DECOUPLING COORDINATE PERMUTATIONS FROM LOW-DIMENSIONAL QUANTILES IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training weight modeling and quantization schemes predominantly treat neural network tensors as monolithic coordinate arrays, simultaneously optimizing value discretization and positional assignment. In this work, we investigate a fundamental question: *How much information truly resides within the scalar values of a Large Language Model (LLM) weight tensor when strictly decoupled from coordinates?* We formalize an orthogonal decomposition of any weight matrix $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ into a sorted value spectrum $\boldsymbol{s} = \mathrm{sort}(\mathrm{vec}(\boldsymbol{W}))$ and a spatial permutation $\pi \in \mathbb{S}_N$. Stripping away positional entropy reveals an unexpected empirical regularity: the position-agnostic value spectrum of trained LLM tensors is universally governed by an ultra-low-dimensional continuum. Across diverse open-weight architectures (Meta-Llama-3.2, Qwen-3, Gemma-4) and across every internal layer type, an unconstrained 5-term Metalog quantile distribution fits the empirical spectrum with near-zero error ($\mathrm{MAE} < 0.0002$, relative $\mathrm{MSE} < 0.002$). This closed-form parameterization decisively outperforms classical Gaussian, extreme-value, and generalized quantile baselines. Crucially, our findings expose an extreme information-theoretic asymmetry: while the value distribution collapses to an $\mathbb{O}(1)$ description length of just 5 scalars (160 bits), spatial permutations retain an irreducible combinatorial entropy of $\lceil \log_2(N!) \rceil \sim \mathbb{O}(N \log N)$ bits. We demonstrate that learned model capacity resides almost entirely within spatial permutations rather than value variance, establishing that future model compression paradigms must pivot toward structured permutation coding.

1 INTRODUCTION

Modern compression and quantization frameworks for Large Language Models (LLMs) operate under the premise that parameter tensors are dense, high-dimensional coordinate arrays whose entries require sophisticated numerical discretization (Dettmers et al., 2023). Significant research effort has been dedicated to constructing sensitivity-weighted codebooks, non-uniform lookup tables, and dedicated sparse handling for heavy-tailed outlier coordinates. Despite these engineering advances, an elementary information-theoretic question remains largely unaddressed: *How much intrinsic information is contained within the raw numerical values of an LLM weight tensor, independent of their spatial positions?*

To answer this, we formalize an exact, loss-free orthogonal decomposition of any weight matrix $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ into two decoupled components:

$$\boldsymbol{W} = \boldsymbol{P}_\pi \boldsymbol{s}, \tag{1}$$

where $\boldsymbol{s} = \mathrm{sort}(\mathrm{vec}(\boldsymbol{W})) \in \mathbb{R}^N$ ($N = m \cdot n$) is a strictly monotonically non-decreasing vector capturing the coordinate-free value spectrum, and $\boldsymbol{P}_\pi \in \{0,1\}^{N \times N}$ is a permutation matrix corresponding to an element $\pi \in \mathbb{S}_N$ that encodes the spatial positioning of each scalar within the tensor (Figure 1). By sorting the entries, spatial coordinates are eliminated, transforming the weights into a direct realization of their underlying empirical quantile function:

$$u_i = \frac{i + 0.5}{N}, \quad s_i = F^{-1}(u_i), \quad i \in \{0, \ldots, N-1\}. \tag{2}$$

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.

