THE LIFE CYCLE OF A MASSIVE ACTIVATION: STOCHASTIC BIRTH, WEIGHT-DECAY-DRIVEN GROWTH, AND COMPETITIVE CONSOLIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Massive activations, residual-stream coordinates with magnitudes far larger than typical activations, are associated with attention sinks in transformers, but how their scale is regulated during training remains incompletely understood. Combining training-trajectory analyses and controlled interventions, we trace their emergence, growth, and consolidation. Sink-carrying channels vary across random seeds but stabilize early within each run. Over longer training, surrounding channels erode and the sink concentrates onto a few redundant carriers. Across ablations, gradient attenuation follows the sink token's collective root-mean-square magnitude rather than any single channel, making collective scale central to understanding their effects. Our central result is that weight decay causally controls the turnover of global activation scale. In controlled continuations, removing decay near the peak allows this scale to keep rising, whereas retaining it produces decline even at constant learning rate. We develop a balance model for the rise and peak of massive-activation magnitude, in which AdamW-preconditioned growth opposes weight decay. Sweeping the decay coefficient λ shifts peak timing approximately log-linearly and yields peak magnitudes scaling approximately as $\lambda^{-1/2}$, consistent with this balance. Optimizer measurements further show that preconditioning sustains the large-channel cohort against decay even when raw maintaining forces are too small to do so. Together, these findings connect the observed life cycle to scale-regulating training dynamics and establish weight decay as a training-time lever on activation magnitude.

1 INTRODUCTION

A trained transformer's hidden states are not uniform. In every model examined so far, a handful of coordinates carry magnitudes hundreds or thousands of times the typical activation, concentrated at a single token position and largely unchanging across inputs Sun et al. (2024). These massive activations are not incidental: they act as implicit attention biases (Xiao et al., 2024), model quality can degrade when they are removed (Yu et al., 2025), and they are the principal obstacle to low-precision training and inference (Dettmers et al., 2022; Bondarenko et al., 2023; He et al., 2024). That obstacle is not a fixed cost of scale: Ahmadian et al. Ahmadian et al. (2023) report that outlier dimensions are sensitive to the optimization conditions under which a model is pre-trained rather than an inherent product of its size. Park et al. Park et al. (2025) take that finding as a design target, replacing Adam with an optimizer chosen to avoid privileged bases and training a 1.4B model on 1T tokens without the usual activation outliers.

Prior work gives a partial account of behaviors of these large outlier activations. Chen et al. (2026) supply a mechanism for why a large magnitude is useful — it attenuates the gradient that accumulates at an attention sink — in the form of a lower bound the magnitude must clear. Gallego-Feliciano et al. (2025) show that the emergence of massive activations over training is highly regular, fitting a five-parameter law to their growth. These are complementary but incomplete: one is a constraint and the other a curve. Neither says what sets the magnitude in practice, what stops it growing, or why one coordinate rather than another comes to carry it. We address those questions, and the answer to the last turns out to govern how the first two should be read.

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.

2 RELATED WORK

Massive activations and attention sinks. Sun et al. (2024) characterised massive activations as a small number of near-constant, extreme-magnitude coordinates appearing at fixed feature dimensions and predominantly at the first token or at delimiters. They function as implicit attention biases: supplying the architecture with explicit learnable bias parameters removes the need for them, and they do not form. The forward-side counterpart is the attention sink (Xiao et al., 2024) — disproportionate attention mass routed to an uninformative early token — which Xiao et al. (2024) exploit for streaming inference by retaining sink tokens in the KV cache. Why a model should want a sink has since received a second answer: Barbero et al. Barbero et al. (2025) argue that attending to the first token limits over-mixing of representations across depth, so the sink guards against representational collapse in long contexts.

Darcet et al. (2023) report the same phenomenon in vision A parallel, static line of work documents outlier dimensions as an obstacle to quantization Dettmers et al. (2022); Kovaleva et al. (2021); Bondarenko et al. (2023). A further line asks why the residual stream should have a privileged basis at all, given that the architecture is close to rotation-symmetric: Elhage et al. Elhage et al. (2023) rule out LayerNorm and finite-precision arithmetic as the cause and attribute the alignment to the per-dimension normalizers in Adam. That places the optimizer rather than the architecture at the origin of coordinate-aligned outliers, a reading our §9 and §11 support from the training-dynamics side.

The gradient-sink account. Chen et al. (2026) give the mechanistic account this paper builds on. Under causal masking, a token attended to by many later positions also aggregates their gradients — a gradient sink. Because the normalisation Jacobian attenuates gradients in proportion to $1/\mathrm{rms}(x)$, a large residual magnitude at that position suppresses the accumulated gradient. Their Theorem gives a sufficient condition,

$$\mathrm{rms}(x) \;\geq\; \frac{\|\gamma\|_\infty}{\tau}\,\|g_y\| \;\Longrightarrow\; \|\nabla_x \mathcal{L}\| \leq \tau, \tag{1}$$

for any target gradient scale τ, where γ is the normalisation gain vector, g_y the gradient arriving at the normalisation output, and $\mathcal{L}$ the training loss. They test the implication with V-scale, a modification attenuating gradient flow on the value path, and report that massive activations are suppressed while attention sinks persist.

Two features of Equation (5) shape what follows. It is stated on $\mathrm{rms}(x)$, a property of the whole token rather than of any coordinate. And it is one-sided: it establishes a floor and penalises nothing above it, so it predicts neither a peak nor a decline during the training cycle and training dynamics.

Emergence over training. Gallego-Feliciano et al. (2025) characterised the developmental side, fitting the max-to-median ratio across 188 layers of the Pythia suite with

$$f(t) = A\, e^{-\lambda_{\mathrm{fit}} x_t} \log(x_t) + K, \qquad x_t = \gamma t + t_0, \tag{2}$$

at a mean R^2 of 0.984, and predicting its parameters from architectural features. The law has a critical point at which the ratio turns over, but the mechanism producing it was left open. On the forward side, Gu et l. Gu et al. (2025) study when the attention sink appears over pre-training and how optimization, data distribution, loss and architecture bear on whether it appears at all, reading the sink as a key bias that absorbs surplus attention without contributing to the value computation. Their subject is the sink's appearance; ours is the trajectory of the massive activation that carries it, including what happens after the sink is established

3 BACKGROUND

We study the life cycle of massive activations by combining measurements across training with controlled perturbations. Our approach pairs densely checkpointed training runs with analyses of released Pythia checkpoints. The experiments used to measure these quantities are setup according to details in Appendix E. Theory needed to understand these experiments is covered below.

3.1 ARCHITECTURE AND HIDDEN STATES

We consider decoder-only transformers of L pre-norm residual blocks. Block ℓ receives a hidden state $h_{\ell-1} \in \mathbb{R}^{S\times d}$, where S is the sequence length and d the model width, and returns

$$h_\ell = h_{\ell-1} + \mathcal{F}_\ell(\mathrm{Norm}(h_{\ell-1}))\,, \tag{3}$$

where $\mathcal{F}_\ell$ comprises multi-head self-attention and an MLP, and Norm is LayerNorm (Ba et al., 2016) or, in the variant of §3, RMSNorm (Zhang & Sennrich, 2019).

We write h_ℓ for the post-residual state — the residual stream — and index its entries as $h_\ell(s, c)$ for token position s and channel c. An *activation* is one such scalar. We do not examine intermediate quantities inside $\mathcal{F}_\ell$ except where stated.

Two consequences of the pre-norm placement matter throughout. Intermediate layers see normalised inputs, so the forward computation depends on the *direction* of h_ℓ at a position far more than on its scale: a coordinate can grow very large without much altering what the next block computes. The backward pass is different. The Jacobian of the normalisation carries a factor of order $1/\mathrm{rms}$, and that factor is not normalised away. Residual magnitude is therefore close to free in the forward direction and consequential in the backward one — the asymmetry that makes large activations useful rather than merely tolerable.

