (Mann–Whitney $p = 0.82$; Appendix I).

The forgetting knee is where the law says it is. Rebuilding $W + s\gamma BA$ in the full model, the knee (the first strength at which WikiText-2 perplexity exceeds $1.5\times$ its base value) falls inside the probed window of 0.5–$3\times$ the median empirical threshold on all 16 decoder-only adapters, which span a $7.8\times$ range of thresholds, and correlates at Spearman 0.84 with the median predicted threshold,

Adapter	hit@×2	Spearman ρ	med. $\lvert \log_2 \rvert$ err
Llama-3.1-8B instr $r16$ / $r256$	0.82 / 0.79	0.76 / 0.80	0.55
Llama-3.1-8B math $r16$ / $r256$	0.80 / 0.79	0.82 / 0.83	0.55
Qwen2.5-7B instr $r16$ / $r256$	0.82 / 0.83	0.50 / 0.66	0.28
Qwen2.5-7B math $r16$ / $r256$	0.80 / 0.82	0.45 / 0.58	0.28
Mistral-7B math $r16$ / $r256$	0.81 / 0.82	0.69 / 0.55	0.55
OLMo-2-7B math $r16$ / $r256$	0.71 / 0.70	0.33 / 0.39	0.55
Mamba-2.8b math $r16$ / $r256$	0.92 / 0.94	0.49 / 0.76	0.55
OLMoE-1B-7B math $r16$ / $r256$	**0.99** / **0.98**	**0.84** / **0.86**	0.28
FLAN-T5-XL math $r16$ / $r256$	0.74 / 0.72	0.82 / 0.77	0.55 / 0.28
pooled, free-position law (18 adapters, 9,840 scans)	**0.822**	**0.661**	—
pooled, geometric mean of both edges	**0.98**	**0.84**	—
reference: oracle constant (needs the scans it predicts)	0.83	—	—
reference: toy $\sigma_{\max}$ condition (Shuttleworth et al., 2025)	0.71	0.72	—

Table 1: The free-position law meets the hit@×2 bar (0.70) on every family and both bars (0.70 / 0.60) when pooled; the low per-family ρ of OLMo-2 and the weaker Qwen adapters is an edge effect that the bulk edge repairs. The oracle constant needs the scans it predicts; the toy $\sigma_{\max}$ condition trails on localization and ranks slightly better (Appendix C); median $\lvert \log_2 \rvert$ errors are quantized by the scan grid (0.28 per step).

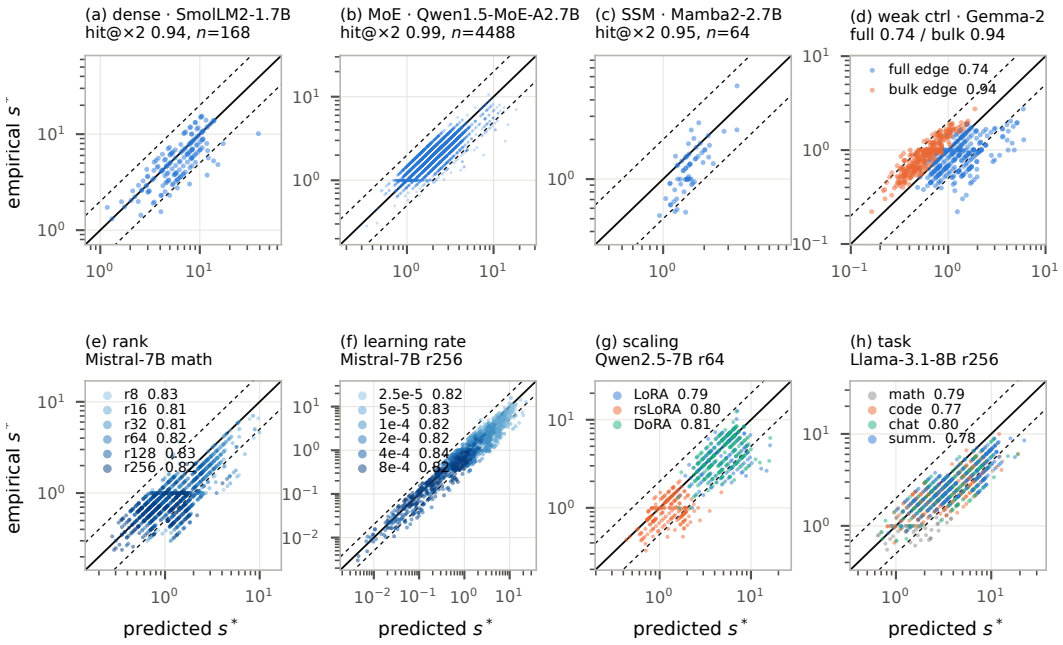

Figure 1: Predicted versus empirical per-layer threshold (log-log; solid, exact; dashed, factor of two). Top: the second-batch r=256 adapter with the highest hit@×2 of each architecture class, and the pre-specified weak control under both edges. Bottom: the rank, learning-rate, parameterization and task series.

which does not enter the probe grid (0.31 with the update magnitude alone; Figure 2, Appendix H). The knee sits at 0.8–2.3× the median threshold: intruders appear first, measurable retention loss follows within a factor of about two in strength, and task perplexity degrades only well above the knee; in the second batch the range widens to 0.3–3× (Appendix H).

The law holds on a state-space model, a mixture of experts, and an encoder-decoder. The law never references attention. On Mamba-2.8b, a selective state-space model, intruder dimensions appear in 45% of layers at deployment (r=256), to our knowledge their first documentation in an

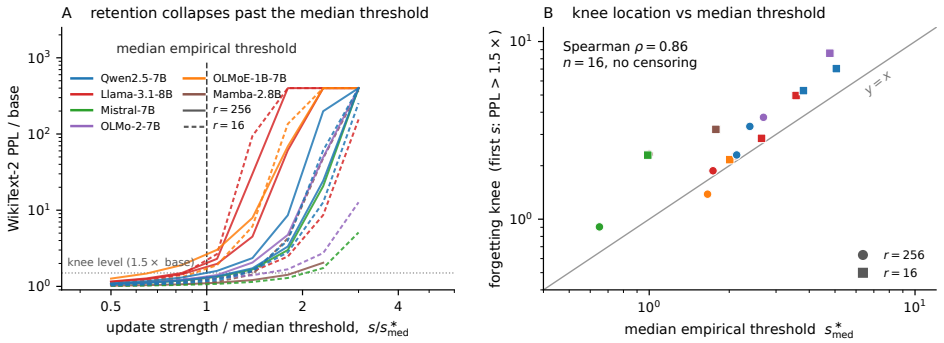

Figure 2: Forgetting begins where the spectrum says it will. Left: WikiText-2 perplexity over base against update strength over the median empirical threshold, 16 adapters. Right: knee (first s with perplexity above $1.5\times$ base) against the median empirical threshold, Spearman 0.86, no censoring.

SSM; the unmodified formula localizes their emergence, and the knee and the anti-alignment bias match the transformer results. On OLMoE-1B-7B, where we adapt attention and all 64 experts per block, the law reaches its best accuracy of the study. On FLAN-T5-XL, an encoder-decoder, the full edge is weakest of the three, consistent with its high edge sensitivity, and the geometric-mean edge restores it (Table 1; Appendix H).

Sensitivity to the intruder criterion, and a failed subclaim. Both pre-specified bars hold for every $\tau \in [0.25, 0.70]$ over all 9,840 scans (Appendix D). The one pre-specified subclaim of this study that failed, a non-vacuous deterministic Wedin rotation control, holds on 0% of the scans because the top spectral gaps of real LLM matrices are too small (Appendix E).

A second pre-specified batch. After the study above we froze a second unit list (59 units: eight further base models from 1.7B to 14B, a rank and learning-rate sweep, rsLoRA and DoRA, three further tasks, seed replicates, six external adapters, two SSMs and two MoEs), each labelled with an expected outcome before any job was submitted (Appendix K). Figure 3 puts every trained family against its label: the batches pool to hit@$\times 2$ of 0.84 (new families), 0.81 (rank, learning rate, parameterization), 0.80 (tasks) and 0.96 (SSM and MoE), all above the bars; the lowest families (OLMo-2, Phi-4, Gemma-2-9B) sit nearest the full-edge bar and are repaired by the bulk edge (Gemma-2-9B $r256$ $0.74 \rightarrow 0.94$), and 12 of its 24 rows meet their label as written, with misses in both directions (Appendix K). Along the knob series of Figure 1, rank and learning rate move the thresholds by more than a decade while localization stays at 0.81–0.84, and four tasks of the same base overlap.

4 From prediction to intervention and control

Predicted supercritical layers carry the forgetting. The law identifies the responsible layers before any intervention is run. Scaling only the predicted supercritical layers of an adapter versus only the subcritical ones, at matched injected Frobenius norm, separates the amount of update from where it lands relative to threshold. On Mistral $r256$ the supercritical-only condition drives WikiText-2 perplexity from 4.9 to $9{,}340$ while the norm-matched subcritical condition reaches only 39, a $270\times$ asymmetry in forgetting at equal injected norm (a $221/3$ layer split, Appendix J); on the balanced splits of Qwen and Llama it is about $1.5\times$, and the differential grows with the supercritical margin, as the law predicts. The ratios are lower bounds, since norm-matching forces the subcritical set past some of its own thresholds; an equal-norm surgery on synthetic updates isolates the sufficient statistic, supercritical strength in generic position, and reproduces the predicted threshold (Appendix F).

Spike-budgeted LoRA: one SVD in place of a validation sweep. Capping the per-layer `lora_alpha` so that the end-of-training spike sits at $\beta\bar{\theta}_\ell$ (the cap rule) removes 66% of the forgetting on the most fragile model, Mistral $r256$, while improving task perplexity, recovering about half of the HellaSwag retention loss, and raising GSM8K accuracy by three points (Table 2; three

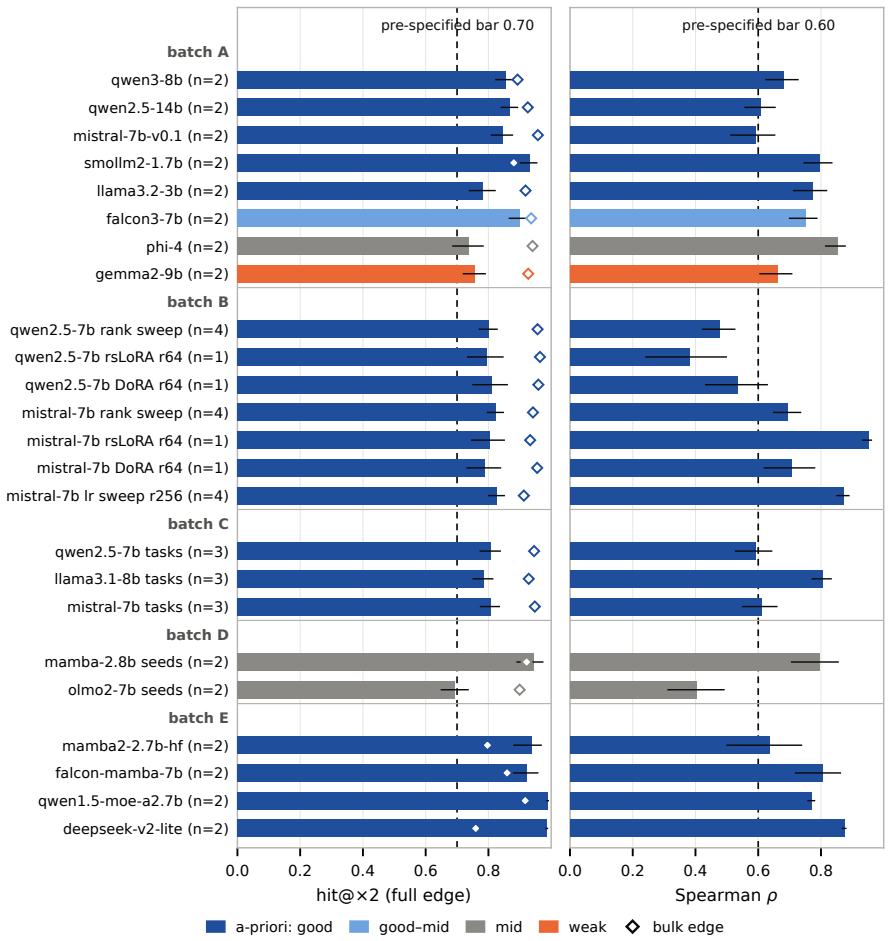

Figure 3: Second batch, pooled per family or configuration: hit@$\times 2$ (bars, full edge; diamonds, bulk edge) and Spearman ρ, coloured by the pre-specified label, with the pre-specified bars (dashed) and 95% intervals.

