Which the Eye Fears: WRITING WITH READ-BLINDNESS EXPLAINS MASSIVE ACTIVATIONS IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Massive activation features (MAs) in Transformers are extreme-value residual-stream features that persist across layers despite the model's ability to suppress them. Why do they survive? Our investigation using an operator-level mechanistic analysis of attention and feed-forward (FFN) blocks reveals that these blocks systematically ignore MA coordinates while reading, but not while writing; creating a read-write asymmetry that blocks corrective feedback while allowing continued accumulation. We find that *both* attention and feed-forward layers have this read-blindness, and contribute to the emergence and persistence of MAs.

To validate prior work that hypothesized that FFN's amplification abilities is the primary reason for MAs (Sun et al., 2026), we analyze the model checkpoints during learning. Contrary to our expectation, read-blindness *emerges before* FFN amplification, suggesting that it acts upstream in the MA mechanism. We further contribute gradient analysis to link this behavior to surprising asymmetries in the loss landscape, concluding that the model actively maintains this read-blindness. Finally, we find that removing read-blocking at different locations induces compensatory shifts elsewhere, but MAs still persist.

1 INTRODUCTION

Massive activation features (MAs) are a small proportion of coordinates in the residual stream of a Transformer whose values exceed the rest of the model features by orders of magnitude (Bondarenko et al., 2023; Sun et al., 2024; 2026). They are central to how the model handles the first token of a sequence (Xiao et al., 2024), make low-bit quantization difficult (Yu et al., 2024), and shape the geometry of key-value cache compression (Ge et al., 2024). The most striking thing about them, however, is how stubbornly they persist. An MA at an early layer typically remains massive through nearly the entire network, even though every later attention head and feed-forward block has, in principle, the linear capacity to subtract it back to a normal range. None of them does. Why?

Several answers to this phenomenon have been proposed, ranging from an attention sink that dominates the BOS token (Xiao et al., 2024; Gu et al., 2025), to low-cost storage on delimiter tokens as compression valleys (Queipo-de-Llano et al., 2026), to more recently, early feed-forward blocks acting as a directional amplifier on a small set of channels (Sun et al., 2026). Though these attempt to identify the origins of massive activations, mechanistic explanations have been either elusive or partial. For example, none of these explain why no subsequent layer reduces the massive activations.

In this paper, we investigate progression of MAs through the key operators (attention, FFN) in the architecture. We first observe that an operator can fail to correct a residual-stream coordinate because either (1) the coordinate lies in its *output null space*: no choice of input produces a non-zero output for that coordinate; or (2) in its *input null space*; the coordinate is not read by the operator, and it produces the same output irrespective of the coordinate's value. We find empirically that MAs are *read-blocked* (read-blindness) by attention's W_V, W_Q, and W_K matrices, and by the feed-forward (FFN) block's gate and up-projections (W_{up}), but they are *not write-blocked*: every layer continue to add a non-zero contribution to MAs. We argue that this read-write asymmetry is the mechanism that makes MAs persistent through the network. Our observations adds several nuances and expands far beyond the simple hypothesis by Sun et al. (2026), which ascribes FFN amplification as the primary

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.

reason for MAs. To investigate this thoroughly, we analyze the model checkpoints during learning and find that read-blindness *emerges before* FFN amplification, suggesting that it acts upstream in the MA mechanism.

We next ask whether the observed asymmetry is actively maintained by training or is merely incidental. At a stable optimum, this question can be probed using local curvature via the Hessian of the loss. We examine the Hessian spectrum of the trained model, along parameters governing read and write pathways. We discover a pronounced asymmetry: directions associated with read operators are *stiff*, exhibiting higher-than-average curvature; while, those associated with write operators are comparatively *sloppy*, with substantially lower curvature. Simply stated, the model appears to actively maintain this read-blindness, and the phenomenon is likely not accidental.

We next attempt two structural interventions to constrain the value projection (W_V) so that no null space can develop. Massive activations attenuate, however, the trained model reconfigures and reinforces its read-blindness through other operators such as FFNs. This suggests that MAs are not tied to a single component, but arise from a collusion of mechanisms. Further analysis of FFNs reveals that the previously identified "amplifier direction" is generic; it can, in principle, amplify many features; but the down-projection W_{down} selects which of these amplified features are actually written back to the residual stream. It selects several MA features to amplify.

In summary, these results collectively represent a substantial advance in our mechanistic understanding of massive activations in Transformers. Our key result is that massive activations lie in the read-null space of various operators (in attention and FFN blocks), but not in their write-null space, that this behavior is actively maintained by the model, and that blocking read-blindness in one component leads to compensatory behavior in other components.

2 Related Work

The Impact and Persistence of Massive Activations: Massive activations (MAs); a small proportion of residual stream coordinates that exceed normal representation ranges by orders of magnitude; are a well-documented phenomenon in Transformers (Bondarenko et al., 2023; Sun et al., 2024; Kaul et al., 2025; An et al., 2025). They significantly constrain model deployment, acting as primary bottlenecks for low-bit quantization (Dettmers et al., 2022; Yu et al., 2024) and shaping the geometry of KV cache compression (Ge et al., 2024). One of the most striking characteristics of MAs is their stubborn persistence across layers. Analyses of their behavior across varied hyperparameters and configurations remain an active area of study (Gallego-Feliciano et al., 2025; Owen et al., 2025), yet the question of why subsequent layers fail to normalize these outliers remains a central mystery.

Hypotheses on Origins and the FFN Amplifier: Several hypotheses have been proposed to explain the origins and functional utility of MAs. They are frequently tied to how models process the beginning-of-sequence (BOS) token, acting as "attention sinks" (Xiao et al., 2024), or viewed as low-cost storage functioning as "compression valleys." More recently, a prominent mechanistic explanation by Sun et al. (2026) posits that early feed-forward (FFN) layers act as directional amplifiers, pushing signals into the massive regime when inputs align with a specific s^* direction. However, our investigation shows that these hypotheses are partial. They primarily address the initiation of MAs but fail to explain their structural persistence; specifically, why subsequent attention heads and FFN blocks do not utilize their linear capacity to subtract these outliers. Furthermore, our empirical findings regarding read-blindness in attention blocks offer a direct counterpoint to the hypothesis that FFN amplifiers are the sole or primary drivers of MAs.

Mechanistic Interpretability and Training Dynamics: To mechanistically understand why models stubbornly maintain these outliers, we build upon the spectral and null-space analysis pioneered by Cancedda (2024). We extend this to formally isolate the *input (read)* and *output (write)* null spaces of key operators across Transformer blocks. This formalized lens allows us to uncover the read-blindness that explains MA persistence. Finally, to determine if this structural asymmetry is merely incidental or actively maintained, we analyze the trained model through the lens of *stiff* and *sloppy* loss landscapes, measured via Hessian curvature (Transtrum et al., 2015). This landscape analysis contextualizes our findings that the model actively defends its read-blindness during training, framing MAs as a resilient, distributed mechanism rather than an isolated architectural quirk.

3 PRELIMINARIES & METHODOLOGY

We investigate the persistence of MAs through the interactions between a Transformer block and the residual stream, which include *reading* from the residual stream through its input projections and *writing* back to it through its output projections. We discover that the persistence of MAs follows from this read–write role asymmetry. As a result, our work creates a framework to systematically evaluate architectural modifications that aim to modify the statistical distribution of feature values, from *massive* to *outlier* and so on, by mechanistically tracking *read* and *write* operations in a Transformer block. Formally, given a weight matrix W, the *read* component of a linear operation $\vec{y} = W\vec{x}$ concerns the vector space of $\vec{x}$. *Read blindness* concerns a subspace of $\vec{x}$ not *read* by W, i.e., it is the (approximate) right null space of W. Similarly, the *write* operation concerns the vector space of $\vec{y}$: *write blindness* implies the presence of an (approximate) left null-space of W preventing it from modifying a subspace of $\vec{y}$. For a general vector space operator, the same concepts hold. The added complexity in Transformers is that the Attention and FFN operators require the residual stream itself as input, before we can study their write operations. We call this the *operator view* (see Sec 4), and general behavior of such operators are reported as average over sampled residual streams.

Transformer architecture. We consider a standard Pre-LN Transformer (Touvron et al., 2023) in which each layer applies multi-head attention (MHA) followed by a SwiGLU FFN, with RMSNorm preceding each block. Given residual input X, a layer computes $Y = X + \mathrm{MHA}(\widetilde{X})$ and $Z = Y + \mathrm{FFN}(\widetilde{Y})$, where tildes denote normalized inputs; full equations are provided in Appendix A. This architecture separates the parameters that *read* the residual stream from those that *write* to it. We focus on the MHA read operations coordinated through matrices W_Q, W_K, and W_V corresponding to the query, key, and value projections because they read the residual stream directly. Likewise, for the the FFN we focus on the reads through the W_{gate} and W_{up} matrices. The W_{down} matrix reads an intermediate vector, namely, the output of SiLU. The *read-write* separation allows us to quantitatively isolate and systematically measure the influence of every residual coordinate on the input and output side of every operation, within every block, of every layer of the Transformer.