3.2 THE SINK TOKEN

Attention weights at each head are a softmax over positions and therefore sum to one. When no position is worth attending to, the mass must still be assigned somewhere, and it accumulates at a token carrying little content — typically the first, which under causal masking is the only position visible to every later one, though delimiters also serve. We write s^* for the sink position and measure all quantities there unless otherwise noted. We identify s^* per model as the position carrying the largest activations rather than assuming position 0, since as we show it is not always position 0 (§4.5).

3.3 THREE MAGNITUDES

Let $x = h_\ell(s^*, \cdot) \in \mathbb{R}^d$ be the sink token's hidden state at layer ℓ and training step t. We use three quantities.

Channel magnitude. $m_{\ell,t}(c) = |x_c|$, the value of a single coordinate. This is what the perturbation experiments manipulate and what the competition among channels is fought over.

Ratio. Following Gallego-Feliciano et al. (2025),

$$r_{\ell,t} = \frac{\max_c |x_c|}{\mathrm{median}_c |x_c|}, \tag{4}$$

the max-to-median ratio, to which Equation (8) was fitted. We retain it for comparability with that work.

Root-mean-square. $\mathrm{rms}_{\ell,t} = \sqrt{\frac{1}{d}\|x\|^2}$, the token's aggregate scale, and the quantity appearing in Equation (5). It is a property of the *token*, not of any channel.

These three do not move together, and conflating them is the easiest way to misread the phenomenon. In particular r has a denominator that is itself changing: over training the median falls, so r can rise while the numerator m of the dominant channel is falling. We observe exactly this (§4.1).

3.4 WHAT COUNTS AS A MASSIVE ACTIVATION

Sun et al. (2024) define an activation as massive when $|a| > 100$ and $|a|/\mathrm{median}(|h_\ell|) \geq 1000$. As noted in Gallego-Feliciano et al. (2025), these absolute thresholds do not transfer to smaller models, which exhibit the same qualitative pattern at lower values, and we follow that work in identifying the massive activation as the largest-magnitude channel at the sink of the layer where r is greatest.

Under a threshold definition the question is settled by fiat; using the argmax instead lets us measure whether the winner is a distinct category or merely the top of a continuum. We therefore also report the *isolation* of the leading channels: the largest ratio between consecutive entries of the sorted sink-magnitude profile. Isolation near 1 indicates a populated continuum; a large isolation indicates a small set standing apart from a collapsed bulk. We use *cohort* for the leading channels above that gap.

3.5 Gradient quantities

The mechanism under test concerns the backward pass, so we measure gradients as well as activations. We write g_x for the gradient with respect to the sink token's representation x at the input to a block's normalisation. Chen et al. (2026) show that, under RMSNorm, this gradient is attenuated by the token's own scale,

$$\|g_x\| \;\le\; \frac{\|\gamma\|_\infty \, \|g_y\|}{\mathrm{rms}(x)}, \tag{5}$$

where γ is the normalisation gain and g_y the gradient at the normalisation output. LayerNorm's Jacobian carries the same $1/\mathrm{rms}(x)$ factor up to centring, so the prediction applies to both variants. Equation (5) is the relation our ablations test in §4.3.

A channel is written into the residual stream by particular rows of the output-projection matrices of the blocks below it. We call these the channel's *writer rows* and collect them as w. A channel's magnitude grows or shrinks according to whether the loss-driven force on w exceeds the pull of weight decay, which we quantify by the *maintaining force*

$$G \;=\; \frac{-\langle w, g_w \rangle}{\|w\|^2}, \tag{6}$$

where g_w is the gradient on the writer rows and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. The numerator measures how strongly the loss pushes w along its own direction; normalising by $\|w\|^2$ makes G a fractional growth rate, directly comparable to the weight-decay coefficient λ, which enters the AdamW (Loshchilov & Hutter, 2019) update as $w \leftarrow w - \eta(g_w + \lambda w)$. The channel therefore grows when $G > \lambda$ and erodes otherwise.

Two versions of Equation (6) appear below and they differ substantially. The *raw* maintaining force G uses g_w directly. The *preconditioned* force $\tilde{G}$ applies AdamW's second-moment normalisation $\tilde{g}_w$ (Kingma & Ba, 2015), dividing each element by the square root of its accumulated squared gradient. Because a small but persistent gradient has a correspondingly small second moment, preconditioning can amplify it by orders of magnitude. As we show (§4.7), the raw force on every large channel falls far below λ — so the sink cannot be sustained by raw gradient magnitude at all — while the preconditioned force does not. Statements about what maintains a channel must therefore be made after preconditioning; instantaneous raw gradients do not read these dynamics.

4 Results

We examine how massive activations develop, how their magnitude is regulated, and how they concentrate onto a small cohort. We first characterise activation trajectories and the role of weight decay in global turnover. We then use ablations and magnitude-matched controls to distinguish effects of channel identity from effects of magnitude and collective scale. Finally, we examine stochastic channel identity, scale, and training maturity, distinguishing optimiser-mediated cohort maintenance from the unresolved mechanism of survivor selection.

4.1 Massive activations emerge and follow a global law

The dominant-layer max/median ratio follows the previously reported five-parameter law (Gallego-Feliciano et al., 2025), while sizeable channels share a rise-and-decline pattern in raw magnitude. These quantities need not move together: the ratio can continue rising while the MA magnitude falls because the median decreases faster. Full trajectories, fit results, and the qualifications on the peak-location analysis are provided in Appendices B and C.

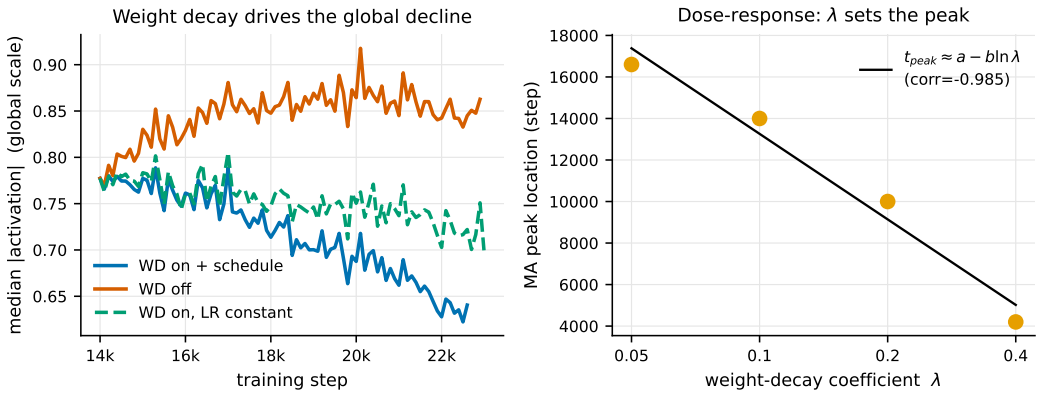

Figure 1: **Weight decay causes the turnover.** (a) Global scale (median $|x|$) rises without weight decay and declines with it, including at constant learning rate. The decaying schedule amplifies the decline. (b) MA peak location varies approximately log-linearly with the decay coefficient ($\mathrm{corr} = -0.985$).

4.2 Weight decay causes the turnover

Continuing from a near-peak checkpoint without weight decay, the global activation scale (median $|x|$) keeps rising. With weight decay, it falls even at constant learning rate: -11% over the compared window, versus -18% under the decaying schedule (Fig. 1a). Thus, in these continuations, weight decay drives the global turnover and the schedule amplifies the decline. This causal comparison concerns the global median: the individual MA is noisier and plateaus rather than declines at constant learning rate.

Sweeping the decay coefficient under the schedule shifts the MA's peak location approximately log-linearly, $t_{\mathrm{peak}} \approx a - b \ln \lambda$ ($\mathrm{corr} = -0.985$; Fig. 1b). At $\lambda = 0$, no peak occurs within the observation window. Peak magnitude scales approximately as $\lambda^{-1/2}$, consistent with the rise-and-peak balance model in §A.