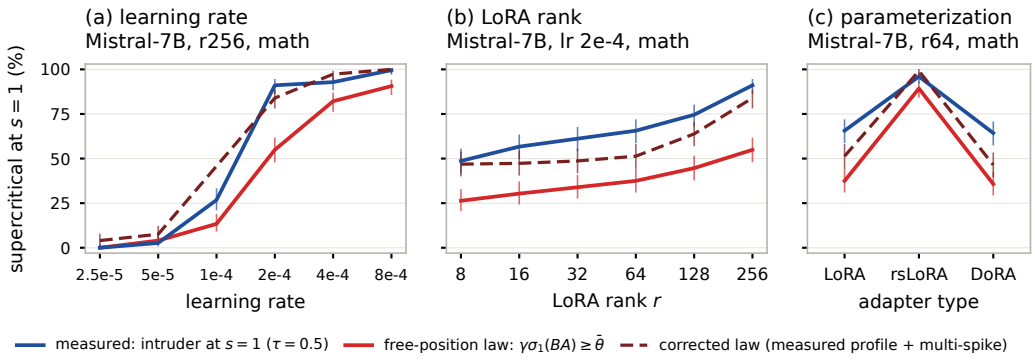

Figure 4: The law reads the knobs. (a–c) Mistral-7B matrices with an intruder at deployment: measured (blue), free-position law (red), law with the corrections of §5 (dark red, dashed).

seeds agree within 0.1 PPL, Appendix M). The rule needs no projections, training-loop changes, or validation sweep: it uses the PEFT `alpha_pattern` interface and one calibration constant, the end-of-training spike scale of the recipe from one prior run, and it never raises a layer toward its

Model	Condition	Forgetting ΔPPL	Task PPL	GSM8K (%)	HellaSwag
Mistral $r256$	base model, no adapter	—	—	—	0.626
	default recipe (baseline)	$+4.71$	1.427	64.6	0.534
	cap rule, β=0.7	$+1.59$	**1.317**	67.6	0.583
	uniform α (matched mean)	$+2.11$	1.327	68.4	0.585
	half learning rate	$+\mathbf{1.12}$	1.318	**68.8**	**0.590**
Qwen $r256$	default recipe (baseline)	$+0.35$	1.206	83.8	0.598
	cap rule, β=0.7	$+0.34$	1.216	82.4	0.598
	half learning rate	$+0.14$	1.206	83.0	0.596

Table 2: Shrinking the end-of-training update below threshold removes most of the forgetting on the fragile model at no task cost, by either knob, and the safe model changes little under any condition. Forgetting is the WikiText-2 perplexity increase over the base model at deployment; GSM8K uses n=500; bold marks the best adapted condition per column in the Mistral block.

threshold. A uniform-α control matched to the mean strength of the uncapped variant of the pattern lands close to that variant ($+2.11$ against $+1.77$ PPL; Appendix J), so the decisive gain is the global scale set by the theory: the spectra determine how strong a fine-tune the model tolerates.

The law reads any knob that shrinks the end-of-training update. Over six learning rates, six ranks, and three parameterizations of Mistral-7B, the fraction of matrices it predicts to be supercritical at deployment tracks the measured fraction, monotone in both knobs (Figure 4); the threshold $\bar{\theta}$ does not depend on the knob, while the deployed spike $\gamma\sigma_1(BA)$ does. The prediction sits below the measurement by the nearly constant margin of §5, whose corrections close about two thirds of the gap (Appendix L). Retraining Mistral at half the learning rate moves the deployment point from 91% to 27% of layers carrying intruders and reduces forgetting at least as much as the cap rule, whereas Qwen, whose deployment point sits below half its median threshold, gains little from either knob (Table 2). Dose responses confirm the mechanism in both directions: a controlled crossing (β=1.3) worsens forgetting by 30%, and raising subcritical layers toward threshold harms an otherwise safe model (Appendix J). The rule gains where the spectra indicate fragility and does no harm where they do not.

Other adapter families, and the rule under continual learning. A third pre-specified batch (Appendix M; three seeds) tests the law and the rule beyond LoRA. On AdaLoRA and MiLoRA the law localizes the threshold on 0.80–0.82 of layers, and on Mistral forgetting follows the intruder count: AdaLoRA keeps the deployed spike $20\times$ below threshold and does not forget (≤ 0.12 PPL), and MiLoRA cuts LoRA's intruders per layer and its forgetting by a third. PiSSA, whose factors start inside the top pretrained subspace, lies outside the generic-position premise: it crosses $2.5\times$ early and, at ten times its recommended learning rate, forgets at least an order of magnitude more than LoRA. On the three-task TRACE sequence (Wang et al., 2023b) the law holds relative to the current weights at every stage (0.81–0.89), intruders accumulate, and the cap rule applied at each stage removes 71–96% of the accumulated WikiText-2 forgetting with overall task accuracy 2.9 points higher, while backward transfer on the earlier tasks is unchanged (Appendix M).

5 The anti-alignment bias

The residual error of the free-position law is a stable early arrival that two corrections computed from the trained update explain. The empirical threshold sits below the prediction by a stable factor on 22 of the 24 adapters, the external ones included and the two OLMoE ones excepted (median early arrival $1.2\times$ on Qwen and Mistral, $1.5\times$ on Llama and OLMo-2, 1.2–$1.5\times$ on Mamba; two-sided sign test $p = 3.6 \times 10^{-5}$). Gradient-trained updates create intruders more readily than a randomly positioned spike of equal strength, an *anti-alignment bias* relative to the free-position null, distinct from the *locking on* to u_1 of Koubbi et al. (2026), which measures alignment of the update itself.

The bias, dissected. The exact secular determinant attributes the deviation across $3{,}992$ scans (the first batch without OLMoE and FLAN-T5, plus the external adapters) through three nested predictors. The free-position law leaves a median emp/pred ratio of 0.80. Substituting the measured energy

profile of the top spike of the update over the pretrained basis closes half the gap (0.90; hit@$\times 2$ $0.80 \to 0.84$). Scoring the eight leading spikes with measured profiles and taking the first crossing, an extreme-value correction, reaches 0.95 and leaves a residual within the scan-grid resolution. The left-right cross-term is refuted as a mechanism (its sign is random, 50.0% positive), so alignment structure and multi-spike order statistics account for the bias.

Why full fine-tuning creates almost no intruders. Full fine-tuning of Qwen2.5-7B and Llama-3.1-8B on the same data and steps spreads the update Δ over several hundred directions, so its largest orthogonal-complement spike sits a median 100–$200\times$ below threshold, and below it on every layer (Appendix I), whereas LoRA concentrates comparable energy into at most r directions and crosses; concentration versus diffusion against the same per-layer threshold accounts for the asymmetry.

6 Related work

Phenomenon. Shuttleworth et al. (2025) discovered and causally grounded intruder dimensions; the toy condition of their Appendix A ($\lambda > \sigma_{\max}$) is not tight, and Eq. (4) gives the critical point. Biderman et al. (2024) established that LoRA forgets less; VeFA (Wang et al., 2025a) and others now treat intruders as standard. **Theories of intruder formation.** Koubbi et al. (2026) prove norm/depth phase transitions in a mean-field attention model (representation dynamics, isotropic adapters, no weight-spectrum statement), and GAIN (Yao et al., 2026) is span containment (cannot separate LoRA from full FT); neither predicts a per-layer threshold. **Spiked-matrix machinery.** BBP (Baik et al., 2005) and BGN (Benaych-Georges and Nadakuditi, 2012) supply the classical tools. BBP-type transitions have been derived for SGD from random initialization (Park et al., 2026), gradient flow in a linear teacher–student model (Coeurdoux et al., 2026), deformed ensembles (Bocchi et al., 2026), rank-one LoRA in a solvable attention model (Duranthon et al., 2026), wide networks without fine-tuning (Lauditi et al., 2026), and one gradient step in two-layer networks (Moniri et al., 2024), on synthetic or limiting spectra. OptShrink (Nadakuditi, 2014) evaluates the D-transform on measured singular values for low-rank denoising; none of these works applies it to a trained update on the spectrum of a pretrained LLM or validates against the intruder criterion. **Hyperparameter rules and forgetting.** OPLoRA (Xiong and Xie, 2026) and rsLoRA (Kalajdzievski, 2023) constrain projections or scaling, PiSSA and MiLoRA (Meng et al., 2024; Wang et al., 2025b) initialize from the largest or smallest singular components, AdaLoRA, LoRA+, and DoRA (Zhang et al., 2023; Hayou et al., 2024; Liu et al., 2024) allocate rank, learning rates, or magnitude during training, and EWC-style regularization, surveys, and scaling laws (Kirkpatrick et al., 2017; Luo et al., 2025; Kalajdzievski, 2024) describe or penalize forgetting; the spike budget is fixed from the pretrained spectrum before training, and the threshold predicts where in the network forgetting originates. **Spectra of trained weights.** Heavy-tailed self-regularization (Martin and Mahoney, 2021; Martin et al., 2021) and the per-layer allocations built on it (Zhou et al., 2023; Qing et al., 2024; Lu et al., 2024) read layer quality from the spectral tail, whereas the threshold is a scale for one update against one layer's edge (further related work in Appendix N).

7 Limitations and future directions

The theory assembles the known results of §2 on measured spectra; on a discrete spectrum the threshold is escape beyond $z_0 = \sigma_1(1+\varepsilon)$ with $\varepsilon = 10^{-3}$ fixed in advance, and deriving the evaluation point from the finite-size edge scale (Bloemendal and Virág, 2013) and a transform for coherent spikes (the PiSSA case) are open. What the transform adds to the location of the spectral edge is the scale: a multiple of the bulk edge fitted on the first batch matches the geometric-mean law out of sample (Appendix C). The evidence consists of weight spectra and behavioural metrics from one recipe and mostly single runs, the knee evidence rests on the absence of censoring and on the predicted threshold, and the $270\times$ asymmetry comes from a $221/3$ layer split. The cap rule needs one calibration run and acts as a global scale that halving the learning rate or post-hoc scaling matches; under continual learning it reduces forgetting of the pretraining distribution while backward transfer is unchanged (Appendix O).

