MAGNITUDE-INDUCED RANK COLLAPSE IN FFN VALUE-VECTOR GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Each column of the FFN down-projection W_{down} maps an intermediate channel to the residual stream in a Transformer model. How individual training sequences contribute to updating these columns is obscured by minibatch aggregation and model- or layer-level gradient summaries. We therefore characterize each column's per-sequence gradient matrix using normalized energy effective rank (NEER), which responds to both directional correlation and gradient-norm imbalance. We first observe that columns within the same layer exhibit a pronounced low-rank tail across optimization, showing that sequence contributions can differ sharply across parameter vectors. We further find that low NEER often reflects gradient-magnitude imbalance rather than directional alignment, with many low-NEER columns retaining substantial directional diversity. To trace the source of these outliers, we use the standard activation–backpropagation factorization. It reveals that large column-specific activations and strong residual-space backpropagated signals coincide at the same token positions, whereas neither factor alone explains the concentration. Together with a simplified optimization analysis, these findings suggest that high-magnitude sequence gradients in some columns can suppress otherwise diverse directions and limit their contribution to optimization. Motivated by this analysis, we test parameter-local selective clipping of dominant sequence gradients in low-NEER columns. Across Llama-3.1-8B, Llama-3.2-3B, and Mistral-7B, selectively rebalancing dominant local contributions yields consistent downstream gains, suggesting that reducing parameter-local gradient imbalance can be beneficial during fine-tuning.

1 INTRODUCTION

Transformer feed-forward networks (FFNs) contain a large fraction of model parameters and have been interpreted as key–value memories (Geva et al., 2021). Their computations have further been implicated in factual association recall and shown to exhibit parameter specialization (Meng et al., 2022; Hong et al., 2025). Each column of W_{down} defines the residual-space output vector of one intermediate channel, making it a natural unit for parameter-local optimization analysis.

Optimization signals, however, are usually examined at a granularity coarser than this structure. Standard minibatch optimization aggregates gradients across sequences before the parameter update is formed, obscuring potentially large differences among individual sequence contributions. Per-example gradient analyses recover this sample-level variation, but often represent each example through a model-wide gradient object (Koh & Liang, 2017; Park et al., 2023; Xia et al., 2024). More structured analyses retain layer- or projection-level information and reveal substantial variation across parameter groups (Li et al., 2025; 2026; Yao et al., 2026), yet still aggregate many individual parameter vectors. Consequently, how different sequences contribute to the same local parameter vector—and how this structure varies across vectors within a layer—remains less understood.

To examine this mismatch more finely, we retain sample and parameter locality by forming a per-sequence gradient matrix for each W_{down} column. We summarize its geometry using the *normalized energy effective rank* (NEER), an energy-weighted and normalized variant of entropy-based effective rank (Roy & Vetterli, 2007). Because it is determined by the singular-value energy spectrum, NEER is sensitive both to how sequence gradients are oriented and to how unevenly their magnitudes are distributed. Figure 1 summarizes this parameter-local perspective and our main findings.

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.

Across Llama and Mistral models, mathematical and code instruction-tuning data, and multiple optimization stages, we observe strong within-layer heterogeneity and persistent low-NEER tails. The key question is what this low rank represents. Although low effective rank is naturally associated with a small number of dominant directions, many low-NEER columns in our experiments remain far from directionally collapsed: their RMS cosine similarities are low, and equalizing sequence-gradient norms produces a large increase in NEER. An exact decomposition of $-\log \mathrm{NEER}$ further separates sequence-energy concentration from directional correlation and shows that magnitude concentration accounts for a substantial part of the rank deficit in many low-NEER columns.

We next ask why a small number of sequences acquire such large local gradient magnitudes. For a fixed W_{down} column, its sequence gradient can be written as a tokenwise sum of residual-space backpropagated signals weighted by the corresponding channel activations, following the standard activation–backpropagation factorization (Martens & Grosse, 2015; Dangel et al., 2020). This decomposition lets us separate sequence-level activation and backpropagation magnitudes from their actual tokenwise pairing. The paired signal consistently explains the exact sequence-energy allocation better than the corresponding marginal null, showing that the fine-grained token correspondence carries information beyond either quantity considered separately.

This parameter-local structure is only partially preserved by coarser layer-level summaries. Layer-level sequence-energy rankings recover broad sequence-contribution trends but miss a substantial fraction of the dominant-sequence relations in low-NEER columns, which are also disproportionately represented in the optimizer-visible gradient energy. This suggests that diverse gradient directions can exist locally while their relative influence is strongly skewed by a few high-magnitude sequence contributions. We therefore test a parameter-local intervention that selectively clips these dominant contributions while preserving their directions. During mathematical fine-tuning, selective local clipping improves downstream reasoning performance on Llama-3.1-8B, Llama-3.2-3B, and Mistral-7B. We interpret this result as evidence that parameter-local contribution balance is a useful optimization perspective; a simplified analysis is provided in Appendix E.3.

- We introduce a parameter-local view of cross-sequence gradient geometry by constructing a per-sequence gradient matrix for each W_{down} column, revealing pronounced within-layer NEER heterogeneity across columns.
- We decompose the NEER deficit into sequence-energy concentration and directional correlation, showing that magnitude imbalance, rather than directional collapse, drives the low-rank regime for many columns.
- We trace anomalous gradient magnitudes to token-level pairing between column-specific FFN activations and backpropagated signals shared across columns.
- We show that coarse layer-level statistics miss energetically significant column-local dominance and use this structure to motivate parameter-local selective clipping, which improves downstream mathematical reasoning on Llama-3.1-8B, Llama-3.2-3B, and Mistral-7B.

2 Related work

Transformer FFNs and parameter-local structure. Interpretability studies characterize FFNs as key–value memories and their outputs as activation-weighted sub-updates associated with columns of the second FFN matrix (Geva et al., 2021; 2022). Neuron- and editing-based studies connect FFNs to factual recall (Dai et al., 2022; Meng et al., 2022). These results make an FFN column a meaningful local unit but do not address its cross-sequence gradient geometry. Activation-outlier studies document large activations (Sun et al., 2024; Lin et al., 2024), whereas our mechanism depends on their token-level coupling with backpropagated signals. Coherent Gradients applies coordinate-wise winsorization to per-example gradients (Chatterjee, 2020), while general analyses characterize clipping bias and convergence (Qian et al., 2021; Koloskova et al., 2023). Our diagnostic instead rescales whole per-sequence gradients in selected low-NEER columns, preserving their directions.

Per-example gradients for attribution and data selection. Influence functions estimate the effect of perturbing a training example (Koh & Liang, 2017). TracIn and TRAK approximate training-data attribution with checkpointed or projected gradients (Pruthi et al., 2020; Park et al., 2023). For language-model post-training, DataInf approximates influence in LoRA-tuned models, LESS selects

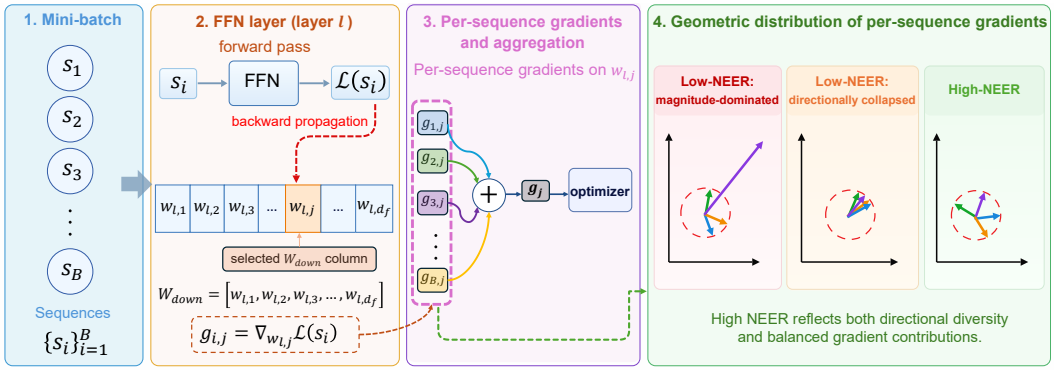


Figure 1: Overview of our parameter-local gradient analysis. For a minibatch, we select one column of W_{down} and retain its per-sequence gradient contributions before optimizer aggregation. Their local geometry distinguishes two sources of low NEER: magnitude dominance by a few sequences despite directional diversity, and directional concentration. High NEER instead reflects both balanced gradient contributions and directional diversity.

instruction data using optimizer-aware gradient similarity, and GAST jointly selects examples and layers (Kwon et al., 2024; Xia et al., 2024; Yao et al., 2026). These methods primarily score data, optionally conditioned on layers. We instead fix a layer and parameter vector and study how an entire minibatch is represented in its gradient geometry.

Gradient spectra and effective rank. A ratio of full-model per-example gradient norms has been related to scalable mini-batching, with a differential variant used for stability (Yin et al., 2018). Entropy-based effective rank instead summarizes a singular-value spectrum (Roy & Vetterli, 2007). It has also been applied to per-example gradients and continual trainability (Lewandowski et al., 2025). Related work studies layer-wise LLM gradients, low-rank weight gradients in two-layer networks, and last-layer gradient scatter under label noise (Li et al., 2026; Sonthalia et al., 2025; Bathula & Joshi, 2026). These analyses use model- or layer-level objects. Our column-local NEER distinguishes energy concentration from direction correlation and, unlike Yin et al.'s ratio, detects norm imbalance even among orthogonal gradients.

3 Parameter-Local Gradient Geometry

We now formalize the parameter-local gradient construction illustrated in Figure 1.

