A Two-Parameter Weibull Framework for Diagnosing Transformer Weight Distributions

Anonymous authors
Paper under double-blind review

Abstract

We apply the Weibull distribution — a two-parameter family from extreme-value theory — as a diagnostic framework for element-wise weight magnitude distributions in transformers. At initialization, i.i.d. Gaussian weights give $|w| \sim$ HalfNormal, which anchors the Weibull shape parameter at $k \approx 1.20$. This makes k a principled, architecture-independent measuring stick for training dynamics; fitting each weight matrix independently at every layer enables diagnostics invisible to aggregate statistics.

Applying this framework to 12 model entries spanning 7 architectural families reveals the following findings. First, FFN modules and the attention output projection (the Transmission Class) fall in a narrow k band $[1.186, 1.204]$ across architectures (CV = 0.51%). Second, the attention input projections (the Selection Class, W_q and W_k) depart from this band in an architecture-dependent manner: separately-stored MHA shows the largest drift, grouped-query attention shows milder drift, and merged storage shows transitional behavior. The scale parameter λ grows during training and tracks $\sqrt{\eta/\lambda_{\text{wd}}}$ as a within-family scaling trend in Pythia ($n = 5$ sizes). The two Weibull parameters carry independent information: k labels the functional class, λ labels training progress.

The framework was further used to diagnose an 11-entry Qwen cohort: shallow-FFN layers exhibit bimodal weight distributions. We release `npm-weibull-py v0.4` and `DATABASE_v9_1` (anonymized for review; URLs in camera-ready).

1 Introduction

1.1 Problem and Motivation

Understanding what happens during transformer training requires quantitative tools for interrogating learned weight distributions. Existing approaches operate in orthogonal spaces: WeightWatcher (Martin & Mahoney, 2019) and HT-SR (Martin & Mahoney, 2020) analyze eigenvalue spectra; AlphaDecay (He et al., 2025) tracks eigenvalue drift; massive activation analysis (Sun et al., 2024) measures activation magnitudes. None of these methods directly characterizes the element-wise distribution of weight magnitudes $|W_{ij}|$ — the most granular representation of what a model has learned.

This gap matters because element-wise statistics reveal structure that spectral methods cannot. Eigenvalue spectra compress all matrix elements into a single distribution, averaging across component types. If different functional components within the same model — say, FFN layers versus attention projections — follow systematically different distributions, the aggregate spectral statistics obscure rather than reveal this distinction.

We apply the Weibull distribution — a two-parameter family from extreme-value theory (Weibull, 1951) — as a diagnostic lens on element-wise $|W|$ distributions. The shape parameter k quantifies distributional tailedness; the scale λ quantifies magnitude. Critically, the shape parameter has an anchor at initialization: i.i.d. Gaussian weights give $|w| \sim$ HalfNormal, which gives $k \approx 1.20$ (via middle-80% probability-plot fit, the protocol used throughout this work). This makes k a precise, dimensionless measuring stick for training dynamics — any departure from $k_0 \approx 1.20$ is attributable entirely to training.

With this measuring stick, element-wise and per-component, the weight distributions of a transformer model present a picture of heterogeneous, depth-dependent evolution. Different layers of the same model specialize at different rates; within each layer, different projection matrices — W_q, W_k, W_o, the FFN gates — evolve along different trajectories. This depth-dependent structure is invisible to aggregate statistics, but becomes measurable when each projection matrix is examined independently at every layer at every checkpoint. Figure 1 summarizes the resulting diagnostic framework; Section 4 documents the per-layer trajectories for the Selection Class.

1.2 Contributions

We make three contributions. (i) A *diagnostic framework*: the initialization anchor ($k_0 \approx 1.20$), the noise-optimal middle-80% trim protocol, and the anti-interference property they confer (Section 2). (ii) *Cross-family empirical evidence* across 12 model entries spanning 7 architectural families, two orders of magnitude in parameter count, two activation patterns (GeLU/SwiGLU), and four normalization placements; the FFN Transmission band (Section 3) and the architecture-dependent Selection drift (Section 4) are reproduced with no tunable parameters. (iii) *Open-source tools*: `npm-weibull-py v0.4` (8 diagnostic functions F1–F8; Appendix B) and the companion benchmark database `DATABASE_v9_1` (12 entries).

1.3 Relationship to Existing Tools

Three existing tools are relevant to this framework, and the relationships are complementary rather than competitive. WeightWatcher (Martin & Mahoney, 2019) and HT-SR (Martin & Mahoney, 2020) analyze eigenvalue spectra; AlphaDecay (He et al., 2025) tracks eigenvalue drift; OrthoAdam (Kaul et al., 2025) reports kurtosis reduction in trained weights. These tools operate in measurement spaces mathematically orthogonal to our element-wise $|W|$ framework (eigenvalue is a quadratic form, kurtosis is a fourth moment, $|W|$ is a linear form), and Weibull k therefore carries information independent of their diagnostics.

2 Framework

2.1 Theoretical Basis

Initialization anchor. The framework begins with an analytical result at initialization. Modern transformers are initialized with i.i.d. Gaussian weight elements: $w \sim \mathcal{N}(0, \sigma^2_{\text{init}})$. The absolute values $|w|$ therefore follow a half-Normal distribution, which is not a closed-form special case of Weibull. Fitting $|w|$ to a Weibull distribution via middle-80% probability-plot least-squares — the protocol used throughout this work — gives shape parameter $k_0 \approx 1.20$ and scale parameter $\lambda_0 \approx 0.8875\,\sigma_{\text{init}}$, analytically derived from the protocol via deterministic special-function integrals (Appendix A.1). The shape k_0 is independent of σ_{init} and therefore universal across initialization schemes; the scale λ_0 is proportional to σ_{init} and varies by initialization recipe and by component (Appendix A.2). We treat $k_0 \approx 1.20$ as the functional zero point: any departure indicates training-induced structural change. Departures of λ from λ_0 reflect both training-induced scaling and the per-component initialization split discussed in Section 5.

Fitting protocol. To measure k in a trained model, we extract all weight matrices from a given layer, flatten the absolute values, and fit a Weibull distribution using least-squares on the Weibull probability plot. A critical design choice is the **middle 80% trim**: before fitting, we discard the smallest and largest 10% of $|W_{ij}|$ values.

This protocol is theoretically principled, not empirically arbitrary. In Weibull probability coordinates $Y = \ln(-\ln(1-F))$, the empirical rank Y_i of the i-th order statistic has sampling-noise variance $\text{Var}(Y_i) \propto p/[(1-p)\cdot(\ln(1-p))^2]$, where $p = i/N$ is the cumulative fraction. This variance **diverges at both ends** ($p \to 0$ and $p \to 1$) and reaches a minimum near $p \approx 0.80$. The bottom 10% of the distribution carries approximately 6.5× more measurement noise than the optimal central region (Figure 2). Empirically, fitting to full data (no trim) systematically underestimates k by 3–7% across architectures — enough to pull 5 of 7 models outside the transmission band. The middle 80% trim is therefore the noise-optimal choice.

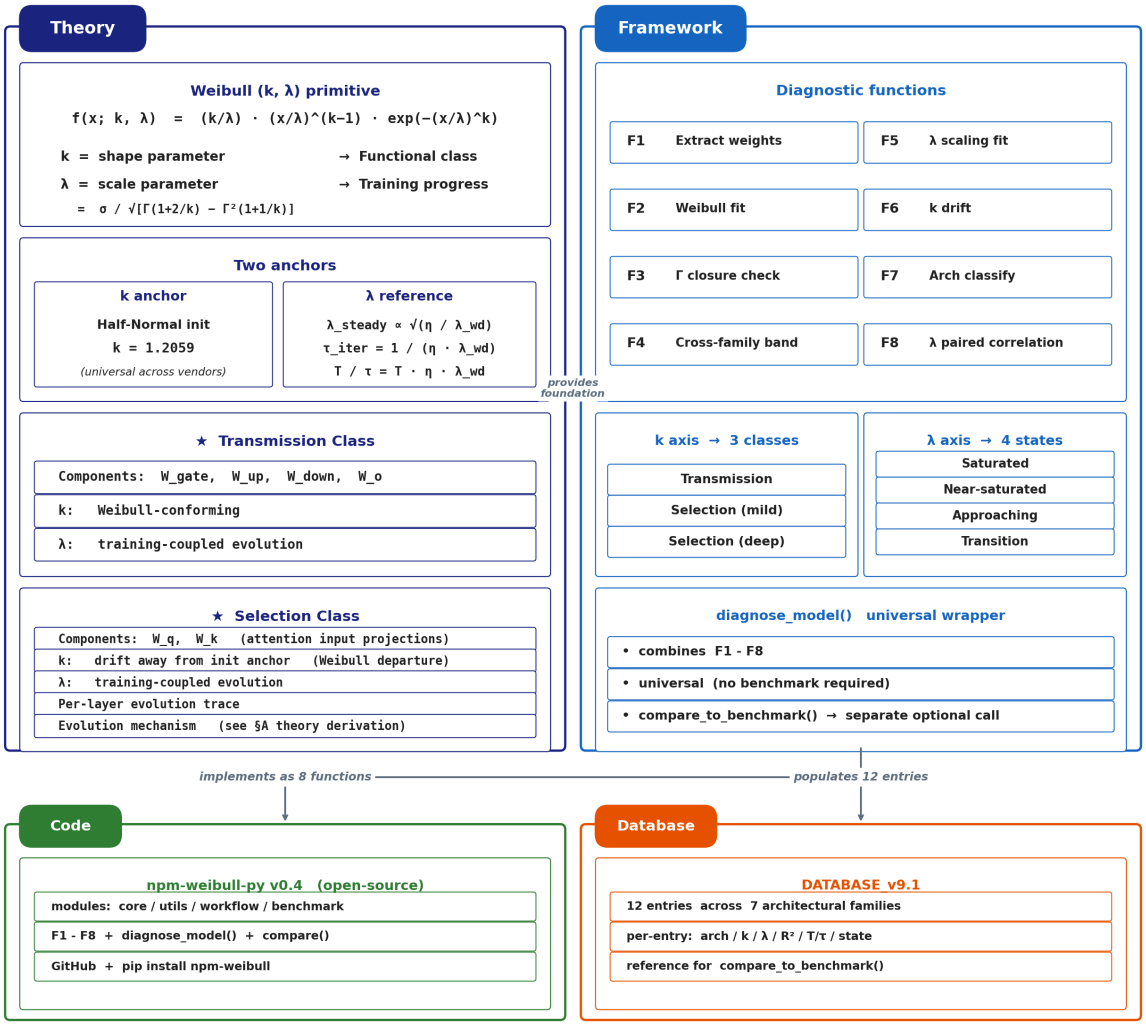

Figure 1: **Diagnostic framework architecture.** The framework comprises four interconnected components organized as a layered architecture. **Theory** (top-left): the Weibull two-parameter primitive $f(x;k,\lambda) = (k/\lambda)(x/\lambda)^{k-1}\exp(-(x/\lambda)^k)$ with its σ-inverse form $\lambda = \sigma/\sqrt{\Gamma(1+2/k) - \Gamma^2(1+1/k)}$, anchored by two reference scales: the k anchor (Half-Normal initialization, $k_0 \approx 1.20$ via middle-80% probability-plot fit — numerically consistent across vendors and σ_{init} scales) and the λ reference (steady-state scaling $\lambda_{\text{steady}} \propto \sqrt{\eta/\lambda_{\text{wd}}}$, $\tau_{\text{iter}} = 1/(\eta\lambda_{\text{wd}})$, T/τ as the dimensionless training progress). Two functional classes are defined: **Transmission Class** ($W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}, W_o$) and **Selection Class** (W_q, W_k). **Framework** (top-right): eight diagnostic functions F1–F8 (Appendix B) project the framework onto two classification axes (k-axis: 3 classes — Transmission, Selection-mild, Selection-deep; λ-axis: 4 states — Saturated, Near-saturated, Approaching, Transition). **Code**: `npm-weibull-py v0.4`, implementing the eight functions as an open-source library. **Database**: `DATABASE_v9_1`, 12 entries across 7 architectural families. The framework's power lies in identifying departures from Weibull: Selection Class departures of W_q, W_k are the primary diagnostic signal, not anomalies.