AI USE STATEMENT

We used generative AI throughout this project: Anthropic's Claude models (Claude Opus 5, Claude Opus 5.5, and Claude Fable 5.1), mainly through the Claude Code agent.

Tasks with required disclosure. We used generative AI tools to propose and refine hypotheses, help develop the theoretical model and formulate mathematical claims, design experiments and give feedback on the methodology, assist in proving the mathematical claims of §2 (including drafting the proof in Appendix B), help interpret results, implement, review, and edit our experiment and analysis code, reformat the training data and the logged results, and assist with translation into English. We have not used generative AI tools to generate synthetic data sets, and qualitative or thematic data analysis is not applicable to this work.

Tasks with recommended disclosure. We also used the tool to suggest experimental parameters, write and edit the plotting scripts, draft and polish text, suggest a structure and candidate titles, check references against their bibliographic records, and adapt the manuscript to the ICLR 2027 template.

Review. The authors reviewed all AI-assisted work, including text, mathematics, code, analyses, figures, and references. The mathematical claims were checked numerically in float64: the core-matrix identity of Proposition 1 and the secular equation against direct SVD on random matrices, the predicted outlier location and overlaps against direct SVD on Marchenko–Pastur spectra, and the scan outputs against the secular equation on all 18 adapters (Appendix A). Every reference was checked against its DOI record or its published version. Tables and figures report values computed by analysis scripts from the logged runs, and the numbers in the text are taken from those outputs. Every plot is drawn by a script from logged results; no figure or image was produced by an image-generation model and no plot was retouched. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

§2 states each rung of the theory with its assumptions, and the protocol of §3 specifies the gates, bars, scan grid, and intruder criterion. Appendix A lists the training recipe, evaluation protocols, strength grids, base models, software versions, and compute; Appendix D reports the sensitivity to the intruder criterion, and Appendix K with Table 4 reports every unit of the second pre-specified batch. The supplementary material contains the analysis code, the pre-registration files of the second and third batches (`PREREGISTRATION_ext2026.md`, `PREREGISTRATION_baselines2026.md`), a per-layer summary table for every scanned matrix, the gate, verdict, probe, and bias-anatomy outputs, the per-unit summaries of Appendix M, and a script that rebuilds Table 3 from these files; the raw per-layer scan archive these are distilled from is available on request.

REFERENCES

A. Abro, M. Tahir. Spectral Unforgetting: Post-Hoc Recovery of Damaged Capabilities Without Retraining. arXiv:2605.20296.

J. Baik, G. Ben Arous, S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann. Probab. 33(5), 2005.

F. Benaych-Georges, R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. J. Multivar. Anal. 111, 2012.

D. Biderman et al. LoRA Learns Less and Forgets Less. TMLR 2024.

A. Bloemendal, B. Virág. Limits of spiked random matrices I. Probab. Theory Relat. Fields 156, 795–825, 2013.

D. Bocchi, G. Biroli, C. Cammarota, F. Ricci-Tersenghi. Discontinuous BBP Transitions. arXiv:2604.27992.

M. Brand. Fast low-rank modifications of the thin singular value decomposition. Linear Algebra Appl. 415(1), 2006.

J. R. Bunch, C. P. Nielsen, D. C. Sorensen. Rank-one modification of the symmetric eigenproblem. Numer. Math. 31, 1978.

H. W. Chung et al. Scaling instruction-finetuned language models. JMLR 25(70), 2024.

K. Cobbe et al. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

F. Coeurdoux, G. Ferré, J.-P. Bouchaud. Random Matrix Theory of Early-Stopped Gradient Flow: A Transient BBP Scenario. arXiv:2604.18450.

T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. NeurIPS 2023.

O. Duranthon, F. Boncoraglio, L. Zdeborová. High-Dimensional Theory of LoRA Fine-Tuning in a Solvable Attention Model. arXiv:2606.05899.

A. Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

A. Gu, T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752; COLM 2024.

S. Hayou, N. Ghosh, B. Yu. LoRA+: Efficient low rank adaptation of large models. ICML 2024.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.

G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, A. Farhadi. Editing models with task arithmetic. ICLR 2023.

A. Q. Jiang et al. Mistral 7B. arXiv:2310.06825, 2023.

D. Kalajdzievski. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. arXiv:2312.03732.

D. Kalajdzievski. Scaling laws for forgetting when fine-tuning large language models. arXiv:2401.05605, 2024.

J. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. PNAS 114(13), 2017.

H. Koubbi, L. Hernandez, M. Boussard. Understanding Catastrophic Forgetting in LoRA via Mean-Field Attention Dynamics. arXiv:2402.15415 (v2, retitled; v1: The Impact of LoRA on the Emergence of Clusters in Transformers); ICML 2026.

C. Lauditi, C. Pehlevan, B. Bordelon. Spectral Dynamics in Deep Networks: Feature Learning, Outlier Escape, and Learning Rate Transfer. arXiv:2605.07870.

Y.-S. Liang, W.-J. Li. InfLoRA: Interference-free low-rank adaptation for continual learning. CVPR 2024.

S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen. DoRA: Weight-decomposed low-rank adaptation. ICML 2024.

H. Lu, Y. Zhou, S. Liu, Z. Wang, M. W. Mahoney, Y. Yang. AlphaPruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. NeurIPS 2024.

Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE Trans. Audio Speech Lang. Process. 33, 2025.

S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, M. Tietz. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft, 2022.

C. H. Martin, M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. JMLR 22(165), 2021.

C. H. Martin, T. Peng, M. W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nature Communications 12, 4122, 2021.

F. Meng, Z. Wang, M. Zhang. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. NeurIPS 2024.

S. Merity, C. Xiong, J. Bradbury, R. Socher. Pointer sentinel mixture models. ICLR 2017.

B. Moniri, D. Lee, H. Hassani, E. Dobriban. A Theory of Non-Linear Feature Learning with One Gradient Step in Two-Layer Neural Networks. ICML 2024.

N. Muennighoff et al. OLMoE: Open mixture-of-experts language models. arXiv:2409.02060; ICLR 2025.

R. R. Nadakuditi. OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. IEEE Trans. Inf. Theory 60(5), 3002–3018, 2014.

Team OLMo et al. 2 OLMo 2 Furious. arXiv:2501.00656, 2025.

A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92(4), 1004–1016, 2010.

C. Park, D. Bocchi, F. D'Amico, B. Lucini, G. Aarts. Spectral Phase Transitions and Trainability in Neural Network Learning Dynamics. arXiv:2606.28486.

P. Qing, C. Gao, Y. Zhou, X. Diao, Y. Yang, S. Vosoughi. AlphaLoRA: Assigning LoRA experts based on layer training quality. EMNLP 2024.

R. Shuttleworth, J. Andreas, A. Torralba, P. Sharma. LoRA vs Full Fine-tuning: An Illusion of Equivalence. arXiv:2410.21228; NeurIPS 2025.

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

M. Thamm, M. Staats, B. Rosenow. Random matrix analysis of deep neural network weight matrices. Phys. Rev. E 106, 054124, 2022.

X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, X. Huang. Orthogonal subspace learning for language model continual learning. Findings of EMNLP 2023.

X. Wang, Y. Zhang, T. Chen, S. Gao, S. Jin, X. Yang, Z. Xi, R. Zheng, Y. Zou, T. Gui, Q. Zhang, X. Huang. TRACE: A Comprehensive Benchmark for Continual Learning in Large Language Models. arXiv:2310.06762, 2023.

P. Wang, M. Gu, Q. Huang. VeFA: Vector-Based Feature Space Adaptation for Robust Model Fine-Tuning. arXiv:2510.19155.

H. Wang, Y. Li, S. Wang, G. Chen, Y. Chen. MiLoRA: Harnessing Minor Singular Components for Parameter-Efficient LLM Finetuning. NAACL 2025; arXiv:2406.09044.

P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. BIT 12, 99–111, 1972.

M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. Gontijo Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, L. Schmidt. Robust fine-tuning of zero-shot models. CVPR 2022.

Y. Xiong, X. Xie. OPLoRA: Orthogonal Projection LoRA Prevents Catastrophic Forgetting during Parameter-Efficient Fine-Tuning. AAAI 2026; arXiv:2510.13003.

A. Yang et al. Qwen2.5 technical report. arXiv:2412.15115, 2024.

Y. Yang, X. Li, Z. Zhou, S. L. Song, J. Wu, L. Nie, B. Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. NeurIPS 2024.

H. Yao, X. Chen, A. Murtadha, G. Wang. GAIN: Multiplicative Modulation for Domain Adaptation. arXiv:2604.04516.

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, W. Liu. MetaMath: Bootstrap your own mathematical questions for large language models. ICLR 2024.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi. HellaSwag: Can a machine really finish your sentence? ACL 2019.

H. Zha, H. D. Simon. On updating problems in latent semantic indexing. SIAM J. Sci. Comput. 21(2), 1999.

Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, T. Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. ICLR 2023.

Y. Zhou, T. Pang, K. Liu, C. H. Martin, M. W. Mahoney, Y. Yang. Temperature balancing, layer-wise weight analysis, and neural network training. NeurIPS 2023.

A EXPERIMENTAL DETAILS

Adapters are trained with AdamW at learning rate 2×10^{-4}, a cosine schedule with 3% warmup, batch size 4 with 16-step accumulation, sequence length $1{,}024$, bf16, $1{,}000$ steps, one seed per configuration except the seed replicates of Appendices I and K and the three seeds of Appendix M; target modules are all linear projections (Mamba: `in_proj`; FLAN-T5: attention and feed-forward projections). WikiText-2 (Merity et al., 2017) perplexity uses sliding windows of $1{,}024$ tokens with stride 512, and task perplexity the same windows on the held-out split of the training task; GSM8K (Cobbe et al., 2021) uses zero-shot greedy decoding with 256 new tokens on 500 test problems; HellaSwag (Zellers et al., 2019) uses mean-NLL choice scoring on $2{,}000$ validation items. Knee locations use an eight-point geometric strength grid (step $1.29\times$) spanning 0.5–$3\times$ each adapter's median empirical threshold; the learning-rate controls retrain Mistral and Qwen $r256$ at 10^{-4} with all other settings fixed, and the retrained adapters are scanned with the standard pipeline. The float64 secular-equation check covers all 18 adapters (see below). Base models: Qwen2.5 (Yang et al., 2024a), Llama 3.1 (Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023), OLMo 2 (OLMo Team, 2025), Mamba (Gu and Dao, 2024), OLMoE (Muennighoff et al., 2025), and FLAN-T5 (Chung et al., 2024). Software: PyTorch 2.8, Transformers 4.57, PEFT 0.17. Total compute is roughly 300 H100/H200 GPU hours for the study and the extension campaign and a further 190 for the batch of Appendix M. The census of the four dense models covers 868 block matrices, and gate G2 declares $\bar{\theta}$ degenerate if it is non-finite, has no ε-plateau, or does not discriminate across layers. Each scan uses 25 auto-centered points over two decades of s plus the deployed point $s{=}1$, and stores the max-cosine value of each of the ten leading singular vectors, from which the sensitivity analysis recomputes the intruder count at any τ.