3.1 Per-sequence gradients of an FFN output column

Consider layer ℓ of a gated Transformer FFN. For sequence s and token position t, let $a^{\ell}_{s,t} \in \mathbb{R}^{d_f}$ denote the intermediate FFN activation, where d_f is the FFN intermediate dimension, and let

$$u^{\ell}_{s,t} = W^{\ell}_{\text{down}} a^{\ell}_{s,t}, \qquad W^{\ell}_{\text{down}} = [w_{\ell,1}, \ldots, w_{\ell,d_f}] \in \mathbb{R}^{d \times d_f}, \tag{1}$$

where d is the model hidden dimension. Here, $u^{\ell}_{s,t} \in \mathbb{R}^{d}$ is the FFN contribution written back to the residual stream. Our analysis starts from the gradient that is actually passed to the optimizer. For a minibatch of B sequences, let $\mathcal{T}^{\text{loss}}_{s}$ denote the token positions of sequence s on which the supervised training loss is evaluated, and let $\mathcal{V}_s$ denote its valid forward positions (i.e., non-padding positions). These sets need not coincide: under response-only supervision, prompt positions are excluded from the loss but can still receive backpropagated signal through later supervised tokens.

Let

$$N = \sum_{s=1}^{B} |\mathcal{T}^{\text{loss}}_{s}| \tag{2}$$

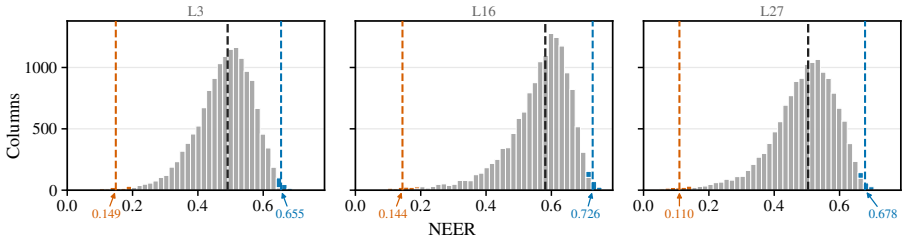


Figure 2: Column-wise NEER distributions of W_{down} in Llama-3.1-8B on MetaMathQA at layers 3, 16, and 27. Orange and blue dashed lines mark the median NEER of the bottom- and top-1% groups, respectively, and the black dashed line marks the overall median. Additional minibatches are reported in Appendix B.2.

denote the total number of loss-bearing tokens in the minibatch. Under the token-averaged training objective, the minibatch loss is

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{N}\sum_{s=1}^{B}\widetilde{\mathcal{L}}_s, \qquad \widetilde{\mathcal{L}}_s = \sum_{t\in\mathcal{T}_s^{\text{loss}}}\ell_{s,t}, \tag{3}$$

where $\ell_{s,t}$ is the token-level loss and $\widetilde{\mathcal{L}}_s$ is the unnormalized loss summed over the loss-bearing tokens of sequence s. We then define

$$g_{s,\ell,i} := \frac{1}{N}\nabla_{w_{\ell,i}}\widetilde{\mathcal{L}}_s$$

as sequence s's contribution to the gradient of column $w_{\ell,i}$ under the token-averaged minibatch objective. With this definition, the minibatch gradient of the column decomposes as

$$\nabla_{w_{\ell,i}}\mathcal{L}_{\mathcal{B}} = \sum_{s=1}^{B} g_{s,\ell,i}.$$

For a minibatch of B sequences, we collect the gradients acting on one column into

$$X_{\ell,i} = [g_{1,\ell,i}, \ldots, g_{B,\ell,i}] \in \mathbb{R}^{d\times B}. \tag{4}$$

3.2 Normalized Energy Effective Rank

Let $\sigma_{\ell,i,1}, \ldots, \sigma_{\ell,i,B}$ denote the singular values of $X_{\ell,i}$, and define the normalized singular-value energies

$$p_{\ell,i,j} = \frac{\sigma_{\ell,i,j}^2}{\sum_k \sigma_{\ell,i,k}^2}. \tag{5}$$

We summarize this spectrum by the *normalized energy effective rank* (NEER),

$$\text{NEER}(X_{\ell,i}) = \frac{\exp\Big(-\sum_{j:p_{\ell,i,j}>0} p_{\ell,i,j}\log p_{\ell,i,j}\Big)}{B}. \tag{6}$$

Our experiments use $B \le d$, so $\text{NEER}(X_{\ell,i}) \in [1/B, 1]$. This is an energy-weighted variant of entropy-based effective rank (Roy & Vetterli, 2007). NEER equals one when all sequence gradients are orthogonal and have equal norm, and approaches $1/B$ when the spectrum is dominated by a single mode. Importantly, such dominance may arise either from directional concentration or from severe gradient-magnitude imbalance, so low NEER alone does not imply a shared update direction.

3.3 Within-Layer Heterogeneity in NEER

Figure 2 shows substantial within-layer variation in NEER, including pronounced low-NEER tails, despite all columns sharing the same minibatch and training objective. We use the layers at approximately 10%, 50%, and 80% relative depth as representative shallow, middle, and deep layers for all models (layers 3, 16, and 27 for Llama-3.1-8B and Mistral-7B, and layers 3, 14, and 22 for Llama-3.2-3B). Unless otherwise specified, main-text analyses use MetaMathQA with $B = 48$ sequences. Sample-size ablations are provided in Appendix B.4.

Table 1: Column-wise NEER statistics across training stages and base models. **(a)** Different fine-tuning steps. **(b)** Different base models. Values are mean ± standard deviation across all FFN layers.

(a) Across training steps

Step	Bottom 1%	Overall	Overall /Bottom
Base	0.148 ± 0.063	0.526 ± 0.059	$(4.095 \pm 1.498)\times$
500	0.090 ± 0.027	0.464 ± 0.019	$(5.662 \pm 1.916)\times$
1000	0.090 ± 0.026	0.457 ± 0.024	$(5.507 \pm 1.879)\times$
1500	0.097 ± 0.027	0.481 ± 0.016	$(5.452 \pm 2.090)\times$
2000	0.096 ± 0.026	0.476 ± 0.016	$(5.438 \pm 2.048)\times$

(b) Across base models

Model	Bottom 1%	Overall	Overall /Bottom
Llama-3.1-8B	0.148 ± 0.063	0.526 ± 0.059	$(4.095 \pm 1.498)\times$
Llama-3.2-3B	0.167 ± 0.076	0.575 ± 0.047	$(4.070 \pm 1.516)\times$
Mistral-7B	0.130 ± 0.068	0.561 ± 0.042	$(5.448 \pm 2.704)\times$

Within-layer rank heterogeneity. For subsequent analyses, we define the bottom and top 1% of the within-layer NEER distribution in each setting as contrastive groups.

Table 1 shows that the low-NEER tail persists across fine-tuning stages and across Llama and Mistral base models. The Overall/Bottom ratio exhibits relatively large variation across layers, partly because layer 0 consistently shows a substantially larger ratio than later layers; full layer-wise statistics are provided in Appendix B.1.

Unless otherwise specified, comparisons within each setting use the same minibatch. Repeating the analysis over independently sampled minibatches recovers the same broad NEER heterogeneity, although the identities of the extreme columns are not stable across batches. Our claims concern this distributional structure rather than persistent column identity; batch-wise stability is analyzed in Appendix B.3. Additional results on CodeFeedback are reported in Appendix B.5. Together, these results indicate that the parameter-local heterogeneity is not specific to a single checkpoint, model family, or data domain. We next ask whether the low-NEER tail arises from unequal sequence energies, correlated gradient directions, or both.

4 Magnitude-Induced Effective-Rank Collapse

The low-NEER tail identified above has two geometrically distinct explanations. Sequence gradients may concentrate around a few shared directions, or they may remain directionally diverse while a few sequences dominate through much larger norms. We next separate the contributions of gradient magnitude and direction to the observed effective-rank collapse.

4.1 Separating energy concentration from direction correlation

For this decomposition, let $K_{\ell,i} = X_{\ell,i}^\top X_{\ell,i}$ be the Gram matrix of the per-sequence gradients for column i at layer ℓ. Its entries encode pairwise inner products between sequence gradients, while its diagonal entries give their squared norms. We define

$$e_{s,\ell,i} = \|g_{s,\ell,i}\|_2^2, \qquad q_{s,\ell,i} = \frac{e_{s,\ell,i}}{\sum_{r=1}^{B} e_{r,\ell,i}}, \tag{7}$$

where $e_{s,\ell,i}$ is the gradient energy contributed by sequence s, and $q_{s,\ell,i}$ is its normalized share of the total sequence-gradient energy for that column. Thus, $q_{\ell,i}$ describes how the column's gradient energy is distributed across sequences. Since $q_{\ell,i}$ and $p_{\ell,i}$ are respectively the normalized diagonal and eigenvalue spectrum of $K_{\ell,i}$, the NEER deficit admits the exact decomposition

$$\begin{aligned} -\log \mathrm{NEER}(X_{\ell,i}) &= \log B - H(p_{\ell,i}) \\ &= \underbrace{\log B - H(q_{\ell,i})}_{C^E_{\ell,i}\text{: energy concentration}} + \underbrace{H(q_{\ell,i}) - H(p_{\ell,i})}_{C^D_{\ell,i}\text{: direction correlation}}, \end{aligned} \tag{8}$$

where H denotes Shannon entropy. Both terms are nonnegative by the Schur–Horn theorem, as shown in Appendix C.1. The energy term $C^E_{\ell,i}$ measures the departure of sequence energies from a uniform allocation and vanishes when all gradients have the same norm. The direction term $C^D_{\ell,i}$ measures the additional spectral concentration contributed by cross-sequence inner products and vanishes when the gradients are mutually orthogonal. Equation 8 therefore distinguishes magnitude dominance from directional collapse without changing the observed gradients.

Table 2: Diagnostics for the low-NEER group across representative layers. $C^E/(C^E + C^D)$, NEER-deficit reduction, and RMS cosine are reported as median $[Q_1, Q_3]$ over low-NEER columns. NEER change reports the median before and after norm equalization, with the difference between the two medians shown in parentheses.