This gives the framework its **anti-interference property**: fitted k is robust to outlier contamination, numerical precision artifacts, and the heavy-tail phenomena that dominate the extreme percentiles of trained weight distributions.

What k measures. The shape parameter k of a Weibull distribution quantifies the **tailedness** of $|W|$: smaller k corresponds to a heavier right tail, while larger k corresponds to a narrower, more uniform body. Because k is a dimensionless shape parameter, it is comparable across weight matrices of different shapes, sizes, and architectures. Figure 3 shows how the $|W|$ distribution evolves from initialization (half-Normal, $k_0 \approx 1.20$) through training: FFN body k stays close to initialization throughout, while attention Q/K projections drift toward smaller k. This pattern motivates the two functional classes — **Transmission** (FFN + attention output) and **Selection** (Q/K) — developed in Section 2.2.

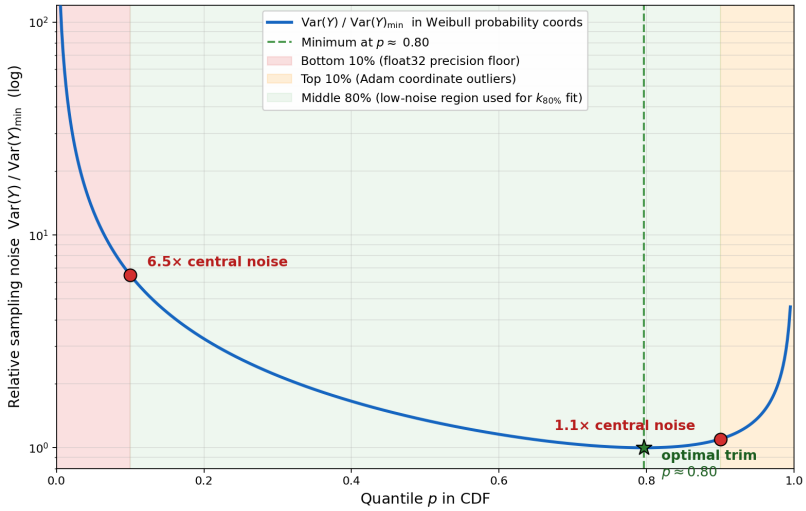

Figure 2: **Middle-80% trim is noise-optimal.** Theoretical curve of $\mathrm{Var}(Y)/\mathrm{Var}(Y)_{\min}$ as a function of quantile p, where $Y = \ln(-\ln(1-F))$ is the Weibull probability transform. Three shaded regions: bottom 10% (red, $p < 0.10$), middle 80% (green, $0.10 \le p \le 0.90$, the low-noise region used for the $k_{80\%}$ fit), top 10% (orange, $p > 0.90$). The variance diverges at both endpoints and reaches its minimum at $p^* \approx 0.797$; the bottom 10% region carries $\sim 6.5\times$ more noise than the optimal region. The middle-80% protocol is the mathematically optimal choice for recovering the body Weibull shape, excluding both the numerical-precision-dominated lower tail and the coordinate-outlier-dominated upper tail.

2.2 Two Functional Classes

Having established the initialization anchor and the measurement protocol, we now apply this diagnostic tool to trained transformer weights. Across 12 model entries spanning 7 architectural families and all training stages examined, transformer components partition into two classes with statistically distinct Weibull k trajectories. We call these **Transmission** and **Selection**.

Transmission Class. The Transmission Class consists of all FFN modules and the attention output projection W_o. We retain the architecture-specific FFN naming convention of each model family rather than imposing a uniform notation: for 3-matrix SwiGLU FFNs (LLaMA-3, Mistral, Qwen2.5/3, OLMo-2), the gate projection W_{gate}, the up projection W_{up}, and the down projection W_{down}; for 2-matrix GeLU FFNs (Pythia, OLMo-1), the input projection $W_{\text{FFN\_in}}$ and the output projection $W_{\text{FFN\_out}}$. Cohort-wide figures (notably Figure 4) adopt the generic cascade labels `FFN_in`, `FFN_out`, `FFN_down` as an alias over both schemes. For these components, k remains remarkably stable throughout training, close to the initialization anchor $k_0 \approx 1.20$. The scale parameter λ varies by more than an order of magnitude across families and layers, yet the shape k remains remarkably stable. We call this empirical stability the Transmission Class signature; the quantitative evidence is presented in Section 3.

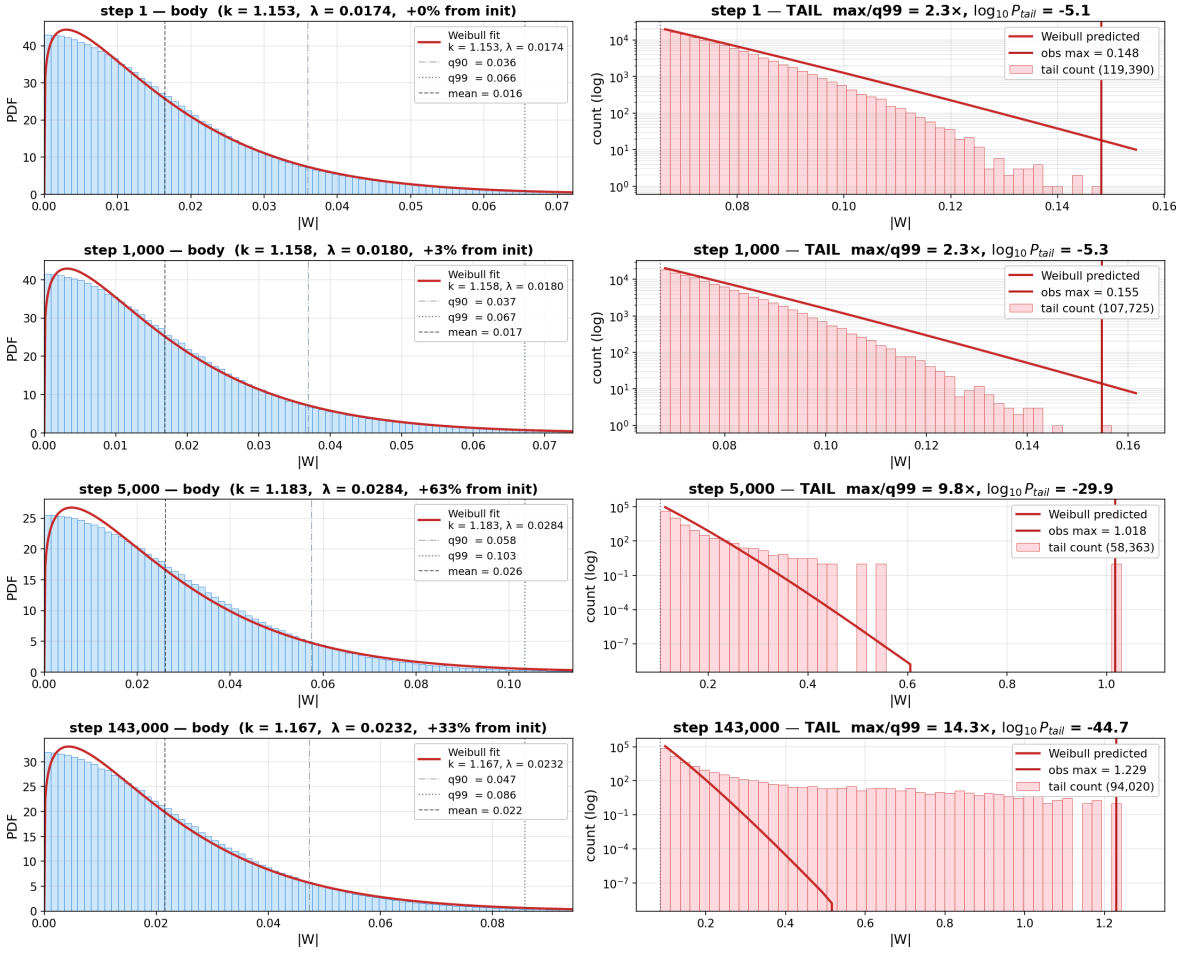

Figure 3: **Pythia-70m $|W|$ Weibull body fit (left column) + tail evolution (right column) across 4 representative training steps (1, 1000, 5000, 143000) — Transmission Class only (W_o + FFN, $W_{\mathbf{qkv}}$ excluded).** The left column shows the middle-80% Weibull fit on the Transmission Class bulk weight distribution (n = 14,155,776 samples per checkpoint). The shape parameter k stays near initialization throughout training (aggregate-fit $k = 1.153 \to 1.183 \to 1.167$) and the aggregate scale parameter λ grows from 0.0174 at step 1 to 0.0232 at step 143000 (+33%). Note that the aggregate-fit $k = 1.153$ at step 1 lies below the per-block anchor $k_0 \approx 1.20$ because pooling $W_o + W_{\text{gate}} + W_{\text{up}} + W_{\text{down}}$ mixes components with different per-component σ_{init} (Appendix A.2 Recipe A: $\sigma_{\text{in}}/\sigma_{\text{out}} = L/\sqrt{10}$); per-block fits at step 0 yield $k = 1.205 \pm 0.001$ across all kinds (Appendix A.1, with the per-component initialization recipes in Appendix A.2, Table 5). The right column shows the tail region beyond q99. Bars are observed counts; the red line is the Weibull-predicted count using the body fit extrapolated. By step 5000, an isolated super-weight emerges at $|w| \approx 1.0$ (max/q99 = 9.8×, $\log_{10} P_{\text{tail}} = -29.9$); at step 143000 the super-weight reaches $|w| = 1.2$ (max/q99 = 14.3×, $\log_{10} P_{\text{tail}} = -44.7$). The body remains Weibull-conforming throughout; only a small number of extreme elements per matrix detach from the bulk. The body-tail gap is a universal trained-model phenomenon; see Table 2 for cross-family quantification and Section 3.3 for discussion.

The QK/OV circuit decomposition of prior work (Elhage et al., 2021) provides the functional reason for this grouping. The OV circuit and FFN modules transmit information without gating; their optimization pressure favors uniform weight distributions that maximize aggregate conductance, preserving the initialization Weibull body. The body-tail decomposition (Section 3) shows that the Weibull body of Transmission Class components remains intact even as isolated super-weights emerge in the tail — only the extreme tail is affected by training. The Transmission Class therefore represents the steady-state that a weight matrix approaches when its primary role is information transmission rather than selective filtering.

Selection Class. The Selection Class consists of the attention input projections W_q and W_k. In contrast to Transmission components, W_q and W_k depart systematically from the initialization anchor $k_0 \approx 1.20$. Three candidate mechanisms collectively drive this departure: functional necessity (sparse attention requires selective W_q, W_k), AdamW sign-descent dynamics, and softmax saturation feedback (no-op heads back-propagating extreme logits into W_k); two further candidate forces (residual-stream coupling and training-budget accumulation) are discussed in Appendix A.3. W_k is the primary site of this pressure, with W_v and W_o exhibiting transitional behavior whose magnitude tracks upstream Q/K drift severity (Figure 4; in OLMo-1-7B, severe MHA Q/K drift propagates downstream to W_o, yielding mid-layer median $k \approx 1.04$ rather than the Transmission band ~ 1.20). Severity and structure differ by architecture; quantitative evidence is presented in Section 4.