Choices fixed before the scans. The evaluation point is $z_0 = \sigma_1(1+\varepsilon)$ with $\varepsilon = 10^{-3}$; the census stores $\bar{\theta}(\varepsilon)$ for $\varepsilon \in \{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}\}$ per layer, from which the sensitivity ratio of §2.3 is computed. The bulk edge trims the top of the spectrum by a gap test, a simplified eigenvalue-gap criterion in the spirit of Onatski (2010): with g the median gap between consecutive singular values ranked between the top 1% and the top 10%, leading singular values are removed one at a time while the gap to the next exceeds $5g$ (at most 64), and the bulk edge is the first value kept; both edges are stored for every layer. The scan grid is 25 log-spaced points on $[0.1, 10] \times s^*_{\mathrm{pred}}$ (full edge) plus the deployed point $s{=}1$; the prediction sets only the window, and a hit requires the transition inside a factor of two of it, so centering cannot manufacture hits. The empirical threshold is the first grid point at which one of the ten leading singular vectors satisfies $\max_i |\langle y_j, u_i\rangle| < 0.5$ over all pretrained u_i.

Numerical accuracy. At the strongest point of each scan, every eighth adapted matrix (every fourth in a recheck of three of the four Llama adapters) is checked: each fp32 outlier of the core SVD above $\sigma_1(1+10^{-3})$ is refined in float64 to a root of the secular determinant inside a relative bracket of 10^{-3} (5×10^{-3} in the recheck). Over the 18 adapters, $5{,}128$ of the $5{,}133$ checked outliers have a root in the bracket; the fp32 value lies above the root in 99.4% of them, by a median relative 2.3×10^{-4} and at most 1.1×10^{-3}. The other five, in four of the $1{,}316$ checks (three in Mistral $r16$, one in OLMo-2 $r16$), have no root inside the bracket and are flagged as 1.0 in the supplementary per-layer table.

Models and evaluation. All first-batch base models are the pretrained checkpoints (`Qwen/Qwen2.5-7B`, `meta-llama/Llama-3.1-8B`, `mistralai/Mistral-7B-v0.3`, `allenai/OLMo-2-1124-7B`, `state-spaces/mamba-2.8b-hf`, `allenai/OLMoE-1B-7B-0924`, `google/flan-t5-xl`); in the second batch Qwen3-8B and Phi-4 are post-trained checkpoints, kept in the analysis. GSM8K uses the prompt `Question: {q}\nAnswer:`, greedy decoding, and the last number in the output against the gold answer. The full fine-tuning runs of §5 use AdamW at 10^{-5}, batch 2×32, and the same $1{,}000$ steps and data as the LoRA runs, one seed. Third-party adapters were downloaded from the Hub under their listed licenses; only their weights were scanned. Per-adapter binomial intervals treat layers as independent; the pooled intervals of §3 bootstrap over adapters.

B Proof of Proposition 1

Let $n \le m$ (the other case is the transpose), so $U \in \mathbb{R}^{n\times n}$ is complete and $V \in \mathbb{R}^{m\times n}$ has orthonormal columns spanning the row space of W. Decompose $A^\top = VP_A + Q_A R_A$ into its components inside and outside the span of V, with $Q_A \in \mathbb{R}^{m\times r}$ orthonormal, $Q_A^\top V = 0$, and $R_A \in \mathbb{R}^{r\times r}$ (the QR factorization of $(I - VV^\top)A^\top$). Since $B = UP_B$ exactly,

$$W + s\gamma BA = U\Big(\begin{bmatrix}\Sigma & 0\end{bmatrix} + s\gamma\, P_B\begin{bmatrix}P_A^\top & R_A^\top\end{bmatrix}\Big)\begin{bmatrix}V^\top \\ Q_A^\top\end{bmatrix} = U\, C(s)\begin{bmatrix}V & Q_A\end{bmatrix}^\top,$$

with $C(s)$ of size $n \times (n+r)$. Both U and $[V\ Q_A]$ have orthonormal columns, so the singular values of $W + s\gamma BA$ are those of $C(s)$, and if $C = U_c\Sigma_c V_c^\top$ the left and right singular vectors of the updated matrix are UU_c and $[V\ Q_A]V_c$; the overlap of the j-th new left singular vector with u_i is $(U_c)_{ij}$, and with v_i it is $(V_c)_{ij}$ for $i \le n$. For $n = m$, Q_A is empty and $C(s) = \Sigma + s\gamma P_B P_A^\top$. □

C Scale baselines and spike-unit ranking

Every predictor of the form $s^* = \kappa X_\ell/(\gamma\sigma_1(BA))$ shares the factor $1/(\gamma\sigma_1(BA))$ with the law, so part of any per-adapter ρ reflects the update rather than the spectrum. Table 3 compares the law with four such predictors, X_ℓ a constant, $\sigma_{\max}$, the bulk edge σ_{bulk} of the gap test, and their geometric mean, each with κ fitted once on the pooled layers of the first batch (median log ratio) and then applied unchanged to the 53 trained adapters of the second batch and to its six external adapters. Two further columns: the spike-unit ranking, which removes the shared factor, is the Spearman correlation, within an adapter, between X_ℓ and the empirical crossing spike $s^*_{\text{emp}}\gamma\sigma_1(BA)$, and the deployment AUC ranks layers by predicted s^* against the presence of an intruder at $s{=}1$. Three readings follow. The update magnitude alone (constant X) ranks layers at $\rho = 0.54$–0.64 and classifies deployment at AUC 0.71–0.80, so the shared factor carries much of the per-adapter ρ; the spectral part of the full-edge law adds 0.05–0.12 in ρ and 0.08–0.18 in AUC, and in spike units it ranks layers at 0.67 (full edge) and 0.82 (geometric mean). A fitted multiple of the bulk edge ($\kappa = 0.97$) matches the geometric-mean law on every metric out of sample, and a fitted multiple of $\sigma_{\max}$ ($\kappa = 0.65$) exceeds the full-edge law on localization: the transform contributes little spectral-shape information beyond the location of the edge. What it contributes is the scale without a fitted constant (the geometric mean has median ratio 1.07 at $\kappa = 1$, bracketed by the full edge's late prediction, 0.83, the anti-alignment bias of §5, and the bulk edge's early one, 1.5) and the identification of the event, escape of a spike beyond the edge, that the interventions of §4 act on. Under the full edge, 7 of 18 first-batch and 15 of 53 second-batch adapters fall below $\rho = 0.6$ individually; the bars of §3 are pooled bars, and the bulk edge lifts the ρ of every family.

		first batch (18 adapters)				second batch (53, out of sample)				external (6)	
X_ℓ	κ	hit	ρ	spike ρ	AUC	hit	ρ	spike ρ	AUC	hit	spike ρ
law, full edge (κ=1)	1	0.82	0.66	0.67	0.89	0.84	0.69	0.70	0.88	0.81	0.84
law, bulk edge (κ=1)	1	0.89	0.90	0.85	0.96	0.92	0.91	0.85	0.97	0.92	0.92
law, geometric mean (κ=1)	1	0.98	0.84	0.82	0.95	1.00	0.87	0.83	0.95	1.00	0.90
constant θ	1.78	0.39	0.54	0.00	0.71	0.33	0.64	0.00	0.80	0.42	0.00
$\sigma_{\max}$	0.65	0.94	0.72	0.68	0.92	0.96	0.76	0.70	0.90	0.96	0.84
σ_{bulk}	0.97	0.98	0.86	0.84	0.95	0.99	0.87	0.84	0.96	0.99	0.91
$\sqrt{\sigma_{\max}\sigma_{\text{bulk}}}$	0.78	0.98	0.84	0.80	0.95	1.00	0.88	0.81	0.95	1.00	0.89

Table 3: The law against scale baselines $s^* = \kappa X_\ell/(\gamma\sigma_1(BA))$. For the baselines κ is fitted once on the pooled layers of the first batch and applied unchanged to the second batch and the external adapters; the law uses $\kappa = 1$. hit is hit@×2 and ρ the per-adapter Spearman correlation, both averaged over adapters; spike ρ correlates X_ℓ with the empirical crossing spike $s^*_{\text{emp}}\gamma\sigma_1(BA)$, removing the shared factor; AUC classifies layers by predicted s^* against the presence of an intruder at $s{=}1$, averaged over adapters with both classes.

D Sensitivity to the intruder criterion

Both pre-specified bars hold for every $\tau \in [0.25, 0.70]$ (Figure 5, recomputed from the stored max-cosine curves at $\Delta\tau = 0.05$ over all 9,840 scans): localization peaks at 0.88 for $\tau \approx 0.30$ and clears

its 0.70 bar for $\tau \lesssim 0.70$, the per-adapter mean ρ clears its 0.60 bar for $\tau \gtrsim 0.25$, and below $\tau \approx 0.10$ few singular vectors qualify at all.

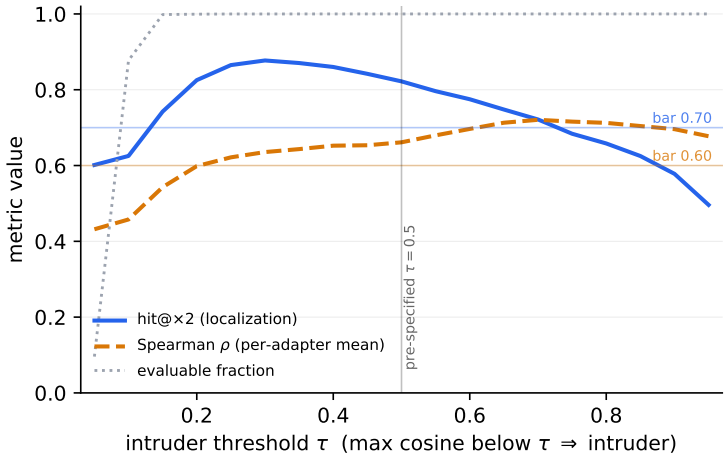

Figure 5: Both pre-specified bars hold for every $\tau \in [0.25, 0.70]$; the pre-specified τ=0.5 sits inside the window (all 9,840 scans).

E Pre-specified negative result: deterministic rotation bounds are vacuous

The plan required $\geq$ 30% of layers to satisfy a non-vacuous deterministic Wedin condition ($\gamma\sigma_1(BA) < \text{gap}_K/2$) at deployed strength for rotation-control claims. The observed fraction is 0% on all 9,840 scans: the top spectral gaps of real LLM matrices are far too small. Classical deterministic perturbation bounds are vacuous in the LoRA regime, and the probabilistic law of Eq. (4) is the operative tool.