Layer	$\frac{C^E}{C^E + C^D}$	NEER change	Deficit reduction	RMS cosine
3	0.591 [0.446, 0.731]	0.149 → 0.728 (↑ 0.579)	84.4% [74.5, 89.5]	0.178 [0.143, 0.216]
16	0.693 [0.579, 0.786]	0.144 → 0.832 (↑ 0.688)	91.1% [84.4, 93.7]	0.116 [0.094, 0.148]
27	0.573 [0.439, 0.691]	0.110 → 0.699 (↑ 0.589)	83.9% [76.6, 90.0]	0.169 [0.133, 0.219]

The additive decomposition is complemented by an intervention that preserves directions exactly. For every nonzero sequence gradient, we define

$$\widehat{g}_{s,\ell,i} = \frac{g_{s,\ell,i}}{\|g_{s,\ell,i}\|_2}, \qquad \widehat{X}_{\ell,i} = [\widehat{g}_{1,\ell,i}, \ldots, \widehat{g}_{B,\ell,i}]. \tag{9}$$

The resulting $\mathrm{NEER}(\widehat{X}_{\ell,i})$ is a direction-only effective rank: norm equalization removes sequence-energy imbalance while retaining every pairwise angle. We quantify the corresponding recovery through the fractional reduction in NEER deficit. We define

$$D^{\mathrm{raw}}_{\ell,i} = -\log \mathrm{NEER}(X_{\ell,i}), \qquad D^{\mathrm{norm}}_{\ell,i} = -\log \mathrm{NEER}(\widehat{X}_{\ell,i}). \tag{10}$$

For columns with $D^{\mathrm{raw}}_{\ell,i} > 0$, we report the deficit reduction $R^{\mathrm{def}}_{\ell,i} = (D^{\mathrm{raw}}_{\ell,i} - D^{\mathrm{norm}}_{\ell,i})/D^{\mathrm{raw}}_{\ell,i} \times 100\%$. A large positive value means that norm imbalance accounts for much of the observed departure from full effective rank. Unlike $C^E_{\ell,i}$, this quantity is an explicit counterfactual: it recomputes the spectrum after replacing the observed norms by a common value.

4.2 Equalizing sequence norms largely restores effective rank

Table 2 and Figure 3(a,b) consistently show that magnitude imbalance is a major source of the low-NEER tail. Across the representative layers, the median deficit reduction within each layer ranges from 83.9% to 91.1% after norm equalization, which substantially restores effective rank, while low RMS cosine similarities indicate that substantial directional diversity is retained. The middle-NEER group is an equally sized set centered at the within-layer NEER median; the same grouping convention is used in subsequent analyses unless otherwise specified. After removing norm differences, the low-, middle-, and high-NEER groups also exhibit much more similar direction-only geometry. Thus, directional concentration alone cannot explain the observed low-NEER tail.

4.3 Low-NEER columns concentrate energy on a few sequences

The same mechanism is visible directly in the sequence-energy distribution. Let $q_{(1),\ell,i} \geq \cdots \geq q_{(B),\ell,i}$ denote the energy shares from Equation 7 sorted in descending order, and define the cumulative energy captured by the largest k sequences as $A_{\ell,i}(k) = \sum_{j=1}^{k} q_{(j),\ell,i}$. Figure 3(c) shows that cumulative energy rises substantially faster for low-NEER columns across the representative layers. Thus, a small number of sequences account for a large fraction of the gradient energy in the low-NEER group, whereas high-NEER columns distribute their energy more broadly across sequences.

Together, the three diagnostics support a magnitude-induced component of effective-rank collapse. Low-NEER columns exhibit strong sequence-energy imbalance despite retaining substantial directional diversity, and removing norm differences largely restores their effective rank. The cumulative-energy analysis further localizes this imbalance to a small set of dominant sequences. We next trace the origin of these anomalous magnitudes to the token-level coupling between FFN activations and backpropagated signals during optimization.

5 Origins of Gradient-Magnitude Dominance

This raises a natural question: why can a sequence that contributes a typical update to one parameter vector dominate another in the same FFN layer? We address this by decomposing each sequence

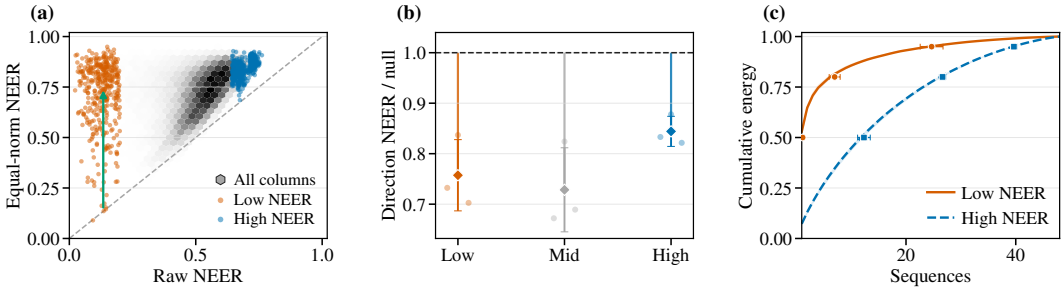


Figure 3: Magnitude and direction diagnostics across three representative layers of Llama-3.1-8B. **(a)** Raw versus norm-equalized NEER; the green arrow indicates the average change in the low-NEER median across layers. **(b)** Direction-only NEER relative to the matched random-direction baseline. Small points denote layer-wise group medians, while diamonds and error bars show the across-layer mean $\pm$ sample standard deviation; vertical lines connect the mean to the random-direction baseline at 1. **(c)** Cumulative gradient-energy concentration. Curves are pointwise averages of the layer-wise median curves, with shaded regions showing the across-layer IQR. Markers and horizontal error bars at 50%, 80%, and 95% energy report the across-layer mean $\pm$ standard deviation of the required number of sequences.

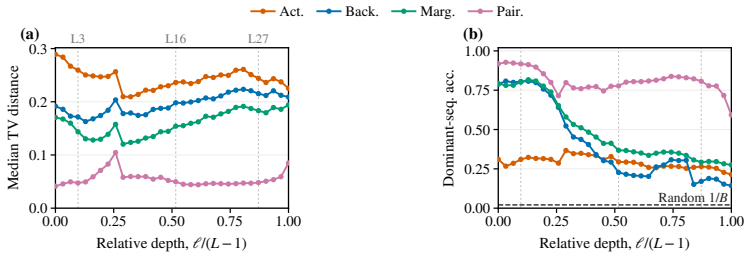


Figure 4: Evaluation of sequence-energy proxies across all layers of Llama-3.1-8B for low-NEER columns. **(a)** Total-variation distance to the exact sequence-energy allocation (lower is better). **(b)** Dominant-sequence identification accuracy (higher is better); the black dashed line denotes the random-selection baseline $1/B$. Vertical dashed lines indicate the representative layers.

gradient into the signal propagated to the FFN output and the activation that routes this signal to a particular down-projection column.

5.1 Activation–Backpropagated-Signal Coupling

Recall that $a^{\ell}_{s,t} \in \mathbb{R}^{d_f}$ denotes the intermediate FFN activation and $u^{\ell}_{s,t} = W^{\ell}_{\text{down}} a^{\ell}_{s,t} \in \mathbb{R}^{d}$ is the corresponding FFN output. We define the backpropagated residual-space signal as $\delta^{\ell}_{s,t} := N^{-1} \nabla_{u^{\ell}_{s,t}} \widetilde{\mathcal{L}}_s \in \mathbb{R}^{d}$. The sequence contribution to the ith W_{down} column then factorizes as

$$g_{s,\ell,i} = \sum_{t \in \mathcal{V}_s} a^{\ell}_{s,t,i}\, \delta^{\ell}_{s,t}, \tag{11}$$

where $\mathcal{V}_s$ denotes the valid forward positions of sequence s. This is the standard activation–backpropagation factorization of a linear-layer gradient (Martens & Grosse, 2015; Dangel et al., 2020). At a fixed layer and token, $\delta^{\ell}_{s,t}$ is shared across all down-projection columns, whereas $a^{\ell}_{s,t,i}$ determines how strongly that signal is routed to column i.

Equation 11 exposes a parameter–sample interaction that is invisible to a sequence-level score. A large activation does not necessarily produce a large column gradient if the associated backpropagated signal is weak, and a strong backpropagated signal has little effect on a column whose activation is small. Gradient-magnitude dominance can therefore depend on how the two factors are paired across tokens rather than on either factor alone.

5.2 Explaining sequence-energy concentration

To test this mechanism, we expand the exact energy of a sequence gradient:

$$\begin{aligned} E^{\mathrm{Exact}}_{s,\ell,i} &:= \|g_{s,\ell,i}\|_2^2 \\ &= \sum_{t\in\mathcal{V}_s} \left(a^{\ell}_{s,t,i}\right)^2 \|\delta^{\ell}_{s,t}\|_2^2 + \sum_{\substack{t,t'\in\mathcal{V}_s \\ t\neq t'}} a^{\ell}_{s,t,i} a^{\ell}_{s,t',i} \langle \delta^{\ell}_{s,t}, \delta^{\ell}_{s,t'} \rangle. \end{aligned} \tag{12}$$

The first term depends only on the observed pairing between activation and backpropagated-signal magnitudes at the same token. The second contains cross-token directional interactions. We construct four energy proxies that differ in the information they retain, where all sums below are over $t \in \mathcal{V}_s$:

$$\begin{aligned} E^{\mathrm{Act}}_{s,\ell,i} &= \sum_t (a^{\ell}_{s,t,i})^2, & E^{\mathrm{Back}}_{s,\ell,i} &= \sum_t \|\delta^{\ell}_{s,t}\|_2^2, \\ E^{\mathrm{Marg}}_{s,\ell,i} &= \frac{E^{\mathrm{Act}}_{s,\ell,i} E^{\mathrm{Back}}_{s,\ell,i}}{|\mathcal{V}_s|}, & E^{\mathrm{Pair}}_{s,\ell,i} &= \sum_t (a^{\ell}_{s,t,i})^2 \|\delta^{\ell}_{s,t}\|_2^2. \end{aligned} \tag{13}$$

The activation-only and backpropagated-signal-only proxies isolate the two factors separately. The marginal-product proxy combines their sequence-level magnitudes while removing token-level pairing; equivalently, it is the expected paired energy under a uniform within-sequence permutation of one factor. The paired proxy preserves the observed tokenwise correspondence while omitting cross-token directional interactions. Comparing E^{Pair} with E^{Marg} therefore isolates the additional information provided by the observed token-level magnitude pairing.