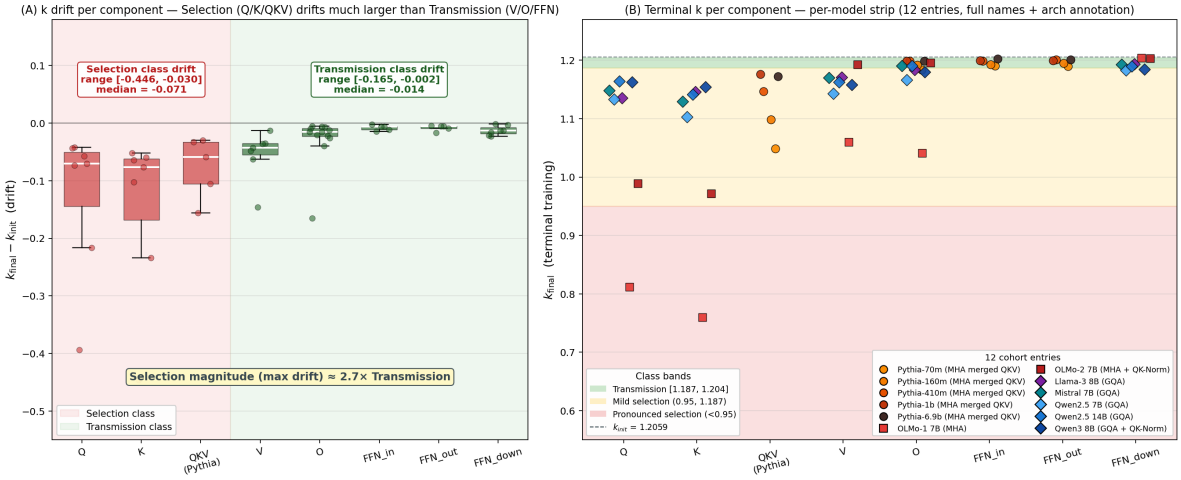

Figure 4: k **drift dichotomy across the 12-entry cohort.** Panel A: k drift magnitude from initialization to terminal checkpoint (Transmission median $= -0.014$, Selection median $= -0.071$). Panel B: terminal k positions relative to the Transmission band $[1.186, 1.204]$ (yellow band). Both panels share the same x-axis: Selection components (W_q, W_k, W_{qkv}) and Transmission components (W_v, W_o, `FFN_in`, `FFN_out`, `FFN_down`); the cascade FFN labels alias per Section 2.2. Transmission components cluster tightly within the band; Selection components fall below, with severity dependent on storage architecture (separately-stored MHA in OLMo-1/2 deepest, GQA milder, Pythia merged W_{qkv} transitional).

3 FFN Transmission Band

Section 2.2 identified the FFN modules and attention output projection W_o as the Transmission Class. This section documents their empirical Weibull signature.

3.1 The k Band

Across 12 model entries spanning 7 architectural families — Pythia, OLMo-1, OLMo-2, LLaMA-3, Mistral, Qwen2.5, Qwen3 — across two orders of magnitude (70M–14B, factor of 200×) in parameter count, three initialization schemes, two activation patterns (GeLU 2-matrix in Pythia, SwiGLU 3-matrix in modern

frontier models), and four norm placements (Pre-LN, Pre-LN + RMSNorm, Peri-LN, QK-Norm), the FFN modules fit a Weibull distribution with median terminal $k \in [1.186, 1.204]$ (Figure 5; per-component drift dichotomy in Figure 4). The coefficient of variation is CV = 0.51% across 12 entries (0.57% on the depth-≥ 12 subset); every fit achieves $R^2 \geq 0.99$ and the Γ closure check passes for 837 FFN fits (relative error $< 2\%$). The band is reproducible with no tunable parameters: while the scale parameter λ varies by more than an order of magnitude across families, the shape k stays within $[1.186, 1.204]$ regardless of activation pattern, normalization placement, or initialization scheme.

3.2 Robustness: Body–Tail Ablation

The transmission band is established using the middle 80% trim fitting protocol. The theoretical justification (sampling-noise divergence at distribution tails) is given in Section 2.1. Re-fitting the same FFN components with three truncation protocols — $k_{80\%}$, $k_{90\%}$, and $k_{100\%}$ — across 12 entries gives:

Protocol	Median k	In band $[1.186, 1.204]$
$k_{80\%}$	1.195	**10 / 12**
$k_{90\%}$	1.182	0 / 12
$k_{100\%}$	1.143	0 / 12

Table 1: Body–tail ablation across the 12-entry cohort. The middle 80% trim isolates the body; full-data fit is dragged by the heavy tail.

The body–tail gap $k_{80\%} - k_{100\%} = 0.0519 \pm 0.0017$ is remarkably stable across all 12 entries spanning 7 families (CV = 3.3%). This $\sim 5\%$ systematic shift quantifies exactly how much the heavy tail — the same phenomenon that HT-SR's spectral exponent $\hat{\alpha}$ characterizes — pulls the full-data fit away from the body. Importantly, this gap is a robustness statement about the fitting protocol — it documents how much full-data fits would deviate from the body fit, validating the choice of middle 80% trim. It does not characterize the trained-model outliers themselves; that is the subject of Section 3.3.

3.3 Cross-Family Super-Weight Observation

The body–tail ablation (Section 3.2) establishes the fitting protocol's robustness. We now turn to the empirical phenomenon that drives the gap — isolated heavy-tailed weight outliers, or *super-weights*, observed in every architectural family of our cohort.

Super-weight emergence within a single training trajectory. Figure 3 (right column) shows the tail evolution of Pythia-70m Transmission Class weights across four training checkpoints. At step 1, the right tail beyond q99 matches the Weibull body prediction within sampling noise. By step 5,000 an isolated outlier emerges at $|w| \approx 1.0$ ($\max/q99 = 9.8\times$, $\log_{10} P_{\text{tail}} = -29.9$). By step 143,000 this outlier reaches $|w| = 1.2$ ($\max/q99 = 14.3\times$, $\log_{10} P_{\text{tail}} = -44.7$). The body of the distribution remains Weibull-conforming throughout; only one — or a small handful of — extreme elements per matrix detach from the bulk.

Cross-family quantification. Re-applying the same per-block tail diagnostic across our 8 measured terminal checkpoints (the 12-entry cohort restricted to checkpoints with per-block weight access; the remaining 4 entries enter the cohort only at the aggregate level used in Section 3) reveals that the Pythia-70m pattern generalizes: every family in our cohort contains isolated outliers (Table 2). The per-block max-to-q99 ratio has cohort medians of approximately 7–28× (range 6.6× at Pythia-160m to 27.7× at OLMo-1-7B; see Table 2) and family-level maxima reaching 107× (OLMo-1-7B). The per-block kurtosis distribution likewise concentrates near $\sim$ 4–10 but exhibits per-family extreme values reaching 446 (OLMo-1-7B) and 257 (Pythia-410m).

These outliers are weight-side counterparts of the *super-weight* (Yu et al., 2024) and *massive activation* (Sun et al., 2024) phenomena documented in adjacent literature. Our contribution is cross-family quantification: every family contains isolated outliers, and the per-family extreme magnitude varies by an order of magnitude.

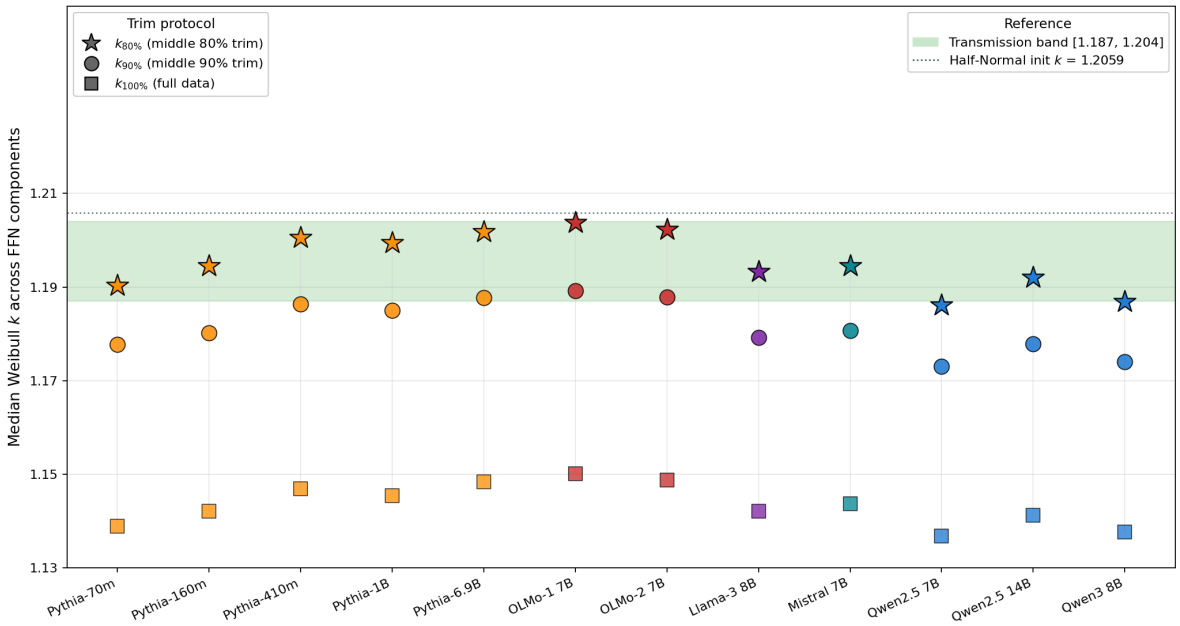

Figure 5: **Body–tail ablation across the 12-entry cohort.** Three trim protocols ($k_{80\%}$, $k_{90\%}$, $k_{100\%}$) applied to FFN + W_o components across 12 entries spanning 7 architectural families. The middle 80% protocol places 10/12 entries inside the Transmission band $[1.186, 1.204]$; full-data fit ($k_{100\%}$) places 0/12 inside, with the body–tail gap $k_{80\%} - k_{100\%} = 0.0519 \pm 0.0017$ (CV = 3.3%) representing the heavy-tail influence on the fit.

Family	# blocks	max/q99 median	max/q99 extreme	kurtosis extreme
Pythia-70m	6	7.2	15.7	196.9
Pythia-160m	12	6.6	13.4	104.2
Pythia-410m	24	8.0	19.6	257.1
Pythia-1B	16	11.3	21.5	21.8
Pythia-6.9B	32	8.6	31.4	27.7
OLMo-1-7B	32	27.7	**107.2**	**445.9**
Qwen2.5-14B	48	17.3	23.7	46.9
Qwen3-8B	36	17.3	40.4	14.5

Table 2: Per-block super-weight signatures across the 8 representative entries with per-block weight access; the full 12-entry cohort is reported in the FFN-band analysis in text (Section 3). "max /q99" is the ratio of the largest $|w_{ij}|$ in a block to the 99th-percentile $|w_{ij}|$ of the same block. Kurtosis is the empirical excess kurtosis. Median values report the typical block in each family; extreme values report the single most outlier-laden block.

For the framework here, the relevant observation is that the body Weibull fit is robust to these outliers (by design of the middle 80% protocol).

3.4 Diagnostic case study: family-specific anomaly detection

We apply the framework to a within-entry deviation flagged by our 12-entry cohort audit. The per-block protocol surfaces three Qwen entries (Qwen2.5-7B/14B, Qwen3-8B) with aggregate k below the Transmission band and a paired λ collapse; disaggregation localizes the deviation to shallow blocks ($\ell \in [1, 6]$). Extending the cascade to eight additional Qwen-family entries (1.5B/3B/7B variants and Math-CPT pairs) yields an 11-entry Qwen cohort that partitions cleanly: all four 1.5B variants stay in-band; all seven 7B-and-above variants exhibit the shallow deviation (with depth varying across 28L/36L/48L). Direct per-block

(k, λ) trajectories and 1D weight-magnitude histograms (Figures 13, 14) confirm the deviation as a bimodal distribution structurally distinct from super-weights (Yu et al., 2024). The framework diagnoses this as a Qwen-family characteristic; the full diagnostic chain is in Appendix C.

4 Q/K Selection Evolution

Section 2.2 established that W_q and W_k constitute the Selection Class and introduced the three mechanisms driving their departure from the initialization anchor. This section documents the empirical evidence along three dimensions: spatial localization across layer depth, temporal evolution with training budget, and architectural modulation of drift severity.