F Surgery: the sufficient statistic, isolated

Five synthetic equal-norm updates ($\|\Delta\|_F = 2\bar{\theta}$ per layer, ~24 Qwen matrices) realize the $2 \times 2(+1)$ design. Concentrated $\times$ generic position: intruders on 100% of layers at median $s = 0.500$ against a predicted 0.5. Concentrated $\times$ top-aligned and diffuse $\times$ top-aligned: no intruders at any strength. Concentrated $\times$ bulk-aligned: a new outlier emerges at $s \approx 0.50$ yet is never an intruder, so spectral escape and novelty are different events, and the criterion distinguishes them. Diffuse $\times$ generic (r=16, per-direction $\theta/\sqrt{r}$): intruders arrive 3$\times$ later (the naive independent-spike prediction is 4$\times$; multi-spike order statistics account for the gap). Intruder formation is governed by a single statistic, supercritical strength in generic position, which explains the immunity of full fine-tuning (diffusion) and of aligned LoRA updates.

G What the law is for

Four uses follow, each tied to a result above: a fragility audit before training (the census, with the end-of-training spike scale of the recipe from one prior run, placed 55% of Mistral layers past threshold, where 91% carry intruders, and OLMo-2 entirely below it, and the measured outcomes, 9.2 HellaSwag points lost against none, matched); strength calibration without sweeps (the cap rule removes 66% of the forgetting and reaches an operating point comparable to post-hoc scaling (Appendix J) from one calibration run and no sweep, while rsLoRA at a fixed rank reduces to a global scale that the uniform-α control already covers); blame assignment before intervention (predicted supercritical layers carry 1.5$\times$ to 270$\times$ the damage of norm-matched subcritical ones);

and adapter auditing (minutes per matrix on one GPU, giving the deployed intruder fraction and a median predicted threshold that anticipates the knee at Spearman 0.84).

H ARCHITECTURE AND KNEE DETAILS

On OLMoE-1B-7B the census shows that the 64 experts of a single block are spectrally heterogeneous, with $\bar{\theta}$ spanning a $2.6\times$ range (coefficient of variation 0.21) within one block, so per-expert budgets are meaningful. On FLAN-T5-XL the geometric-mean edge reaches $0.88/0.91$ with $\rho \approx 0.88$ and median ratio 1.08, and the encoder-decoder carries few intruders at deployment (1–4% of matrices). Two representative forgetting curves: Llama math $r256$ (median $s^*_{\text{emp}} = 1.74$) has WikiText-2 PPL $6.3 \to 7.2 \to 13.7 \to 15{,}562$ at $s = 0.5, 1, 2, 4$, whereas Qwen instr $r16$ (median 3.80) stays at $6.6 \to 6.7 \to 6.8 \to 7.8$. The extremes bracket this range: Mistral $r256$ is already past threshold at deployment (median $s^*_{\text{emp}} = 0.65$; 91% of layers carry intruders; PPL $9.7 \to 9{,}092$ between s=1 and s=2), while OLMo-2 $r16$ sits far below it (median 4.78; no deployed intruders; PPL $6.0 \to 7.5$ through s=4).

Knee: censoring, nulls, and the predicted threshold. Knees are probed on the 16 decoder-only adapters (the probe computes causal-LM perplexity) on a geometric grid spanning 0.5–$3\times$ each adapter's median empirical threshold. Because the grid is relative to that threshold, a Spearman correlation between knee and threshold is high under any hypothesis that keeps the knee inside the window (a knee at a random multiple in 0.8–$2.3\times$ of the threshold, unrelated to it otherwise, gives $\rho \approx 0.86$ for these 16 adapters), so ρ alone is not the evidence. Three checks are. All 16 knees fall inside the window, whereas a knee at a fixed absolute strength, drawn from the observed range of knees, would leave 32% of adapters outside it ($p = 0.002$ for $16/16$). The knee correlates with the median *predicted* threshold, a pre-scan quantity that does not define the grid, at $\rho = 0.84$ (knee over predicted median 0.7–1.8, median 1.0), and with the update magnitude alone, $\text{median}\, 1/(\gamma\sigma_1(BA))$, at only 0.31. In the second batch, the 53 trained adapters were probed on the same relative grid extended by the absolute points 0.25–4: none is right-censored, three are left-censored at the grid floor (rsLoRA and the two largest learning rates on Mistral, whose deployment points lie far past threshold), and on the remaining 50 the knee sits at 0.28–$2.97\times$ the median empirical threshold (median 1.38; below $0.8\times$ on nine units: both SmolLM2-1.7B and both DeepSeek-V2-Lite units, the two lowest learning rates, Llama-3.2-3B $r256$, Falcon3-7B $r16$, and Gemma-2-9B $r16$), with $\rho = 0.81$ against the empirical median, 0.79 against the predicted median, and 0.08 against the update magnitude. The first-batch correlation is insensitive to the knee definition (0.86 under a $1.2\times$ and 0.84 under a $2.0\times$ criterion), and the median knee sits at $1.4\times$ the median threshold.

On Mamba-2.8b (LoRA on `in_proj`, $r \in \{16, 256\}$, math) the knee again tracks the median threshold (PPL $9.8 \to 14.2 \to 2{,}478$ for $s = 1, 2, 4$ at median $s^*_{\text{emp}} = 1.00$), the anti-alignment bias has the same sign and magnitude (0.68–$0.83\times$), and the float64 secular check passes ($\leq 6.4 \times 10^{-4}$). On OLMoE-1B-7B we adapt 3,136 matrices per adapter; intruders appear in 9% and 19% of matrices at deployment for $r16$ and $r256$, and the forgetting knee sits at the predicted location.

I ADDITIONAL VALIDATION DETAILS

Grid quantization and uncertainty. Because the empirical threshold is the first grid point at which an intruder is present, grid quantization can only delay it by at most one step (0.28 in $\log_2$), which biases against rather than toward the early-arrival finding of §5. Per-adapter binomial half-widths on hit@$\times$2 are at most 0.066.

Training seeds. Two further seeds on the Mistral and Qwen $r256$ configurations move hit@$\times$2 by at most 0.02 and ρ by at most 0.07; the median per-layer cross-seed range of the empirical threshold equals the scan-grid step (0.26–0.27 against 0.28 in $\log_2$), and the layer ranking agrees between seeds at $\rho = 0.88$–0.93.

Third-party adapters. The six external LoRA adapters were downloaded from the HuggingFace Hub and trained with unknown, heterogeneous protocols (ranks 16–64; two QLoRA (Dettmers et al., 2023), one Bengali instruction adapter, one medical QA adapter). Under the unmodified full edge the

law localizes their thresholds with a per-adapter mean hit@$\times$2 of 0.81 and $\rho = 0.74$ (pooled over $1{,}288$ scans: 0.81/0.84), and a Mann–Whitney test on per-adapter hit rates finds no difference from our own adapters ($p = 0.82$).

Edge effects and full fine-tuning. The full edge errs most on `o_proj` in OLMo-2, where $\sigma_{\max}$ exceeds the bulk edge $3.7\times$, the median emp/pred ratio is 0.32 (a late prediction), and the within-type ρ is negative. Full fine-tuning spreads its update over roughly 385 (Qwen) and 345 (Llama) directions, measured by the participation ratio of the delta spectrum; its largest orthogonal-complement spike sits at a median of $0.005\,\bar{\theta}$ (Qwen) and $0.008\,\bar{\theta}$ (Llama), all 420 layers stay below threshold, and the intruder fractions are 0.5% and 0.0%.

J Intervention details

On Mistral $r256$, 91% of layers carry intruders at deployment, and the norm-matched comparison at s=2 scales 221 supercritical layers, raising WikiText-2 perplexity from 4.9 to $9{,}340$; the matched subcritical condition places the same Frobenius norm on the remaining three matrices, each scaled $21\times$, and reaches 39. Both conditions lie in the collapse regime, so the $270\times$ ratio measures how far past threshold three matrices can be pushed before the model breaks; a matched-layer-count design is future work. On Qwen at s=4 (a $143/53$ split) the matched asymmetry is $1.56\times$ ($+4.20$ versus $+2.69$ PPL), and on Llama at s=2 (an $89/135$ split) it is $1.52\times$ ($+1.71$ versus $+1.12$). The uniform-α control of Table 2 is matched to the mean strength of the uncapped β=0.7 pattern, whose own forgetting is $+1.77$ PPL: per-layer allocation is 16% better in perplexity than that matched uniform scale, a difference the accuracy metrics (GSM8K on n=500) do not resolve. Capping the pattern so that no layer is raised, the rule as stated in §4 and the row of Table 2, improves forgetting further to $+1.59$ PPL. The deployed spike $\gamma\sigma_1(BA)$ ranges from 0.72 to 1.14 of the median threshold across ranks, and rsLoRA at α=$2r$ multiplies it by $\sqrt{r}$. Given the spike scale of the recipe, deployment sits at $1.5\times$ the median empirical threshold for Mistral $r256$ ($1.1\times$ the predicted one) and below half of either for Qwen; halving the learning rate of Mistral $r256$ moves its median threshold from 0.65 to 1.49. The calibration constant is that spike scale, taken from one prior run under the recipe (the baseline row of Table 2); the rule replaces the sweep over α, not the first training run. Post-hoc downscaling of the trained Mistral $r256$ adapter, which needs no retraining, gives $+1.65$ PPL at s=0.70 with task perplexity 1.38 and $+0.93$ at s=0.54 with 1.41 (cap rule: $+1.59$ at 1.32); GSM8K and HellaSwag were not measured for scaled adapters. The rule does not find a better operating point than scaling does; it sets the scale from the spectrum before training. In the dose responses, the pattern at β=1.3, which targets every layer's end-of-training spike at $1.3\,\bar{\theta}_\ell$, raises the Mistral forgetting from $+4.71$ to $+6.14$ PPL (30%), and a target-mode ablation that raises subcritical layers toward threshold worsens the safe Qwen adapter from $+0.35$ to $+0.99$ PPL.

K Extension campaign: every unit

After the pre-specified study of the main text, a second pre-specified batch of 59 units (eight new base models, a rank and learning-rate sweep, rsLoRA and DoRA, three further tasks, seed replicates, six external adapters, and two state-space and two mixture-of-experts models) was run under the recipe of §3; its unit list, order, a-priori labels and bars were frozen before any job was submitted (`PREREGISTRATION_ext2026.md` in the supplementary code). Figure 1 draws representative panels by the rule stated in its caption; Table 4 reports every unit. Units are labelled with their pre-specified expectation (good, good–mid, mid, weak) so that the reader can check each prediction against its outcome.

Labels against outcomes. Each label keeps the operational definition of `PREREGISTRATION_ext2026.md` (good: pooled hit@$\times$2 $\geq$ 0.80 and $\rho \geq 0.60$ under the full edge; good–mid: 0.70 and 0.50; mid: hit@$\times$2 ≥ 0.70; weak: $\rho < 0.50$), and batches B to E add their stated predictions (B: monotone deployment fractions and a deployment AUC of at least 0.8 on every unit with both classes; C: every task within ± 0.10 of the math values of its base model; D: seeds within ± 0.03 in hit@$\times$2 and ± 0.10 in ρ of seed 0, with a cross-seed layer ranking of at least 0.80; E: hit@$\times$2 ≥ 0.90). Scored this way, 12 of the 24 rows of Figure 3 meet their label, and the external adapters, which the figure omits, meet theirs. Ten labels were too optimistic: the

three Qwen2.5-7B rows of batch B keep the low full-edge ρ of that model (0.38–0.53), and so, at $\rho = 0.59$, do its tasks; Mistral-7B-v0.1 reaches $\rho = 0.59$; Llama-3.2-3B, Mistral-7B DoRA and the Llama-3.1-8B tasks reach hit@$\times$2 of 0.78–0.79, and the OLMo-2 seeds 0.69; and the Mistral learning-rate sweep has a deployment AUC of 0.77 on its 8×10^{-4} unit, in which a single layer is free of intruders. Two were too cautious: Gemma-2-9B, labelled weak, reaches $\rho = 0.66$, and Mistral-7B on code exceeds its math ρ by 0.14. Falcon3-7B and Phi-4 meet their labels with ρ above the good threshold. Scoring batches C and D by their batch predictions alone gives 15 of 24.