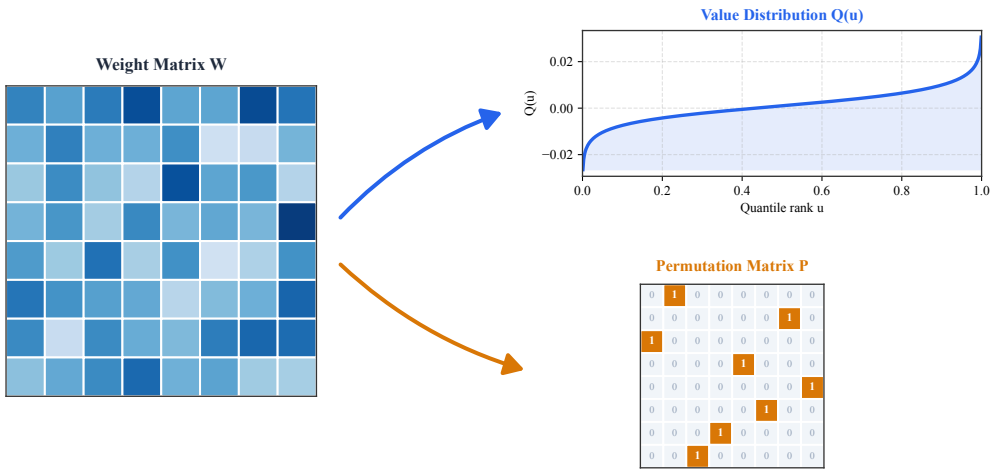


Figure 1: Schematic overview of the orthogonal value-position decomposition. A 2D weight matrix $\boldsymbol{W}$ is flattened and sorted into an empirical quantile spectrum $\boldsymbol{s} = \mathrm{sort}(\mathrm{vec}(\boldsymbol{W}))$, completely isolating value magnitudes from positional entropy. The spatial assignment is governed by a permutation operator $\boldsymbol{P}_\pi$ requiring $\lceil \log_2(N!) \rceil$ bits, whereas the continuous spectrum $\boldsymbol{s}$ collapses to a 5-parameter Metalog quantile function $Q(u)$.

Our primary empirical finding is that this coordinate-free value spectrum is remarkably simple. While unconstrained weight matrices span millions of parameters per layer, their quantile spectra lie on an ultra-low-dimensional manifold. Specifically, we demonstrate that a 5-term Metalog distribution (Keelin, 2016) approximates the sorted empirical weights across entire models with near-perfect fidelity, yielding a mean absolute error (MAE) below 0.0002 and a relative mean squared error (rel-MSE) below 0.002.

This discovery uncovers an acute information-theoretic asymmetry between value distribution and coordinate placement:

- **Value Entropy is $\mathbb{O}(1)$:** Modeling the continuous quantile spectrum $\boldsymbol{s}$ using a 5-term Metalog distribution requires storing exactly five scalar parameters $\boldsymbol{c} = [c_1, c_2, c_3, c_4, c_5]^\top$. In single-precision floating point, this corresponds to $5 \times 32 = 160$ bits of total metadata per weight matrix, regardless of whether $N = 10^6$ or $N = 10^8$.
- **Permutation Entropy is $\mathbb{O}(N \log N)$:** Encoding the spatial permutation $\pi \in \mathbb{S}_N$ combinatorially demands $\lceil \log_2(N!) \rceil \approx N \log_2 N - 1.4427N$ bits.

This asymmetry demonstrates that the learned capacity and functional complexity of an LLM are almost entirely concentrated within its combinatorial permutation structure. The continuous values themselves contribute virtually negligible dimensional complexity.

Furthermore, our empirical observations resolve a prominent theoretical tension regarding weight distributions. Popular quantization priors such as NormalFloat4 (NF4) in QLoRA (Dettmers et al., 2023) assume idealized Gaussian weights. However, theoretical treatments of stochastic gradient descent dynamics (Barsbey et al., 2021) and random matrix analyses of trained deep networks (Martin & Mahoney, 2019; Vladimirova et al., 2019) prove that weights necessarily develop heavy-tailed, leptokurtic spectra as an outcome of implicit self-regularization. Standard heavy-tailed distributions (e.g., Student-t, Pareto) lack analytically tractable, closed-form quantile functions, preventing accurate closed-form spectrum fitting.

The 5-term Metalog distribution reconciles these regimes in closed form: the first two terms establish central location and scale, the third term captures asymmetric skewness, and the fourth and fifth terms absorb the heavy tails and excess kurtosis. Because the basis functions are linear with respect to the coefficients, optimal parameters are computable directly via chunked linear least squares within milliseconds, avoiding non-convex iterative optimization.