Massive Activation Coordinates. We now formally identify the set $\mathcal{M}$ of massively activated coordinates. Let $Z^l_{t,k}(s)$ be the residual coordinate k in the output of layer l and token position t for an input sequence s from a held-out corpus. We define the set of massive coordinates as $\mathcal{M} = \left\{ k : \max_{s,\,l,\,t} \frac{|Z^l_{t,k}(s)|}{\mathrm{median}_j |Z^l_{t,j}(s)|} > R \right\}$. A coordinate belongs to $\mathcal{M}$ if it exceeds the threshold at any evaluated sequence, layer, or token position. Unless otherwise specified, we use $R = 5$; our results are robust to changes in this threshold (Appendix H). We then compare the read and write behavior of Transformer blocks on coordinates in $\mathcal{M}$ and its complement set $\neg\mathcal{M}$.

Measuring Read-blindness. Let A be a given read-side matrix. A is blind to coordinate k when values in coordinate k cause a weak (near-zero) response. For instance, if the standard vector e_k corresponding to coordinate k has Ae_k close to zero, we say that A is blind to coordinate k. Based on this intuition, we define the read-blindness (erasure) of a matrix A to coordinate k as follows:

$$\mathrm{era}_k(G) = 1/(1 + d_m\, G_{kk}/\,\mathrm{tr}(G)) \;\in\; (0, 1]. \tag{1}$$

where $G = A^\top A$ is the input-side Gram matrix of A. We use Gram matrix instead of the matrix itself as both G and A have the same null space; G being positive semi-definite makes it a more stable measure for assessing read-blindness. The read-blindness $\mathrm{era}_k(G)$ is close to 1 when G_{kk} is small relative to the trace, and close to 0 when G_{kk} is large. Under uniform treatment across coordinates, the expected score is 0.5.

To assess whether $\mathcal{M}$ has systematically higher erasure scores than a control set $\neg\mathcal{M}$, we use a one-sided Mann-Whitney U test on the layer-wise maximum of era_k for each feature k. We use either a size-matched sample or the full complement as specified in each analysis. We define U to count pairwise wins of features in $\mathcal{M}$ over features in $\neg\mathcal{M}$, assigning half a win to ties, and report the rank-biserial correlation $r_b = 2U/(|\mathcal{M}|\,|\neg\mathcal{M}|) - 1 \in [-1, +1]$ as an effect-size measure, together with the standardized z-score of the U statistic. A value $r_b > 0.95$ with $p \ll 0.05$ indicates that virtually every feature in $\mathcal{M}$ has a higher maximum erasure score than virtually every feature in the control set on that Gram matrix. As a complementary spectral diagnostic, we also measure how strongly each coordinate aligns with the approximate null subspace of G. We define this *null occupancy* measure and report its results in Appendix C.

4 READ BLINDNESS IN TRANSFORMERS

We now investigate whether massive coordinates exhibit the read–write asymmetry described above, and whether this behavior persists across model scale and architecture. We analyze three Llama models spanning 135M, 1.28B, and 2.56B parameters, together with a Qwen3 1.7B model. All four models use the Pre-LN residual architecture introduced in Section 3, with attention and SwiGLU FFN blocks that expose distinct read and write pathways. The Llama models allow us to test whether the phenomenon is stable over an order-of-magnitude change in scale, while Qwen3 tests whether it extends beyond a single model family.

For each model and operator role, Table 1 reports the median over $k \in \mathcal{M}$ of the layer-wise maximum erasure score, with the rank-biserial effect against the full complement $\neg\mathcal{M}$ in parentheses. We focus the main-text analysis on this erasure statistic and report the complementary null-occupancy results in Appendix C. We analyze the learned projections (*Weight View*) and the block inputs and residual updates observed on held-out data (*Operator View*).

Weight View. On the read side, we apply the input-Gram construction from Section 3 to W_V, W_Q, W_K, and the FFN gate/up projections. The write side instead requires a Gram over residual output coordinates. For attention, let h denote the concatenated head output, so that the residual update is $W_O h$. We use $G_O^{\text{write}} = W_O W_O^\top$, whose kth diagonal entry is $(G_O^{\text{write}})_{kk} = \|W_O^\top e_k\|_2^2$, the total squared weight through which the attention output can modify residual coordinate k. High write-side erasure indicates weak structural access to that coordinate, while low erasure indicates an open write pathway. We apply the same construction to W_{down}; precise definitions of all weight-side Grams are given in Appendix B. Note that Weight View on the write side is an approximation: the actual write operations are dependent on the Attention and FFN operators which are nonlinear functions of the residual.

Operator View. As we mentioned above, the write-side weight Grams are an approximation of the actual write operations by the operators of Attention and FFN. The exact results are empirical second moments over a corpus. For attention, $G^{a,oi} = \mathbb{E}[\widetilde{X}\widetilde{X}^\top]$ measures mean squared value (we use the shorthand *energy* from here on) of each coordinate at the normalized input, while $G^{a,oo} = \mathbb{E}[(Y-X)(Y-X)^\top]$ measures the energy of the attention residual update at each coordinate. We analogously use $G^{f,oi} = \mathbb{E}[\widetilde{Y}\widetilde{Y}^\top]$ and $G^{f,oo} = \mathbb{E}[(Z-Y)(Z-Y)^\top]$ for the FFN in the same way. High erasure on an operator-input Gram means a coordinate carries little magnitude at the input; this differs from erasure on a weight Gram, which means the following block cannot read the coordinate at all. High erasure on an operator-output Gram means the block writes little to that coordinate. The input Grams show whether massive coordinates are present to be read; the output Grams show whether the write pathways identified above are used in practice.

4.1 MAIN RESULTS

Massive coordinates are present in the normalized inputs to both blocks. On the operator-input Grams, the $\mathcal{M}$-versus-$\neg\mathcal{M}$ rank-biserial is negative in every model, for both $\widetilde{X}$ and $\widetilde{Y}$: massive coordinates are erased less than the non-massive ones, so they carry a larger mean squared value (*energy*) at the input. Their downstream suppression therefore comes from the learned read projections. The signal is present at the input after normalization.

This suppression is strongest and most consistent in W_V, W_Q, and the FFN gate/up projections. Across all four models, these projections show near-complete separation between massive coordinates and the full complement: $r_b \geq 0.99$ for W_V and the FFN projections, and $r_b \geq 0.96$ for W_Q. The coordinate-wise erasure evidence is weaker for W_K, where r_b ranges from 0.11 to 0.31. The spectral diagnostic in Appendix C, however, shows that massive coordinates consistently have greater occupancy in the approximate null subspace of W_K, with effects ranging from $r_b = 0.86$ for Llama 135M to $r_b = 1.00$ for Qwen3 1.7B (Table 4). Thus, W_K exhibits strong null-subspace alignment, although its total response to these coordinates is only modestly attenuated.

Both blocks can still write to massive coordinates. For the write matrices W_O and W_{down}, erasure stays near the uniform baseline and the $\mathcal{M}$-versus-$\neg\mathcal{M}$ rank-biserial is negative in every model, so massive coordinates sit inside the write range of attention and FFN even though the read side suppresses them. The operator-output Grams agree: the rank-biserial is negative in every model, so

Table 1: Erasure results. Cells report the median $\mathrm{era}_{\mathcal{M}}$ with rank-biserial r_b against the full complement $\neg\mathcal{M}$ in parentheses. Bold cells have $r_b > 0.95$; stars denote one-sided Mann–Whitney tests of $\mathrm{era}_{\mathcal{M}} > \mathrm{era}_{\neg\mathcal{M}}$ with $p < 0.05$.

	Llama 135M	Llama 1.28B	Llama 2.56B	Qwen3 1.7B
Massive coordinates $\lvert\mathcal{M}\rvert$	10	10	5	7
Maximum activation magnitude	280	2.27×10^3	3.92×10^3	2.23×10^3
Input availability:				
Attention normalized input, $\tilde{X}$	0.58 (-0.60)	0.68 (-0.73)	0.55 (-0.64)	0.58 (-0.37)
FFN normalized input, $\tilde{Y}$	0.56 (-0.65)	0.56 (-0.13)	0.50 (-0.63)	0.54 (-0.48)
Read side weight view:				
Value projection W_V	**0.70 (+0.99)***	**0.75 (+0.99)***	**0.78 (+1.00)***	**0.76 (+1.00)***
Query projection W_Q	**0.63 (+0.99)***	**0.65 (+0.99)***	**0.64 (+0.99)***	**0.66 (+0.96)***
Key projection W_K	0.56 (+0.29)	0.55 (+0.31)*	0.53 (+0.11)	0.59 (+0.25)
FFN gate/up $W_{\mathrm{gate}}, W_{\mathrm{up}}$	**0.66 (+0.99)***	**0.63 (+1.00)***	**0.59 (+0.99)***	**0.62 (+0.99)***
Write side weight view:				
Attention output W_O	0.51 (-0.82)	0.51 (-0.61)	0.49 (-0.59)	0.51 (-0.15)
FFN down projection W_{down}	0.49 (-0.86)	0.49 (-0.64)	0.49 (-0.46)	0.49 (-0.88)
Write side operator view:				
Attention update Δ^{attn}	0.61 (-0.87)	0.85 (-0.44)	0.62 (-0.64)	0.64 (-0.36)
FFN residual update Δ^{ffn}	0.81 (-0.29)	0.69 (-0.77)	0.59 (-0.59)	0.56 (-0.60)

Table 2: Signed-alignment results. For each coordinate, S is median-aggregated across layers. Cells report the median $S_{\mathcal{M}}$ with rank-biserial r_b comparing $\mathcal{M}$ and $\neg\mathcal{M}$ in parentheses.