For each proxy m, we normalize its energies across sequences as $p^{m}_{s,\ell,i} = E^{m}_{s,\ell,i} / \sum_{r=1}^{B} E^{m}_{r,\ell,i}$, and compare this allocation with the exact one, p^{Exact}. We evaluate the match using the per-column total-variation (TV) distance between the two energy distributions and the accuracy of identifying the exact dominant sequence. Lower TV and higher accuracy indicate a closer match; random selection gives expected dominant-sequence accuracy $1/B$.

5.3 Token-level pairing explains the dominant energy allocation

Figure 4 evaluates the energy proxies across all FFN layers. Preserving the observed token-level pairing consistently outperforms the marginal null across depth. On the three representative layers, relative to E^{Marg}, E^{Pair} reduces the per-layer median TV distance by 0.112 on average, yielding a mean median TV distance of 0.048 and a dominant-sequence accuracy of 0.834. Thus, sequence-level activation and backpropagated-signal magnitudes alone do not explain the observed energy allocation; their fine-grained tokenwise pairing provides substantial additional information.

We additionally perform a conditional within-sequence vector shuffle that breaks the fine-grained correspondence between activations and backpropagated signals while controlling for coarse position and supervision status. Across the representative layers, the shuffle reduces the original–shuffled NEER rank correlation to 0.263–0.521 and preserves the dominant sequence for only 23.5%–63.3% of columns. Full details and additional diagnostics are provided in Appendix D.

This coupling also explains the parameter-local nature of the phenomenon. Backpropagated signals are shared across W_{down} columns at a fixed token, whereas activations are column-specific and therefore route these signals differently. A sequence can consequently dominate some parameter vectors without dominating the entire layer.

6 Optimization implications of magnitude dominance

The preceding results show that low-NEER columns often retain directional diversity despite strong sequence-level magnitude imbalance. Before intervening on this structure, we ask whether it is visible from coarser layer-level gradient statistics. We rank sequences by their leave-one-column-out layer-level gradient energy and test whether this ranking recovers the dominant sequence of individual W_{down} columns. Layer-level statistics retain broad predictive information, but systematically miss much of the low-NEER structure: the top 10% of layer-ranked sequences recover only 38.7%

Table 3: Downstream performance with and without selective local clipping. All values are accuracies (%). Improvements and decreases relative to the corresponding baseline are shown in blue and red, respectively.

Model	Method	GSM8K	MATH-500	SVAMP	AQuA-RAT	Avg.
Llama-3.1-8B	Baseline	77.71	26.80	79.30	32.28	54.02
	Clip	78.85	27.40	79.40	34.65	55.08
Llama-3.2-3B	Baseline	65.50	16.00	65.90	29.13	44.13
	Clip	65.66	17.20	66.80	31.10	45.19
Mistral-7B	Baseline	71.42	22.00	73.90	30.31	49.41
	Clip	71.65	23.60	74.30	29.13	49.67

of low-NEER dominant-sequence relations, compared with 70.1% and 70.1% for the middle- and high-NEER groups. Full agreement diagnostics are reported in Appendix E.1 and Table 16.

Importantly, these local structures are not confined to columns with negligible gradients. Although the lowest-NEER group contains only 1% of W_{down} columns, it accounts for $4.8\% \pm 1.8\%$ of the optimizer-aggregated gradient energy across the representative layers, representing a $4.73\times$ enrichment over its column fraction. Thus, coarse layer-level statistics capture the overall trend in sequence contributions but can miss energetically consequential dominance at the column level.

Together with the directional diagnostics in Section 4, this suggests a simple optimization motivation: many low-NEER columns already contain diverse gradient directions, but a few large-magnitude sequence contributions can dominate their aggregation and suppress the relative contribution of the remaining directions. We therefore test a parameter-local rebalancing intervention that selectively reduces these dominant contributions while preserving their directions.

At each intervention step, we rank the W_{down} columns within each layer by NEER and select the lowest fraction ρ for rebalancing. Let $\mathcal{C}_\ell$ denote the selected columns, and let $\mathcal{A}_{\ell,i}$ denote the indices of the k sequences with the largest gradient norms for column i in layer ℓ. For each $i \in \mathcal{C}_\ell$, we clip these selected gradients to the within-column median norm $m_{\ell,i} = \text{median}_s \|g_{s,\ell,i}\|_2$:

$$\widetilde{g}_{s,\ell,i} = \min\left(1, \frac{m_{\ell,i}}{\|g_{s,\ell,i}\|_2}\right) g_{s,\ell,i}, \qquad i \in \mathcal{C}_\ell, \quad s \in \mathcal{A}_{\ell,i}, \tag{14}$$

while all other contributions are left unchanged. Unless otherwise specified, we use $\rho = 5\%$ and $k = 1$. A simplified analysis in Appendix E.3 shows that, under suitable conditions, reducing sequence-level energy concentration simultaneously increases parameter-local NEER and improves the alignment of the aggregated update with a balanced target direction.

We evaluate selective local clipping on Llama-3.1-8B, Llama-3.2-3B, and Mistral-7B using MetaMathQA. Table 3 reports the downstream results. Since the current implementation requires explicit parameter-local per-sequence statistics, we view it primarily as a mechanism test for parameter-local contribution rebalancing rather than as a computationally optimized training algorithm.

7 Conclusion

We identify a parameter-local source of low-rank gradient structure during language-model fine-tuning. For many FFN down-projection columns, low NEER is driven by a few sequences with large gradient magnitudes, despite substantial directional diversity. This dominance arises from token-level coupling between column-specific activations and shared backpropagated signals. Selective local clipping rebalances dominant sequence contributions and improves downstream mathematical reasoning, supporting the view that directional diversity in low-NEER gradients can be underutilized during aggregation. More broadly, our results show that coarse gradient statistics can miss sample–parameter structure and motivate parameter-local contribution balance as a useful perspective on optimization. Future work should develop efficient online surrogates for local dominance and test whether this principle extends beyond FFN columns and supervised fine-tuning.

AI USE STATEMENT

During this research, LLMs were used to assist with the review of relevant literature, parts of the mathematical derivations and proofs, the design of supplementary experiments building on the established core results, and language polishing of the manuscript. The core methodology and experiments of this work were independently conceived, developed, and carried out by the authors.

REFERENCES

Satwik Bathula and Anand A. Joshi. Fisher rank inflation: A spectral signature of memorization under label noise. *arXiv preprint arXiv:2607.12438*, 2026. doi: 10.48550/arXiv.2607.12438. URL https://arxiv.org/abs/2607.12438.

Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryeFY0EFwS.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Felix Dangel, Frederik Kunstner, and Philipp Hennig. BackPACK: Packing more into backprop. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJlrF24twB.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Yihuai Hong, Yiran Zhao, Wei Tang, Yang Deng, Yu Rong, and Wenxuan Zhang. The rise of parameter specialization for knowledge storage in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 98940–98965. Curran Associates, Inc., 2025. doi: 10.52202/085713-3305. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8e8388eca45355776f164f00b84c3815-Abstract-Conference.html.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363. PMLR, 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Y Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations*, pp. 21921–21942, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e84a0f233611a1dc8fb794dc52415a3-Abstract-Conference.html.

Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, András György, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C. Machado. Learning continually by spectral regularization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5565ab682d6c7f8d9da34ba0919974b0-Abstract-Conference.html.

Ming Li, Yanhong Li, and Tianyi Zhou. What happened in LLMs layers when trained for fast vs. slow thinking: A gradient perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32017–32154. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1545. URL https://aclanthology.org/2025.acl-long.1545/.

Ming Li, Yanhong Li, Ziyue Li, and Tianyi Zhou. How instruction and reasoning data shape post-training: Data quality through the lens of layer-wise gradients. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33249–33299. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1536. URL https://aclanthology.org/2026.acl-long.1536/.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87766–87800. Curran Associates, Inc., 2024. doi: 10.52202/079017-2786. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9febda1c8344cc5f2d51713964864e93-Abstract-Conference.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

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

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1504–1512. PMLR, 2021. URL https://proceedings.mlr.press/v130/qian21a.html.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://www.eurasip.org/Proceedings/Eusipco/Eusipco2007/Papers/a5p-h05.pdf.

Rishi Sonthalia, Michael Murray, and Guido Montufar. Low rank gradients and where to find them. In *Advances in Neural Information Processing Systems*, volume 38, pp. 152341–152385. Curran Associates, Inc., 2025. doi: 10.52202/085713-5093. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/df7ac32d352db77aad1f355ad43206a9-Abstract-Conference.html.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Kai Yao, Zhenghan Song, Kaixin Wu, Mingjie Zhong, Danzhao Cheng, Zhaorui Tan, Yixin Ji, and Penglei Gao. GAST: Gradient-aligned sparse tuning of large language models with data-layer selection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4401–4416. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.206. URL https://aclanthology.org/2026.eacl-long.206/.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: A key ingredient for scalable distributed learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1998–2007. PMLR, 2018. URL https://proceedings.mlr.press/v84/yin18a.html.

A Additional Experimental Details

A.1 Models, datasets, and probing setup

Models. We evaluate three decoder-only Transformer models: Llama-3.1-8B, Llama-3.2-3B, and Mistral-7B-v0.3. All three models use gated FFN blocks with a down-projection matrix $W_{\mathrm{down}} \in \mathbb{R}^{d \times d_f}$, which is the primary object of our parameter-local gradient analysis. Unless otherwise specified, Llama-3.1-8B serves as the primary model in the main text, while Llama-3.2-3B and Mistral-7B-v0.3 are used to test whether the observed gradient structure generalizes across model scales and architectures.