4.1 Spatial Localization

Figure 6 shows the per-layer per-component $|W|$ distribution heatmap for OLMo-1-7B. Q/K distributions start near the initialization body and progressively develop heavy tails — lower k — concentrated in the **mid-to-deep layers** (approximately layers 8–20 of 32). This layer-wise pattern is consistent with prior findings that attention heads develop specialized roles (Voita et al., 2019), with induction heads and other middle-layer-concentrated patterns characterized by Elhage et al. (2021) and Olsson et al. (2022). The heavy-tail signal in W_q/W_k is not uniform across depth but forms clusters in specific layer ranges, indicating that Selection is a spatially localized specialization rather than a global property of attention.

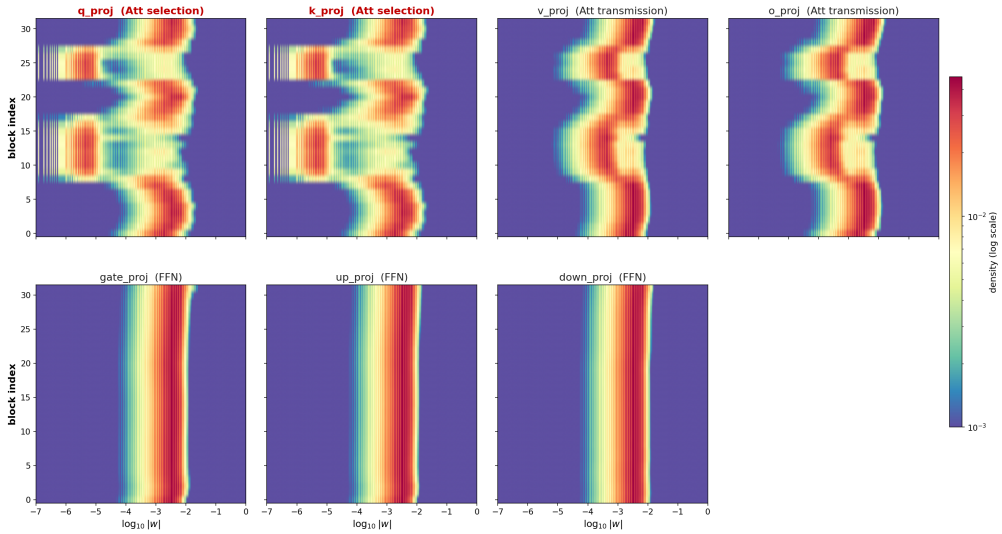

Figure 6: **OLMo-1 (7B, terminal) per-layer $|w|$ distribution heatmap — 7 components × 32 blocks.** Q/K (Selection Class) show specialization tail extending to $\log_{10}|w| \approx -5$ in blocks 8–17 and 23–27; V/O (Transmission Class) and FFN gate/up/down show tight distributions across blocks. Consistent with the QK/OV circuit decomposition (Elhage et al., 2021).

Qwen3-8B (with QK-Norm) shows visibly tighter Q/K distributions in shallow-to-mid blocks than Qwen2.5-14B (no QK-Norm; Figure 7). The two models also differ in size, training data, and other recipe details, so QK-Norm can only be noted as one architectural difference among several; a controlled QK-Norm ablation on otherwise-identical training would be required to isolate the effect.

4.2 Temporal Evolution

The dimensionless training budget is $T/\tau = T \cdot \eta \cdot \lambda_{\text{wd}}$, where T is total training steps, η is the peak learning rate, and λ_{wd} is the weight-decay coefficient; equivalently $T/\tau = T/\tau_{\text{iter}}$ with $\tau_{\text{iter}} = 1/(\eta\lambda_{\text{wd}})$ the AdamW characteristic timescale (Fan et al., 2025; Wang & Aitchison, 2024). This quantity measures how

many optimizer timescales have elapsed during training. Within the Pythia family (5 sizes), the severity of Selection drift tracks T/τ monotonically: larger T/τ produces more severe drift (Figure 8). All five Pythia sizes used identical 143k-step training budgets, but the 6.9B model uses a lower $\eta_{\text{peak}} = 1.2 \times 10^{-4}$ versus $\eta \in [3.0 \times 10^{-4}, 1.0 \times 10^{-3}]$ for the others, yielding $T/\tau = 0.17$ (Transition regime) instead of 0.43–1.43. This confirms that Selection drift is driven by the cumulative training signal $\eta \cdot \lambda_{\text{wd}} \cdot T$, not by model size per se. This observation is consistent with candidate mechanism D5 (Appendix A.3: cumulative training signal T/τ).

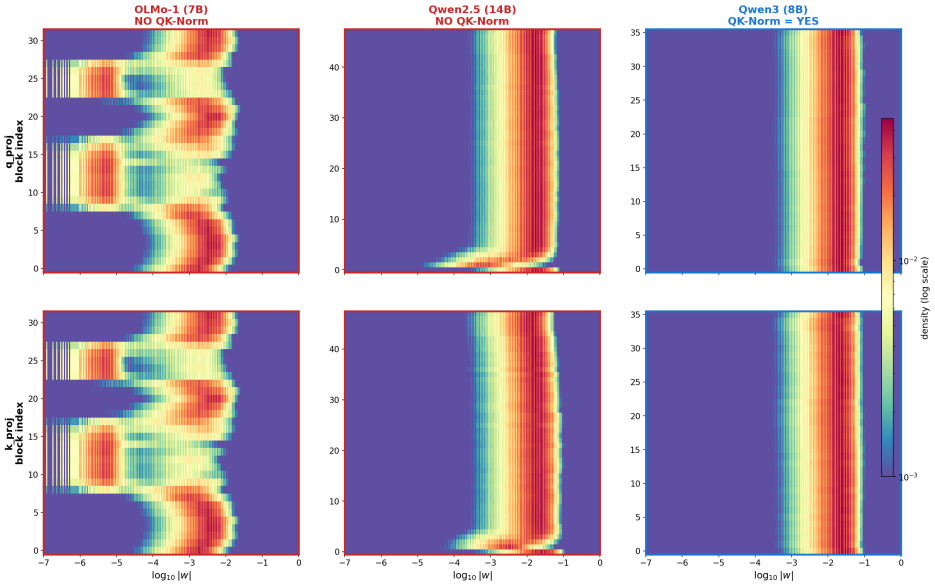

Figure 7: **QK-Norm contrast: cross-family terminal Q/K per-layer heatmap.** OLMo-1 7B (red, no QK-Norm), Qwen2.5 14B (red, no QK-Norm), Qwen3 8B (blue, QK-Norm). NO-QK-Norm models show tail extending to $\log_{10}|w| \approx -5$ in shallow-to-mid blocks; Qwen3 (QK-Norm) tail compressed and confined to fewer blocks.

4.3 Mechanism and Architecture

The three candidate mechanisms introduced in Section 2.2 — functional necessity, AdamW sign-descent, and softmax saturation feedback — are consistent with Selection drift; their formal statement and supporting references are given in Appendix A.3 (D1, D2, D3). The combined pressure manifests as k drifting below the Transmission band.

The severity of this drift is modulated by attention architecture (Figure 9). Models with separately-stored Q/K projections (OLMo-1, OLMo-2; multi-head attention) show the most severe drift: median terminal $k_q, k_k \in [0.76, 0.99]$, with individual blocks as low as $k = 0.28$ (OLMo-1, block 15, q-proj). Grouped-query attention models (LLaMA-3, Mistral, Qwen2.5, Qwen3) show substantially less drift: median $k_q, k_k \in [1.10, 1.16]$, consistently below the Transmission band but far less extreme. The five Pythia checkpoints, using a merged W_{qkv} tensor, occupy a transitional zone: median $k_{\text{qkv}} \in [1.05, 1.18]$, tracking T/τ monotonically. The architectural constraint of K-head sharing in GQA mechanically limits the specialization freedom available to individual K-head projections, consistent with the observed MHA/GQA dichotomy — an observational correlation; causal derivation requires controlled ablation (Section 6).

5 λ Evolution

The Weibull scale parameter λ measures the magnitude of weight matrices. Unlike the shape k, which is nearly invariant for Transmission Class components, λ increases substantially during training.

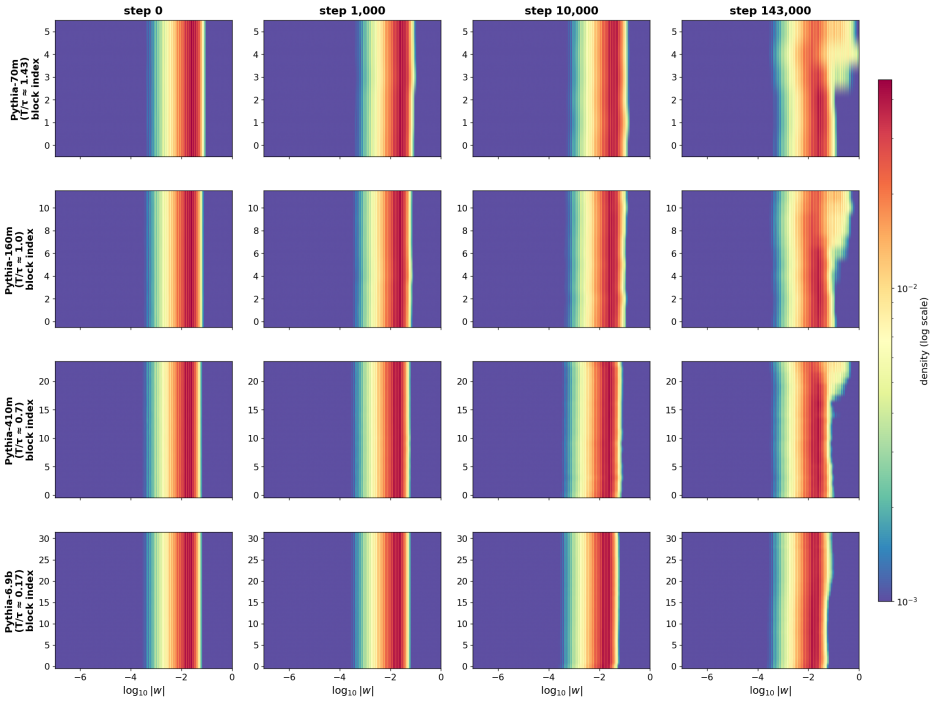

Figure 8: **Pythia merged $W_{\mathbf{qkv}}$ per-layer $|w|$ distribution — 4 sizes × 4 training steps.** Heatmap density of $\log_{10}|w|$ across blocks, with columns = training step and rows = Pythia size. The bulk distribution stays narrow; the left tail extends progressively, signaling Selection-class specialization within the merged W_{qkv} tensor. The dimensionless $T/\tau = T \cdot \eta \cdot \lambda_{\text{wd}}$ partitions the 5 Pythia sizes into Physical States: Pythia-70m ($T/\tau = 1.43$, Saturated), Pythia-160m ($T/\tau = 0.86$, Near-saturated), Pythia-410m ($T/\tau = 0.43$, Approaching), Pythia-6.9B ($T/\tau = 0.17$, Transition). Selection drift severity tracks T/τ monotonically.