4.3 The gradient floor is on the token's collective rms

The gradient-sink bound predicts $\|g_x\| \propto 1/\mathrm{rms}(x)$ (Equation (5)). We ablate the MA and other large channels individually, measuring $(\mathrm{rms}(x), \|g_x\|)$ with one forward and backward pass per condition and no training. The product $\|g_x\| \cdot \mathrm{rms}(x)$ remains approximately constant across ablation conditions: 0.94 when the MA is ablated, within the 0.89–0.95 range spanned by ablations of the other large channels (Fig. 2a). Ablating the MA changes the token's rms by only approximately 5%. These measurements support attenuation governed by the token's collective rms, with the MA one of several contributors rather than an exceptional channel.

4.4 Apparent MA-specific regulation resolves to magnitude

In perturb-and-restore experiments, an ablated MA is rebuilt (a floor), while an over-driven MA is pulled back during decline (a ceiling). These responses exceed those of a matched-*norm* random channel and occur independently of weight decay, initially suggesting MA-specific regulation. However, the random channel differs in activation magnitude, motivating magnitude-matched controls.

The ceiling tracks collective rms. We compare scaling the MA alone with scaling the top-16 channels proportionally, matched for the induced rms displacement. Despite an $11\times$ difference in direction distortion, the two perturbations remove similar fractions of the rms excess: 73% and 71%, respectively (Fig. 2b). A small random channel shows only approximately 20% pullback. The measured ceiling therefore tracks aggregate magnitude rather than uniquely responding to the MA or its direction distortion. These comparisons concern the continuum regime of the 51M model; §4.6 examines when comparable-magnitude peers exist.

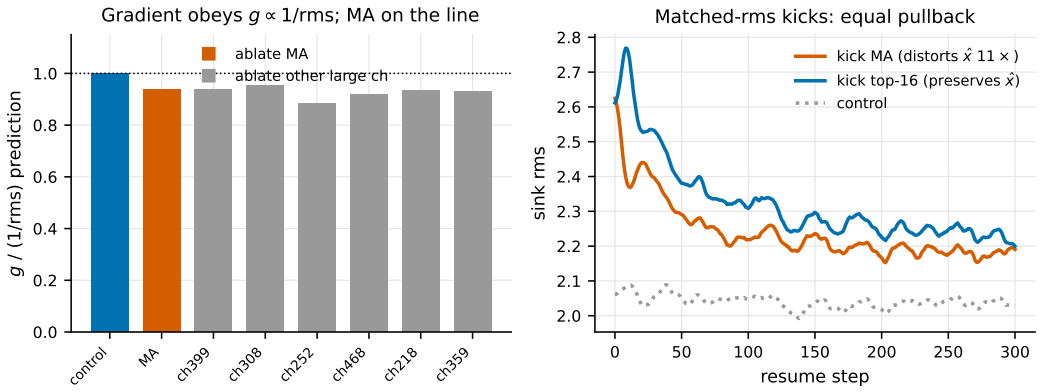

Figure 2: **Controls: apparent MA-specificity resolves to magnitude.** (a) Across individual-channel ablations, the downstream sink gradient follows $\|g_x\| \propto 1/\mathrm{rms}(x)$; the MA lies with the other large channels. (b) Matched-rms displacements show similar pullback despite different direction distortions, supporting a collective-scale interpretation of the declining-regime ceiling.

Model	Seed	Top-three sink channels
Ours (51M)	1337	207, 399, 308
Ours (51M)	2024	263, 114, 399
Pythia-410m	Default	357, 130, 966
Pythia-410m	Seed 1	110, 345, 27
Pythia-410m	Seed 2	424, 584, 216

Table 1: **Sink identity varies across seeds.** Pythia-410m has no pairwise top-three overlap across the three seeds; the two 51M runs share one secondary channel.

Gain suppression is also a large-channel effect. The MA's LayerNorm gain is comparable to that of other large channels rather than uniquely suppressed (Appendix D, Fig. 9).

4.5 The massive activation is a stochastic role, not a fixed dimension

Changing the training seed while holding architecture and data fixed changes the sink-carrying coordinates. In our 51M model, the dominant channel is 207 under one seed and 263 under another; three seeds of Pythia-410m share no top-three sink channels (Table 1). The number of sink channels (one to three), birth layer, and sink position also vary across seeds. Channel identity is therefore a stochastic outcome of training rather than a fixed architectural coordinate.

The identity stabilises early (Fig. 3). In Pythia-410m, the eventual top three occupy ranks 1–3 from approximately step 4k through the end of the 143k-step run. Ranks at 8k already correlate with final ranks (Spearman $\approx$ 0.5–0.7), while a channel remaining at rank 4 later collapses in magnitude. In our 51M model, the argmax migrates from channel 221 to 308 to 207, then locks by approximately step 5k. The top-three identity also remains unchanged across ten unrelated text passages. Together, these observations support early stochastic symmetry-breaking followed by stable, input-robust sink identity.

4.6 Scale and training maturity set whether the MA stands apart

Magnitude-matched controls require channels comparable to the MA. We examine this condition using ranked sink-channel magnitudes in our from-scratch models and Pythia (70M–6.9B; Fig. 4). Separation from the typical channel ($\max/\mathrm{median}$) increases with scale in both families, and our models fall within Pythia's range. However, the from-scratch models are trained much less, so this comparison mixes scale with training maturity. Their magnitude profiles remain populated continua, whereas fully trained Pythia models have a few large channels separated from a collapsed bulk.

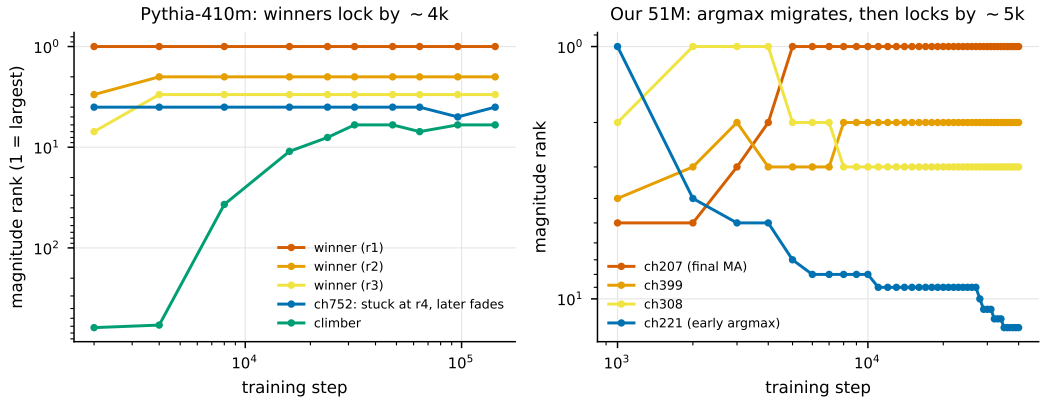

Figure 3: **Early lock-in of sink channels.** Magnitude rank versus training step (rank 1 is largest; log axes). (a) Pythia-410m's winners settle at ranks 1–3 by approximately 4k steps; the rank-4 channel later collapses (Fig. 10). (b) The 51M model's argmax migrates $221 \rightarrow 308 \rightarrow 207$ before locking by approximately 5k steps.