Signed alignment	Llama 135M	Llama 1.28B	Llama 2.56B	Qwen3 1.7B
Attention, X_k with Δ_k^{attn}	+0.03 (+0.45)	+0.09 (+0.29)	+0.02 (-0.27)	+0.06 (+0.16)
FFN, Y_k with Δ_k^{ffn}	+0.05 (+0.57)	+0.25 (+0.59)	-0.02 (-0.20)	+0.10 (+0.14)

both blocks deposit a larger energy to massive coordinates than to the non-massive ones on held-out data. Note that because the second moment is not signed, we cannot tell from it if the deposits coordinate in being additive across layers. We investigate that next.

Signed deposits. To investigate signed direction of the deposits acorss layers, we compute $S^b_{\ell,k} = \mathbb{E}_{s,t}[\mathrm{sign}(R^b_{\ell,k})\Delta^b_{\ell,k}]$, using $(R^b, \Delta^b) = (X, Y - X)$ for attention and $(Y, Z - Y)$ for the FFN. Positive values mean that the coordinate's current sign is reinforced whereas negative values mean that they partially cancel. Table 2 reports the median over layers and then over $k \in \mathcal{M}$, with the rank-biserial effect against the full complement $\neg\mathcal{M}$.

4.2 Probing on Read-Blindness

The previous section showed that several read pathways are blind to massive coordinates. Is blindness in any one of these pathways necessary, or can the model preserve the asymmetry when that pathway is forced open? We test this by preventing either the attention value projection or the FFN input projections from becoming read-blind during training. We then ask whether massive activations disappear and, if not, where the read-blindness reconfigures.

Forcing a read pathway open. We train three variants of the Llama 1.28B model using the same hyperparameters as the base model. In the W_V *Frozen* variant, the combined W_V is initialized as a random orthogonal matrix and held fixed throughout training. It therefore has no null space, but also no freedom to learn. The W_V *Reparam* variant instead uses a Cayley parametrization, which allows W_V to learn during training while keeping it orthogonal. In the third variant, *FFN Frozen*, we initialize W_{gate} and W_{up} orthogonally and freeze them. We do not use a Cayley parametrization for these large rectangular matrices because of its computational cost and large number of strategies available in such a construction with nothing to motivate a particular choice. Relative to the base-

model perplexity of 14.21, perplexity changes to 14.67 (+0.46) for W_V Frozen, 14.24 (+0.03) for W_V Reparam, and 13.67 (−0.54) for FFN Frozen. Table 5 reports the erasure results; the complementary null-occupancy results are reported in Table 6 in the appendix.

Read-blindness re-configures elsewhere. Both W_V interventions work as intended. The erasure and null-occupancy scores of W_V are 0.50, with no separation between massive and control coordinates. In the Frozen variant, number of massive coordinates is reduced to 7, and the maximum activation magnitude falls to 2.08×10^3. In the Reparam variant, the number of massive coordinates is 10, and the maximum magnitude is 2.12×10^3, slightly lower compared to the base model. Nevertheless, massive activations remain. With W_V forced open, read-blindness remains in the other read pathways. In both variants, W_Q and the FFN gate/up projections almost completely separate massive coordinates from their controls, with $r_b \geq 0.99$ under both diagnostics. W_K also remains strongly aligned with its approximate null subspace, with null-occupancy effects of $r_b = 0.98$ and 0.99. Thus, removing read-blindness from W_V re-distributes its role across the remaining read operators rather than eliminating it from the model.

The FFN intervention shows the same pattern in the opposite direction. Freezing W_{gate} and W_{up} brings both of their blindness scores back to the uniform baseline of approximately 0.50. Read-blindness remains strong in attention. The effects are $r_b = 0.98$ for W_V and $r_b = 0.99$ for W_Q, while W_K has a null-occupancy effect of $r_b = 0.97$. The model still develops 11 massive coordinates, although their maximum magnitude falls to 681. Opening the FFN read pathway therefore weakens the largest activation, but does not remove massive activations. Because these FFN projections are frozen, part of this reduction may also come from their loss of learning capacity (reduced number of trainable parameters).

The re-distribution is confined to the read side. Across all three variants, W_O, W_{down}, and the realized attention and FFN updates show no preferential suppression of massive coordinates. The write pathways remain open and continue to deposit energy into these coordinates.

These interventions demonstrate that no single read projection is responsible for the read–write asymmetry. When one pathway is forced open, read-blindness is re-distributed across the pathways that remain available, while the write side stays open. MAs are therefore maintained and reconfigured in a collaboration between Attention and FFN operators. This result was a surprise to us.

5 SUN'S AMPLIFIER AND THE GEOMETRY OF FFN READ-BLINDNESS

We now investigate the relationship of our observations so far to previous work on FFN in Sun et al. (2026). They blame the origins of massive activations exclusively on a directional FFN amplifier. They formulate FFN's write to coordinate k as a quadratic form in the normalized residual $\widetilde{h}$.

$$y_k(\widetilde{h}) \approx \widetilde{h}^\top S_k \widetilde{h}, \qquad S_k = \tfrac{1}{2}(U_k + U_k^\top), \qquad U_k = \sum_i W_{\text{down}}[k,i]\, W_{\text{gate}}[i,:]^\top W_{\text{up}}[i,:].$$

Sun et al. (2026) show that under approximation (Appendix F), $\text{FFN}(\widetilde{h})_k \approx \lambda_\star^{(k)} (s_\star^{(k)\top} \widetilde{h})^2$, where $s_\star^{(k)}$ is the unit input direction (eigen-vector of S_k) that drives output coordinate k and $\lambda_\star^{(k)}$ is its signed gain (corresponding eigen-value). Note that since $s_\star^{(k)}$ need not align with the residual coordinate direction e_k, the FFN can be *blind* to the value stored in k while still reading $s_\star^{(k)}$ and writing a large update to k proportional to the signed-gain $\lambda_\star^{(k)}$. This will become important in our analysis.

Amplifier specificity lies in gain and spectral concentration Each output coordinate k has an amplifier direction $s_\star^{(k)}$ and a corresponding gain $\lambda_\star^{(k)}$. We ask which of these is distinctive for massive coordinates: do they share a special amplifier direction, or do they instead have unusually strong amplification along their respective directions? We test direction using the within-set absolute cosine, which measures how strongly the amplifier directions for different coordinates align. We test amplifier strength using $\|U_k\|_F$, the overall scale of the quadratic construction, and $|\lambda_\star^{(k)}|$, the gain along $s_\star^{(k)}$. Finally, $|\lambda_\star^{(k)}|/\|S_k\|_F$ measures how strongly the quadratic form is concentrated in that dominant direction. We report an additional subspace-alignment diagnostic in Appendix F.

Table 3 shows that direction is not the main distinction. The within-set absolute cosine is only moderately higher for $\mathcal{M}$ than for the size-matched non-massive set $c\mathcal{M} \subset \neg\mathcal{M}$ (0.53 versus 0.46), so massive coordinates do not share a clearly distinctive amplifier direction. Gain provides a much sharper distinction: $\|U_k\|_F$ and $|\lambda_\star^{(k)}|$ have rank-biserial effects of 0.78 and 0.98, respectively. The strongest discriminator is the concentration $|\lambda_\star^{(k)}|/\|S_k\|_F$, which separates the two groups perfectly in this sample ($r_b = 1.00$). Massive coordinates are therefore characterized primarily by unusually strong gain concentrated in a dominant direction, rather than by the identity of that direction.

Table 3: $s_\star$ amplifier diagnostics for Llama 1.28 Base. Per-coordinate entries are medians over $\mathcal{M}$ and a deterministic, size-matched $\neg\mathcal{M}$; the final row is the layer-mean of the within-set median pairwise absolute cosine. $r_b > 0$ means larger values on $\mathcal{M}$.

Diagnostic	value $\mathcal{M}$	r_b
$\|U_k\|_F$	7.74	+0.78
$\|\lambda_\star^{(k)}\|$	2.19	+0.98
$\|\lambda_\star^{(k)}\|/\|S_k\|_F$	0.39	+1.00
Within-set $\|\cos(s_\star)\|$	0.53/0.46	—

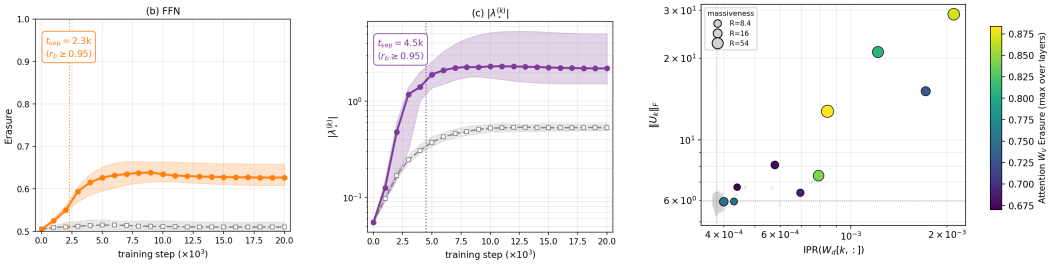
Overall, massive coordinates are strongly distinguished by having higher quadratic gain concentrated in a dominant direction, indicating a shift toward approximately rank-one behavior of the amplifier. However, the specific direction in the FFN's internal space is coordinate-specific and is not shared among the massive coordinates. Intuitively, this lack of collusion is perhaps unsurprising because the FFN has a separate quadratic form S_k for each output coordinate. What is surprising is the collapse of these forms toward a single direction with high gain. What occurs first: read-blindness or this collapse? We answer that next.