The contributions of this paper are summarized as follows:

1. **Orthogonal Spectrum Decomposition:** We formalize the exact decoupling of LLM weight matrices into coordinate-free empirical quantile spectra and combinatorial permutations, framing parameter analysis strictly through an information-theoretic lens.
2. **Universality of the 5-Term Metalog Fit:** We conduct an exhaustive evaluation across Meta-Llama-3.2-3B, Qwen3-8B, and Gemma4-12B, demonstrating that every 2D weight matrix is accurately approximated by a 5-term Metalog distribution (Table 1, Figure 2).
3. **Distributional Benchmark:** We benchmark the Metalog distribution against seven classical parametric and generalized quantile families—including Gaussian, Lognormal, Gumbel (Gumbel, 1935), Weibull (Weibull, 1951), Exponential, Generalized Lambda Distributions (GLD) (Ramberg & Schmeiser, 1974), and Tukey Lambda (Tukey, 1962)—proving that Metalog provides the strictly superior fit across transformer weight spectra (Table 2).
4. **Scalable Chunked Formulation:** We derive a numerically stable, chunked accumulation algorithm with Tikhonov regularization that evaluates parameter matrices containing tens of millions of weights without memory bottlenecks.
5. **A New Frontier for Compression:** We establish that value discretization is effectively solved by low-dimensional quantile parameterizations, identifying the compression of the $\mathbb{O}(N \log N)$ permutation bottleneck as the critical challenge for next-generation model compression.

2 Related Work

2.1 Weight Distribution Modeling and Quantization Priors

Modern post-training quantization methods rely on structural assumptions regarding the statistical distribution of weights across deep architectures. A widely adopted paradigm is the NormalFloat4 (NF4) data type introduced in QLoRA (Dettmers et al., 2023), which constructs an information-theoretically optimal quantile grid under the explicit assumption that pre-trained transformer parameters $\boldsymbol{W}$ follow an idealized zero-mean Gaussian distribution $\mathcal{N}(0, \sigma^2)$. While this parametric prior provides a computationally convenient discretization grid for 4-bit representation, it treats weight tensors as coordinate-homogeneous Gaussian fields.

2.2 Heavy Tails and Implicit Self-Regularization

A growing body of statistical learning theory refutes the assumption that trained deep network weights are normally distributed, demonstrating instead that overparameterized models inherently develop heavy-tailed, leptokurtic distributions.

First, from an optimization standpoint, Barsbey et al. (2021) show that the stochastic gradient noise arising from mini-batch sampling does not conform to classical Brownian motion. Instead, gradient updates follow heavy-tailed α-stable Lévy processes, driving the parameters into heavy-tailed stationary distributions and governing the intrinsic compressibility of overparameterized networks.

Second, from the perspective of random matrix theory (RMT), Martin & Mahoney (2019) analyze the empirical spectral density (ESD) of weight matrices across diverse architectures. If weight entries were independent and identically distributed Gaussians, the eigenvalues of the correlation matrix $\boldsymbol{W}^\top \boldsymbol{W}$ would strictly conform to the Marchenko-Pastur distribution with bounded support. Instead, Martin & Mahoney (2019) prove that well-trained models universally violate Marchenko-Pastur predictions, exhibiting power-law spectral tails ($\rho(\lambda) \sim \lambda^{-\alpha}$) driven by implicit self-regularization during training. These tail characteristics are further amplified through deep network composition, where non-linear activation layers systematically induce heavy-tailed behavior across layer units (Vladimirova et al., 2019).

2.3 Quantile-Parameterized Distributions and the Metalog Family

When empirical data exhibit non-standard skewness and heavy tails, classical parametric density functions often prove too rigid. To provide flexible shape modeling, statistical theory established the

framework of Quantile-Parameterized Distributions (QPDs), most prominently the Tukey Lambda distribution (Tukey, 1962) and its four-parameter generalization, the Generalized Lambda Distribution (GLD) (Ramberg & Schmeiser, 1974). By modeling the inverse cumulative distribution $Q(u) = F^{-1}(u)$ directly, QPDs can accommodate wide ranges of kurtosis and asymmetry. However, classical QPD formulations define their shape parameters non-linearly (e.g., as exponents of probability terms). Consequently, parameter estimation requires non-convex numerical optimization, grid searches, or method-of-moments approximations, which are computationally expensive, prone to non-convergence, and unstable when scaled to datasets with millions of observations.