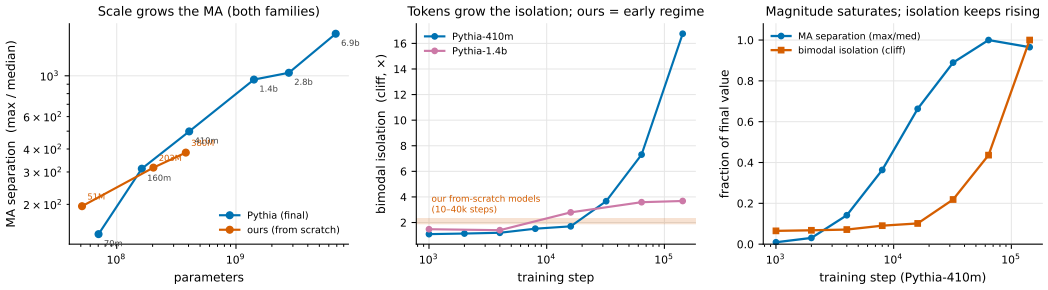

Figure 4: **Scale, training maturity, and isolation.** (a) Separation ($\max/\text{median}$) versus parameter count. Our models fall within Pythia's range but are trained much less. (b) Isolation, the largest consecutive-rank magnitude ratio (§3), across Pythia training. The shaded band shows our models over their 10–40k-step horizons. (c) Separation and isolation normalised by their final values in Pythia-410m: separation settles earlier, while isolation continues growing.

Isolation also develops during training at fixed model size. Pythia-410m and 1.4B begin near isolation 1, with Pythia-410m reaching approximately 17 by 143k steps (Fig. 4b). Our models remain in the low-isolation regime over their 10–40k-step training horizons, where magnitude-matched controls are available; those comparisons do not establish the same conclusions in the separated regime.

Separation and isolation evolve differently. In Pythia-410m, separation largely settles by approximately 10^4 steps, while isolation continues increasing through the end of training (Fig. 4c). Magnitude growth and the emergence of a gap between the leading cohort and the bulk are distinct processes.

4.7 Birth and Death: Competitive Consolidation of the Sink

Concentration without conservation. From step 16k to 143k in Pythia-410m, the surviving channels rise and then decline mildly, while mid-rank channels erode gradually on staggered schedules (Fig. 10). The top-three's share of the sink token's squared norm $\|x\|^2$ grows from 0.55 to 0.99. This is not redistribution of a conserved quantity: total sink rms peaks and declines, and winner gain is approximately 2.9 times field loss. In the 51M continuum regime, decline is graded rather than selective: peak-to-end retention ranges from 0.78 to 0.89 across the top eight channels, with the MA part of this pattern rather than separate from it.

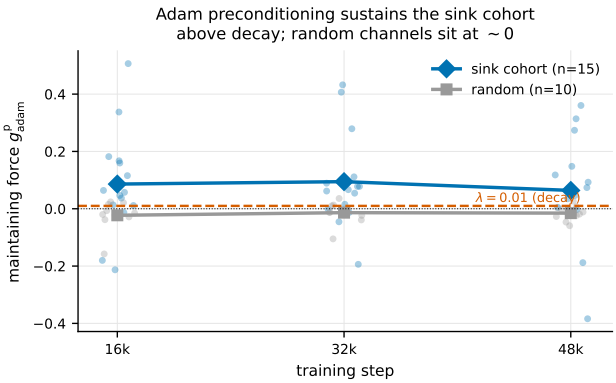

Figure 5: **AdamW preconditioning sustains the sink cohort.** The preconditioned maintaining force $\tilde{G}$ of Equation (6), over three checkpoints. The cohort of winners and collapsers lies above Pythia's decay threshold $\lambda = 0.01$ (dashed), while random channels lie near zero. Raw forces are approximately 10^{-4}. Winners and collapsers overlap, leaving intra-cohort selection unresolved.

The sink is a redundant collective. In Pythia-410m at step 16k, ablating a single surviving channel reduces attention to the sink by only 0.004–0.015, whereas ablating the top three jointly produces a drop of 0.08, an approximately $3\times$ superadditive effect. Random-channel ablations have negligible effect (Fig. 11). This supports a redundant implementation of the sink, consistent with the collective-rms result (§4.3) and attention-sink accounts (Sun et al., 2024; Xiao et al., 2024).

AdamW preconditioning sustains the cohort. Raw maintaining forces are of order 10^{-4}, below Pythia's $\lambda = 10^{-2}$ decay coefficient in the optimiser-force comparison. After AdamW second-moment normalization, the large-channel cohort lies above the decay threshold, while random channels remain near zero (Fig. 5). Thus the measured cohort maintenance depends on optimiser preconditioning rather than raw gradient magnitude. This does not identify a mechanism selecting survivors: winners and collapsers overlap in preconditioned maintaining force.

Tested mechanisms and unresolved selection. The measurements do not support several candidate explanations. Collapsing channels' writer-row gradients do not fall as $1/\mathrm{rms}$: the product of writer-row gradient norm and rms rises 1.5–$2\times$, and their gradients remain comparable to those of survivors. Across seed re-runs at 70M and 160M, survivor count does not track total sink attention (correlation ≈ 0), providing no support for a count set by attention-sink saturation. Total sink rms is not conserved, as shown above, and matched-rms perturbations do not support an MA-specific ceiling (§4.4).

Early identity lock-in (§4.5) and cohort-level maintenance do not explain why some cohort members survive and others collapse. A survivor can even exhibit negative maintaining force on held-out text. These post-hoc gradient measurements do not identify an intra-cohort selector; that mechanism remains unresolved.

5 Discussion and Conclusion

Together, our results describe a life cycle of stochastic birth, decay-controlled turnover, and competitive consolidation. Sink-carrying channels vary across seeds but stabilise early within each run; continued training erodes the surrounding field and concentrates the sink onto a small, redundant cohort. Weight decay causally drives turnover in global activation scale, while the massive activation's rise and peak are consistent with a balance between optimiser-preconditioned growth and decay (Appendix A). Its subsequent decline reflects reduced activation per unit writer-weight norm, not necessarily shrinking weights. AdamW's second-moment normalisation sustains the cohort against decay, but the measured maintaining force does not distinguish eventual survivors.

The central distinction is between channel identity and collective magnitude. Across the tested ablations, gradient attenuation follows $\|g_x\| \propto 1/\mathrm{rms}(x)$, locating the effect in the sink token's aggregate scale rather than any single coordinate. In the continuum regime, magnitude-matched controls resolve apparent MA-specific regulation into magnitude effects. A functional role for massive activations and magnitude-organised training dynamics are therefore compatible descriptions at different levels.

These findings have consequences for how massive activations are interpreted and targeted. Because the role is stochastic — filled by whichever channels win an early symmetry-breaking competition — interventions that zero or suppress individual channels may leave the underlying role intact: the sink will reconstitute on different coordinates, as the seed experiments confirm. Because gradient attenuation is governed by the sink token's collective RMS rather than any privileged dimension, the effect is distributed across the cohort and not uniquely attributable to the largest channel. Mitigation strategies that target the token's aggregate scale, rather than individual outliers, are therefore better motivated by the present account.

These conclusions are limited by experimental scale and training regime. Identity and life-cycle analyses include Pythia models spanning 70M-6.9B parameters, whereas perturb-and-restore controls remain at approximately 51M. Magnitude-matched comparisons require a populated tail and do not establish the same behavior in the bimodal, late-training regime. We do not test the attention-bias function (Sun et al., 2024) or rebuilding when denied a home (Vemula, 2026); intra-cohort survivor selection also remains unresolved. Within this scope, a massive activation is not a privileged architectural coordinate, but a training-selected contributor to a collective sink.

Ethics statement

An institutional review process determined that this work is exempt human-subjects research; identifying details, including the institution and protocol number, are withheld during double-blind review and will be disclosed in the camera-ready version if the paper is accepted.

AI use statement

The authors used generative AI tools to assist with spelling and grammar checking and with drafting and editing portions of the manuscript, including improving clarity, readability, and consistency. Generative AI tools were not used for the scientific design, execution, analysis, or interpretation of the work. All AI-assisted text was reviewed, edited, and verified by the authors. The authors take full responsibility for the final content of this work, including all claims, results, citations, and artifacts.

References

Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. Intriguing properties of quantization at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2305.19268.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.02732.

Stella Biderman, Hailey Schoelkopf, Quentin G. Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. arXiv:2304.01373.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2306.12929.