Read-blindness is detected before amplifier specialization. We track FFN read-blindness and amplifier gain through training. For each checkpoint and coordinate, we take the maximum across layers of the FFN erasure era_k derived from FFN's read-side matrices and the amplifier gain $|\lambda_\star^{(k)}|$. We assign $\mathcal{M}$ and the non-massive control set using the final checkpoint, and then trace these fixed sets backward through training. This lets us ask which FFN signature first distinguishes the coordinates that eventually become massive. We concede that this analysis tests temporal ordering, not causal dependence.

Figure 1(a–b) summarizes the comparison. We define the separation time t_{sep} as the first checkpoint at which the rank-biserial effect between eventual massive and control coordinates reaches $r_b \geq 0.95$. FFN read-blindness separates the groups at 2.3k steps (Figure 1(a)), whereas amplifier gain separates them later, at 4.5k steps (Figure 1(b)). Thus, FFN read-blindness is already detectable before the amplifier specializes. This temporal precedence does not establish that read-blindness causes amplifier specialization; it shows that read-blindness is the earlier structural signature of the coordinates that eventually become massive. Together with Section 4, the results support complementary roles: the amplifier explains how selected coordinates receive exceptional writes, while FFN read-blindness explains why subsequent FFN blocks do not condition corrective updates on the massive values already present.

W_{down} concentration and the read-blindness leak. The FFN amplifier poses a puzzle. The FFN is read-blind to massive coordinates, so how does it nevertheless produce large writes to them? The blindness is not total. Attention erases massive coordinates more strongly than the FFN: in the 1.28B model, W_V erasure is 0.75, compared with 0.63 for $W_{\text{gate}}, W_{\text{up}}$ (Table 1); the same gap appears through training in Figure 3. This difference is plausible since, unlike attention, the FFN acts independently at each token. The weaker FFN erasure therefore leaves a partial read pathway through which some FFN intermediate channels can still respond to massive residual coordinates.

Under the quadratic approximation, the FFN writes strongly to residual coordinate k when the input $\widetilde{h}$ aligns with $s_\star^{(k)}$, the top eigenvector of S_k. From the sum defining U_k, this direction is determined by the read directions $W_{\text{gate}}[i,:]$ and $W_{\text{up}}[i,:]$ of the intermediate channels i weighted by $W_{\text{down}}[k,i]$. For the partial read pathway to produce a strong amplifier, two conditions are needed. First, some intermediate channels that contribute to output coordinate k must retain read access to the massive residual coordinates. Second, the write row $W_{\text{down}}[k,:]$ must concentrate on those channels. Then S_k is dominated by a few read directions and can develop a high-gain dominant mode; if $W_{\text{down}}[k,:]$ is spread across many unrelated units, their contributions are diluted.



(a) FFN read-blindness **(b)** Amplifier gain **(c)** Write-row concentration vs. amplifier norm

Figure 1: FFN read-blindness precedes amplifier specialization. We assign massive and control coordinates at the final checkpoint and trace these fixed sets through 21 checkpoints from 0 to 20k training steps. **(a)** Maximum-over-layer FFN erasure for eventual massive coordinates (orange) and non-massive controls (gray). **(b)** Maximum-over-layer leading amplifier gain for the same sets. Lines and shaded regions show medians and interquartile ranges; vertical dotted lines mark t_{sep}, the first checkpoint at which the rank-biserial effect reaches $r_b \geq 0.95$. FFN erasure separates at 2.3k steps, before amplifier gain at 4.5k steps. **(c)** Final-checkpoint amplifier norm $\|U_k\|_F$ versus the inverse participation ratio of the corresponding FFN write row $W_{\text{down}}[k, :]$. Colored points are massive coordinates, with size denoting massiveness and color denoting attention-input erasure; gray points are non-massive coordinates.

The hypothesis predicts that FFN gain on coordinate k grows with the concentration of $W_{\text{down}}[k, :]$. We measure concentration using the inverse participation ratio (IPR) metric defined below.

$$\text{IPR}\big(W_{\text{down}}[k, :]\big) = \frac{\sum_i W_{\text{down}}[k, i]^4}{\big(\sum_i W_{\text{down}}[k, i]^2\big)^2} \in \big[\, 1/d_{\text{ffn}},\ 1 \,\big]$$

A higher IPR means that the write row for residual coordinate k is concentrated on fewer FFN intermediate units. Figure 1(c) reports the result. Within $\mathcal{M}$, residual coordinates with more concentrated W_{down} rows have larger amplifier norm $\|U_k\|_F$, consistent with concentration turning the partial FFN read pathway into a strong amplifier. Why does $W_{\text{down}}[k, :]$ concentrate? We suspect a token-independent coordinate is served by a few dedicated FFN intermediate channels, with weight decay and small-initialization gradient descent favoring the low-norm row that uses only those units.

6 Training Actively Reinforces Massive Activations

The intervention experiments show that the model can reorganize its read pathways to preserve read-blindness. We now ask whether this configuration is also visible in the local loss geometry, and whether the next optimizer step would preserve or reduce the resulting massive activations.

The loss geometry follows the read–write asymmetry. We borrow the language of *stiff* and *sloppy* directions from Transtrum et al. (2015). A direction is stiff when moving away from trained parameters increases the loss, and sloppy when the parameters can move without incurring much penalty. We measure this sensitivity using the directional Hessian curvature $v^\top H v$. For each parameter slice associated with a residual coordinate, we sample multiple random unit directions v and average their curvature. Positive curvature indicates a locally stable stiff direction, curvature near zero suggests a flat direction, and negative curvature indicates that the parameters do not lie in a locally convex basin along that direction. Appendix G.1 gives the full definition and sampling procedure.

Figure 2 shows that this geometry separates the two operator roles. Curvature is higher on $\mathcal{M}$ than on $\neg\mathcal{M}$ for the read-side matrices W_V, W_{gate}, and W_{up}. Moving these parameters is therefore more costly. The pattern reverses for the write-side matrices W_O and W_{down}, whose curvature is lower on $\mathcal{M}$. Thus, the loss constrains how the model reads MAs more strongly than how it writes to them. This geometric asymmetry matches the read-blind, write-open structure found in Section 4.

For deep strong-amplifier $\mathcal{M}$ rows of W_{down}, the cross-entropy gradient locally favors increasing the row norm, while weight decay opposes this tendency (Appendix G.1 and Figure 5). This partial force balance does not determine the net effect once AdamW's saved state is included.

The full AdamW step increases massive-activation magnitude. We therefore measure the end effect directly. At the trained checkpoint, we apply a counterfactual next AdamW step to the block that produces each target activation and compute its first-order effect on the activation's L_2 norm. We call this *gradient pressure*; its definition and the exact AdamW decomposition are given in Appendix G.3. As shown in figure 6, the full update has positive pressure on $\mathcal{M}$ at every layer, with a larger effect in later layers, while its pressure on matched non-massive coordinates stays near zero. Decomposing the step shows that the positive pressure comes mainly from AdamW's saved moment state. The new batch gradient has a much smaller effect, and weight decay reduces rather than increases the massive activations.

Together, the two analyses answer different parts of the question. The curvature analysis shows that massive coordinates occupy a distinct read–write geometry in parameter space. The optimizer analysis shows the next AdamW step is poised to reinforce their activation magnitude, with the accumulated optimizer state overcoming weight decay. This is local evidence that training actively maintains massive activations rather than merely tolerating them.

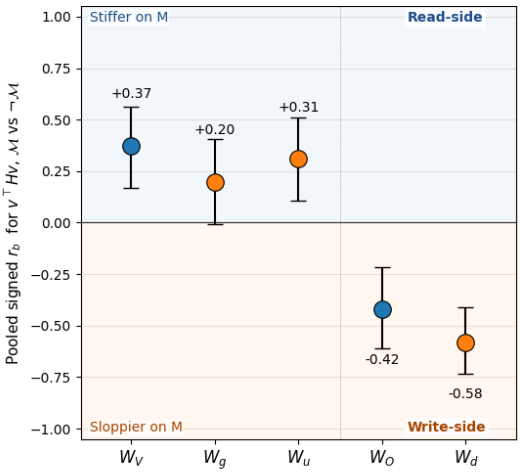


Figure 2: Directional curvature on $\mathcal{M}$ versus $c\mathcal{M}$ in the 1.28B base model (30 batches, 64 directions). Read-side slices are stiffer on $\mathcal{M}$; write-side slices are sloppier. Error bars are bootstrap 95% confidence intervals.

7 Conclusion

Massive activations persist because the transformer architecture systematically learns to ignore them while reading, yet continues to amplify them while writing. Our findings reveal that this read-write asymmetry is not an architectural quirk, but a systems-level configuration strictly enforced by the training dynamics and loss landscape. By constraining the W_V, W_{up} and W_{gate} matrices, we demonstrated that MAs are not tied to a single component; rather, the network actively reorganizes read-blindness to other operators like the FFN. We speculate that read-blindness is a cheap route for the model to maintain token-independent features, which are perhaps necessary for modeling language. However, the model architecture, especially the quadratic amplifier in FFN and lack of feedback control in write operations for all operators, can amplify a subset of these features to massive values as an unintended consequence.

Specifically, our analysis refines the directional-amplifier hypothesis. Massive coordinates do not share a distinctive amplifier direction; instead, they are distinguished by unusually high quadratic gain concentrated in their respective dominant directions. FFN read-blindness becomes detectable before this amplifier specialization, suggesting complementary rather than competing roles: the amplifier explains how selected coordinates receive exceptional FFN writes, while read-blindness explains why subsequent blocks do not condition corrective updates on the massive values already present. Within $\mathcal{M}$, stronger amplification is associated with greater concentration of the corresponding W_{down} row, consistent with a partial FFN read pathway being concentrated into a dominant mode.