To overcome these estimation bottlenecks, Keelin (2016) introduced the Metalog distribution family. Unlike classical QPDs, the Metalog quantile function is parameterized as an expansion over deterministic basis functions—combining a central logistic kernel with polynomial transforms of the centered probability $z = u - 0.5$. Because the formulation is strictly linear with respect to its coefficients, parameter estimation reduces to an ordinary linear least squares problem. This permits closed-form, deterministic fitting of arbitrary degrees of skewness and heavy-tailed kurtosis in milliseconds.

3 METHODOLOGY

3.1 QUANTILE FUNCTIONS IN DISTRIBUTION THEORY

Let X be a real-valued continuous random variable characterized by a cumulative distribution function (CDF) $F_X(x) = P(X \leq x)$. The quantile function (or inverse CDF) $Q_X : (0, 1) \to \mathbb{R}$ is defined as:

$$Q_X(u) = F_X^{-1}(u) = \inf \{x \in \mathbb{R} : F_X(x) \geq u\}, \quad u \in (0, 1). \tag{3}$$

While standard density modeling operates in the probability space via the probability density function (PDF) $f_X(x) = \frac{dF_X(x)}{dx}$, quantile representations invert this relationship: the input domain is the standardized probability interval $u \in (0, 1)$, and the codomain is the data space $\mathbb{R}$.

Quantile functions offer distinct structural advantages when modeling empirical parameters. First, sorting a discrete set of observations directly constructs the non-parametric empirical quantile function without requiring continuous kernel bandwidth estimation, spatial binning, or numerical integration. Second, quantile functions model asymmetric skewness and leptokurtic tails directly as asymptotic properties of $Q(u)$ as $u \to 0^+$ and $u \to 1^-$, avoiding the numerical instability often encountered during maximum likelihood estimation of heavy-tailed density functions.

3.2 THE METALOG DISTRIBUTION FAMILY

The Metalog distribution (Keelin, 2016) is an unconstrained, highly flexible system of continuous probability distributions defined directly by its quantile function. Unlike classical distributions whose quantile functions are non-linear or lack closed-form representations (e.g., the Gaussian error function or Student-t integrals), the Metalog quantile function is parameterized as a linear combination of polynomial and logistic basis functions:

$$Q_K(u) = \sum_{j=1}^{K} c_j b_j(u), \tag{4}$$

where $\boldsymbol{c} = [c_1, \ldots, c_K]^\top \in \mathbb{R}^K$ is a vector of expansion coefficients and $b_j(u)$ denote deterministic basis functions. Following Keelin (2016), we center the probability domain by defining $z = u - 0.5$ and the canonical logit transform:

$$\text{logit}(u) = \ln\left(\frac{u}{1-u}\right). \tag{5}$$

The basis terms $b_j(u)$ are constructed recursively:

$$b_1(u) = 1, \tag{6}$$

$$b_2(u) = \text{logit}(u), \tag{7}$$

$$b_j(u) = z^{(j-1)/2} \,\text{logit}(u) \quad \text{for odd } j \geq 3, \tag{8}$$

$$b_j(u) = z^{(j/2)-1} \quad \text{for even } j \geq 4. \tag{9}$$

Each coefficient in this expansion governs an interpretable geometric transformation of the resulting distribution:

- c_1 controls the **location** (median shift, as $z = 0$ at $u = 0.5$).
- c_2 controls the **scale** (dispersion of the central bulk).
- c_3 controls the **skewness** (asymmetry between upper and lower quantiles).
- c_4, c_5 govern **kurtosis and tail heaviness** (rate of tail divergence relative to the center).

For a 5-term Metalog specification ($K = 5$), the closed-form quantile function simplifies to:

$$Q_5(u) = c_1 + c_2 \operatorname{logit}(u) + c_3 \left(u - \frac{1}{2}\right) \operatorname{logit}(u) + c_4 \left(u - \frac{1}{2}\right) + c_5 \left(u - \frac{1}{2}\right)^2 \operatorname{logit}(u). \quad (10)$$

Crucially, because Equation 4 is strictly linear in the coefficients $\boldsymbol{c}$, fitting the distribution to empirical data reduces to a convex linear least squares problem, eliminating the need for iterative numerical optimizations.