L THE DEPLOYMENT TRANSITION, POOLED OVER ADAPTERS

Figure 7 plots the event of Figure 4, an intruder at deployment, against the one variable the free-position law uses: the deployed spike over the layer's threshold. The law predicts a step at 1. The measured transition is a logistic in the log ratio with midpoint 0.61 over the 16 Mistral-7B-v0.3 adapters of both studies (the 13 of the second batch and the $r16$, $r256$ and half-learning-rate adapters of the first; 3,584 matrices) and 0.85 over the 53 trained adapters of the second batch (29,170 matrices). The corrections of §5, computed here with 64 spikes over the full pretrained basis (measured energy profile of each spike, first crossing among the spikes), move the Mistral-7B midpoint to 0.81 and the median threshold ratio from 0.83 to 1.05, and they close about two thirds of the gap along the rank series of Figure 4 (49–92% per adapter; free law 26–55%, corrected 47–84%, measured 49–91%) and overshoot at the largest learning rates (12 adapters, 2,688 matrices; the 10^{-4} adapter of the main study has no stored weights). The remainder is not truncation: 64 spikes give the same result as 8, and the full basis moves the fractions by only 1–3 points relative to the census top-512. It is the independent-spike, escape-equals-intruder approximation itself; the exact core matrix of Proposition 1 couples every spike and reproduces the measurement by construction.

M PEFT VARIANTS AND CONTINUAL LEARNING

A third pre-specified batch (unit list, recipe, and eight predictions frozen in `PREREGISTRATION_baselines2026.md` before any job ran; three training seeds throughout) extends the study to three PEFT families and to sequential fine-tuning.

Protocol. PiSSA (Meng et al., 2024), MiLoRA (Wang et al., 2025b), and AdaLoRA (Zhang et al., 2023) are trained on Mistral-7B and Qwen2.5-7B with the recipe of Appendix A and Table 2 (MetaMathQA, r=256, α=512; AdaLoRA with initial rank 384 pruned to a budget of 256 under the PEFT defaults for its schedule and orthogonality weight). Each trained adapter is re-expressed exactly as a standard LoRA update over the *pretrained* weights (PiSSA and MiLoRA: $[B, -B_0]\,(\gamma[A; A_0])$ of rank 512, since their initialization subtracts $\gamma B_0 A_0$ from the base; AdaLoRA: $B(A \odot E)\,\alpha/r$ of rank 384) and verified against the PEFT deployment (maximum relative error below 10^{-6} on the update; for the PiSSA units the perplexities of PEFT's own unmerged forward pass are reproduced to 2×10^{-3}), so that the scan, the WikiText-2 probe, and the GSM8K/HellaSwag evaluation run unchanged and the threshold is evaluated on the same pretrained spectrum as for every other adapter. The cap rule of §4 is rerun with the cap enforced (β=0.7); the Mistral row of Table 2 reports its seed-0 run. TRACE (Wang et al., 2023b) is run from the official release (MeetingBank, Py150, and ScienceQA in that order, 5,000 training examples per task, the official epochs 7/5/3, loss on answer tokens, prompts left-truncated to 1,024 tokens, the official metrics ROUGE-L, edit similarity, and first-token accuracy on the released test sets, greedy decoding) on Mistral-7B with the LoRA configuration above. After each stage the adapter is merged, the model is evaluated on the three tasks and on WikiText-2, HellaSwag, and GSM8K, the stage update is scanned relative to the *current* weights (so $\bar{\theta}$ comes from the current spectrum), and the cumulative update is scanned relative to W_0. Two chains are run: uniform α, and the cap rule applied at every stage with $\bar{\theta}_\ell$ from the chain's current weights and the calibration spike from the uniform chain's adapter of the same stage.

The law holds beyond LoRA. On AdaLoRA and MiLoRA the full edge localizes the threshold on 0.80–0.82 of layers and the bulk edge on 0.92–0.98 (Table 5), with an early arrival of 1.21$\times$ for MiLoRA and the cap rule and of 1.47$\times$ (Mistral) and 1.21$\times$ (Qwen) for AdaLoRA, whose flat update spectrum (σ_{10}/σ_1=0.84, σ_{100}/σ_1=0.62 against 0.46 and 0.18 for LoRA) brings many comparable spikes to the edge at once, the multi-spike order statistic of §5.

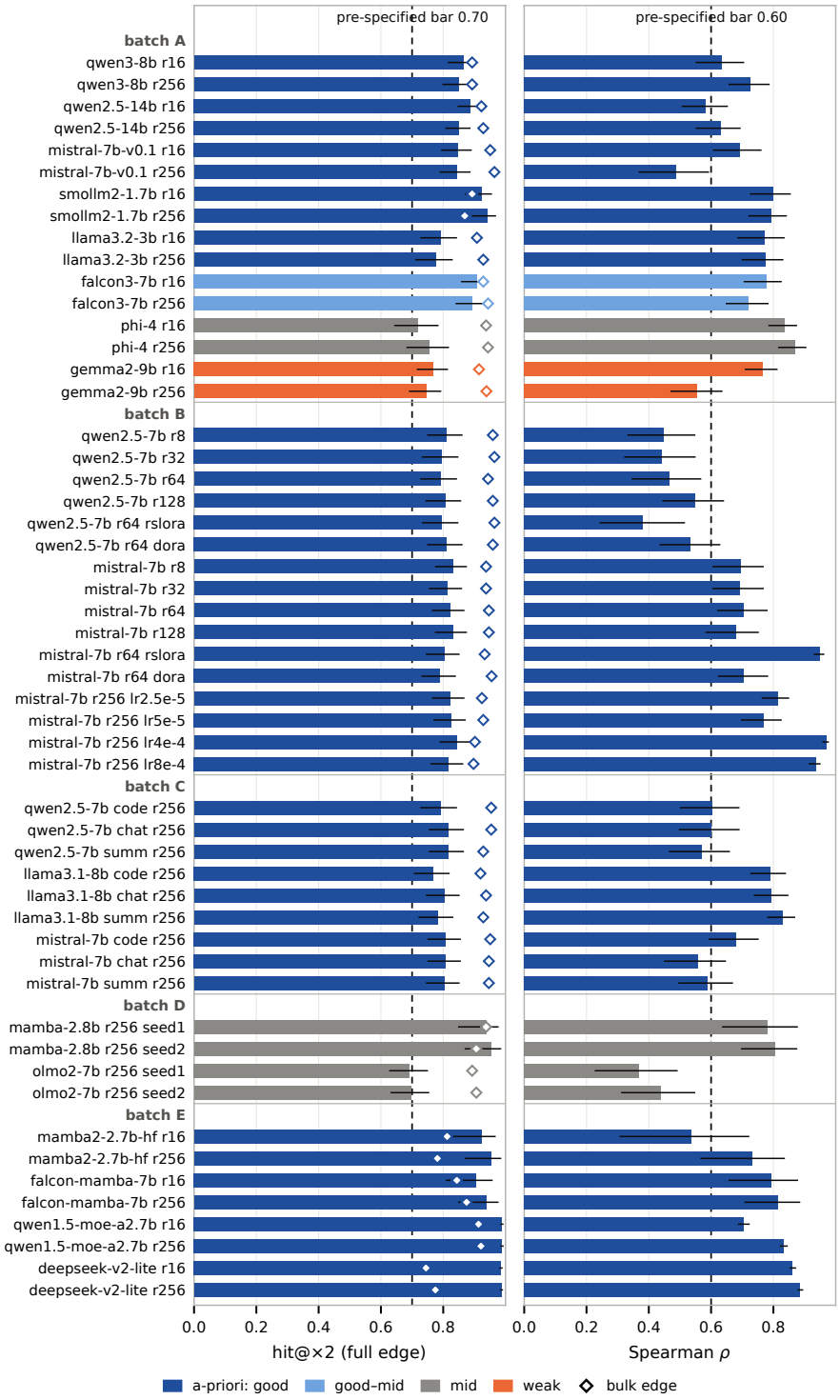

Figure 6: Every scanned unit of the second batch, as Figure 3 without pooling.