Our work provides strong evidence that the MA state is a fixed point actively maintained by the optimizer. Read-side rows exhibit high curvature, heavily penalizing deviations from the read-blind state, while write-side rows sit at a saddle point stabilized by a delicate force balance: the cross-entropy loss pulls the weights outward, while weight decay pushes them inward. AdamW optimizer actively maintains the MA via its momentum state. By identifying read-blindness as the core mechanism, we provide a compelling mechanistic explanation for why subsequent layers in a transformer fail to normalize these extreme outliers.

AI USE STATEMENT

The ideas presented in this paper were developed by the authors. We used an AI Assistant to provide feedback on those ideas, suggest high-level section outlines, and proofread drafts for language and fluency. AI also assisted with coding; we validated the resulting code through manual inspection and tests. We reviewed the final manuscript and take full responsibility for its text, claims, and artifacts, including those produced with AI assistance.

ETHICS STATEMENT

This work began with our curiosity about how transformers, which are used in many modern AI systems, work internally. We use publicly available datasets and libraries. We provide their details and attributions in Appendix E.

We did not create a new dataset or involve human subjects. We hope this work helps improve our understanding of transformers. We do not expect it to raise any specific concerns about discrimination, bias, or fairness.

REPRODUCIBILITY STATEMENT

To support reproducibility, we attach our code in the supplementary material. The datasets used in this work are publicly available. We will release both the code and the data publicly upon acceptance. All training details are reported in Appendix E.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=rLX7Vyyzus.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/edbcb7583fd8921dad78adecfe06a99b-Abstract-Conference.html.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 4792–4808. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.263. URL https://doi.org/10.18653/v1/2024.acl-long.263.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332. Curran Associates, Inc., 2022. doi: 10.52202/068431-2198. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

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

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=uNrFpDPMyo.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=78Nn4QJTEN.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unraveling the enigmas of large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=IjduZQK8gM.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Louis Owen, Nilabhra Roy Chowdhury, Abhay Kumar, and Fabian Güra. A refined analysis of massive activations in llms. *CoRR*, abs/2503.22329, 2025. doi: 10.48550/ARXIV.2503.22329. URL https://doi.org/10.48550/arXiv.2503.22329.

Enrique Queipo-de-Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael M. Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1734b19d9afe7d2c7f1154954eaf0d5a-Abstract-Conference.html.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. *CoRR*, abs/2402.17762, 2024. doi: 10.48550/ARXIV.2402.17762. URL https://doi.org/10.48550/arXiv.2402.17762.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *CoRR*, abs/2603.05498, 2026. doi: 10.48550/ARXIV.2603.05498. URL https://doi.org/10.48550/arXiv.2603.05498.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, 143(1), 2015.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Mengxia Yu, De Wang, Qi Shan, Colorado Reed, and Alvin Wan. The super weight in large language models. *CoRR*, abs/2411.07191, 2024.

A Definitions and Transformer Architecture

Let d_m be the residual dimension, H the number of attention heads, $d_h = d_m/H$ the dimension of each head, and d_{ff} the FFN hidden dimension. We represent the residual stream after layer l as $Z^l = [z_1^l, \ldots, z_T^l] \in \mathbb{R}^{d_m \times T}$, with one token per column. For a single Pre-LN layer with input X, we write

$$Y = X + \mathrm{MHA}(\widetilde{X}), \qquad Z = Y + \mathrm{FFN}(\widetilde{Y}),$$

where X is the layer input and $\widetilde{X} = \mathrm{RMSNorm}(X)$ and $\widetilde{Y} = \mathrm{RMSNorm}(Y)$ are computed independently for each token. Below, we also write $Z_0 = X$, $Z_1 = Y$, and $Z_2 = Z$.

We use column-vector linear maps of the form $y = Ax$. Each attention head h therefore has read matrices $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d_h \times d_m}$ and an output matrix $W_O^{(h)} \in \mathbb{R}^{d_m \times d_h}$. Defining the row-normalized attention matrix

$$A^{(h)}(X) = \mathrm{softmax}\left(\frac{(W_Q^{(h)}X)^\top (W_K^{(h)}X)}{\sqrt{d_h}}\right),$$

the full multi-head output is

$$\mathrm{MHA}(X) = \sum_{h=1}^{H} W_O^{(h)} W_V^{(h)} X A^{(h)}(X)^\top.$$

We use $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d_m \times d_m}$ to denote the corresponding stacked or concatenated MHA matrices.

The SwiGLU feed-forward block uses read matrices $W_{\text{gate}}, W_{\text{up}} \in \mathbb{R}^{d_{ff} \times d_m}$ and an output matrix $W_{\text{down}} \in \mathbb{R}^{d_m \times d_{ff}}$:

$$\mathrm{FFN}(X) = W_{\text{down}}\big(\mathrm{SiLU}(W_{\text{gate}}X) \odot (W_{\text{up}}X)\big).$$

Finally, we denote the additive contributions of each module as $\Delta^{\text{attn}} = Y - X$ and $\Delta^{\text{ffn}} = Z - Y$.

B Gram Matrices for Read and Write Analysis

To apply the read-blindness scores η_k and era_k to a transformer, we form one Gram matrix per operator role. We use two complementary views.

Weight View. The Weight View asks what the trained weights, taken alone, permit each block to read or write. On the *read side*, a feature in the null space of a weight matrix is provably invisible to that operator regardless of input. On the *write side*, the Gram measures structural access to each residual output coordinate; what is actually deposited also depends on the operator's internal activations. The six Weight-View Grams, summing over heads h, are:

$$\begin{aligned}
G^{a,wi} &= \textstyle\sum_h W_V^{(h)\top} W_V^{(h)} && (W_V \text{ read — value projection}) \\
G^{a,wo} &= \textstyle\sum_h W_O^{(h)} W_O^{(h)\top} && (W_O \text{ write — output projection}) \\
G^{f,wi} &= W_{\text{gate}}^\top W_{\text{gate}} + W_{\text{up}}^\top W_{\text{up}} && (\text{FFN gate/up read}) \\
G^{f,wo} &= W_{\text{down}}\, W_{\text{down}}^\top && (\text{FFN down write}) \\
G^{Q} &= \textstyle\sum_h W_Q^{(h)\top} W_Q^{(h)} && (W_Q \text{ read — query projection}) \\
G^{K} &= \textstyle\sum_h W_K^{(h)\top} W_K^{(h)}. && (W_K \text{ read — key projection})
\end{aligned}$$

We list G^Q and G^K separately from $G^{a,wi}$ because, while all three read from the residual stream, they play different computational roles: W_V determines which information is routed into the attention output, whereas W_Q and W_K determine the attention pattern itself. Read-blindness in W_V blocks information flow directly; read-blindness in W_Q/W_K distorts where attention looks.

Operator View. The Operator View uses empirical second-moment matrices computed on a held-out corpus. Its input Grams measure which residual coordinates are available at the input to each block, while its output Grams measure which coordinates receive energy from the block's realized residual update. The four Operator-View Grams are:

$$G^{a,oi} = \mathbb{E}\big[\tilde{Z}_0\,\tilde{Z}_0^\top\big] \qquad \text{(Attention input second moment)}$$
$$G^{a,oo} = \mathbb{E}\big[(Z_1 - Z_0)(Z_1 - Z_0)^\top\big] \qquad \text{(Attention residual deposit)}$$
$$G^{f,oi} = \mathbb{E}\big[\tilde{Z}_1\,\tilde{Z}_1^\top\big] \qquad \text{(FFN input second moment)}$$
$$G^{f,oo} = \mathbb{E}\big[(Z_2 - Z_1)(Z_2 - Z_1)^\top\big]. \qquad \text{(FFN residual deposit)}$$

Expectations are taken over tokens and sequences, with each token represented as a column vector. $G^{a,oi}$ and $G^{f,oi}$ measure which features the normalized residual stream emphasizes at the input to each block; they do not by themselves measure whether the following operator reads those features. $G^{a,oo}$ and $G^{f,oo}$ measure which residual coordinates actually receive energy from each block's output.

C Null Occupancy

A component of the input in the null space of A has no effect on its output, so approximate null directions reveal features that the matrix effectively ignores. We use the input-side Gram matrix $G = A^\top A$, which has the same null space as A and is symmetric positive semidefinite. Because the strict algebraic null space is brittle for learned matrices, we instead use an *energy-thresholded* approximate null subspace.

Let $G = \sum_i \sigma_i v_i v_i^\top$ be the eigendecomposition of G, with eigenvalues in descending order. We define

$$\mathcal{N}_\tau(G) = \operatorname{span}\left\{ v_i \;:\; \sigma_i \le \varepsilon\,\sigma_{\max} \;\text{ and }\; \sum_{j\ge i}\sigma_j \le (1-\tau)\sum_j \sigma_j \right\},$$

where $\tau \in (0,1)$ controls the retained energy and ε prevents spurious null detection under nearly uniform spectra. We use $\tau = 0.99$ and $\varepsilon = 0.01$. Let $P_{\mathcal{N}_\tau(G)}$ denote the orthogonal projector onto this subspace.