Datasets. MetaMathQA is used as the primary probing dataset in the main-text experiments. To test whether the observed behavior is specific to mathematical instruction tuning, we additionally repeat the analysis on CodeFeedback. For each probing run, we randomly sample a minibatch from the corresponding training split. The supervised loss is evaluated only at unmasked target positions, while the token-level gradient factorization retains all valid forward positions, including prompt positions that can receive backpropagated signal from later supervised tokens.

Probing protocol. Unless otherwise specified, each per-sequence gradient matrix is constructed from $B = 48$ sequences sampled from a fixed minibatch. For comparisons within the same experimental setting, the minibatch is held fixed so that changes in NEER can be attributed to the model checkpoint, layer, or intervention being compared rather than to batch composition. Robustness to independently sampled minibatches is evaluated separately in Appendix B.2, and the effect of B is studied in Appendix B.4.

To compare layers across models with different depths, we use relative network depth rather than absolute layer indices. The layers at approximately 10%, 50%, and 80% of total depth are used as representative shallow, middle, and deep layers, respectively. Full layer-wise results are reported in Appendix B.1. For analyses over the course of fine-tuning, we probe the base model and checkpoints at steps 500, 1000, 1500, and 2000.

A.2 Fine-tuning setup

We fine-tune the models using supervised next-token prediction with a token-averaged cross-entropy objective. For instruction-tuning examples, the loss is computed only on supervised response tokens, with prompt positions masked from the loss. Prompt tokens nevertheless remain in the forward computation and may receive nonzero backpropagated signals from later supervised positions. For MetaMathQA, we use at most $100{,}000$ training examples. Unless otherwise specified, all non-FFN parameters are frozen, and the gate, up-, and down-projection matrices are trained in every Transformer layer.

Training is performed for one epoch using fused AdamW with a cosine learning-rate schedule. We use 8 NVIDIA RTX A6000 GPUs (48 GB memory each) with a per-device batch size of 6 and no gradient accumulation, giving an effective batch size of 48 sequences. Computation uses bfloat16 precision with float32 master parameters. Checkpoints at optimizer steps 500, 1000, 1500, and 2000 are used for the training-stage analyses in the main text.

The default configuration is summarized in Table 4.

B Additional Results on Parameter-Local NEER

B.1 Full layer-wise results

The main text reports representative shallow, middle, and deep layers selected according to normalized network depth. Here we report the corresponding NEER statistics for every FFN layer of Llama-3.1-8B, Llama-3.2-3B, and Mistral-7B on MetaMathQA using the same probing protocol. The results are reported in Table 5, Table 6, Table 7.

For each layer, we report the median NEER within the bottom-1% group, the overall median NEER, and their ratio (O/B). Across all three models, the low-NEER tail appears throughout the network,

Table 4: Default fine-tuning configuration used in our experiments.

Hyperparameter	Setting
Trainable parameters	FFN gate/up/down, all layers
Training examples	up to 100,000
Epochs	1
Maximum sequence length	1024
Optimizer	AdamW (fused)
Learning rate	5×10^{-6}
LR schedule	Cosine
Warmup ratio	0.03
Weight decay	0
Per-device batch size	6
Gradient accumulation	1
Number of GPUs	8
GPU	NVIDIA RTX A6000
GPU memory	48 GB per GPU
Effective batch size	48
Maximum gradient norm	1.0
Compute precision	bfloat16
Master-parameter precision	float32
Analyzed checkpoints	500, 1000, 1500, 2000

although its strength varies with network depth. These results verify that the representative layers used in the main text are not isolated cases.

Across the three models, low-NEER columns are observed throughout network depth rather than only at the representative layers highlighted in the main text. The strength of the low-NEER tail nevertheless varies across layers and architectures. In particular, unusually large O/B ratios in some early layers contribute to the cross-layer variability reported in Table 1.

B.2 Robustness across random minibatches

The main-text analyses use fixed minibatches to isolate differences caused by model checkpoints, layers, and diagnostic interventions. To examine whether the observed NEER heterogeneity depends on a particular batch composition, we repeat the probing procedure over independently sampled minibatches from MetaMathQA. Unless otherwise specified, each minibatch contains $B = 48$ sequences, and the bottom-1% group is recomputed independently for each batch.

Table 8 summarizes the results at representative shallow, middle, and deep layers of Llama-3.1-8B. Across random minibatches, the bottom-1% NEER remains substantially below the overall median, and the Overall/Bottom gap persists across all three depths. This indicates that the low-NEER tail is not specific to a single sampled minibatch.

Figure 5 shows the corresponding variation across individual minibatches. Although the exact NEER values fluctuate with batch composition, the separation between the low-NEER tail and the overall distribution remains qualitatively stable across repeated samples.

Importantly, these results establish robustness of the distributional pattern rather than persistence of the identities of individual low-NEER columns. The stability of column identities across minibatches is examined separately in Appendix B.3.

B.3 Batch-to-batch stability of extreme-column identities

The random-minibatch experiment above concerns the stability of the *distribution* of column-wise NEER. We separately quantify whether the same columns repeatedly enter the extreme groups. Let $\mathcal{L}_\ell^{(b)}$ and $\mathcal{H}_\ell^{(b)}$ denote the bottom- and top-1% column sets, respectively, in minibatch b. For two independently sampled minibatches, we measure set agreement by the Jaccard index

$$J(\mathcal{S}^{(b)}, \mathcal{S}^{(b')}) = \frac{|\mathcal{S}^{(b)} \cap \mathcal{S}^{(b')}|}{|\mathcal{S}^{(b)} \cup \mathcal{S}^{(b')}|}, \tag{15}$$

Table 5: Full layer-wise NEER statistics for Llama-3.1-8B on MetaMathQA. Bottom 1% denotes the median NEER among the lowest-NEER 1% of columns within each layer. O/B denotes the ratio of the overall median to the bottom-1% median.

Layers 0–15

Layer	Bottom 1%	Overall	O/B
0	0.040	0.402	10.059
1	0.096	0.415	4.303
2	0.083	0.446	5.366
3	0.149	0.491	3.302
4	0.154	0.520	3.378
5	0.153	0.515	3.371
6	0.175	0.521	2.973
7	0.196	0.526	2.686
8	0.171	0.490	2.864
9	0.263	0.614	2.334
10	0.280	0.614	2.194
11	0.289	0.614	2.126
12	0.276	0.614	2.221
13	0.245	0.607	2.474
14	0.211	0.603	2.859
15	0.162	0.597	3.693

Layers 16–31

Layer	Bottom 1%	Overall	O/B
16	0.144	0.581	4.034
17	0.128	0.565	4.399
18	0.119	0.559	4.704
19	0.127	0.547	4.309
20	0.142	0.536	3.770
21	0.126	0.534	4.242
22	0.111	0.523	4.726
23	0.102	0.515	5.041
24	0.106	0.508	4.784
25	0.106	0.505	4.749
26	0.105	0.508	4.820
27	0.110	0.504	4.589
28	0.113	0.499	4.408
29	0.078	0.480	6.122
30	0.090	0.468	5.224
31	0.085	0.418	4.927

Table 6: Full layer-wise NEER statistics for Llama-3.2-3B on MetaMathQA, using the same definitions as Table 5.

Layers 0–13

Layer	Bottom 1%	Overall	O/B
0	0.063	0.510	8.148
1	0.096	0.518	5.389
2	0.094	0.565	6.033
3	0.189	0.598	3.161
4	0.170	0.615	3.620
5	0.178	0.616	3.454
6	0.219	0.612	2.796
7	0.272	0.623	2.291
8	0.267	0.614	2.298
9	0.281	0.634	2.252
10	0.339	0.637	1.878
11	0.293	0.631	2.152
12	0.276	0.625	2.267
13	0.247	0.620	2.505

Layers 14–27

Layer	Bottom 1%	Overall	O/B
14	0.187	0.611	3.266
15	0.163	0.602	3.692
16	0.117	0.591	5.057
17	0.139	0.582	4.200
18	0.133	0.563	4.242
19	0.114	0.560	4.918
20	0.130	0.543	4.184
21	0.122	0.539	4.419
22	0.123	0.543	4.404
23	0.113	0.542	4.802
24	0.116	0.531	4.592
25	0.081	0.507	6.251
26	0.081	0.516	6.393
27	0.087	0.460	5.306

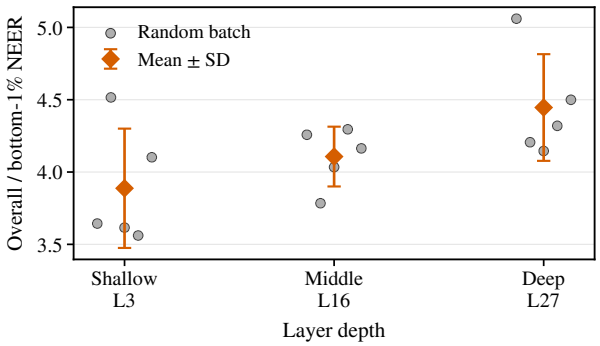


Figure 5: Robustness of the Overall/Bottom NEER gap across independently sampled minibatches at representative layers of Llama-3.1-8B. Each point corresponds to one minibatch.