Yihong Chen, Zhouchen Lin, and Quanming Yao. Attention sinks induce gradient sinks: Massive activations as gradient regulators in transformers. *arXiv preprint arXiv:2603.17771*, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30318–30332, 2022. arXiv:2208.07339.

Nelson Elhage, Robert Lasenby, and Christopher Olah. Privileged bases in the transformer residual stream. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/privileged-basis/index.html`.

Jorge Gallego-Feliciano, S. Aaron McClendon, Juan Morinelli, Stavros Zervoudakis, and Antonios Saravanos. Hidden dynamics of massive activations in transformer training. *arXiv preprint arXiv:2508.03616*, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10781.

Bobby He, Lorenzo Noci, Daniele Paliotta, Imanol Schlag, and Thomas Hofmann. Understanding and minimising outlier features in transformer training. *Advances in Neural Information Processing Systems*, 37:83786–83846, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, 2024. arXiv:2305.17212.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3392–3405, 2021. arXiv:2105.06990.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

Jungwoo Park, Taewhoo Lee, Chanwoong Yoon, Hyeon Hwang, and Jaewoo Kang. Outlier-safe pre-training for robust 4-bit quantization of large language models. *arXiv preprint arXiv:2506.19697*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024. doi: 10.18653/v1/2024.acl-long.840.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2402.17762.

Maruthi Vemula. Massive activations are architecturally robust: A controlled scratch/commitment residual stream test. *arXiv preprint arXiv:2606.20743*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Mengxia Yu, De Wang, Qi Shan, Colorado J Reed, and Alvin Wan. The super weight in large language models, 2025. URL https://arxiv.org/abs/2411.07191.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. arXiv:1910.07467.

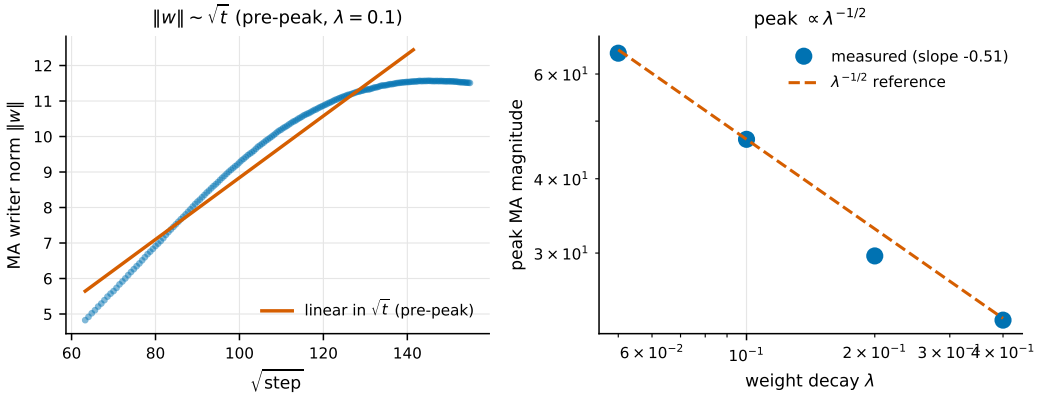

Figure 6: **The balance-condition theory.** (a) The MA writer norm $\|w\|$ grows linearly in $\sqrt{t}$ pre-peak (not exponentially), then bends over as the balance $G = \lambda\|w\|$ is reached ($\lambda = 0.1$). (b) The peak MA magnitude scales as $\lambda^{-1/2}$ across a weight-decay sweep (measured slope -0.51), fixing the drive exponent $q = -1$.

A THEORY: A BALANCE CONDITION, AND WHERE THE DECLINE LIVES

Let w be the aggregate writer weights feeding the MA channel; to leading order the MA magnitude $m \propto \|w\|$. Under AdamW, $w_{t+1} = w_t - \eta(\tilde{g}_w + \lambda w_t)$, where $\tilde{g}_w \equiv \hat{\mu}_t/(\sqrt{\hat{\nu}_t} + \epsilon)$ is the *preconditioned* update — not the raw gradient, from which it differs by one to two orders of magnitude on these weights (§4.7). Squaring and discarding the $O(\eta^2)$ term gives $\frac{d}{dt}\|w\|^2 \approx -2\eta(\langle w, \tilde{g}_w\rangle + \lambda\|w\|^2)$, so $\|w\|$ grows while the update's projection onto w opposes decay, and turns over when the preconditioned maintaining force of Equation (6) falls to the decay coefficient:

$$\tilde{G}(t^*) = \lambda. \tag{7}$$

This is the growth criterion of §3 evaluated at the peak, with the preconditioned drive balancing the decay pull, an equality complementing (Chen et al., 2026). This is the single-channel form of a general property of training under decoupled weight decay. Kosson et al. (2024) show that AdamW drives weight vectors towards a rotational equilibrium in which the expected growth contributed by the update balances the shrinkage contributed by decay, so the equilibrium scale is set by λ and the drive rather than by the loss alone. Their analysis concerns units made scale-invariant by a following normalisation, whereas our w are writer rows into the residual stream and are only partially scale-invariant under pre-norm; we therefore take (7) as calibrated by the λ-sweep of §4.2 rather than inherited from that result. What the connection supplies is the expectation that a balance exists at all.

Assuming the un-normalised preconditioned drive $\|w\|\,\tilde{G} = -\langle \hat{w}, \tilde{g}_w\rangle$, with $\hat{w} = w/\|w\|$, follows a power law in magnitude, $\|w\|\,\tilde{G} = c\,m^q$, substituting into Equation (7) with $\|w\| \propto m$ gives $c\,m^{q-1} \propto \lambda$, hence $m_{\text{peak}} \propto \lambda^{-1/(1-q)}$ for $q \neq 1$. The measured peak-magnitude scaling ($\propto \lambda^{-1/2}$; slope -0.51 to -0.57 across estimators, Fig. 6b) fixes $q = -1$: the drive falls as $1/m$, so each further increment of magnitude is worth progressively less. Because $\tilde{G}$ is defined on the preconditioned update, this characterises the optimiser's effective drive rather than the loss landscape directly; the two need not share an exponent, and we do not assume they do. Directly measured, $\|w\|$ grows as $\sqrt{t}$ pre-peak (not exponentially), the balance $G = \lambda\|w^*\|$ holds at the peak, and G falls over training (Fig. 6a).

The decline is not weight shrinkage. At baseline weight decay the MA magnitude turns over *while its writer weights keep growing*: at $\lambda = 0.1$, $\|w\|$ rises $10.8 \to 11.5$ across the activation's decline while the ratio $m/\|w\|$ falls $4.2 \to 3.2$. Only at strong decay does $\|w\|$ itself contract. The same weights therefore produce less activation than before, and the decline is a change in that relationship rather than in the weights' size. We do not claim that weight decay shrinks the MA weights. What causes the ratio to fall we do not establish.

Scope of the balance law. Equation (7) rests on $m \propto \|w\|$ and describes the *rise and peak*, where it is calibrated via the λ-sweep. It does *not* govern the decline, precisely because that proportionality breaks there. Two consequences follow. The fractional-decline prediction of the single-channel $G(m) = c\,m^q$ does not carry over to the declining, multi-channel population: the cross-channel retention ordering during the decline (larger channels retained more; §4.7) is a competition phenomenon of that regime, not a consequence of this equation. And $q = -1$ is fixed by the MA's *own* peak scaling rather than being a universal cross-channel law. The balance condition and the competitive consolidation are accounts of two stages of the life cycle, and we do not conflate them.

Honest negatives, and their positive form. We could not read the regulation off instantaneous *raw* gradients (activation- or weight-space, multi-batch averaged); the signal is below chance and entangled with Adam preconditioning. §4.7 turns this into a quantitative statement: the raw maintaining gradient is one-to-two orders below the weight-decay pull at every large channel, so the sink is held up not by raw gradient magnitude but by AdamW's second-moment normalization—the regulation is an emergent optimiser-preconditioning effect, visible under displacement or after preconditioning but not in the raw gradient. A direction test did not isolate a representational cost for the ceiling.