The *null occupancy* of feature k is the fraction of its coordinate direction e_k that lies in the approximate null subspace:

$$\eta_k(G) = \frac{\|P_{\mathcal{N}_\tau(G)} e_k\|^2}{\|e_k\|^2} \;\in\; [0,1].$$

When $\eta_k(G) \approx 1$, coordinate k lies predominantly in directions that A effectively ignores; when $\eta_k(G) \approx 0$, it lies outside the approximate null subspace and is actively read.

D Probing Results on Read-Blindness

E Implementation and Model Details

We follow the data and hyperparameter setup of Gu et al. (2025) for all models. Each model is trained on the 5-billion-token RegMix dataset for 20,000 optimization steps, using a context length of 2,048 tokens. We use the GPT-NeoX tokenizer adopted by Gu et al. (2025), whose vocabulary contains 50,257 tokens. Table 7 summarizes the architectural configurations evaluated in our experiments. Training is done under a fixed seed.

We optimize every model with AdamW (Loshchilov & Hutter, 2019), using a peak learning rate of 10^{-4}, a weight decay of 0.01, and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is warmed up linearly for the first 100 steps and then decayed according to a cosine schedule for the remainder of training. Models are trained on 8x8 Nvidia A100 80GB accelerator hardward.

Table 4: Null occupancy results. Cells report the median $\eta_{\mathcal{M}}$ with rank-biserial r_b in parentheses. Bold cells have $r_b > 0.95$; stars denote one-sided Mann–Whitney tests of $\eta_{\mathcal{M}} > \eta_{\neg\mathcal{M}}$ with $p < 0.05$.

	Llama 135M	Llama 1.28B	Llama 2.56B	Qwen3 1.7B
Massive coordinates $\lvert\mathcal{M}\rvert$	10	10	5	7
Maximum activation magnitude	280	2.27×10^3	3.92×10^3	2.23×10^3
Input availability:				
Attention normalized input, $\tilde{X}$	0.52 (-0.40)	0.52 (-0.24)	0.52 (-0.56)	0.76 (+0.19)
FFN normalized input, $\tilde{Y}$	0.51 (-0.41)	**0.80 (+1.00)***	0.65 (+0.93)*	0.80 (+0.68)*
Read side weight view:				
Value projection W_V	**0.69 (+0.99)***	**0.72 (+1.00)***	**0.73 (+1.00)***	**0.74 (+1.00)***
Query projection W_Q	**0.69 (+0.98)***	**0.70 (+0.99)***	**0.71 (+1.00)***	**0.72 (+1.00)***
Key projection W_K	0.60 (+0.84)*	**0.63 (+0.96)***	**0.61 (+0.98)***	**0.67 (+1.00)***
FFN gate/up $W_{\text{gate}}, W_{\text{up}}$	**0.79 (+0.99)***	**0.87 (+1.00)***	**0.89 (+1.00)***	**0.85 (+1.00)***
Write side weight view:				
Attention output W_O	0.51 (-0.71)	0.50 (-0.58)	0.49 (-0.50)	0.51 (-0.11)
FFN down projection W_{down}	0.49 (-0.84)	0.50 (-0.26)	0.50 (-0.49)	0.50 (-0.16)
Write side operator view:				
Attention update Δ^{attn}	0.51 (-0.74)	0.50 (-0.61)	0.49 (-0.57)	0.51 (-0.13)
FFN residual update Δ^{ffn}	0.49 (-0.83)	0.49 (-0.63)	0.49 (-0.49)	0.50 (-0.41)

Table 5: Llama 1.28B Erasure results for intervention settings. Bold cells have $r_b > 0.95$; stars denote one-sided Mann–Whitney tests of $\text{era}_{\mathcal{M}} > \text{era}_{\neg\mathcal{M}}$ with $p < 0.05$.

	W_V Frozen	W_V Reparam	$W_{\text{up}}, W_{\text{gate}}$ Frozen
Massive coordinates $\lvert\mathcal{M}\rvert$	7	10	11
Maximum activation magnitude	2.08×10^3	2.12×10^3	681
Input availability:			
Attention normalized input, $\tilde{X}$	0.51 (-0.52)	0.56 (-0.57)	0.61 (-0.17)
FFN normalized input, $\tilde{Y}$	0.58 (+0.70)*	0.52 (-0.28)	0.57 (-0.17)
Read side weight view:			
Value projection W_V	0.50 (+0.22)	0.50 (+0.09)	**0.69 (+0.98)***
Query projection W_Q	**0.61 (+1.00)***	**0.64 (+0.99)***	**0.66 (+0.99)***
Key projection W_K	0.55 (+0.42)*	0.55 (+0.24)	0.63 (+0.93)*
FFN gate/up $W_{\text{gate}}, W_{\text{up}}$	**0.64 (+1.00)***	**0.64 (+0.99)***	0.50 (-0.21)
Write side weight view:			
Attention output W_O	0.52 (-0.16)	0.51 (-0.41)	0.51 (-0.37)
FFN down projection W_{down}	0.49 (-1.00)	0.50 (-0.95)	0.50 (-0.77)
Write side operator view:			
Attention update Δ^{attn}	0.84 (-0.68)	0.83 (-0.43)	0.60 (-0.75)
FFN residual update Δ^{ffn}	0.59 (-0.96)	0.61 (-0.48)	0.60 (-0.70)

Huggingface Transformers[1] framework is used for the model training. To speed up the training process, liger-kernel[2] library is used. A typical training run concludes in 8 hours.

F Directional FFN amplifier diagnostics

Directional quadratic amplifier. We compare our results with the directional quadratic amplifier of Sun et al. (2026). Under the near-identity approximation to the SiLU gate, let $g_i = (W_{\text{gate}})_{i,:}^\top$ and

[1] https://github.com/huggingface/transformers

[2] https://github.com/linkedin/Liger-Kernel

Table 6: Llama 1.28B Null occupancy results for intervention settings. Bold cells have $r_b > 0.95$; stars denote one-sided Mann–Whitney tests of $\mathrm{era}_{\mathcal{M}} > \mathrm{era}_{\neg\mathcal{M}}$ with $p < 0.05$.

	W_V Frozen	W_V Reparam	$W_{\mathrm{up}}, W_{\mathrm{gate}}$ Frozen
Massive coordinates $\lvert\mathcal{M}\rvert$	7	10	11
Maximum activation magnitude	2.08×10^3	2.12×10^3	681
Input availability:			
Attention normalized input, $\tilde{X}$	0.57 (+0.30)	0.54 (-0.13)	0.55 (+0.02)
FFN normalized input, $\tilde{Y}$	**0.89 (+1.00)***	**0.80 (+1.00)***	0.68 (+0.77)*
Read side weight view:			
Value projection W_V	0.50 (+0.00)	0.50 (+0.00)	**0.67 (+0.99)***
Query projection W_Q	**0.66 (+0.99)***	**0.69 (+1.00)***	**0.66 (+0.99)***
Key projection W_K	**0.59 (+0.98)***	**0.61 (+0.99)***	**0.65 (+0.97)***
FFN gate/up $W_{\mathrm{gate}}, W_{\mathrm{up}}$	**0.87 (+1.00)***	**0.90 (+1.00)***	0.50 (+0.00)
Write side weight view:			
Attention output W_O	0.52 (+0.00)	0.51 (-0.12)	0.51 (-0.38)
FFN down projection W_{down}	0.50 (-0.42)	0.51 (-0.43)	0.50 (-0.76)
Write side operator view:			
Attention update Δ^{attn}	0.52 (-0.07)	0.52 (-0.10)	0.51 (-0.48)
FFN residual update Δ^{ffn}	0.49 (-0.99)	0.50 (-0.54)	0.49 (-0.78)

Table 7: Llama (Touvron et al., 2023) and Qwen3 (Team, 2025) Architectures of the models used in our experiments.

Property	Llama 135M	Llama 1.28B	Llama 2.56B	Qwen 1.7B
Number of Transformer layers (L)	10	16	27	28
Residual-stream dimension (d_{model})	768	2,048	2,560	2,048
Feed-forward dimension (d_{ff})	1,536	8,192	7,680	6,144
Number of query heads (n_h)	8	32	40	16
Number of key–value heads (n_{kv})	8	32	40	16
Attention-head dimension (d_h)	96	64	64	128

$u_i = (W_{\mathrm{up}})_{i,:}^\top$ denote the input directions of intermediate unit i. For output coordinate k, define

$$U_k = \sum_i W_{\mathrm{down}}[k,i] g_i u_i^\top = W_{\mathrm{gate}}^\top \operatorname{diag}\big(W_{\mathrm{down}}[k,:]\big) W_{\mathrm{up}}, \qquad S_k = \tfrac{1}{2}\big(U_k + U_k^\top\big). \tag{2}$$

The FFN contribution to coordinate k is then approximately

$$\mathcal{F}_{\mathrm{ffn}}(\widetilde{h})_k \approx \widetilde{h}^\top S_k \widetilde{h}.$$

Let $(\lambda_\star^{(k)}, s_\star^{(k)})$ be the eigenpair of S_k with largest eigenvalue magnitude, with $\|s_\star^{(k)}\|_2 = 1$. When this eigenpair dominates the spectrum,

$$\mathcal{F}_{\mathrm{ffn}}(\widetilde{h})_k \approx \lambda_\star^{(k)} \big(s_\star^{(k)\top} \widetilde{h}\big)^2.$$

Thus, $s_\star^{(k)}$ is the residual-stream direction with the largest quadratic gain into output coordinate k, and $\lambda_\star^{(k)}$ is its signed gain.