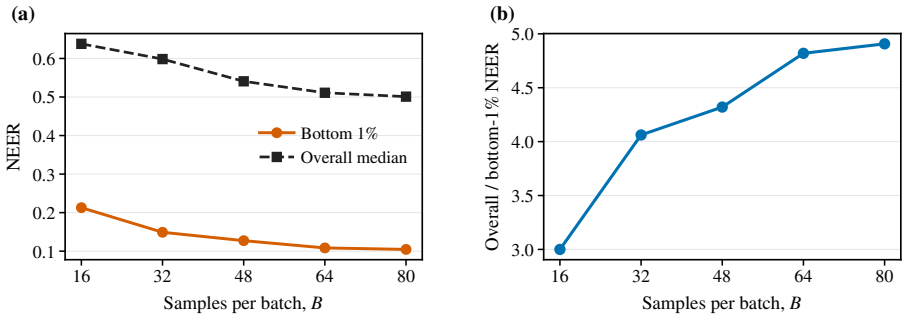


Figure 6: Sample-size sensitivity of the low-NEER tail.

and report the Spearman correlation between the full column-wise NEER rankings in Table 9. All statistics are averaged over distinct minibatch pairs.

The results show that the identities of the extreme columns are not stable across minibatches, even though the low-NEER tail itself persists. This does not affect our subsequent analyses or intervention design, since neither assumes a fixed set of intrinsically low-NEER columns: the relevant groups are identified from the current minibatch, and the proposed intervention selects columns adaptively according to their current NEER.

B.4 Effect of the number of sampled sequences

Because NEER is normalized by the number of sampled sequences, we test whether the separation between the low-NEER tail and the overall column distribution persists as B changes. For each independently sampled pool, we form nested subsets whenever possible and recompute the bottom-1% median, the overall median, and their ratio. The results are reported in Figure 6 and Table 10. Repeating the construction over multiple pools separates sensitivity to B from ordinary batch-composition variation.

B.5 Results on an additional data domain

We repeat the parameter-local probing analysis on CodeFeedback without changing the definitions of the per-sequence gradient matrix or NEER. Results are reported in Table 11 and Figure 7. The bottom-1% group is recomputed within every layer and experimental setting.

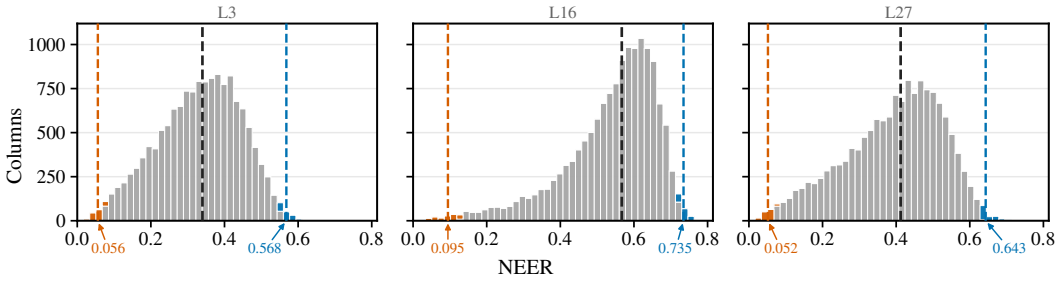


Figure 7: Column-wise NEER distributions of W_{down} in Llama-3.1-8B on CodeFeedback at the representative shallow, middle, and deep layers (layers 3, 16, and 27). Orange and blue dashed lines mark the median NEER of the bottom- and top-1% groups, respectively, and black dashed lines mark the overall median. The axes and histogram binning follow the corresponding MetaMathQA figure in the main text.

C Additional Analysis of Magnitude-Induced Rank Collapse

C.1 Proof of the NEER decomposition

Proposition 1. *Let $X \in \mathbb{R}^{d\times B}$ satisfy $\|X\|_F > 0$, and let $K = X^\top X$. Let $p \in \mathbb{R}^B$ be the eigenvalues of K normalized by $\operatorname{tr}(K)$, including zero eigenvalues, and let $q = \operatorname{diag}(K)/\operatorname{tr}(K)$. Then*

$$-\log \mathrm{NEER}(X) = \big(\log B - H(q)\big) + \big(H(q) - H(p)\big), \tag{16}$$

and both terms on the right-hand side are nonnegative.

Proof. Let

$$\overline{K} = \frac{K}{\operatorname{tr}(K)}.$$

By construction, the eigenvalue vector of $\overline{K}$ is $p = (p_1, \ldots, p_B)$, while its diagonal is $q = (q_1, \ldots, q_B)$. Since $\overline{K}$ is Hermitian, the Schur–Horn theorem gives

$$q \prec p.$$

Because Shannon entropy is Schur-concave, it follows that

$$H(q) \geq H(p),$$

and hence

$$H(q) - H(p) \geq 0.$$

Moreover, since q is a probability vector on the B-simplex,

$$H(q) \leq \log B,$$

so

$$\log B - H(q) \geq 0.$$

Finally, using

$$\mathrm{NEER}(X) = \frac{\exp(H(p))}{B},$$

we obtain

$$-\log \mathrm{NEER}(X) = \log B - H(p) \tag{17}$$

$$= \big(\log B - H(q)\big) + \big(H(q) - H(p)\big). \tag{18}$$

Thus, with

$$C^E = \log B - H(q), \qquad C^D = H(q) - H(p),$$

we obtain

$$-\log \mathrm{NEER}(X) = C^E + C^D, \qquad C^E \geq 0, \qquad C^D \geq 0,$$

which proves the stated decomposition. □

The first term is zero when the sequence-gradient energies are equal, so q is uniform. If the sequence gradients are mutually orthogonal, then K is diagonal and its eigenvalues equal its diagonal entries up to a permutation; hence $H(p) = H(q)$ and the second term is zero. These two cases recover the magnitude and direction interpretations used in the main text.

C.2 Additional magnitude–direction diagnostics

We extend the decomposition, norm-equalization intervention, and direction-only diagnostics beyond the representative Llama-3.1-8B setting in Table 12. For every model–dataset pair, the reported statistics are computed over the bottom-1% columns selected independently within each layer.

D Additional Results on Gradient-Magnitude Origins

Table 13 extends the proxy comparison to all model–dataset pairs. Each entry aggregates per-column statistics over the low-NEER group, using the same representative relative depths and the same bottom-1% NEER selection rule as in the main text.

Conditional within-sequence vector shuffle. For each sequence, we partition valid token positions by supervision status and coarse relative-position stratum. Within each stratum, we hold the activation sequence fixed and uniformly permute the backpropagated vectors $\delta^{\ell}_{s,t}$. We then reconstruct the shuffled column gradient

$$g^{\mathrm{shuf}}_{s,\ell,i} = \sum_{t \in \mathcal{V}_s} a^{\ell}_{s,t,i}\, \delta^{\ell}_{s,\pi_s(t)}. \tag{19}$$

This intervention preserves the activation values and the multiset of backpropagated vectors within every stratum while disrupting their observed token-level pairing. Table 14 records the exact shuffle configuration, and Table 15 reports how the intervention changes column-wise NEER rankings and dominant-sequence identities.

E Additional Optimization Analysis

E.1 Layer-to-column dominant-sequence agreement

For a target column i, define the leave-one-column-out layer-level energy of sequence s by

$$E^{(-i)}_{s,\ell} = \sum_{j \neq i} \|g_{s,\ell,j}\|_2^2. \tag{20}$$

Let $s^{\star}_{\ell,i} = \arg\max_s \|g_{s,\ell,i}\|_2^2$ be the dominant sequence for column i. We report whether $s^{\star}_{\ell,i}$ lies among the top-r fraction of sequences ranked by $E^{(-i)}_{s,\ell}$ in Table 16. Excluding the target column prevents the layer-level score from trivially containing the quantity it is asked to predict.

E.2 Optimizer-visible energy of low-NEER columns

For a set of columns $\mathcal{S}_\ell$, we measure its share of the optimizer-aggregated W_{down} gradient energy as

$$F_\ell(\mathcal{S}_\ell) = \frac{\sum_{i \in \mathcal{S}_\ell} \|\sum_s g_{s,\ell,i}\|_2^2}{\sum_{i=1}^{d_f} \|\sum_s g_{s,\ell,i}\|_2^2}. \tag{21}$$

The enrichment relative to the fraction of selected columns is $F_\ell(\mathcal{S}_\ell)/(|\mathcal{S}_\ell|/d_f)$. The results are reported in Table 17.

E.3 Energy rebalancing and first-order alignment

We provide a simplified local analysis connecting sequence-level energy imbalance, parameter-local NEER, and the direction of the aggregated update. The analysis idealizes the low-NEER regime observed empirically, where sequence-gradient directions remain diverse and often close to orthogonal while their magnitudes are highly imbalanced.

Proposition 2 (Energy rebalancing and local NEER)**.** *Let the per-sequence gradients be*

$$g_s = \alpha_s u_s, \qquad s = 1, \dots, B,$$

where $\alpha_s \geq 0$ and $u_1, \dots, u_B \in \mathbb{R}^d$ are orthonormal. Define the normalized sequence-energy distribution

$$p_s = \frac{\alpha_s^2}{\sum_{j=1}^{B} \alpha_j^2},$$

and let p' denote the corresponding distribution after a rebalancing operation. If

$$p' \prec p,$$

so that p' is more balanced than p in the majorization order, then:

1. *the parameter-local NEER weakly increases,*

$$\mathrm{NEER}(p') \geq \mathrm{NEER}(p);$$

2. *the normalized aggregated gradient becomes at least as well aligned with the balanced direction*

$$v = \frac{1}{\sqrt{B}} \sum_{s=1}^{B} u_s.$$

Furthermore, if the local validation gradient satisfies

$$\nabla \mathcal{L}_{\mathrm{val}}(\theta) = \gamma v, \qquad \gamma > 0,$$

then, after rescaling both aggregate updates to the same norm, the rebalanced update gives a weakly larger first-order decrease in the validation objective.