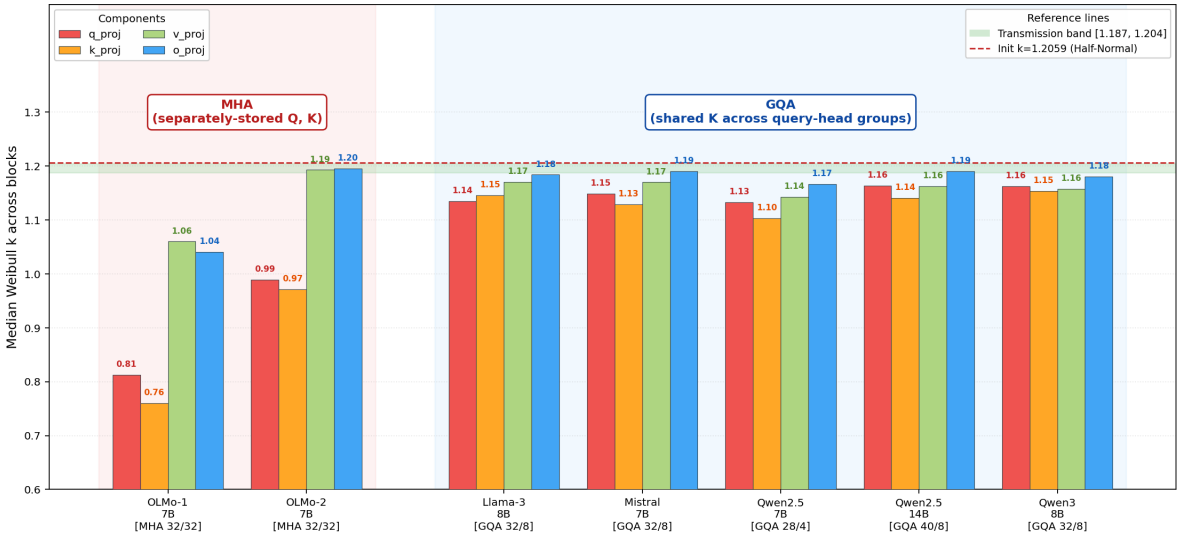

Figure 9: **MHA vs GQA dichotomy: terminal k_q and k_k across the 12-entry cohort.** Three architectural groups: separately-stored MHA (OLMo-1, OLMo-2; deep Selection, $k \in [0.76, 0.99]$), grouped-query attention (LLaMA-3, Mistral, Qwen2.5-7B/14B, Qwen3; mild Selection, $k \in [1.10, 1.16]$), and Pythia merged W_{qkv} (transitional, $k_{\text{qkv}} \in [1.05, 1.18]$, T/τ-monotonic across 70M–6.9B). The architecture-dependent severity is consistent with K-head sharing constraints in GQA limiting specialization freedom.

5.1 Initialization Reference

At initialization, λ is fully determined by the per-component initializer: the four Transmission Class kinds (W_{qkv}, W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$) do not share a single σ_{init} but split into input-side and output-side groups with distinct recipes. Within Pythia, the input/output ratio $\lambda_{\text{in}}/\lambda_{\text{out}}$ follows $L/\sqrt{10}$ for sizes 70m through 1B and $\sqrt{2L}$ for the 6.9B (Appendix A.2, Table 5; the closed-form derivation in Appendix A.1 and 5-size verification give agreement within 0.13% across all kinds). Training then collapses this recipe-specific initial ratio toward $\sim 1.2\times$ through the paired growth described below (for example, 6.9B terminal $\lambda_{W_{\text{qkv}}}/\lambda_O \approx 1.29$ and $\lambda_{W_{\text{FFN\_in}}}/\lambda_{W_{\text{FFN\_out}}} \approx 1.14$, down from the initial $8\times$).

5.2 Per-Component-Type Trajectories

Figure 10 shows the median λ trajectory for each of the four Transmission Class kinds (W_{qkv}, W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$) across the 5 Pythia sizes. Each component type exhibits a characteristic trajectory shape. The λ_O trajectory is non-monotonic: it rises through learning-rate warmup, peaks near step 10k (warmup completion), then retreats during cosine decay. The degree of post-peak retreat depends on the Physical State of the model. Models with $T/\tau > 1$ (Pythia-70m, $T/\tau = 1.43$, Saturated) show strong post-peak retreat (λ_O retreats 46% from its peak); models with $T/\tau \approx 0.4$–0.9 (160m–1B) show moderate retreat or near-flat trajectories; models with $T/\tau < 0.2$ (6.9B, Transition) are still rising at the terminal checkpoint, indicating unsaturated training. The growth itself reflects increased weight magnitude under AdamW, directionally consistent with the steady-state scaling analysis of Fan et al. (2025). Table 3 reports initialization-to-terminal growth across all five sizes.

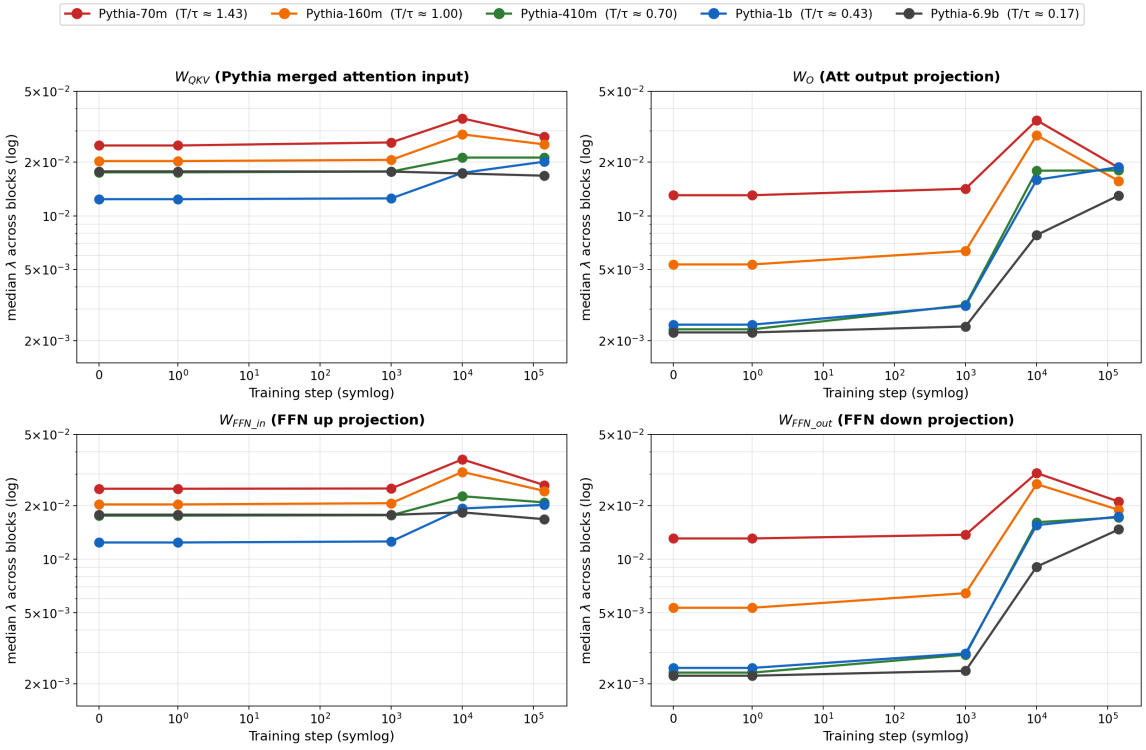

Figure 10: **4-component λ trajectory across Pythia 5 sizes.** Four subplots show the median λ per block across training steps for the four Transmission Class kinds (W_{qkv}, W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$), with each subplot overlaying the 5 Pythia sizes (70m–6.9B) color-coded by T/τ Physical State. Subplot ordering follows the transformer forward-pass: $W_{\text{qkv}} \to W_o \to W_{\text{FFN\_in}} \to W_{\text{FFN\_out}}$. The paired growth across W_o and $W_{\text{FFN\_out}}$ (Pearson $r = 0.9967$ on 25 size–step combinations) is reported in the main text (Section 5.3).

Size	T/τ	λ_O init	λ_O terminal	Mean λ (3 Trans. kinds, terminal)	Growth
70m	1.43 (Saturated)	0.0131	0.0186	0.0224	+43%
160m	0.86 (Near-sat.)	0.0053	0.0157	0.0200	+194%
410m	0.43 (Approaching)	0.0023	0.0180	0.0190	+678%
1B	0.43 (Approaching)	0.0025	0.0188	0.0191	+665%
6.9B	0.17 (Transition)	0.0022	0.0130	0.0149	+487%

Table 3: λ_O growth across the Pythia family. The "Mean λ (3 Trans. kinds, terminal)" column reports the terminal-checkpoint mean λ across the three Transmission Class kinds W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$ (mean per kind across blocks, then mean over the three kinds; W_{qkv} excluded as Selection class per Section 2.2). This is the quantity used in the within-Pythia scaling fit of Section 5.4 and reported again for Pythia in Table 4.

5.3 Per-Component-Type Coupling

The scale increase is not localized to the attention output. Across all 25 size–step combinations, $\log(\lambda_O)$ vs. $\log(\lambda_{\text{FFN\_out}})$ yields Pearson $r = 0.9967$ (log–log, $n = 25$). This near-perfect correlation indicates that both component types scale in lockstep throughout training: as W_o scales, $\lambda_{W_{\text{FFN\_out}}}$ tracks it because both write into the same residual stream. This paired behavior contrasts with the non-monotonic single-trajectory of λ_O (peaking at warmup completion, then cosine-decaying, Section 5.2).

The component-paired uniformity does not, however, extend to per-block uniformity within a model. Figure 11 shows the per-block (k, λ) profile for Pythia-410m terminal across 24 blocks: λ exhibits a modest depth-dependent rise (max/min $\approx 1.22\times$, CV $\approx 6.5\%$), with deep blocks (16–21) systematically larger — consistent with residual-stream magnitude accumulation. Block-aggregate k stays within the Transmission band $[1.186, 1.204]$ across most blocks, with a slight drop at the deepest 2 blocks. The two phenomena — component-paired uniformity and per-block depth heterogeneity — coexist.

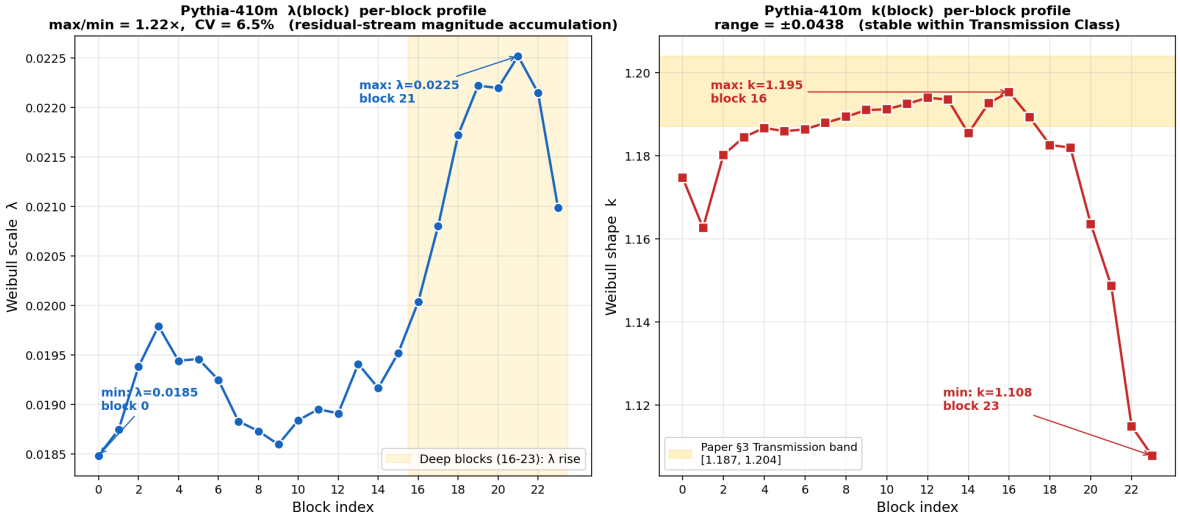

Figure 11: **Pythia-410m terminal per-block (k, λ) profile (24 blocks, aggregate fit per block).** λ (left) shows depth-dependent rise in deep blocks (16–23, max $\sim 1.22\times$ shallow); k (right) stays within the Transmission band $[1.186, 1.204]$ for most blocks, with slight drop at the deepest 2 blocks (super-weight tail effect). Per-block depth-heterogeneity complements component-paired uniformity (Pearson $r = 0.9967$ between λ_O and $\lambda_{\text{FFN\_out}}$): the two phenomena coexist.