Per-coordinate diagnostics. We derive four diagnostics from this construction. Except where the layer aggregation is shown explicitly, the quantities below are computed separately at each layer.

- The overall quadratic magnitude is $\|U_k\|_F$, computed without explicitly assembling U_k as

$$\|U_k\|_F^2 = W_{\mathrm{down}}[k,:]^\top \big(W_{\mathrm{gate}} W_{\mathrm{gate}}^\top \odot W_{\mathrm{up}} W_{\mathrm{up}}^\top\big) W_{\mathrm{down}}[k,:],$$

  where $\odot$ is the element-wise (Hadamard) product.

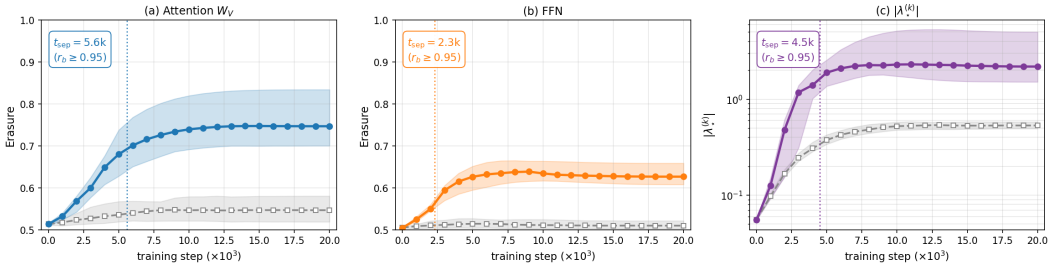


Figure 3: Emergence of $\mathcal{M}$ discrimination during pre-training of the 1.28B model (21 checkpoints, 0–20k steps; per-coordinate maximum across 16 layers). Solid colored and dashed gray lines show medians and interquartile ranges for $\mathcal{M}$ ($n = 10$) and sampled $c\mathcal{M}$ controls ($n = 100$), respectively. The dotted line marks t_{sep}, the first checkpoint with rank-biserial effect > 0.95. *FFN read-blindness preceeds its amplifier.*

- The massive-subspace mass of the amplifier direction is

$$\rho_{\mathcal{M}}^{(k)} = \operatorname*{mean}_{L} \sum_{j \in \mathcal{M}} \left(s_\star^{(k,L)}[j]\right)^2. \tag{3}$$

  Because $s_\star^{(k,L)}$ is a unit vector, this is the fraction of its squared norm lying in the coordinate subspace $\operatorname{span}\{e_j : j \in \mathcal{M}\}$, averaged across layers. It is bounded between zero and one. An isotropically oriented unit vector has expected mass $|\mathcal{M}|/d_m$; larger $\rho_{\mathcal{M}}^{(k)}$ therefore indicates stronger alignment with the massive-coordinate subspace.

  In the base 1.28B model, massive output coordinates have median $\rho_{\mathcal{M}}^{(k)} = 0.031$, about six times the isotropic baseline $10/2048 \approx 0.0049$. This quantity is larger than in the deterministic, size-matched non-massive set with rank-biserial effect $r_b = 0.60$. Thus, their amplifier directions are preferentially aligned with the massive-coordinate subspace, but are not confined to it: only 3.1% of their squared norm lies in that subspace.

- The leading-gain magnitude is $|\lambda_\star^{(k)}|$. Given the previously identified unit eigenvector, it can be computed as

$$\lambda_\star^{(k)} = s_\star^{(k)\top} U_k s_\star^{(k)} = W_{\text{down}}[k,:]^\top \left((W_{\text{gate}} s_\star^{(k)}) \odot (W_{\text{up}} s_\star^{(k)}) \right).$$

  We report $\max_L |\lambda_\star^{(k,L)}|$ for each coordinate.

- The rank-one purity is

$$r^{(k)} = \frac{|\lambda_\star^{(k)}|}{\|S_k\|_F} \in [0,1],$$

  where

$$\|S_k\|_F^2 = \tfrac{1}{2}\left(\|U_k\|_F^2 + \operatorname{tr}(U_k^2)\right),$$

  and

$$\operatorname{tr}(U_k^2) = W_{\text{down}}[k,:]^\top \left(W_{\text{gate}} W_{\text{up}}^\top \odot W_{\text{up}} W_{\text{gate}}^\top\right) W_{\text{down}}[k,:].$$

  Following Sun et al. (2026), $r^{(k)} \to 1$ indicates rank-one dominance. The random-symmetric baseline scales as $\mathbb{E}[r^{(k)}] \sim 2/\sqrt{d_m}$ (approximately 0.044 for $d_m = 2048$). We report $\max_L r^{(k,L)}$ for each coordinate.

We additionally measure direction sharing within a coordinate set $\mathcal{S}$ using the sign-invariant statistic

$$\overline{|\cos|}_{\mathcal{S}} = \operatorname*{mean}_{L} \operatorname*{mean}_{\substack{i,j \in \mathcal{S} \\ i \neq j}} \left| \left\langle s_\star^{(i,L)}, s_\star^{(j,L)} \right\rangle \right|.$$

For two independent isotropic directions in d_m dimensions, the large-d_m baseline is approximately $\sqrt{2/(\pi d_m)}$.

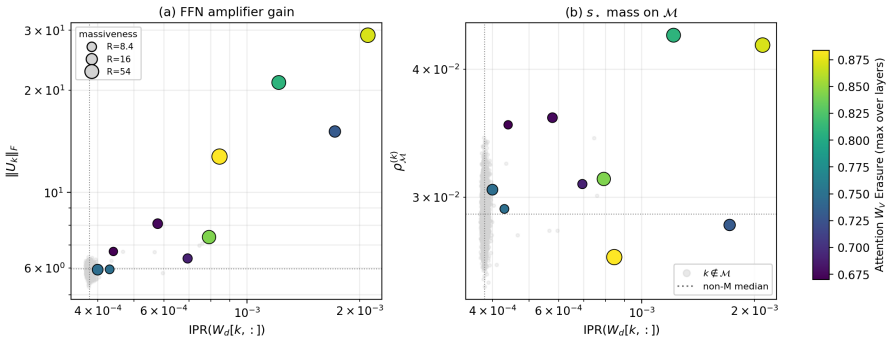


Figure 4: FFN write-row concentration vs. amplifier gain (left) and $s_\star$ mass on $\mathcal{M}$ (right). Color denotes attention-input erasure, size denotes massiveness R, and grey points are non-MA coordinates.

G Loss geometry and local AdamW pressure

This appendix gives the definitions behind the two analyses in Section 6. The curvature analysis describes the local geometry of parameter space. The AdamW-pressure analysis instead asks how a complete next optimizer step changes activation magnitude. These are different measurements and we do not identify one with the other.

G.1 Directional curvature and radial gradient

For each parameter slice associated with residual coordinate k, we sample unit directions v within that slice and measure the signed directional curvature

$$\kappa(v) = v^\top H v, \qquad H = \nabla^2 \mathcal{L}_{\mathrm{CE}}. \tag{4}$$

We average over 64 sampled directions and compare the resulting per-coordinate values for $\mathcal{M}$ and $\neg\mathcal{M}$ over 30 batches. Positive rank-biserial effects mean that the signed curvature tends to be higher on $\mathcal{M}$; negative effects mean that it tends to be lower.

The read-side matrices W_V, W_{gate}, and W_{up} have higher curvature on $\mathcal{M}$, whereas the write-side matrices W_O and W_{down} have lower curvature (Figure 2). For the strong-amplifier $\mathcal{M}$ rows of W_{down} in deep layers, the average signed curvature is negative. These rows therefore do not lie in an ordinary locally convex basin along the measured directions.

For the same W_{down} rows, we also report the radial projection

$$r_k = \nabla_{w_k} \mathcal{L}_{\mathrm{CE}} \cdot \hat{w}_k, \qquad \hat{w}_k = \frac{w_k}{\|w_k\|}. \tag{5}$$

The strong-amplifier coordinates largely have $r_k < 0$. This sign requires care: the gradient vector itself points inward, but a gradient-descent update uses the opposite direction and therefore points outward. By contrast, AdamW's decoupled weight-decay update $-\eta\lambda w_k$ always points inward in parameter space. The two tendencies oppose one another. Because Figure 5 uses the raw loss gradient rather than the complete preconditioned update, it establishes this local geometry but not an exact AdamW force balance.