Proof. Let

$$X = [g_1, \dots, g_B].$$

Since the sequence directions are orthonormal,

$$X^\top X = \mathrm{diag}\left(\alpha_1^2, \dots, \alpha_B^2\right).$$

Therefore, the normalized eigenvalues of the Gram matrix are precisely $p_1, \dots, p_B$, and hence

$$\mathrm{NEER}(p) = \frac{\exp(H(p))}{B}. \tag{22}$$

Because Shannon entropy is Schur-concave, the majorization relation $p' \prec p$ implies

$$H(p') \geq H(p).$$

Using Equation 22, we obtain

$$\mathrm{NEER}(p') = \frac{\exp(H(p'))}{B} \geq \frac{\exp(H(p))}{B} = \mathrm{NEER}(p).$$

Now consider the aggregated gradient

$$G = \sum_{s=1}^{B} \alpha_s u_s.$$

Orthogonality gives

$$\|G\|_2^2 = \sum_{s=1}^{B} \alpha_s^2.$$

Its alignment with the balanced direction v is therefore

$$A(p) := \left\langle \frac{G}{\|G\|_2}, v \right\rangle \tag{23}$$

$$= \frac{\sum_{s=1}^{B} \alpha_s}{\sqrt{B \sum_{s=1}^{B} \alpha_s^2}} \tag{24}$$

$$= \frac{1}{\sqrt{B}} \sum_{s=1}^{B} \sqrt{p_s}. \tag{25}$$

Since $x \mapsto \sqrt{x}$ is concave, the majorization relation $p' \prec p$ implies

$$\sum_{s=1}^{B} \sqrt{p'_s} \geq \sum_{s=1}^{B} \sqrt{p_s}.$$

Combining this inequality with Equation 25 yields

$$A(p') \geq A(p).$$

Finally, let the fixed-norm update be

$$\overline{G} = r \frac{G}{\|G\|_2}, \qquad r > 0.$$

For a sufficiently small step size η, a first-order Taylor expansion gives

$$\mathcal{L}_{\text{val}}(\theta - \eta \overline{G}) = \mathcal{L}_{\text{val}}(\theta) - \eta \left\langle \nabla \mathcal{L}_{\text{val}}(\theta), \overline{G} \right\rangle + O(\eta^2).$$

Under $\nabla \mathcal{L}_{\text{val}}(\theta) = \gamma v$,

$$\left\langle \nabla \mathcal{L}_{\text{val}}(\theta), \overline{G} \right\rangle = \gamma r A(p).$$

Since $A(p') \geq A(p)$, the rebalanced update has a weakly larger first-order descent term at the same update norm. □

The proposition formalizes a limited but useful connection between NEER and contribution rebalancing. In the orthogonal-direction regime, parameter-local NEER and aggregate-update alignment are both controlled by the same normalized sequence-energy distribution. Making this distribution more balanced in the majorization sense simultaneously increases NEER and the representation of the balanced target direction in the aggregate update.

E.4 Evaluation setup

We fine-tune all three models on MetaMathQA using the training configuration in Table 4. For downstream evaluation, we use GSM8K and MATH-500 as in-domain (ID) benchmarks, and SVAMP and AQuA-RAT as out-of-domain (OOD) benchmarks. The generation limits used for each benchmark are summarized in Table 18.

We use flexible answer matching for all benchmarks rather than requiring an exact match to a single output format. For AQuA-RAT, a prediction is counted as correct when it identifies the correct answer either through the corresponding multiple-choice option or through the correct numerical answer. The same evaluation protocol is applied to the baseline and selective-clipping models.

F Sequence–Column Atlases Across Representative Layers

To complement the aggregate magnitude–direction diagnostics, we visualize how per-sequence gradient energy is distributed across individual W_{down} columns. Figure 8 shows sequence–column atlases for representative shallow, middle, and deep layers of Llama-3.1-8B on MetaMathQA.

The heatmap visualizes the normalized sequence-gradient energy shares $q_{s,\ell,i}$ defined in equation 7, so that each heatmap column sums to one. Rows are ordered by the total gradient energy of each

sequence within the layer. Columns are divided into equal-sized low-, middle-, and high-NEER groups, corresponding respectively to the bottom, central, and top 1% of the layer-wise NEER distribution. Within each group, columns are ordered by their dominant-sequence position to expose shared sequence-level structure.

The staircase-like bands arise from the display ordering and should be distinguished from the underlying energy values. After sequences are ranked by total layer energy, columns within each NEER group are sorted by the display rank of their dominant sequence, with ties broken by decreasing top-1 energy share and then by the original column index. Columns sharing the same dominant sequence therefore form horizontal plateaus, whereas a change in dominant-sequence rank produces a discrete vertical jump, yielding the observed staircase pattern. This ordering does not alter $q_{s,\ell,i}$ or create energy concentration; it makes the persistence and transitions of dominant-sequence identities easier to see. The precise step geometry should therefore not be interpreted as an intrinsic adjacency among W_{down} columns.

Across the representative layers, the low-NEER group exhibits markedly more localized sequence-energy allocation than the middle- and high-NEER groups: a small subset of sequences accounts for a large fraction of the energy of many low-NEER columns, producing concentrated horizontal and block-like patterns in the atlas. In contrast, energy in higher-NEER columns is distributed more diffusely across the minibatch. This visualization is consistent with the magnitude-based interpretation of low NEER developed in the main text: the reduced effective rank is associated with strongly unequal sequence-wise energy rather than requiring the sequence gradients to collapse onto a common direction.

The annotation tracks provide two complementary views of this structure. The *Pair gain* reports the reduction in TV distance obtained by retaining same-token activation–backpropagation pairing relative to the marginal-product null,

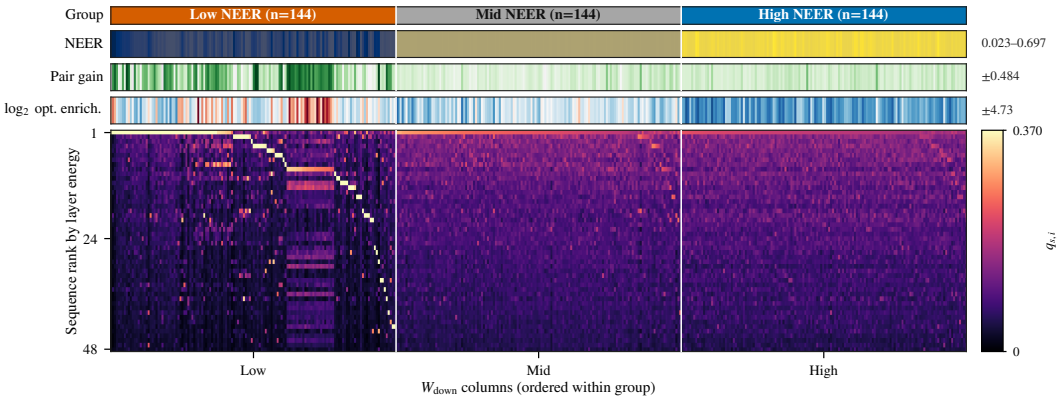
$$\Delta_i^{\mathrm{Pair}} = \mathrm{TV}\Big(q_i^{\mathrm{Exact}}, q_i^{\mathrm{Marg}}\Big) - \mathrm{TV}\big(q_i^{\mathrm{Exact}}, q_i^{\mathrm{Pair}}\big)\,.$$

Positive values therefore indicate that token-level pairing contains information about the observed sequence-energy allocation beyond the sequence-level activation and backpropagated-signal marginals alone. The optimizer-enrichment track further reports

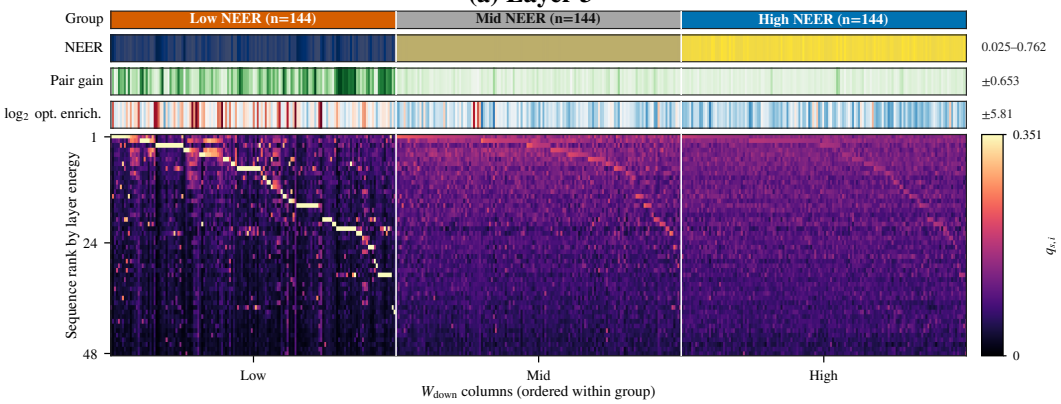
$$\log_2 R_{\ell,i}, \qquad R_{\ell,i} = d_f \frac{\|\sum_s g_{s,\ell,i}\|_2^2}{\sum_j \|\sum_s g_{s,\ell,j}\|_2^2},$$

which measures the aggregated gradient energy of a column relative to the layer-wise average.

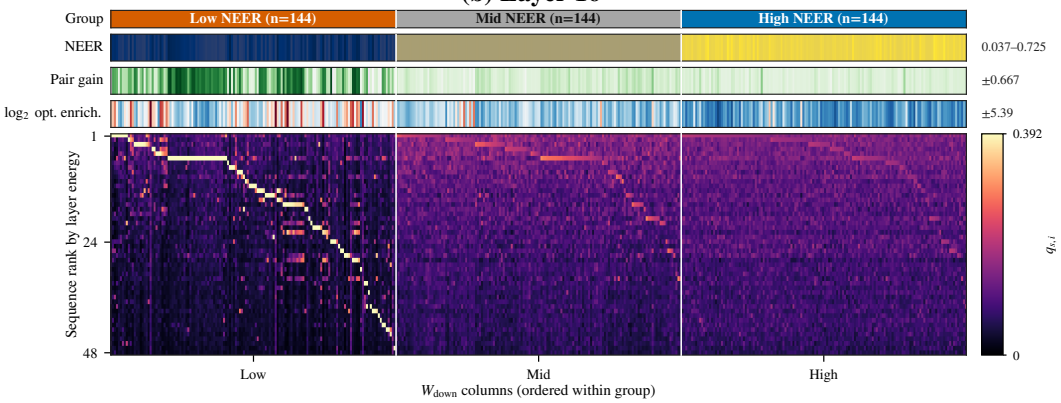
Together, the atlases provide a column-resolved view of the mechanism underlying the aggregate statistics in the main text. Low-NEER columns exhibit localized sequence-energy concentration together with pronounced regions of positive Pair gain and positive optimizer-energy enrichment, visible as green and red bands in the corresponding annotation tracks. These patterns are substantially weaker in the middle- and high-NEER groups, suggesting that the sequence-energy concentration characteristic of low NEER is accompanied by stronger sensitivity to token-level activation–backpropagation pairing and, for a subset of columns, elevated optimizer-visible gradient energy. These visualizations are intended as structural diagnostics; the quantitative Pair-vs.-Marginal comparisons and optimizer-energy analyses provide the corresponding aggregate evidence.