B EMERGENCE-LAW FIT

The max/median *ratio* of the dominant layer follows the five-parameter emergence law of Gallego-Feliciano et al. (2025),

$$r(t) = A\,e^{-\kappa x_t}\log x_t + K, \qquad x_t = \tau t + t_0, \tag{8}$$

with amplitude A, decay rate κ, time scaling τ, time offset t_0 and asymptotic baseline K. We rename the last two from the λ, γ of Gallego-Feliciano et al., which here denote weight decay and the normalisation gain. The fit attains $R^2 = 0.999$ over 40k steps in the 51M model (Fig. 7). The ratio and the raw *magnitude* can disagree in direction: at layer 5, the ratio rises monotonically to approximately 400, while the underlying MA magnitude peaks near step 20k and then declines, because the median falls faster than the MA. The ratio therefore reflects changes in both the dominant channel's magnitude and the median, rather than the dominant channel's magnitude alone.

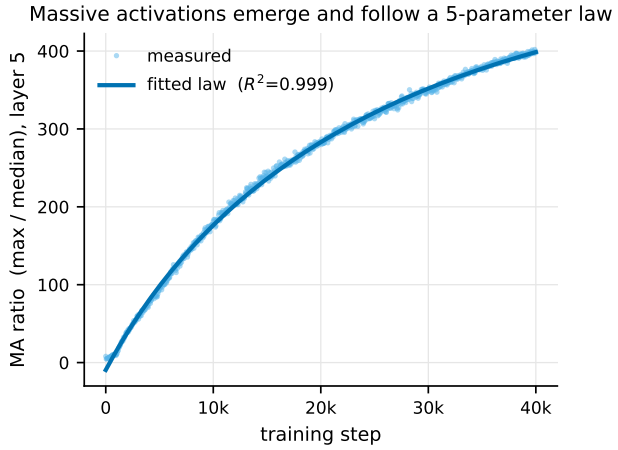

Figure 7: MAs emerge and fit the five-parameter law of (Gallego-Feliciano et al., 2025) ($R^2 = 0.999$; layer 5, 51M model).

C GLOBAL CHANNEL-MAGNITUDE DYNAMICS

Measuring per-channel *magnitudes* directly (Fig. 8a, layer 2), essentially every sizeable channel rises and then declines: 98/98 at layer 2 and 382/398 at layer 5. The peak *location* tends to increase

with amplitude (Fig. 8b), with small channels peaking early and large channels peaking late, and the MA at the extreme.

This relationship is noisy and right-censored. Channels whose peaks occur beyond the 40k-step observation window cannot appear in the peak-location analysis, biasing its high-amplitude end downward. We therefore interpret the relationship as a tendency rather than a clean law. The rise-and-decline shape is shared across the sizeable-channel population rather than being specific to the MA; the five-parameter ratio law reflects both this magnitude dynamic and the falling median.

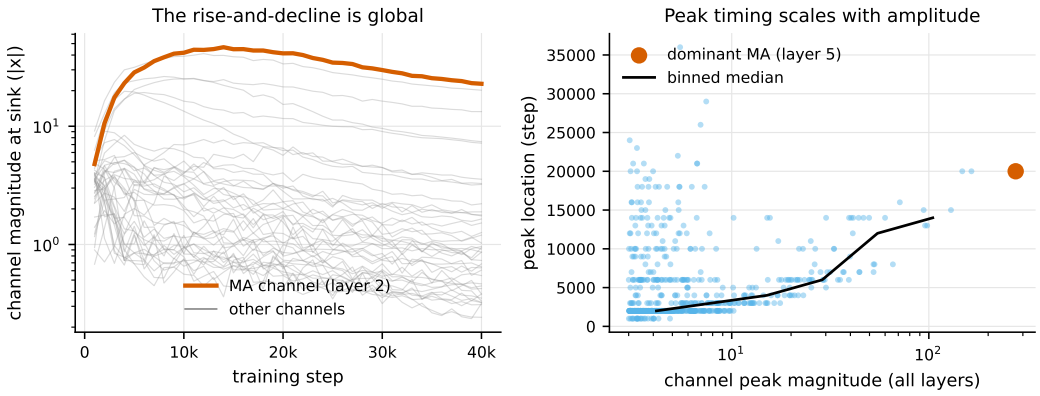

Figure 8: **The rise-and-decline is global.** (a) Per-channel sink magnitude (layer 2): nearly all sizeable channels rise then decline; the MA is the large-amplitude tail. (b) Peak location vs. channel peak magnitude, pooled over layers; the binned median (line) rises with amplitude, but the per-channel scatter is wide at any amplitude and *right-censored*—channels peaking beyond the 40k-step window cannot appear, biasing the high-amplitude end downward—so we read a tendency, not a clean law.

D LayerNorm gain suppression

On the unperturbed trajectory, the massive activation's magnitude rises and persists while its normalisation gain γ falls from approximately 1.0 early in training to 0.69 by the end of training (Fig. 9a).

The suppression is not peculiar to the massive activation. At the layer where it is largest, gain decreases with magnitude across the whole population, and every one of the top eight channels sits at $\gamma \approx 0.65$–0.84, against a bulk median of approximately 1.25. The massive activation, at $\gamma = 0.69$, lies among its equal-magnitude peers rather than below them (Fig. 9b). As with the floor and the ceiling, what looks like an MA-specific property is a property of being large.

E Experimental Setup

E.1 Models and training

We train decoder-only transformers from scratch at three scales, which we label by their parameter counts: 51M (8 layers, width 512, 8 heads), 203M (12 layers, width 1024, 16 heads), and 380M (16 layers, width 1280, 16 heads). All use context length 512, the GPT-2 (Radford et al., 2019) byte-pair vocabulary (50,304 tokens), no bias terms, learned positional embeddings, and tied input/output embeddings, and are trained on the Dolma corpus (Soldaini et al., 2024). The primary model uses LayerNorm, with an RMSNorm variant at matched configuration for architecture-generality. Optimization uses AdamW (Loshchilov & Hutter, 2019) ($\beta = (0.9, 0.95)$, gradient clipping at 1.0) with weight decay $\lambda = 0.1$ and a peak learning rate of 6×10^{-4} decayed by a cosine schedule to 6×10^{-5} after a linear warmup of 300 steps (400 for the two larger models). The effective batch is 96 sequences of length 512 for the 51M model (64 and 40 for the 203M and 380M models), in bfloat16. The three models train for 40k, 20k, and 10k steps, corresponding to approximately

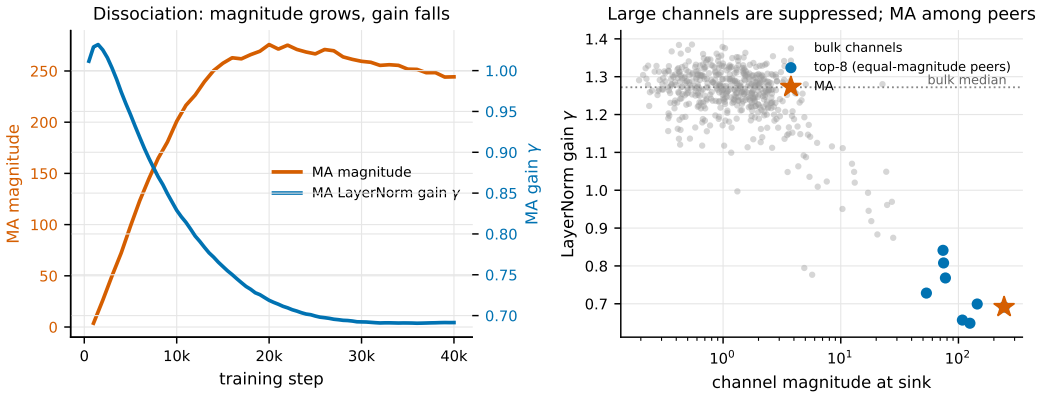

Figure 9: **Gain suppression tracks magnitude, not identity.** (a) Over training the massive activation's magnitude rises and holds while its LayerNorm gain γ falls from approximately 1.0 to 0.69. (b) At the layer where the massive activation is largest, γ decreases with channel magnitude across the population; the top eight channels all sit far below the bulk median, and the massive activation lies among them.