5.4 Cross-Size Scaling

Within the Pythia training family — identical recipe, varying model size — the terminal mean λ across the three Transmission Class kinds (W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$; W_{qkv} excluded) shows a within-family scaling trend tracking $\sqrt{\eta_{\text{peak}}/\lambda_{\text{wd}}}$ ($n = 5$, linear fit through origin: $\lambda = 0.087 \cdot \sqrt{\eta/\lambda_{\text{wd}}}$, Pearson $r = 0.94$; Figure 12). The scaling direction — larger learning rate or lower weight decay produces larger terminal λ — shows **directional consistency** with the AdamW steady-state $\sqrt{\eta/\lambda_{\text{wd}}}$ scaling analysis of Fan et al. (2025) within their validated regime (LLaMA, $d \leq 2048$). Per-size deviations from the linear fit range from 7% to 36%, indicating directional consistency rather than quantitative magnitude match.

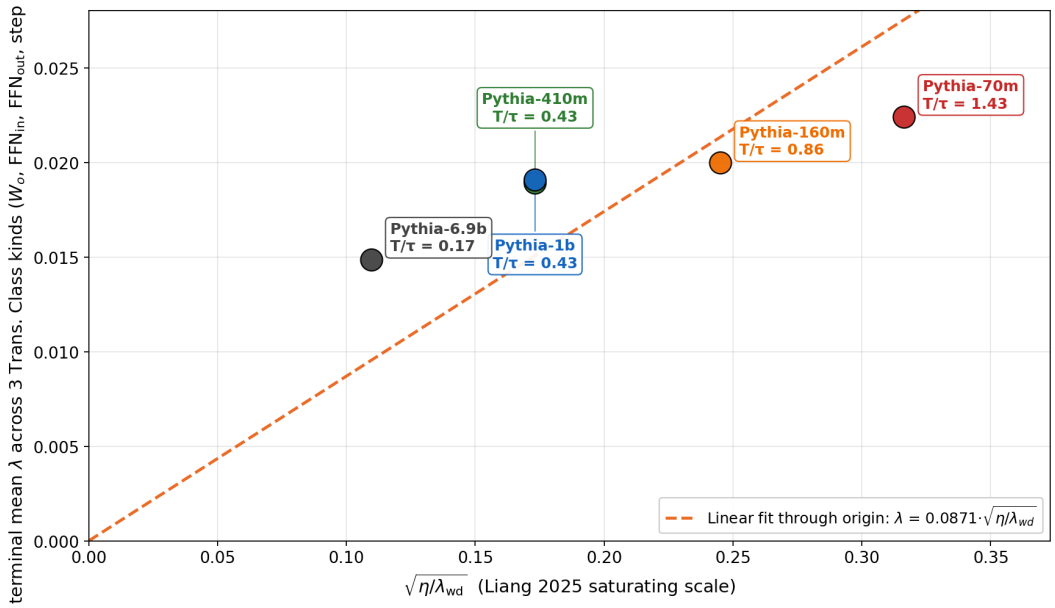

Figure 12: **Within-Pythia λ scaling vs. $\sqrt{\eta/\lambda_{\text{wd}}}$.** Terminal mean λ across the three Transmission Class kinds (W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$; W_{qkv} excluded as Selection per Section 2.2) plotted against $\sqrt{\eta_{\text{peak}}/\lambda_{\text{wd}}}$ for the 5 Pythia sizes. Linear fit through origin gives slope 0.087, Pearson $r = 0.94$. Per-size deviations of 7–36% indicate directional rather than quantitative match with the Fan et al. (2025) scaling law.

For completeness, Table 4 reports the terminal mean λ (across Transmission Class kinds) for all 12 cohort entries alongside the published η_{peak} and λ_{wd} used to compute $\sqrt{\eta/\lambda_{\text{wd}}}$. The cross-family slope ratio $\lambda/\sqrt{\eta/\lambda_{\text{wd}}}$ is shown in the final column and is the quantity discussed in Section 6 (cross-family scaling limitation).

5.5 Relationship to k

While λ_O grows by factors of 1.4×–7.8× across the Pythia family (Table 3), Transmission Class k stays in $[1.05, 1.20]$. AdamW scales magnitude (λ) while preserving distributional shape (k); W_q/W_k (the Selection Class) develop heavy tails without destroying the Weibull body. The two parameters therefore carry independent information.

6 Limitations and Future Work

Four limitations of the current work deserve explicit acknowledgment.

Selection mechanism. We document that W_q and W_k depart from the Weibull family during training and identify five candidate driving forces (Appendix A.3, D1–D5). However, we do not provide controlled ablation

Family / size	η_{peak}	λ_{wd}	$\sqrt{\eta/\lambda_{\text{wd}}}$	Terminal mean λ	$\lambda/\sqrt{\eta/\lambda_{\text{wd}}}$
Pythia-70m	1.0×10^{-3}	0.01	0.316	0.0224	0.071
Pythia-160m	6.0×10^{-4}	0.01	0.245	0.0200	0.082
Pythia-410m	3.0×10^{-4}	0.01	0.173	0.0190	0.110
Pythia-1B	3.0×10^{-4}	0.01	0.173	0.0191	0.110
Pythia-6.9B	1.2×10^{-4}	0.01	0.110	0.0149	0.136
OLMo-1-7B	3.0×10^{-4}	0.10	0.055	0.0032	0.058
OLMo-2-7B	3.0×10^{-4}	0.10	0.055	0.0167	0.304
LLaMA-3-8B	3.0×10^{-4}	0.10	0.055	0.0097	0.178
Mistral-7B	3.0×10^{-4}	0.10	0.055	0.0026	0.048
Qwen2.5-7B	3.0×10^{-4}	0.10	0.055	0.0138	0.253
Qwen2.5-14B	3.0×10^{-4}	0.10	0.055	0.0171	0.312
Qwen3-8B	5.0×10^{-4}	0.10	0.071	0.0225	0.319

Table 4: Cross-family terminal mean λ across the Transmission Class kinds (for Pythia: W_o, $W_{\text{FFN\_in}}$, $W_{\text{FFN\_out}}$; for SwiGLU families LLaMA-3/Mistral/Qwen2.5/Qwen3/OLMo-2: W_o, W_{gate}, W_{up}, W_{down}; W_q, W_k, W_{qkv} excluded as Selection class per Section 2.2) and the slope ratio $\lambda/\sqrt{\eta/\lambda_{\text{wd}}}$ across the 12 cohort entries. Aggregation is mean per kind across blocks then mean over kinds. Within Pythia (top block, identical recipe), the ratio spans $\sim 1.9\times$ (0.071–0.136). Across the 7 non-Pythia 7B–14B entries (bottom block), the ratio spans $\sim 6.6\times$ (Mistral-7B at 0.048 to Qwen3-8B at 0.319). The non-Pythia η_{peak} and λ_{wd} values are taken from published training configurations and may differ from end-of-training schedules; this is an additional source of uncertainty discussed in Section 6.

experiments isolating the contribution of each force. The observational correlation between the MHA/GQA architecture and Selection drift severity (Section 4) suggests K-head sharing as a plausible modulating factor, but a definitive causal claim requires an ablation study — specifically, training an otherwise-identical model in MHA versus GQA configuration from scratch. This is a direction we leave to future work.

Cross-family λ scaling. The $\lambda \propto \sqrt{\eta/\lambda_{\text{wd}}}$ scaling within the Pythia family (Section 5.4) is validated within Fan et al.'s regime ($d \leq 2048$). Cross-family generalization — applying the same formula to the 7B–14B modern cohort ($d \in [3584, 5120]$) — is observational only. Cross-family comparison of $\lambda/\sqrt{\eta/\lambda_{\text{wd}}}$ is confounded by two factors beyond η and λ_{wd}: (1) different initialization recipes (small_init vs Kaiming vs scaled_normal) set different σ_{init} baselines, directly shifting λ_{init} (Appendix A.2); and (2) different layer normalization placements (Pre-LN vs Post-LN vs QK-Norm) alter the residual-stream dynamics that govern λ evolution. Both factors introduce systematic shifts in λ that are orthogonal to the η/λ_{wd} scaling law. Across the 7 non-Pythia entries in our cohort, the per-family slope $\lambda/\sqrt{\eta/\lambda_{\text{wd}}}$ spans $\sim 6.6\times$ (Mistral-7B at 0.048 to Qwen3-8B at 0.319), substantially larger than the $\sim 1.9\times$ Pythia-internal range, consistent with these confounding factors (the non-Pythia η and λ_{wd} values are taken from published training configurations; the actual end-of-training schedules may differ from these reported values, which is an additional source of uncertainty). The observed scatter across families is therefore reported as a qualitative trend, not a quantitative law.

Downstream performance. The framework characterizes the statistical structure of weight distributions. We do not establish a direct correlation between the (k, λ) signature of a checkpoint and its downstream task performance. Such a correlation, if it exists, would make k and λ useful as early-stopping or model-selection proxies — a direction we leave to future work.

A Theory Details

A.1 Half-Normal Initialization Anchor

Modern transformers initialize weights from i.i.d. Gaussian distributions: $w_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{init}})$. The absolute value $|w_{ij}|$ therefore follows a half-Normal distribution. Fitting half-Normal samples to a Weibull distribution via least-squares on the Weibull probability plot with middle-80% trim — the protocol used throughout this work — yields a fit pair (k_0, λ_0) that is fully determined by the fit protocol and σ_{init}.

Closed-form derivation of the fit constants. The half-Normal CDF $F(x) = \operatorname{erf}\big(x/(\sigma_{\text{init}}\sqrt{2})\big)$ has inverse $x(F) = \sigma_{\text{init}} \cdot \sqrt{2} \cdot \operatorname{erf}^{-1}(F)$. Writing $g(F) = \ln(\sqrt{2}\operatorname{erf}^{-1}(F))$ and $Y(F) = \ln(-\ln(1-F))$, the probability-plot coordinates become $X = \ln\sigma_{\text{init}} + g(F)$ and $Y = Y(F)$. Least-squares regression of Y on X over $F \in [0.1, 0.9]$ gives

$$k_0 = \frac{\operatorname{Cov}_{[0.1,0.9]}(g, Y)}{\operatorname{Var}_{[0.1,0.9]}(g)}, \qquad \frac{\lambda_0}{\sigma_{\text{init}}} = \exp\Big(\overline{g}_{[0.1,0.9]} - \overline{Y}_{[0.1,0.9]}/k_0\Big). \tag{1}$$

The integrals are deterministic numerical integrals of special functions over the trim interval. They evaluate to

$$k_0 \approx 1.2054, \qquad \lambda_0 \approx 0.8875\,\sigma_{\text{init}}. \tag{2}$$

The constant 0.8875 is specific to this fit protocol and is not interchangeable with moment-matching values, which would yield different conversion factors. We confirm 0.8875 by direct measurement at the step-0 checkpoint across all 5 Pythia sizes and 4 Transmission Class kinds, which yield $\lambda_{\text{init}}/\sigma_{\text{init}} \in [0.887, 0.889]$, agreeing with the deterministic closed-form derivation above to within 0.13%.

Empirical zero-mean verification. The half-Normal anchor presumes symmetric initialization. We verify this assumption directly on Pythia-70m raw weight matrices: at step 0 the per-matrix means satisfy $|\mu|/\sigma < 0.3\%$ across all 4 Transmission Class kinds, and at the terminal checkpoint (step 143,000) the ratio remains below 0.6%, indicating that the AdamW with weight-decay training schedule preserves zero-mean weights to within finite-sample noise. Beyond this empirical check, the diagnostic pipeline (Appendix B) subtracts the per-matrix sample mean before histogramming, providing a universal safeguard for models that may not share Pythia's symmetric-init property.