B	unit	hit@×2	bulk	geo. mean	ρ	dep. %	knee	knee/s^*_{med}	a-priori	status
A	qwen3-8b_math_r16	0.87	0.89	0.99	0.63	0	6.45	1.38	good	done
A	qwen3-8b_math_r256	0.85	0.89	1.00	0.73	6	3.48	1.38	good	done
A	qwen2.5-14b_math_r16	0.89	0.92	0.99	0.58	0	11.66	1.79	good	done
A	qwen2.5-14b_math_r256	0.85	0.93	1.00	0.63	3	4.00	1.69	good	done
A	mistral-7b-v0.1_math_r16	0.85	0.95	1.00	0.69	56	2.00	2.00	good	done
A	mistral-7b-v0.1_math_r256	0.84	0.96	1.00	0.49	92	0.91	1.38	good	done
B	qwen2.5-7b_math_r8	0.81	0.96	1.00	0.45	1	8.52	1.38	good	done
B	qwen2.5-7b_math_r32	0.80	0.96	1.00	0.44	1	8.29	1.79	good	done
B	qwen2.5-7b_math_r64	0.79	0.94	0.99	0.47	1	6.69	1.79	good	done
B	qwen2.5-7b_math_r128	0.81	0.96	1.00	0.55	1	5.27	1.79	good	done
B	qwen2.5-7b_math_r64_rslora	0.80	0.96	1.00	0.38	66	1.64	1.79	good	done
B	qwen2.5-7b_math_r64_dora	0.81	0.96	0.99	0.53	1	6.83	1.79	good	done
C	qwen2.5-7b_code_r256	0.79	0.95	1.00	0.60	5	3.90	1.38	good	done
C	qwen2.5-7b_chat_r256	0.82	0.95	0.99	0.60	4	3.28	1.07	good	done
C	qwen2.5-7b_summ_r256	0.82	0.93	1.00	0.57	4	2.62	1.07	good	done
D	DIA-MVP__qwen2.5-7b-lora-a100	0.84	0.92	0.99	0.55	1	5.54	1.38	mid	done
D	Niazi__Qwen2.5-7B-qlora-bengali	0.81	0.92	0.99	0.68	2	4.00	1.07	mid	done
D	mamba-2.8b_math_r256_seed1	0.94	0.94	1.00	0.78	45	2.00	2.00	mid	done
D	mamba-2.8b_math_r256_seed2	0.95	0.91	1.00	0.81	44	2.00	1.90	mid	done
A	smollm2-1.7b_math_r16	0.92	0.89	1.00	0.80	0	4.00	0.50	good	done
A	smollm2-1.7b_math_r256	0.94	0.87	0.99	0.79	0	2.67	0.50	good	done
A	llama3.2-3b_math_r16	0.79	0.91	1.00	0.77	1	2.58	0.83	good	done
A	llama3.2-3b_math_r256	0.78	0.93	1.00	0.78	5	1.55	0.64	good	done
A	falcon3-7b_math_r16	0.91	0.93	1.00	0.78	1	4.00	0.77	good-mid	done
A	falcon3-7b_math_r256	0.89	0.94	0.99	0.72	4	3.27	1.07	good-mid	done
A	gemma2-9b_math_r16	0.77	0.91	1.00	0.77	26	0.50	0.28	weak	done
A	gemma2-9b_math_r256	0.74	0.94	1.00	0.55	66	1.56	1.79	weak	done
A	phi-4_math_r16	0.72	0.94	1.00	0.84	0	5.58	1.38	mid	done
A	phi-4_math_r256	0.76	0.94	1.00	0.87	9	3.45	1.79	mid	done
B	mistral-7b_math_r8	0.83	0.94	1.00	0.70	49	2.30	2.30	good	done
B	mistral-7b_math_r32	0.81	0.94	1.00	0.69	61	2.00	2.09	good	done
B	mistral-7b_math_r64	0.82	0.95	1.00	0.71	66	1.61	1.79	good	done
B	mistral-7b_math_r128	0.83	0.95	1.00	0.68	75	1.46	1.79	good	done
B	mistral-7b_math_r64_rslora	0.80	0.93	1.00	0.95	96	0.08	0.50	good	done
B	mistral-7b_math_r64_dora	0.79	0.96	1.00	0.71	64	1.68	1.79	good	done
B	mistral-7b_math_r256_lr2.5e-5	0.82	0.92	1.00	0.81	0	2.73	0.65	good	done
B	mistral-7b_math_r256_lr5e-5	0.83	0.93	1.00	0.77	3	1.82	0.65	good	done
B	mistral-7b_math_r256_lr4e-4	0.84	0.90	1.00	0.97	93	0.08	0.50	good	done
B	mistral-7b_math_r256_lr8e-4	0.82	0.90	1.00	0.94	100	0.10	0.50	good	done
C	llama3.1-8b_code_r256	0.77	0.92	1.00	0.79	3	3.29	1.38	good	done
C	llama3.1-8b_chat_r256	0.80	0.94	1.00	0.79	5	2.83	1.07	good	done
C	llama3.1-8b_summ_r256	0.78	0.93	1.00	0.83	1	2.48	0.83	good	done
C	mistral-7b_code_r256	0.81	0.95	1.00	0.68	79	1.37	1.79	good	done
C	mistral-7b_chat_r256	0.81	0.95	1.00	0.56	68	1.34	1.38	good	done
C	mistral-7b_summ_r256	0.80	0.95	1.00	0.59	61	1.38	1.38	good	done
D	olmo2-7b_math_r256_seed1	0.69	0.89	0.98	0.37	1	2.89	1.07	mid	done
D	olmo2-7b_math_r256_seed2	0.70	0.91	0.98	0.44	1	3.71	1.38	mid	done
D	chloeli__llama-3.1-8b-baseline	0.81	0.90	1.00	0.83	0	4.90	0.64	mid	done
D	chloeli__llama-3.1-8b-pro-america-spec-msm	0.79	0.91	1.00	0.83	2	4.00	0.89	mid	done
D	Krishna27s__medical-qa-mistral-7b-qlora	0.81	0.94	1.00	0.76	33	2.95	2.30	mid	done
D	byzhsm__mistral3	0.82	0.91	1.00	0.77	0	2.77	0.83	mid	done
E	mamba2-2.7b-hf_math_r16	0.92	0.81	0.97	0.54	8	2.58	1.38	good	done
E	mamba2-2.7b-hf_math_r256	0.95	0.78	0.97	0.73	38	1.71	1.38	good	done
E	falcon-mamba-7b_math_r16	0.91	0.84	1.00	0.79	11	3.28	1.79	good	done
E	falcon-mamba-7b_math_r256	0.94	0.88	1.00	0.82	38	3.25	2.97	good	done
E	qwen1.5-moe-a2.7b_math_r16	0.99	0.91	1.00	0.71	0	10.66	1.79	good	done
E	qwen1.5-moe-a2.7b_math_r256	0.99	0.92	1.00	0.83	5	2.00	1.00	good	done
E	deepseek-v2-lite_math_r16	0.99	0.74	0.97	0.86	0	4.00	0.42	good	done
E	deepseek-v2-lite_math_r256	0.99	0.77	0.98	0.89	0	3.62	0.50	good	done

Table 4: Every unit of the pre-specified extension campaign (§K): hit@×2 under the full edge, the bulk edge and their geometric mean, Spearman ρ, fraction of layers carrying intruders at deployment, WikiText-2 knee and its ratio to the median empirical threshold, the pre-specified label, and the unit's status. Column B is the batch (A new families, B rank and configuration, C tasks, D stability, E SSM and MoE); the six external adapters, the batch-D rows named `owner__adapter`, were downloaded rather than trained by us and have no task held-out set; their scan and knee metrics are computed as for every other unit.

Forgetting follows the number of displaced directions. The intruder count summed over layers orders the WikiText-2 increase exactly (Table 7); the fraction of layers carrying an intruder ties LoRA and MiLoRA at 0.90–0.91, whereas the count per layer (5.7 against 3.9, a third fewer) separates them, as does the forgetting of MiLoRA, 37% lower over three seeds and 41% in the seed-0 run of

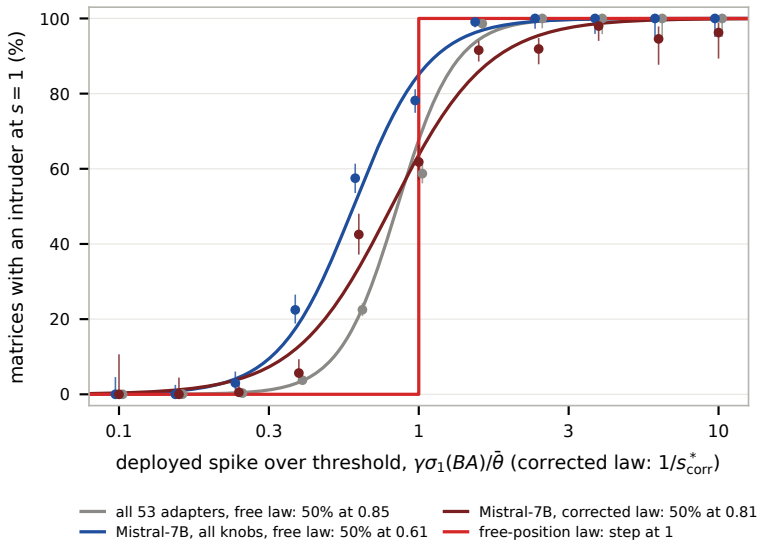

Figure 7: Fraction of matrices with an intruder at $s{=}1$ against the deployed spike over the threshold (for the corrected law, against $1/s^*_{\text{corr}}$), pooled over the 53 trained adapters of the second batch (grey), over the 16 Mistral-7B adapters (blue), and over the 12 Mistral-7B adapters with stored weights under the corrected law (dark red). Points are binned fractions with 95% Wilson intervals, curves are logistic fits, and the free-position law is the step at 1.

Mistral $r256$	LoRA (seed 0)	PiSSA	MiLoRA	AdaLoRA	cap rule
hit@×2 (full edge)	0.82	0.19 (0.17, 0.22)	0.82 (0.81, 0.82)	0.80 (0.79, 0.80)	0.82 (0.80, 0.84)
hit@×2 (bulk edge)	0.95	0.85 (0.82, 0.88)	0.93 (0.93, 0.93)	0.98 (0.97, 0.99)	0.96 (0.94, 0.97)
early arrival	1.21	2.61 (2.61, 2.61)	1.21 (1.21, 1.21)	1.47 (1.47, 1.47)	1.21 (1.21, 1.21)
intruder fraction	0.91	0.99 (0.99, 1.00)	0.90 (0.88, 0.91)	0.00 (0.00, 0.00)	0.69 (0.68, 0.71)
ΔPPL	+4.71	+2035 (52, 3046)	+2.97 (2.79, 3.25)	+0.12 (0.12, 0.12)	+1.62 (1.59, 1.67)
task perplexity	1.43	10.8 (1.9, 17.5)	1.38 (1.37, 1.39)	1.56 (1.55, 1.57)	1.33 (1.32, 1.33)
GSM8K (%)	64.6	45.4 (44.0, 46.4)	63.3 (62.2, 64.6)	67.5 (67.0, 68.0)	67.1 (65.8, 67.8)
HellaSwag	0.534	0.404 (0.397, 0.411)	0.561 (0.550, 0.570)	0.627 (0.627, 0.627)	0.584 (0.577, 0.593)
Qwen $r256$	LoRA (seed 0)	PiSSA	MiLoRA	AdaLoRA	
hit@×2 (full edge)	0.82	0.32 (0.31, 0.34)	0.82 (0.80, 0.84)	0.81 (0.81, 0.82)	
hit@×2 (bulk edge)	0.96	0.85 (0.83, 0.86)	0.92 (0.90, 0.93)	0.97 (0.96, 0.97)	
early arrival	1.21	2.47 (2.29, 2.61)	1.21 (1.21, 1.21)	1.21 (1.21, 1.21)	
intruder fraction	0.08	0.69 (0.67, 0.71)	0.02 (0.01, 0.03)	0.00 (0.00, 0.00)	
ΔPPL	+0.35	+16.2 (14.8, 17.3)	+1.46 (1.45, 1.47)	+0.04 (0.04, 0.04)	
task perplexity	1.21	1.55 (1.50, 1.59)	1.25 (1.24, 1.27)	1.21 (1.21, 1.21)	
GSM8K (%)	83.8	72.8 (69.8, 74.6)	79.7 (79.2, 80.0)	82.2 (81.6, 83.2)	
HellaSwag	0.598	0.494 (0.487, 0.502)	0.590 (0.585, 0.593)	0.600 (0.599, 0.601)	

Table 5: PEFT variants under the recipe of Table 2 (math task, rank 256), re-expressed as standard LoRA updates over the pretrained weights and scanned, probed, and benchmarked with the unchanged pipeline. Early arrival is the median ratio of the predicted to the empirical threshold $s^*_{\text{pred}}/s^*_{\text{emp}}$; the intruder fraction is the fraction of layers carrying an intruder at deployment; ΔPPL is the WikiText-2 perplexity increase over the base model. Means over three seeds with the seed range in parentheses; the LoRA column is the seed-0 run of Table 2; the other columns are three seeds each.

Table 7. On Qwen every unit except PiSSA carries at most 0.1 intruders per layer, and their residual drift ($+0.04$ to $+1.46$ PPL) lies below what the mechanism resolves.

AdaLoRA. The parameterization $B\,\mathrm{diag}(E)\,A$ with orthonormalized factors makes the singular values of the update the entries of E, which start at zero and move by at most the Adam step per iteration: after $1{,}000$ steps at 2×10^{-4}, $|E| \leq 0.08$ and σ_1 of the update is 0.06, one twentieth of the median threshold. Hence no intruders, no forgetting, and a slower task fit (task perplexity 1.56 against 1.43 for LoRA on Mistral), as the law predicts for an update this far below threshold.