2.0B, 0.7B, and 0.2B tokens. Runs use seeds 1337 (primary) and 2024 (second seed). Checkpoints including optimiser state are saved every 500 steps and activation probes recorded every 50 steps.

For the identity, scale, and consolidation analyses we additionally use released Pythia checkpoints (Biderman et al., 2023) at 70M, 160M, 410M, 1.4B, 2.8B, and 6.9B parameters, including the seed re-runs at 70M, 160M, and 410M, and the ∼150 intermediate training checkpoints released per model. These are analysed by inference only; no Pythia model is trained or fine-tuned here.

E.2 Probes

Activations are measured at the sink position of each layer over a fixed batch of 24 sequences of length 512 from the validation split, held constant across checkpoints so that trajectories are comparable (for Pythia, a fixed passage of held-out text). At each probe we record the per-layer ratio r, the rms of the sink token, the full ranked profile of channel magnitudes — not only the argmax, so that near-degeneracy between the leading channels is visible — the identity of the argmax channel and its layer, and the norms of the writer rows for the leading channels.

E.3 Perturbation protocol

To ask whether a channel is actively maintained rather than passively accumulated, we displace it and observe what training does. From a checkpoint we create forks sharing batch order and a learning rate frozen at its value at the fork. In each fork the target channel's writer rows are scaled by a factor α at every layer — single-layer intervention is insufficient, since the channel is written redundantly across depth (§4.7). We use $\alpha = 1$ as an unperturbed control, which establishes how far two runs drift from data ordering alone; $\alpha = 0$ to test for a restoring force from below; and $\alpha = 2$ to test for one from above. Forks resume for 200–300 steps, with replicate resume seeds where error bars are reported.

We report the perturbed fork's absolute trajectory alongside its gap to the control. This matters because the control is itself moving during both the rising and declining phases, so a closing gap does not by itself demonstrate that anything is being restored. Recovery is summarised by the fraction of the initial gap closed within the window and by the half-recovery time, which is more robust than an exponential time constant when the return is not single-exponential.

E.4 CONTROLLING FOR MAGNITUDE

The perturbation protocol answers whether a channel responds; it does not answer whether the response is peculiar to *that* channel. A randomly chosen channel differs from the massive activation in magnitude, so comparing the two confounds identity with size. Since almost every quantity we measure varies with magnitude, that confound is not incidental.

We therefore compare against equal magnitude wherever the comparison is available, by three means. First, we regress the property of interest on channel magnitude and ask whether the massive activation lies above the value its own magnitude predicts. The trend must be fitted on channels well below the massive activation and the leading channels scored against it out-of-sample; a trend fitted on all remaining channels is contaminated by the runners-up, which are close in magnitude and can depress the high end enough to manufacture an apparent outlier where none exists (§4.4). Second, we match perturbations for the change they induce in rms: a single-channel displacement and one distributed proportionally across the leading channels are scaled to produce the same increase in the token's aggregate magnitude, so that the comparison isolates what distinguishes them — how far each rotates the token's direction — rather than how far each moved rms. Third, we pool across layers, architectures, and seeds, and require at least 10 channels per group before reporting any differentiator, having twice found clean separations at smaller samples dissolve under proper sampling.

This control has a precondition, examined in §4.6: there must exist channels of comparable magnitude to serve as the reference. Whether they do is a property of the training regime rather than a fixed fact about transformers.

Weight-decay causal test (§4.2). Starting from a near-peak checkpoint of the 51 M model, we resume training under two conditions sharing an identical batch order: weight decay enabled at the checkpoint's λ, and weight decay disabled ($\lambda = 0$). The optimizer's accumulated first and second moments are carried over from the checkpoint; only the `weight_decay` field of each parameter group is overridden. Both runs continue for 8,000 steps with gradient accumulation over four microbatches of 24 sequences, and activations are probed every 100 steps on the fixed validation batch.

To isolate the effect of weight decay from that of the cosine learning-rate schedule, one variant holds the learning rate fixed at the checkpoint value; the comparison in §4.2 uses this constant rate fork. A second variant follows the schedule, which amplifies the decline but is not required for the qualitative result.

The dose-response curve (Fig. 1b) is obtained by re-running the same protocol at $\lambda \in \{0.05, 0.1, 0.2, 0.4\}$ under the cosine schedule and recording the training step at which the MA channel reaches its maximum magnitude in each run. Peak locations are fit to $t_{\mathrm{peak}} \approx a - b \ln \lambda$ by ordinary least squares.

Gradient floor test (§4.3). We test whether the product $\|g_x\| \cdot \mathrm{rms}(x)$ is constant across ablation conditions, as predicted by Eq. 5, or whether ablating the MA channel produces an outlier.

Layer and channel selection. The target layer L is the layer with the highest max/median ratio, identified automatically via the activation probe on the fixed validation batch. The MA channel is the coordinate of highest mean absolute activation at position s^* over the probe batch; six comparison channels are taken as the next-highest by the same criterion, excluding the MA.

Ablation. Each channel is ablated by zeroing its writer rows across all layers ($\alpha = 0$ in the notation of §E.3), using a fresh copy of the checkpoint loaded for each condition.

Gradient extraction. To obtain $g_x = \partial \mathcal{L} / \partial h_L$ at the sink position, we register a forward hook on block L and call `retain_grad()` on its output tensor, which PyTorch would otherwise discard as a non-leaf intermediate. The model is set to training mode to enable gradient tracking, but no optimizer step is taken. A single forward and backward pass through the cross-entropy loss is run on the fixed probe batch of 24 sequences. We extract the residual-stream vector $x = h_L[s^*, :]$ and its gradient $g = \partial \mathcal{L} / \partial h_L[s^*, :]$ at sink position s^*, then average $\mathrm{rms}(x) = \sqrt{\frac{1}{d}\|x\|^2}$ and $\|g\|$ over the batch.

Test statistic. For each ablation condition i we report $g_i \cdot \mathrm{rms}_i$ (constant under the pure $1/\mathrm{rms}$ law) and the ratio $g_i \,/\, (g_0\, \mathrm{rms}_0/\mathrm{rms}_i)$, where subscript 0 denotes the unablated control. A ratio near 1.0 across all conditions indicates that attenuation is governed by the collective token scale rather than the identity of any individual channel.

Apparent MA-specific regulation (§4.4). Section 4.4 reports two classes of perturbation: a floor experiment (ablated MA is rebuilt by continued training) and a ceiling experiment (over-driven MA relaxes back), followed by a direction-controlled comparison that resolves the apparent specificity to a magnitude effect.

Floor and ceiling (perturb-and-restore). Following the protocol of §E.3, we create three forks from each checkpoint: an unperturbed control ($\alpha = 1$), a perturbed MA fork ($\alpha = 0$ for the floor; $\alpha = 2$ for the ceiling), and a fork perturbing a random channel matched by writer-row L2 norm to the MA. All forks share batch order and resume with the learning rate frozen at its checkpoint value. Forks run for 200–300 steps and recovery is summarised by the fraction of the initial perturbation gap closed within the window and the model-free half-recovery time $t_{1/2}$.

Ceiling with matched-norm random channel. For the ceiling comparison, the random channel is selected as the channel whose aggregate writer-row L2 norm $\|\mathbf{w}\|$ is closest to the MA's, excluding the MA itself. Both the MA and the random channel are scaled by $\alpha = 2$ in separate forks. Recovery is measured as the fraction of the initial excess magnitude $(m_0 - m_0/\alpha)$ removed by the end of the window: $(m_0 - m_{\mathrm{end}})/(m_0 - m_0/\alpha)$.