G.2 AdamW update breakdown

We next measure the end effect of the optimizer on the activations. Let (m_t, v_t) be the saved AdamW state, let g be the new local loss gradient, and let $s = t + 1$ be the next optimizer step. With $c_1 = 1 - \beta_1^s$ and $c_2 = 1 - \beta_2^s$, define the adaptive update without weight decay as

$$F(g; m_t, v_t) = -\eta \frac{[\beta_1 m_t + (1-\beta_1)g]/c_1}{\sqrt{[\beta_2 v_t + (1-\beta_2)g^2]/c_2} + \epsilon}. \tag{6}$$

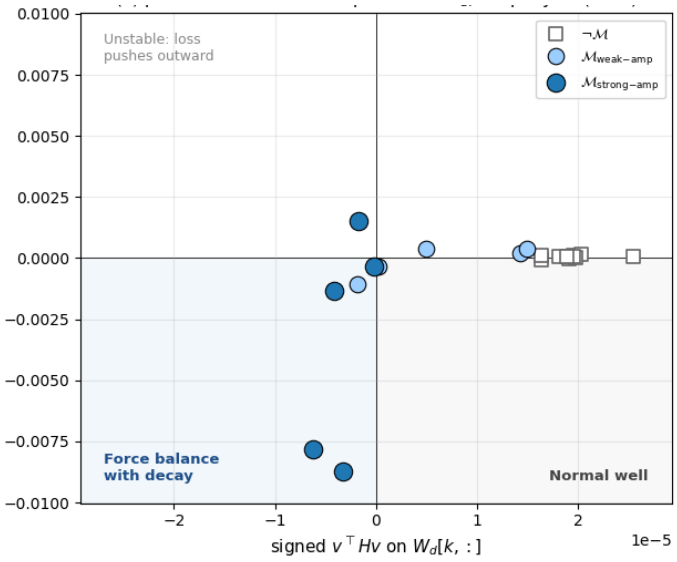


Figure 5: Curvature and radial loss-gradient projection for $W_{\text{down}}[k,:]$ in deep layers $L \in \{9, 12, 15\}$. Negative $\nabla\mathcal{L}_{\text{CE}} \cdot \hat{w}$ means that the gradient vector points inward, so the loss-descent direction points outward. Strong-amplifier $\mathcal{M}$ coordinates concentrate in the negative-curvature, negative-gradient quadrant, unlike the matched non-massive coordinates.

We decompose the next step into

$$\Delta\theta_{\text{state}} = F(0; m_t, v_t), \tag{7}$$

$$\Delta\theta_{g|\text{state}} = F(g; m_t, v_t) - F(0; m_t, v_t), \tag{8}$$

$$\Delta\theta_{\text{WD}} = -\eta\lambda\theta, \tag{9}$$

$$\Delta\theta_{\text{full}} = F(g; m_t, v_t) - \eta\lambda\theta. \tag{10}$$

The components add exactly:

$$\Delta\theta_{\text{full}} = \Delta\theta_{\text{state}} + \Delta\theta_{g|\text{state}} + \Delta\theta_{\text{WD}}. \tag{11}$$

This decomposition separates the contribution already stored in the optimizer state from the additional contribution of the current gradient and from decoupled weight decay.

G.3 Activation pressure

For coordinate k at the output of block l, define its activation magnitude on the selected token positions V_i of batch i as

$$A_k^{(l,i)} = \left(\sum_{u \in V_i} h_{u,k}^{(l,i)2} \right)^{1/2}. \tag{12}$$

For update component c, its local first-order pressure is

$$P_{k,c}^{(l,i)} = \left\langle \nabla_{\theta^{(l)}} A_k^{(l,i)}, \Delta\theta_c^{(l,i)} \right\rangle. \tag{13}$$

Positive pressure predicts an increase in activation magnitude under that component; negative pressure predicts a decrease. By linearity,

$$P_{\text{full}} = P_{\text{state}} + P_{g|\text{state}} + P_{\text{WD}}. \tag{14}$$

The implementation computes the Jacobian–vector product for each additive direction once and then evaluates all selected coordinates together.

Figure 6 shows the layerwise result. The full AdamW step has positive mean pressure on $\mathcal{M}$ at every layer, and the pressure generally grows in the deeper half of the model. The matched non-massive

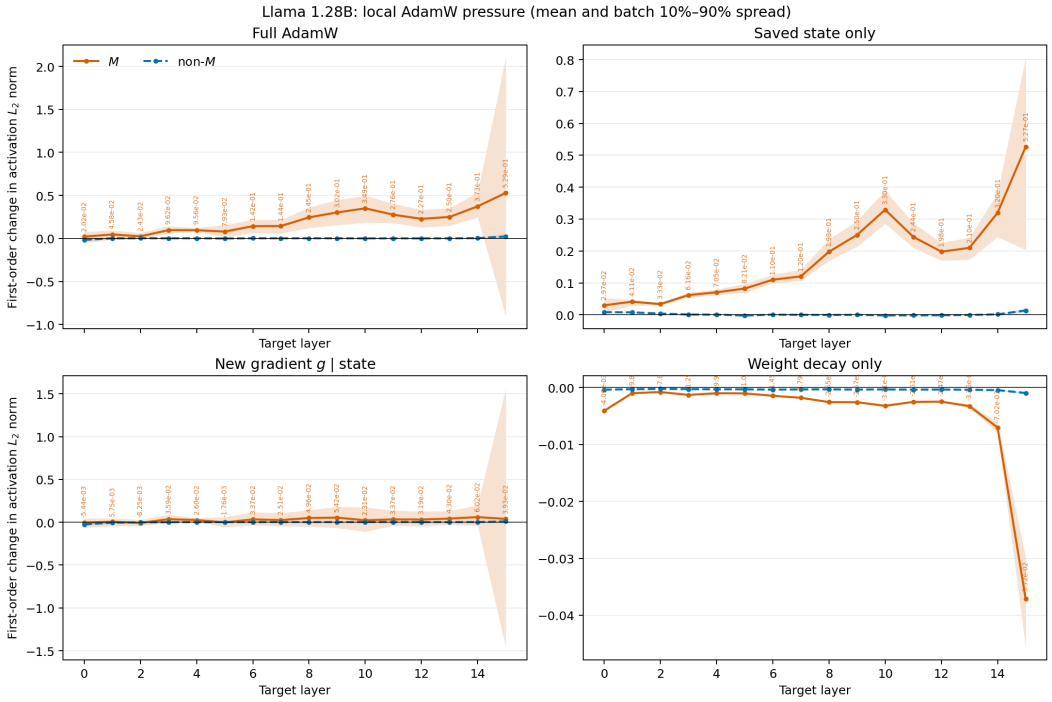


Figure 6: Local first-order AdamW pressure on activation magnitude in the Llama 1.28B model. Lines show the batch mean and shaded regions show the batch 10th–90th percentile spread. Orange denotes $\mathcal{M}$ and dashed blue denotes the matched non-massive control. The full AdamW step increases the magnitude of $\mathcal{M}$ while leaving the control near zero. This effect is dominated by the saved optimizer state. The current gradient conditioned on that state is smaller, and weight decay acts in the opposite direction.

control remains near zero. The saved optimizer state explains most of the positive pressure; the additional contribution of the current gradient is much smaller. Weight decay has negative pressure on $\mathcal{M}$, especially in the final layer. Thus, weight decay acts as a brake on massive activations, while the accumulated adaptive state more than compensates for it.

Scope. This is a local, first-order, one-step counterfactual at a fixed checkpoint. We update only the decoder block that produces the target activation and do not include prefix parameters. The result therefore shows the direction in which the next local AdamW step is poised to move the activations; it does not claim that every step throughout training increases them.

H Sensitivity to MA Selection Threshold

To evaluate sensitivity to the MA Selection threshold R discussed in section 3, we additionally repeat erasure analysis on the Llama 1.28B model using progressively stricter thresholds. Results are reported in table 8.

The main observations remain consistent across thresholds. The value, query, and FFN gate/up read projections continue to show strong positive separation between M and cM coordinates. At the more permissive threshold (R=2), the separation is slightly weaker. This is expected since M set here includes less-extreme coordinates. As the threshold increases, the read-side separation becomes stronger. Thus, the reported read-blind, write-open asymmetry is not specific to the choice and becomes more pronounced for more extreme massive activations.

Table 8: Erasure results at different MA selection thresholds for Llama 1.28B base model. Cells report the median $\mathrm{era}_{\mathcal{M}}$ with rank-biserial r_b in parentheses. Bold cells have $r_b > 0.95$; stars denote one-sided Mann–Whitney tests of $\mathrm{era}_{\mathcal{M}} > \mathrm{era}_{\neg\mathcal{M}}$ with $p < 0.05$.

	R=2	R=5	R=10	R=20
Massive coordinates $\lvert\mathcal{M}\rvert$	15	10	6	4
Maximum activation magnitude	2.27×10^3	2.27×10^3	2.27×10^3	2.27×10^3
Input availability:				
Attention normalized input, $\tilde{X}$	0.68 (-0.66)	0.68 (-0.73)	0.62 (-0.74)	0.61 (-0.77)
FFN normalized input, $\tilde{Y}$	0.56 (-0.14)	0.56 (-0.13)	0.56 (-0.04)	0.56 (-0.12)
Read side weight view:				
Value projection W_V	0.72 (+0.81)*	**0.75 (+0.99)***	**0.83 (+1.00)***	**0.86 (+1.00)***
Query projection W_Q	0.63 (+0.84)*	**0.65 (+0.99)***	**0.69 (+1.00)***	**0.71 (+1.00)***
Key projection W_K	0.54 (+0.32)*	0.55 (+0.31)*	0.53 (-0.04)	0.52 (-0.08)
FFN gate/up $W_{\mathrm{gate}}, W_{\mathrm{up}}$	**0.61 (+0.96)***	**0.63 (+1.00)***	**0.65 (+1.00)***	**0.69 (+1.00)***
Write side weight view:				
Attention output W_O	0.52 (-0.31)	0.51 (-0.61)	0.50 (-0.98)	0.42 (-0.99)
FFN down projection W_{down}	0.50 (-0.57)	0.49 (-0.64)	0.49 (-0.96)	0.46 (-1.00)
Write side operator view:				
Attention update Δ^{attn}	0.86 (-0.34)	0.85 (-0.44)	0.86 (-0.34)	0.67 (-0.50)
FFN residual update Δ^{ffn}	0.69 (-0.80)	0.69 (-0.77)	0.67 (-0.90)	0.66 (-0.97)