Per-matrix zero-mean verification (cross-architecture). Beyond the Pythia-70m verification above, we cross-validate the zero-mean assumption on Qwen2.5-14B: across 48 layers × 4 Transmission Class kinds (192 matrices), the median $|\mu|/\sigma$ per kind ranges from 0.008% to 0.059%, with a maximum of 0.319% (W_{gate}). Both Pythia-70m (70 M, MHA-merged, small_init) and Qwen2.5-14B (14 B, GQA-7:1, scaled_normal) — spanning 200× parameter scale, two architecture generations, and two initialization recipes — satisfy $|\mu|/\sigma < 1\%$, confirming the Half-Normal anchor ($k_0 \approx 1.205$) is architecture- and scale-independent.

Weibull fit on $|w|$ robustness. To verify that fitting Weibull on $|w|$ (combined positive and negative weights) does not introduce asymmetry bias, we fit positive-W and $|-W|$ separately on 6 representative Qwen2.5-14B matrices (layers 0, 24, 47 × W_{down}, W_q). The shape parameter difference Δk between $+W$ and $|-W|$ fits is below 0.5% relative across all 6 matrices, and Δk between $|W|$ (combined) and $+W$ (positive only) is below 1% relative. Exact zero counts are 0.0001%–0.0002% (fewer than 200 exact zeros per matrix), confirming $|w|$ histogramming has negligible numerical impact.

Γ closure consistency. The fit pair (k, λ) obeys an internal self-consistency relation: for any Weibull(k, λ) distribution, the theoretical second raw moment is $E[W^2] = \lambda^2\Gamma(1 + 2/k)$, so the empirical sample standard deviation satisfies $\hat{\sigma} = \sqrt{\hat{\lambda}^2\Gamma(1 + 2/\hat{k}) - \hat{\lambda}^2\Gamma^2(1 + 1/\hat{k})}$. All 837 FFN fits in our cohort pass this Γ closure check with relative error below 2%, confirming that the middle-80% probability-plot fit recovers a self-consistent Weibull parameterization rather than a numerical artifact.

A.2 Initialization Reference for Pythia

Together, the closed-form anchor $(k_0, \lambda_0) \approx (1.205,\ 0.8875\,\sigma_{\text{init}})$ provides a complete reference: both depend only on σ_{init} and the fit protocol. The shape $k_0 \approx 1.20$ is universal across vendors; the scale λ_0 is initialization-scheme-specific and depends on which component of the model the weight matrix belongs to.

Component-specific σ_{init} in Pythia. Within Pythia, σ_{init} varies by component because input-side and output-side projections use different initializers in the GPT-NeoX codebase (Black et al., 2022). Two distinct recipes appear across the five sizes:

Recipe A (70m / 160m / 410m / 1B). Input-side projections (W_{qkv}, $W_{\text{FFN\_in}}$) use `small_init` with $\sigma_{\text{small\_init}} = \sqrt{2/(5d)}$ (Nguyen & Salazar, 2019); output-side projections (W_o, $W_{\text{FFN\_out}}$) use `wang_init` with $\sigma_{\text{wang\_init}} = 2/(L\sqrt{d})$ (Black et al., 2022). The input-to-output ratio simplifies to $L/\sqrt{10}$.

Recipe B (6.9B). Falls back to GPT-NeoX defaults: `normal` ($\sigma_{\text{in}} = 0.02$) input-side and `scaled_normal` ($\sigma_{\text{out}} = 0.02/\sqrt{2L}$) output-side. The ratio simplifies to $\sqrt{2L}$.

Step-0 verification across 5 sizes. Since the Weibull fit gives $\lambda = c(k) \cdot \sigma$ with a kind-independent constant $c(k)$ at fixed $k_0 \approx 1.20$, the measured λ ratio at the step-0 checkpoint should match the predicted σ ratio. Table 5 reports the verification across all 5 Pythia sizes.

Size	L	d	Recipe	Predicted ratio	Measured ratio	Error
70m	6	512	A	$6/\sqrt{10} = 1.897$	1.900	0.12%
160m	12	768	A	$12/\sqrt{10} = 3.795$	3.798	0.07%
410m	24	1024	A	$24/\sqrt{10} = 7.589$	7.589	0.00%
1B	16	2048	A	$16/\sqrt{10} = 5.060$	5.053	0.13%
6.9B	32	4096	B	$\sqrt{64} = 8.000$	8.000	0.00%

Table 5: Component-specific λ_{init} ratio verification across the Pythia family. "Measured ratio" is the mean of the two input-to-output pairs at the step-0 checkpoint: attention-side $\lambda_{W_{\text{qkv}}}/\lambda_O$ and FFN-side $\lambda_{W_{\text{FFN\_in}}}/\lambda_{W_{\text{FFN\_out}}}$. All 5 sizes agree with the closed-form prediction to within 0.13%.

The two recipes produce distinct initial scaling laws: Recipe A scales the ratio linearly in L, Recipe B as $\sqrt{L}$. During training the input/output ratio collapses from these recipe-specific initial values toward $\sim 1.2\times$ at the terminal checkpoint, consistent with the component-paired growth documented in Section 5 (Pearson $r = 0.9967$ between λ_O and $\lambda_{\text{FFN\_out}}$).

A.3 The Five Driving Forces for Selection Evolution

Five candidate mechanisms have been hypothesized to drive W_q/W_k out of the Weibull family:

D1 — Functional necessity. Elhage et al. (2021) showed that functional transformers require sparse attention patterns. Producing sparse patterns forces W_q and W_k to selectively amplify some embedding directions and suppress others — a configuration that manifests as heavier tails (lower k).

D2 — AdamW sign-descent dynamics. Adam's gradient-sign updates push individual weight elements away from zero, amplifying the heavy-tail signal that D1 introduces (Kunstner et al., 2023); empirically implicated by Kaul et al. (2025).

D3 — Softmax saturation feedback. Bondarenko et al. (2023) documented "no-op" attention heads that push logits toward $\pm\infty$, backpropagating extreme values into W_k.

Two further candidate forces — D4: residual-stream coupling (Elhage et al., 2021); D5: cumulative training signal T/τ (Pythia k drift tracks T/τ monotonically, observational) — may modulate the above; neither has controlled-experiment isolation.

A.4 K-Head Architecture Constraint

In multi-head attention (MHA), each (Q_i, K_i) pair is independently stored, maximizing specialization freedom. In grouped-query attention (GQA) (Ainslie et al., 2023), a single K_j head serves 4 to 7 query heads simultaneously (4-to-1 for LLaMA-3/Mistral, 7-to-1 for Qwen2.5-7B), mechanically constraining how selectively W_k can respond to any single query direction. This architectural constraint is consistent with the observed MHA/GQA dichotomy in k values.

B Software and Data

B.1 `npm-weibull-py v0.4`

The `npm-weibull-py` library provides eight diagnostic functions for fitting and benchmarking Weibull parameters on transformer weight matrices:

Function	Description
`F1_extract_weights`	Extract all weight matrices from a specified layer
`F2_fit_weibull`	Fit Weibull(k, λ) via least-squares on the Weibull probability plot
`F3_gamma_closure`	Verify Γ closure consistency
`F4_cross_family_band`	Compute per-entry median k, aggregate to cross-family CV and band
`F5_lambda_scaling`	Fit $\lambda \sim \sqrt{\eta/\lambda_{\text{wd}}}$ within the Pythia family
`F6_k_drift`	Compute k drift magnitude: $\Delta k = k_{\text{terminal}} - k_{\text{init}}$
`F7_attention_arch_classify`	Classify attention architecture: MHA / GQA / MQA
`F8_lambda_paired_correlation`	Pearson correlation of λ_O vs. $\lambda_{\text{FFN\_out}}$

Table 6: The eight diagnostic functions of `npm-weibull-py v0.4`.

The library is pip-installable; source at
[GitHub URL withheld for double-blind review; code and database publicly released and will be linked in camera-ready version].

B.2 DATABASE_v9_1

The companion benchmark database contains per-component Weibull fits for 12 model entries across 7 architectural families:

- **Models**: Pythia-70m/160m/410m/1B/6.9B, OLMo-1-7B, OLMo-2-7B, LLaMA-3-8B, Mistral-7B, Qwen2.5-7B/14B, Qwen3-8B
- **Components**: $W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}, W_o, W_q, W_k, W_{\text{qkv}}$
- **Metrics per component**: $k_{80\%}$, $k_{90\%}$, $k_{100\%}$, λ, R^2, Γ closure relative error
- **Format**: JSON, one file per model/checkpoint

Released alongside the paper at *[GitHub URL withheld for double-blind review; code and database publicly released and will be linked in camera-ready version].*

Model sources. All weights analyzed in this work were extracted from publicly-available open-source checkpoints on the Hugging Face Hub under their original licenses:

- Pythia 70m/160m/410m/1B/6.9B: `EleutherAI/pythia-{70m,160m,410m,1b,6.9b}`
- OLMo-1-7B: `allenai/OLMo-7B`

- OLMo-2-7B: `allenai/OLMo-2-1124-7B`
- LLaMA-3-8B: `meta-llama/Meta-Llama-3-8B`
- Mistral-7B: `mistralai/Mistral-7B-v0.1`
- Qwen2.5-7B/14B: `Qwen/Qwen2.5-7B`, `Qwen/Qwen2.5-14B`
- Qwen3-8B: `Qwen/Qwen3-8B`

Pythia releases all training checkpoints from step 0 to step 143000; we use 14 log-spaced revisions for trajectory analyses and the terminal checkpoint for cross-family comparisons. No model was retrained or fine-tuned in this study; all analyses are run on the released weights without modification.

C Diagnostic Framework Application Cases

This appendix documents worked application cases of the diagnostic framework. The first case — expanding the within-entry deviation flagged in §3.4 — studies a Qwen-family shallow FFN signature; future application cases (e.g., from forthcoming follow-up studies) will be added here as additional subsections.

C.1 Cohort scope

This appendix expands the diagnostic case study (§3.4). The 11-entry Qwen cohort comprises the three Qwen entries from our main cohort plus eight additional Qwen-family entries (Qwen2-1.5B/7B, Qwen2.5-1.5B/3B, four base–Math-CPT pairs), spanning three generations (Qwen2, Qwen2.5, Qwen3), five sizes (1.5B/3B/7B/8B/14B), and three depths (28L/36L/48L); the full enumeration is in Table 7.

C.2 Anomaly detection and regime partition

We classify each entry by the minimum shallow-block k across $W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}$ ($\ell \in [0, 5]$): *Regime A (clean)* if ≥ 1.0, *Regime B (bimodal)* if ≤ 0.7. The partition is clean: 4 Regime A entries (all 1.5B variants, independent of Math-CPT) and 7 Regime B entries (all 7B/8B/14B; Table 7). The same protocol applied to the 9 non-Qwen entries in our main cohort (Pythia ×5, OLMo-1, OLMo-2, LLaMA-3, Mistral) yields zero Regime B detections; the deviation is Qwen-family-specific within the cohort surveyed.

Entry	Layers	Params	Tokens	Shal. med. k_{g}	Shal. min k	Deep med. k_{g}	Regime
Qwen2-1.5B-base	28	1.5B	7T	1.19	1.18	1.18	A
Qwen2.5-1.5B	28	1.5B	18T	1.18	1.17	1.17	A
Qwen2-Math-1.5B	28	1.5B	7T + Math CPT	1.19	1.19	1.19	A
Qwen2.5-Math-1.5B	28	1.5B	18T + Math CPT	1.19	1.12	1.19	A
Qwen2-7B	28	7B	7T	0.96	0.53	1.19	B
Qwen2.5-7B	28	7B	18T	0.97	0.57	1.19	B
Qwen2-Math-7B	28	7B	7T + 200B Math	0.96	0.52	1.20	B
Qwen2.5-Math-7B	28	7B	18T + 700B Math	1.01	0.62	1.20	B
Qwen2.5-3B	36	3B	18T	0.65	0.57	1.18	B
Qwen3-8B	36	8B	36T	0.69	0.34	1.19	B
Qwen2.5-14B	48	14B	18T	0.44	0.40	1.19	B

Table 7: 11-entry Qwen-family cohort: shallow vs deep block k medians, regime classification. Shal. = shallow (blocks 0–5); Deep = blocks 10 onwards; $k_{\text{g}} \equiv k$ for W_{gate}. "Shal. min k" is the minimum block-level k across $W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}$ within shallow blocks. Tokens: base = pretraining; "+ Math CPT" / "+ 200B Math" = math-domain continued pretraining. Regime A: shallow min $k \geq 1.0$ (clean). Regime B: shallow min $k \leq 0.7$ (bimodal).