3.3 Orthogonal Weight Decoupling and Empirical Quantile Mapping

Given a two-dimensional weight tensor $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ representing an arbitrary linear projection within a transformer layer, we flatten the tensor into a one-dimensional coordinate array:

$$\boldsymbol{w} = \operatorname{vec}(\boldsymbol{W}) \in \mathbb{R}^N, \quad N = m \cdot n. \quad (11)$$

To eliminate all spatial coordinate information, we apply an ascending sort permutation operator $\operatorname{sort}(\cdot)$ to yield the ordered spectrum:

$$\boldsymbol{w}_{\text{sorted}} = \operatorname{sort}(\boldsymbol{w}), \quad \text{such that } s_0 \leq s_1 \leq \cdots \leq s_{N-1}. \quad (12)$$

This operation formally factors the weight tensor into:

$$\boldsymbol{w} = \boldsymbol{P}_\pi \boldsymbol{w}_{\text{sorted}}, \quad (13)$$

where $\boldsymbol{P}_\pi \in \{0, 1\}^{N \times N}$ is a permutation matrix governed by $\pi \in \mathbb{S}_N$, satisfying $\boldsymbol{P}_\pi \boldsymbol{e}_i = \boldsymbol{e}_{\pi(i)}$.

The sorted vector $\boldsymbol{w}_{\text{sorted}}$ represents the empirical quantile spectrum of the parameter distribution. To align discrete parameter indices with the continuous support $u \in (0, 1)$, we assign each element s_i to its continuous probability quantile via the standard midpoint formulation:

$$u_i = \frac{i + 0.5}{N}, \quad \text{for } i \in \{0, \ldots, N - 1\}. \quad (14)$$

The $+0.5$ continuity correction strictly avoids evaluating the logit function at boundary singularities $u \in \{0, 1\}$, where $\operatorname{logit}(u)$ diverges to $\pm\infty$, ensuring well-behaved numerical evaluation across all N coordinates.

3.4 Chunked Normal Equations with Tikhonov Regularization

Evaluating the linear basis functions across all N weights defines a design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times K}$, where the entry at row i and column j is given by $X_{i,j} = b_j(u_i)$. The parameter vector $\boldsymbol{c} \in \mathbb{R}^K$ is then obtained by solving the ordinary least squares objective:

$$\min_{\boldsymbol{c} \in \mathbb{R}^K} L^2 \boldsymbol{X}\boldsymbol{c} - \boldsymbol{w}_{\text{sorted}}{}^2. \quad (15)$$

For modern LLM layers, N frequently ranges from 10^6 to 10^8. Materializing the complete design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ directly into main memory requires gigabytes of allocation per layer and triggers Out-Of-Memory (OOM) failures when scaled across entire models.

To overcome this, we reformulate the optimization through chunked accumulation of the normal equations:

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{c} = \boldsymbol{X}^\top \boldsymbol{w}_{\text{sorted}}. \quad (16)$$

We partition the index set $\{0, \ldots, N-1\}$ into $C = \lceil N/B \rceil$ contiguous blocks of fixed chunk size B (e.g., $B = 10^6$). For chunk $c \in \{1, \ldots, C\}$, we compute the local design block $\boldsymbol{X}_c \in \mathbb{R}^{B_c \times K}$ and the corresponding sorted sub-vector $\boldsymbol{w}_{\text{sorted},c} \in \mathbb{R}^{B_c}$, accumulating the inner products online:

$$\boldsymbol{X}^\top \boldsymbol{X} = \sum_{c=1}^{C} \boldsymbol{X}_c^\top \boldsymbol{X}_c, \quad \boldsymbol{X}^\top \boldsymbol{w}_{\text{sorted}} = \sum_{c=1}^{C} \boldsymbol{X}_c^\top \boldsymbol{w}_{\text{sorted},c}. \tag{17}$$

Since $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{K \times K}$ (e.g., 5×5) and $\boldsymbol{X}^\top \boldsymbol{w}_{\text{sorted}} \in \mathbb{R}^{K \times 1}$, the total resident memory footprint remains $\mathbb{O}(K^2)$, completely independent of tensor size N.