Direction-controlled comparison (Fig. 2b). To test whether the ceiling reflects MA identity or aggregate token scale, we compare two perturbations from the same checkpoint that induce equal rms displacement at the sink but differ in direction distortion by a factor of $11\times$:

- **Kick A**: scale the MA channel alone by $\alpha = 2$, strongly distorting the sink direction.
- **Kick B**: scale the top-16 channels by a factor β, calibrated iteratively (up to three passes, convergence tolerance 3%) so that the resulting $\mathrm{rms}(x)$ matches Kick A, preserving the sink direction.

Both forks resume for 300 steps. We track $\mathrm{rms}(x_{\mathrm{sink}})$ and the cosine distortion $1 - \cos(\hat{x}_t,\, \hat{x}_t^{\mathrm{ctrl}})$ at each step, where $\hat{x}_t^{\mathrm{ctrl}}$ is the unit sink vector of the unperturbed control at the same step.

Scale and isolation profiles (§4.6). Section 4.6 reports two measurements: a cross-model snapshot at the end of training (Fig. 4a), and a trajectory of isolation over training in Pythia-410m (Fig. 4b–c).

Cross-model scale profile. For the 51 M from-scratch model we use the final checkpoint (step 40k) with the fixed probe batch (§E.2), hooking all layers simultaneously to obtain residual-stream activations. For Pythia models (70M–6.9B) we load the final released weights from Hugging Face, pass a single fixed passage of 512 tokens (batch size 4) through `output_hidden_states`, and collect all layer outputs in one forward pass. In both cases the target layer is identified automatically as the layer with the highest max/median ratio across all layers and token positions.

We report two complementary measures of how far the leading channels stand apart from the rest:

- **Separation** (max / median): the ratio of the largest channel magnitude to the median across channels at the sink position, following Eq. 4.
- **Isolation** (cliff): the largest ratio between adjacent entries in the top-15 sorted magnitude profile, $\max_{i=0}^{13}(s_i/s_{i+1})$, where $s_0 \geq s_1 \geq \cdots$ are the sorted magnitudes. A large cliff indicates that the leading cohort has pulled away from the remainder; a cliff near 1 indicates a populated continuum.

Isolation trajectory (Pythia-410m). We load Pythia-410m at checkpoints $\{1\mathrm{k}, 2\mathrm{k}, 4\mathrm{k}, 8\mathrm{k}, 16\mathrm{k}, 32\mathrm{k}, 64\mathrm{k}, 143\mathrm{k}\}$ steps, passing the same fixed passage (batch size 4, sequence length 512) at each. Separation and isolation are computed identically to the cross-model profile above, with the target layer re-identified independently at each checkpoint. Figure 4c normalises each quantity by its value at the final checkpoint to place them on a common scale.

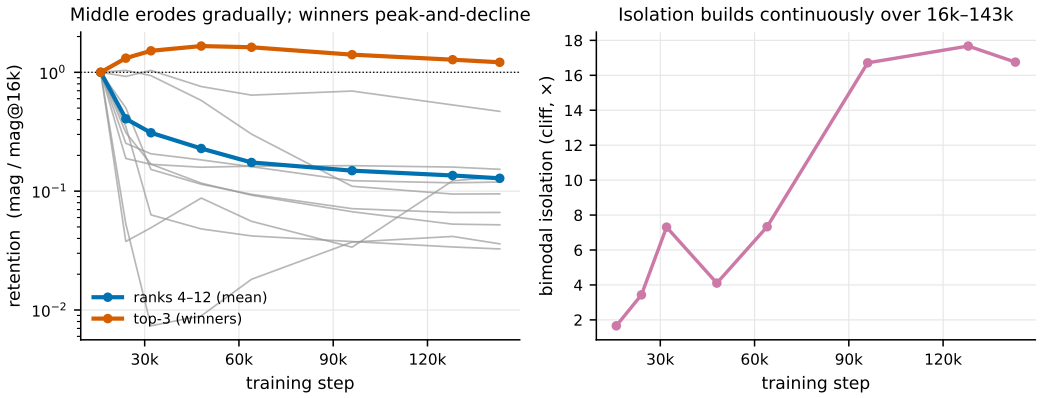

Figure 10: **Competitive consolidation in Pythia-410m (16k–143k steps).** (a) The top-three winners rise and then decline mildly, while mid-rank channels erode on staggered schedules. (b) Isolation builds continuously rather than appearing at a single transition step.

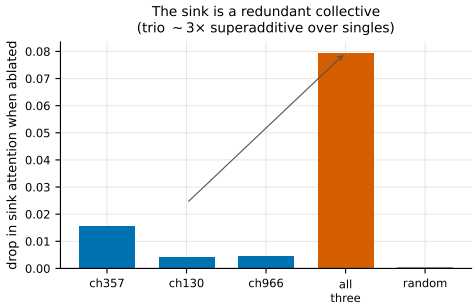

Figure 11: **The sink is a redundant collective.** Drop in sink attention after channel ablation (Pythia-410m, step 16k). Individual surviving-channel ablations have small effects; joint top-three ablation is approximately 3× superadditive. A random-channel ablation has negligible effect.

Competitive consolidation (§4.7). Section 4.7 reports three separate measurements on Pythia-410m: a retention analysis of channel magnitudes over training, an attention-ablation test of redundancy, and an optimizer-state measurement of cohort maintenance. All three use the released Pythia-410m checkpoints loaded from Hugging Face.

Retention and concentration (Fig. 10). Starting from the checkpoint at step 16k (near the emergence onset), we track the magnitude of each channel at the sink position across checkpoints at steps $\{16\text{k}, 30\text{k}, 60\text{k}, 90\text{k}, 120\text{k}, 143\text{k}\}$. Retention for rank k at step t is defined as $\text{mag}(t) \,/\, \text{mag}(16\text{k})$, normalising each channel's trajectory to its value at the reference checkpoint. We report the mean retention for the top-3 winners and for the mid-tier (ranks 4–12 at step 16k) separately.

Redundancy test (Fig. 11). To measure how much each channel contributes to the attention sink, we ablate channels by zeroing them via forward hooks registered on the embedding layer and on every transformer layer output, so that the ablated channel carries no signal at any depth. We then measure the mean attention weight assigned to position 0 from all query positions $s \geq 1$, averaged across all heads and all layers, on a single 256-token sequence. Survivors (channels 357, 130, 966) and a collapser (channel 752) are ablated individually and jointly; three randomly sampled channels serve as a baseline. Superadditivity is the ratio of the joint top-three drop to the sum of the individual drops.

AdamW preconditioning (Fig. 5). We measure the preconditioned maintaining force $\tilde{G} = -\langle \mathbf{w}, \hat{g}_w \rangle \,/\, \|\mathbf{w}\|^2$ (Eq. 6) for three groups of channels at steps 16k, 32k, and 48k. Channel groups are defined from the step-16k and step-143k checkpoints: *winners* are the top-10 channels by magnitude at step 143k; *collapsers* are channels in the top-40 at step 16k

whose retention $\mathrm{mag}(143\mathrm{k})/\mathrm{mag}(16\mathrm{k}) < 0.2$ and that are not winners; *random* channels are drawn uniformly from the remainder. At each checkpoint we accumulate per-sample gradients over 200 batches of 128 tokens from WikiText-2, computing the mean gradient m and mean squared gradient v per writer-row entry. The preconditioned gradient is $\hat{g}_w = m\,/\,(\sqrt{v} + \varepsilon)$, approximating AdamW's second-moment normalisation. Writer rows are the rows of the attention output projection (`attention.dense.weight`) and the MLP down-projection (`mlp.dense_4h_to_h.weight`) indexed by channel. The decay threshold $\lambda = 0.01$ is Pythia's weight-decay coefficient.