After stage 3	Base model	LoRA, uniform α (3 seeds)	cap rule, β=0.7 (3 seeds)
MeetingBank	17.6	40.9 (38.7, 42.7)	45.5 (41.9, 47.8)
Py150	14.6	24.7 (16.3, 29.3)	24.7 (16.0, 29.3)
ScienceQA	50.9	88.3 (87.5, 88.8)	92.4 (91.7, 92.8)
OP (mean final score)	—	51.3 (47.8, 53.1)	54.2 (50.2, 56.4)
BWT	—	−23.4 (−28.5, −20.7)	−23.2 (−29.5, −19.8)
WikiText-2 perplexity	4.9	241 (67, 569)	22.8 (19.4, 26.2)
HellaSwag	0.626	0.453 (0.443, 0.459)	0.535 (0.523, 0.548)
GSM8K	0.086	0.017 (0.014, 0.022)	0.039 (0.036, 0.044)

Table 6: TRACE continual learning on Mistral-7B (MeetingBank → Py150 → ScienceQA, official data and metrics, LoRA rank 256): task scores of the final model, overall performance (OP), backward transfer (BWT), WikiText-2 perplexity, HellaSwag, and GSM8K accuracy. Means over three seeds with the seed range in parentheses. The cap rule is applied at every stage to the current weights.

Method	Intruders	Per layer	σ_1 of update	σ_{100}/σ_1	Top-256 coherence (L / R)	ΔPPL
LoRA	1,284	5.7	1.52	0.18	0.14 / 0.07	+4.71
PiSSA	1,966	8.8	1.19	0.44	0.53 / 0.39	+3046
MiLoRA	871	3.9	1.09	0.23	0.10 / 0.06	+2.79
cap rule, β=0.7	459	2.0	0.94	0.12	—	+1.59
AdaLoRA	0	0.0	0.06	0.62	0.12 / 0.06	+0.12

Table 7: Mechanistic diagnostics on Mistral $r256$ (seed 0). Intruders among the ten leading singular vectors at s=1, summed over the 224 adapted matrices (2,240 vectors) and per layer; σ_1 and the decay of the update spectrum (median over eight modules); energy of the update's ten leading left/right singular vectors inside the pretrained top-256 subspace (a random direction gives 0.06); WikiText-2 increase at deployment.

PiSSA lies outside the premise. The free-position law assumes an update in generic position relative to the pretrained singular vectors. PiSSA's factors start as the top-256 singular components, and after training the leading directions of its update still keep 0.4–0.7 of their energy inside the pretrained top-256 subspace on both models (LoRA, MiLoRA, and AdaLoRA: 0.05–0.24, the random-position value being 0.06; Table 7 gives the Mistral values). Coherent with existing large components, the update lifts rotated directions over the edge below the strength a random spike would need: intruders arrive at 0.38–0.44 of the predicted strength, with a new outlier above $\sigma_1(W)$ already present at the first crossing on 88–92% of layers, so this is emergence rather than mixing, and 98% of the intruders are new on both sides. The alignment-corrected rung, which removes the top-K components, moves the wrong way (hit@×2 0.00 at K=256), while the bulk edge still reaches 0.85. The update is 40–53% of the weight norm on Mistral (15–22% on Qwen), because the Adam displacement of the factors is amplified by the large initial factors; it replaces 8.8 of the ten leading directions per Mistral layer and forgets at least an order of magnitude more than LoRA on both models, with HellaSwag 22 and 11 points below the base model. The models remain intact on short single-example contexts (task perplexity 1.35 per example on Mistral) but break on 1,024-token windows, which we report without a tested mechanism. PiSSA's own recipe uses a ten times smaller learning rate; we ran the common recipe by design and did not tune it.

Continual learning. The law holds at every stage (hit@×2 0.81–0.89 under the full edge, 0.91–0.96 under the bulk edge, all seeds; the scans of the cumulative update reach 0.79–0.82 and 0.95–0.98), because a rank-256 merge leaves the bulk of the current spectrum essentially unchanged. In the uniform chain intruders accumulate (cumulative fraction 0.92–0.95 after stages 2 and 3) and WikiText-2 perplexity rises at every stage in every seed (17 → 50 → 67, 18 → 62 → 87, 18 → 90 → 569). Under the cap rule the predicted supercritical fraction of the first two stages is 0.16–0.21 against 0.32–0.53, the deployed intruder fraction is 0.55–0.65 and 0.23–0.36 against 0.83–0.88 and 0.42–0.48, and the cumulative fraction is 0.76–0.79 against 0.92–0.95; the final perplexity is 19–26 against 67–569, HellaSwag 0.52–0.55 against 0.44–0.46, and overall task accuracy is 2.4–3.6 points higher (Table 6), while backward transfer on the earlier tasks is unchanged (−23.4 against −23.2): the rule reduces forgetting of the pretraining distribution, not the net interference between tasks. The task

gain arises because smaller updates fit each task at least as well ($5{,}000$ examples over seven epochs are far past the point where extra strength helps; Py150 and ScienceQA score 3–5 points higher right after they are learned), and because MeetingBank loses less in the later stages (3–5 points higher after stage 3) while Py150 loses more. One interference is not spectral: Py150 drops to 16–29 after ScienceQA in both chains alike, an output-format conflict that the rule does not address.

Pre-specified outcomes. Scored as the labels of Appendix K are, four of the eight predictions held (AdaLoRA follows the law; the cap rule; the law and the accumulation under continual learning). The rule under continual learning held on every clause except that OP should stay within 2 points of the uniform chain, which it exceeds in the favourable direction ($+2.9$), and its clause on backward transfer holds by only 0.2, which the text reports as unchanged. Three failed and are reported as such: MiLoRA was predicted to arrive late, with fewer deployed intruders than predicted supercritical layers (it arrives early, as LoRA does), PiSSA was predicted to arrive late with a better $K{=}256$ rung (the opposite holds), and units on Qwen with a predicted supercritical fraction below 0.3 were predicted to forget less than 1 PPL (MiLoRA $+1.46$, PiSSA $+16$).

N Further related work

Hyperparameter rules. OPLoRA (Xiong and Xie, 2026) (projection constraints) and rsLoRA (Kalajdzievski, 2023) ($\alpha/\sqrt{r}$ stability) are baselines rather than threshold-derived budgets. PiSSA and MiLoRA (Meng et al., 2024; Wang et al., 2025b) initialize the factors from the largest or smallest singular components; MiLoRA then arrives as early as LoRA and carries intruders on as many layers, whereas PiSSA stays coherent with the pretrained top subspace and crosses $2.5\times$ early, outside the law's premise (Appendix M); Spectral Unforgetting (Abro and Tahir, 2026) repairs damage post hoc. **Budget allocation and forgetting.** AdaLoRA (Zhang et al., 2023) allocates rank across layers from training-time importance scores (its SVD parameterization keeps the spike $20\times$ below threshold under the LoRA recipe; Appendix M), LoRA+ (Hayou et al., 2024) sets asymmetric learning rates from infinite-width arguments, and DoRA (Liu et al., 2024) decomposes updates into magnitude and direction; all act during training, whereas the spike budget is fixed from the pretrained spectrum before training starts. QLoRA (Dettmers et al., 2023) appears among our external adapters, and the PEFT library (Mangrulkar et al., 2022) exposes the per-module α interface the rule uses. EWC-style regularization (Kirkpatrick et al., 2017), empirical surveys (Luo et al., 2025), and scaling laws in parameters and steps (Kalajdzievski, 2024) describe or penalize forgetting. **Spectra of trained weights.** Heavy-tailed self-regularization (Martin and Mahoney, 2021; Martin et al., 2021) reads layer quality from the tail of the weight spectrum, and TempBalance, AlphaLoRA, and AlphaPruning (Zhou et al., 2023; Qing et al., 2024; Lu et al., 2024) turn per-layer spectral statistics into learning-rate, LoRA-expert, and sparsity allocations; Thamm et al. (2022) find that most singular values and vectors of trained weights follow random-matrix statistics and that the learned structure sits in the largest singular values. These works allocate by spectral shape, whereas the threshold here is a scale for one update against one layer's edge. The anti-alignment observation has an early precedent: Hu et al. (2022) (their §7.3) observed that, instead of repeating the top singular directions of W, ΔW amplifies directions not emphasized in W. Scanning $W + s\Delta$ is the weight-interpolation axis of WiSE-FT (Wortsman et al., 2022), used here as a diagnostic. O-LoRA and InfLoRA (Wang et al., 2023a; Liang and Li, 2024) constrain adapters to subspaces that protect earlier tasks in continual learning, and CorDA (Yang et al., 2024b) builds adapters from context-oriented decompositions of the weights that can preserve pretrained knowledge, whereas the cap rule constrains only the scale. Deterministic subspace perturbation bounds go back to Wedin (1972); Appendix E shows they are vacuous in this regime.

O Scope of the claims in detail

Theory. The ladder assembles known results, the SVD-update identity (Zha and Simon, 1999; Brand, 2006), the secular determinant (Bunch et al., 1978), and the rectangular transform (Benaych-Georges and Nadakuditi, 2012); the contribution is their assembly on measured spectra and the falsification test. On a discrete spectrum the critical strength is defined relative to the evaluation point $z_0 = \sigma_1(1+\varepsilon)$, fixed at $\varepsilon = 10^{-3}$ and bounded in influence by the sensitivity ratio. A derivation of the evaluation point from the finite-size edge scale (Bloemendal and Virág, 2013), an edge chosen per module type

from the spectrum, and a transform for coherent spikes (the PiSSA case, Appendix M) are the next theoretical steps.

Evidence. What the transform adds to the location of the spectral edge is the scale: a multiple of the bulk edge fitted on the first batch matches the geometric-mean law out of sample, and the update magnitude alone ranks layers at $\rho = 0.54$ (Appendix C). Under the full edge, 7 of 18 first-batch and 15 of 53 second-batch adapters fall below $\rho = 0.6$ individually; the bars are pooled bars, and the bulk edge lifts the ρ of every family. The evidence is at the level of weight spectra and behavioural metrics (WikiText-2 perplexity, HellaSwag, GSM8K, the TRACE tasks), from one training recipe (α=2r) and mostly single runs (three seeds in Appendix M); representation-level probes, instruction following, and safety behaviour are open. The knee evidence rests on the absence of censoring and on the predicted threshold, since a correlation computed on a threshold-relative grid is high by construction (Appendix H); the $270\times$ asymmetry comes from a $221/3$ layer split in the collapse regime, with $1.5\times$ on balanced splits.

Control. The cap rule needs the end-of-training spike scale of the recipe, obtained from one calibration run, and its gain is that of a global scale set from the spectrum: halving the learning rate matches it, post-hoc downscaling of the trained adapter reaches a similar perplexity trade-off without retraining (Appendix J), the per-layer allocation adds little over the uniform scale, and under continual learning the rule reduces forgetting of the pretraining distribution while backward transfer on earlier tasks is unchanged. Below threshold, drift of up to 1.5 PPL (MiLoRA on Qwen) lies outside what the intruder count resolves. Transferring the calibration constant across families, a Pareto comparison over α, learning rate, and post-hoc scale, and merge interference (Ilharco et al., 2023) against the per-layer thresholds are future work.