C.3 Family-wide phenomenology

Figure 13 decomposes the cohort along the block axis; Figure 14 drills down into representative blocks. The two views together characterize where the deviation occurs and what it looks like at the weight level. The block-axis view exposes spatial localization: in all 7 Regime B entries, the k drop and paired λ collapse are confined to shallow blocks ($\ell \in [1, 6]$) and appear simultaneously in all three SwiGLU sub-components ($W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}$); deep blocks ($\ell \geq 10$) return to the Transmission band. The match across the three sub-components is exact at the per-block level, ruling out any single-component anomaly story.

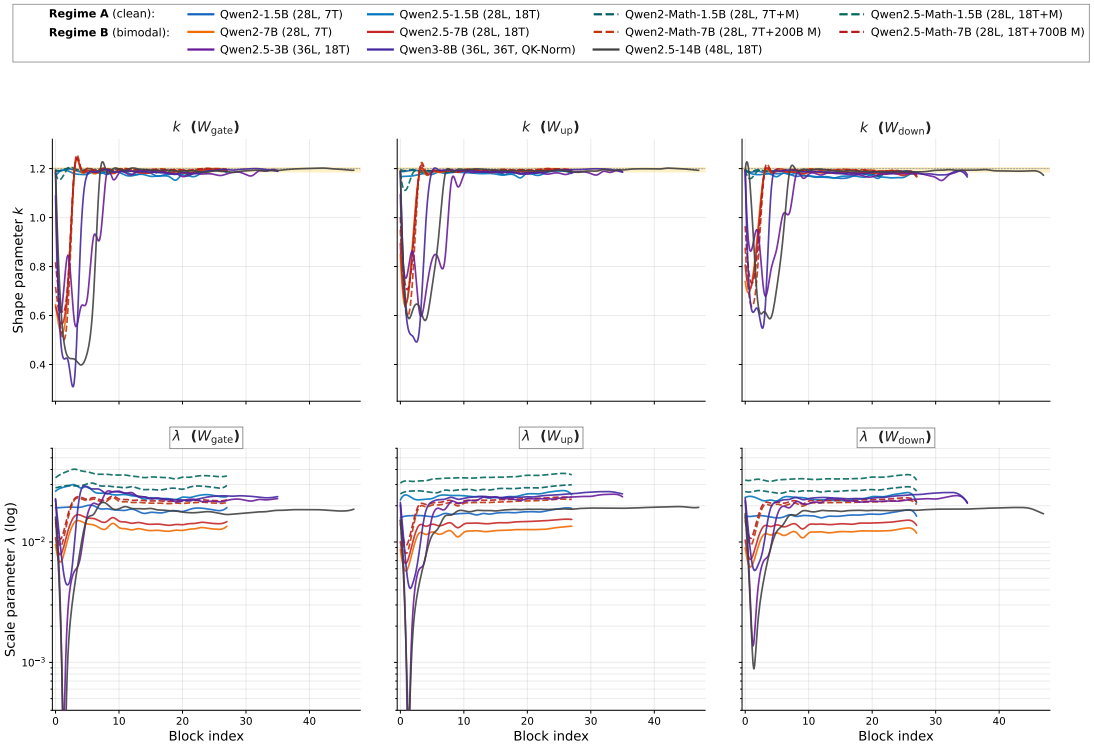

Figure 13: Per-block FFN shape parameter k (top row) and scale parameter λ (bottom row) across the 11-entry Qwen-family cohort, for the three SwiGLU FFN sub-components W_{gate}, W_{up}, W_{down}. The amber shaded band in the k panels marks the Transmission band $k \in [1.186, 1.204]$ established in Section 3 across the 12-entry main cohort. **Regime A** (1.5B class, N=4, cool blue/teal lines): every block remains within the Transmission band; λ trajectory is monotonic across depth. **Regime B** (7B+ class, N=7, warm red/orange/purple/gray lines): shallow blocks ($\ell \in [1, 6]$) show paired k drop and λ collapse across all three FFN sub-components, while deep blocks ($\ell \geq 10$) return to the Transmission band. The Regime A/B partition aligns with model parameter count and is independent of base vs. Math-CPT distinction (Table 7).

The drill-down view (Figure 14) reveals the microstructure underlying the per-block k drop: in shallow blocks, Regime B entries show a main body at $\log_{10}|w| \approx -2$ plus a secondary small-magnitude population extending into $\log_{10}|w| \in [-7, -3]$. In deep blocks, all 11 entries — including those with severe shallow drift — collapse onto a unimodal main body, confirming the deviation is a genuine bimodal microstructure spatially confined to shallow blocks rather than a global property of the Regime B entries.

This bimodal microstructure is structurally distinct from the super-weight phenomenon (Yu et al., 2024), which describes a small number of large-magnitude outliers. Per-block $\max/q_{99}$ ratios on shallow FFN blocks (Table 2) show Qwen shallow-FFN ratios (median 10.4×, max 19.3×) *lower* than the SwiGLU controls (OLMo-1/2, LLaMA-3, Mistral; max 107.2× at OLMo-1) that nevertheless maintain single-mode bodies. By this max/q99 metric, the Qwen signature is body-level bimodal (a separate population near $|w| \approx 10^{-5}$)

rather than tail-level outlier — a distinction the diagnostic framework surfaces directly through the k value rather than through ad-hoc outlier counting.

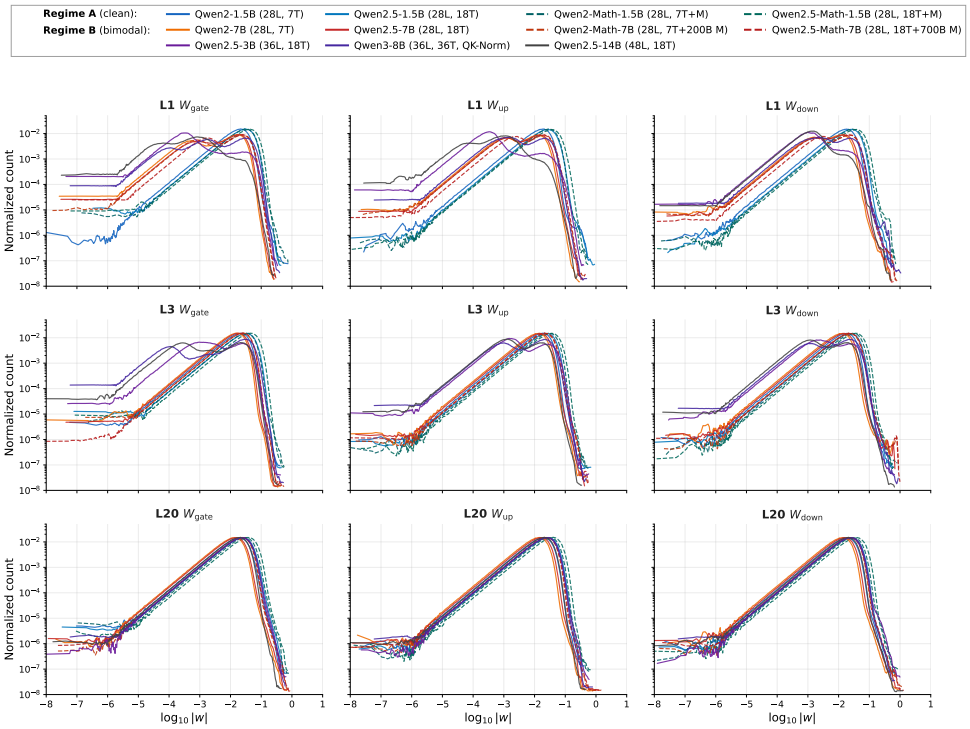

Figure 14: FFN weight magnitude $\log_{10}|w|$ distributions across the 11-entry Qwen-family cohort at three representative layer depths (rows: L1 shallow, L3 shallow, L20 deep) and three SwiGLU sub-components (columns: W_{gate}, W_{up}, W_{down}). **L1 / L3 rows (shallow):** Regime A entries (cool blue/teal, 4 entries, all 1.5B) show a single unimodal main body at $\log_{10}|w| \approx -2$; Regime B entries (warm colors, 7 entries, 7B+) show the main body plus a secondary small-magnitude population extending into $\log_{10}|w| \in [-7, -3]$. **L20 row (deep):** all 11 entries collapse onto a single unimodal main body — the regime-B secondary population is absent. This demonstrates the bimodal signature is spatially localized to shallow blocks rather than a global property of the Regime B entries.

C.4 Independent corroboration and deferred mechanism

To verify that the framework caught a real anomaly rather than a measurement artifact, we cross-checked against independent published findings. Wong et al. (2025) reported secondary attention sinks exclusively in the Qwen family (an activation-side match) and documented emergence following math-mid-training (Qwen2 → Qwen2-Math), noting that LLaMA-3.1, Phi-4, Mathstral, and CodeLlama do not exhibit the signature. Zhang et al. (2026) reported localized Projected Adapter Gradient Energy (PAGE) peaks at shallow FFN down-projection layers across multiple model families. Voita et al. (2023) documented dead neurons in the first half of large language models. Convergence across activation, gradient, and dead-neuron modalities on the same family, location, and time-scale supports the framework's diagnostic reasonableness. Disentangling the causal mechanism requires controlled experiments — observational cohort analysis cannot isolate causes — and is reserved for a focused follow-up study (e.g., matched-architecture pretraining with vs. without late-stage recipe shocks).

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4895–4901, Singapore, 2023. Association for Computational Linguistics.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. Published Dec 22, 2021. `https://transformer-circuits.pub/2021/framework/index.html`.

Zhiyuan Fan, Yifeng Liu, Qingyue Zhao, Angela Yuan, and Quanquan Gu. Robust layerwise scaling rules by proper weight decay tuning. *arXiv preprint arXiv:2510.15262*, 2025.

Di He, Songjun Tu, Ajay Jaiswal, Li Shen, Ganzhao Yuan, Shiwei Liu, and Lu Yin. AlphaDecay: Module-wise weight decay for heavy-tailed balancing in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unravelling the enigmas of large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Frederik Kunstner, Jacques Chen, J. Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *International Conference on Learning Representations (ICLR)*, 2023.

Charles H. Martin and Michael W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4284–4293. PMLR, 2019.

Charles H. Martin and Michael W. Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*, pp. 505–513. SIAM, 2020.

Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. In *International Conference on Spoken Language Translation (IWSLT)*, 2019.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread, arXiv preprint arXiv:2209.11895*, 2022. Published Mar 8, 2022.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024. COLM 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5797–5808, Florence, Italy, 2019. Association for Computational Linguistics.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*, 2023.

Xi Wang and Laurence Aitchison. How to set AdamW's weight decay as you scale model and dataset size. *arXiv preprint arXiv:2405.13698*, 2024. Preprint; v3 released 1 Jun 2025.

Waloddi Weibull. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18 (3):293–297, 1951. doi: 10.1115/1.4010337.

Jeffrey T. H. Wong, Cheng Zhang, Louis Mahon, Wayne Luk, Anton Isopoussu, and Yiren Zhao. On the existence and behavior of secondary attention sinks. *arXiv preprint arXiv:2512.22213*, 2025.

Mengxia Yu, De Wang, Qi Shan, Colorado J. Reed, and Alvin Wan. The super weight in large language models. *arXiv preprint arXiv:2411.07191*, 2024.

Suoxin Zhang, Run He, Di Fang, Xiang Tan, Kaixuan Chen, and Huiping Zhuang. Rethinking adapter placement: A dominant adaptation module perspective. *arXiv preprint arXiv:2605.06183*, 2026.