(a) Layer 3



(b) Layer 16



(c) Layer 27

Figure 8: Sequence–column atlases of W_{down} gradient-energy allocation for representative layers of Llama-3.1-8B on MetaMathQA: **(a)** layer 3, **(b)** layer 16, and **(c)** layer 27. Within each panel, columns are partitioned into equal-sized low-, middle-, and high-NEER groups, and rows are ordered by full-layer sequence gradient energy. Each heatmap entry shows the normalized energy share $q_{s,\ell,i} = \|g_{s,\ell,i}\|_2^2 / \sum_r \|g_{r,\ell,i}\|_2^2$. The annotation tracks report column-wise NEER, the improvement of the same-token Pair proxy over the marginal-product null, and the log-scaled enrichment of optimizer-aggregated gradient energy.

Table 7: Full layer-wise NEER statistics for Mistral-7B on MetaMathQA, using the same definitions as Table 5.

Layers 0–15

Layer	Bottom 1%	Overall	O/B
0	0.032	0.519	16.252
1	0.064	0.539	8.367
2	0.064	0.554	8.719
3	0.090	0.529	5.858
4	0.116	0.542	4.679
5	0.140	0.552	3.937
6	0.151	0.558	3.706
7	0.136	0.561	4.138
8	0.170	0.609	3.591
9	0.234	0.616	2.631
10	0.268	0.631	2.350
11	0.279	0.636	2.285
12	0.269	0.625	2.326
13	0.247	0.609	2.462
14	0.219	0.616	2.811
15	0.176	0.608	3.453

Layers 16–31

Layer	Bottom 1%	Overall	O/B
16	0.149	0.604	4.062
17	0.127	0.599	4.702
18	0.108	0.580	5.383
19	0.087	0.555	6.400
20	0.090	0.552	6.170
21	0.073	0.548	7.509
22	0.063	0.531	8.465
23	0.068	0.523	7.695
24	0.069	0.538	7.835
25	0.086	0.539	6.245
26	0.092	0.539	5.834
27	0.103	0.546	5.283
28	0.095	0.533	5.635
29	0.090	0.517	5.720
30	0.095	0.498	5.245
31	0.099	0.454	4.583

Table 8: Robustness of column-wise NEER statistics across independently sampled MetaMathQA minibatches for Llama-3.1-8B. Values are mean $\pm$ standard deviation across minibatches.

Layer	Bottom 1%	Overall	Overall / Bottom
3	0.138 ± 0.017	0.533 ± 0.031	$(3.89 \pm 0.41)\times$
16	0.137 ± 0.011	0.560 ± 0.023	$(4.11 \pm 0.21)\times$
27	0.114 ± 0.008	0.504 ± 0.017	$(4.45 \pm 0.37)\times$

Table 9: Stability of column identities across independently sampled MetaMathQA minibatches.

Layer	Bottom-1% Jaccard	Top-1% Jaccard	Rank corr.
3	0.163 ± 0.040	0.074 ± 0.028	0.592 ± 0.067
16	0.183 ± 0.044	0.075 ± 0.020	0.633 ± 0.040
27	0.212 ± 0.022	0.087 ± 0.019	0.629 ± 0.035

Table 10: Sensitivity of column-wise NEER to the number of sampled sequences for Llama-3.1-8B on MetaMathQA. Values are averaged across independently sampled pools and the three representative FFN layers.

B	Bottom 1%	Overall	Overall / Bottom
16	0.213	0.638	3.000
32	0.149	0.598	4.062
48	0.127	0.541	4.321
64	0.108	0.511	4.819
80	0.105	0.501	4.907

Table 11: Column-wise NEER statistics on CodeFeedback, reported as mean $\pm$ standard deviation across the three representative FFN layers.

Model	Bottom 1%	Overall	Overall / Bottom
Llama-3.1-8B	0.067 ± 0.019	0.440 ± 0.095	6.682 ± 0.917
Llama-3.2-3B	0.091 ± 0.023	0.498 ± 0.088	5.555 ± 0.473
Mistral-7B	0.088 ± 0.030	0.688 ± 0.035	8.682 ± 2.624

Table 12: Magnitude–direction diagnostics across models and data domains. Within each representative layer, we compute the median over low-NEER columns. Entries report the mean of these layer-wise medians across the three representative layers.

Model	Dataset	$C^E/(C^E + C^D)$	NEER change	Deficit reduction	RMS cosine
Llama-3.1-8B	MetaMathQA	0.619	$0.134 \rightarrow 0.753$ ($\uparrow$ 0.619)	86.4%	0.154
Llama-3.1-8B	CodeFeedback	0.949	$0.068 \rightarrow 0.947$ ($\uparrow$ 0.879)	98.0%	0.052
Llama-3.2-3B	MetaMathQA	0.604	$0.166 \rightarrow 0.761$ ($\uparrow$ 0.595)	85.8%	0.141
Llama-3.2-3B	CodeFeedback	0.936	$0.091 \rightarrow 0.950$ ($\uparrow$ 0.859)	98.0%	0.051
Mistral-7B	MetaMathQA	0.557	$0.114 \rightarrow 0.753$ ($\uparrow$ 0.639)	87.8%	0.154
Mistral-7B	CodeFeedback	0.934	$0.088 \rightarrow 0.968$ ($\uparrow$ 0.880)	98.7%	0.039

Table 13: Sequence-energy proxy quality across models and datasets. TV denotes total-variation distance to the exact energy allocation; Acc. denotes dominant-sequence identification accuracy.

		TV distance ↓				Dominant-sequence acc. ↑			
Model	Dataset	Act.	Back.	Marg.	Pair	Act.	Back.	Marg.	Pair
Llama-3.1-8B	MetaMathQA	0.245	0.195	0.160	0.048	29.0%	40.1%	48.7%	83.4%
Llama-3.1-8B	CodeFeedback	0.414	0.245	0.184	0.060	27.1%	43.1%	58.2%	84.0%
Llama-3.2-3B	MetaMathQA	0.199	0.179	0.140	0.050	32.3%	24.7%	39.1%	76.8%
Llama-3.2-3B	CodeFeedback	0.381	0.213	0.153	0.061	23.8%	35.0%	54.4%	79.1%
Mistral-7B	MetaMathQA	0.222	0.188	0.157	0.047	24.6%	16.2%	34.1%	78.6%
Mistral-7B	CodeFeedback	0.261	0.223	0.167	0.048	29.6%	11.9%	41.4%	81.8%

Table 14: Configuration of the conditional vector-shuffle intervention for Llama-3.1-8B.

Setting	Value
Supervision-status strata	2: supervised / non-supervised next-token positions
Relative-position strata	4 equal-width normalized-position bins per sequence
Independent shuffle repeats	8 per layer and minibatch
Layers evaluated	All 32 layers; Layers 3, 16, and 27 reported
Column groups evaluated	All columns

Table 15: Effect of conditional token-level shuffling on MetaMathQA, aggregated across the three representative FFN layers for each model. Rank corr. is the Spearman correlation between original and shuffled column-wise NEER; dominant-sequence preservation is the fraction of columns whose maximum-energy sequence remains unchanged. Values are reported as mean $\pm$ standard deviation across shuffle repetitions and representative layers.

Model	NEER rank corr.	Dominant-seq. preservation	Mean absolute NEER change
Llama-3.1-8B	0.396 ± 0.110	0.397 ± 0.176	0.111 ± 0.016
Llama-3.2-3B	0.447 ± 0.099	0.360 ± 0.049	0.105 ± 0.021
Mistral-7B	0.422 ± 0.137	0.287 ± 0.050	0.125 ± 0.028

Table 16: Recovery of column-local dominant sequences from leave-one-column-out layer-level energy rankings for Llama-3.1-8B on MetaMathQA. We report the fraction of columns recovered within each NEER group, aggregated over representative layers.

Column group	Hit@1	Hit@5%	Hit@10%
Low NEER	$19.9 \pm 20.2\%$	$29.6 \pm 18.3\%$	$38.7 \pm 19.1\%$
Mid NEER	$44.0 \pm 35.8\%$	$60.6 \pm 26.3\%$	$70.1 \pm 21.0\%$
High NEER	$46.1 \pm 32.5\%$	$65.0 \pm 19.5\%$	$70.1 \pm 18.0\%$

Table 17: Optimizer-visible gradient-energy share of the bottom-1% NEER columns for Llama-3.1-8B on MetaMathQA.

Layer	Aggregated-energy share	Enrichment
3	2.7%	2.73×
16	5.2%	5.15×
27	6.4%	6.32×
Aggregate	4.8 ± 1.8%	4.73 ± 1.83×

Table 18: Evaluation configuration for the downstream benchmarks. GSM8K and MATH-500 are treated as ID benchmarks, while SVAMP and AQuA-RAT are treated as OOD benchmarks.

Dataset	Domain	Max input length	Max new tokens
GSM8K	ID	1024	1024
MATH-500	ID	2048	2048
SVAMP	OOD	1024	512
AQuA-RAT	OOD	2048	512