Numerical Stability via Tikhonov Regularization. Because polynomial powers of z combined with the logit transform can introduce severe multicollinearity across basis columns, the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is frequently ill-conditioned, with condition numbers $\kappa(\boldsymbol{X}^\top \boldsymbol{X}) \gg 10^{12}$. To guarantee invertibility and prevent catastrophic numerical cancellation during matrix inversion, we impose a lightweight Tikhonov regularization (Ridge penalty):

$$\boldsymbol{c}^* = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_K\right)^{-1} \boldsymbol{X}^\top \boldsymbol{w}_{\text{sorted}}, \tag{18}$$

where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix and λ is a damping factor fixed to $\lambda = 10^{-6}$. This regularization ensures that the closed-form solution remains stable across all transformer projections without perturbing the fitted quantiles.

4 Experiments and Results

To validate our findings, we conduct an extensive empirical evaluation designed to answer two central questions: (1) *Is the low-dimensional compressibility of the value spectrum universal across modern LLM architectures and scales?* and (2) *How does the Metalog quantile formulation compare against classical parametric and generalized quantile distributions?*

4.1 Experimental Setup

Evaluated Model Suite. We benchmark across a diverse selection of state-of-the-art open-weights causal language models spanning distinct model families, parameter scales, and training recipes:

- **Meta-Llama-3.2-3B** (Dubey et al., 2024)
- **Qwen3-8B** (Qwen Team, 2025)
- **Gemma4-12B** (Gemma Team, 2026)

Scope of Evaluated Layers. Rather than isolating specific sub-modules, our evaluation covers the *entire* parameter space of each model. We extract and evaluate every two-dimensional weight matrix across all transformer blocks:

- **Multi-Head Attention:** Query ($\boldsymbol{W}_q$), Key ($\boldsymbol{W}_k$), Value ($\boldsymbol{W}_v$), and Output Projection ($\boldsymbol{W}_o$) matrices.
- **Feed-Forward Networks (MLP):** Gate ($\boldsymbol{W}_{\text{gate}}$), Up ($\boldsymbol{W}_{\text{up}}$), and Down ($\boldsymbol{W}_{\text{down}}$) projections.
- **Input/Output Projections:** The input token embedding table ($\boldsymbol{W}_{\text{emb}}$) and the final unembedding language model head ($\boldsymbol{W}_{\text{head}}$).

Basis Sweep (K). To examine how incremental basis expansions capture higher-order statistical properties, we vary the number of Metalog basis terms systematically from $K = 2$ to $K = 5$:

- $K = 2$: Captures only location (c_1) and scale (c_2), directly analogous to a two-parameter Gaussian or logistic fit.
- $K = 3$: Introduces the linear skewness term (c_3), accounting for distributional asymmetry between positive and negative parameter tails.
- $K = 4, 5$: Introduces higher-order polynomial log-odds terms (c_4, c_5), providing the degrees of freedom required to absorb excess kurtosis and heavy-tailed divergence.

Table 1: Cross-architecture reconstruction distortion (rel-MSE) as a function of Metalog basis terms K. Values represent layer-averaged metrics across all 2D weight tensors within each model.

MODEL	$K = 2$	$K = 3$	$K = 4$	$K = 5$
META-LLAMA-3.2-3B	0.0035	0.0035	0.0009	**0.0007**
QWEN3-8B	0.0063	0.0063	0.0016	**0.0008**
GEMMA4-12B-IT	0.0052	0.0052	0.0027	**0.0019**

Evaluation Metrics. Let $\boldsymbol{s} = \boldsymbol{w}_{\text{sorted}} \in \mathbb{R}^N$ denote the true empirical quantile spectrum and $\hat{\boldsymbol{s}} \in \mathbb{R}^N$ denote the reconstructed spectrum generated via $\hat{s}_i = Q_K(u_i)$. We evaluate reconstruction accuracy via two complementary metrics:

1. **Mean Absolute Error (MAE):**
$$\text{MAE}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{1}{N} \sum_{i=0}^{N-1} |s_i - \hat{s}_i|. \tag{19}$$
MAE provides a robust, linear measurement of average coordinate displacement across the central probability mass (the bulk).
2. **Relative Mean Squared Error (rel-MSE):**
$$\text{rel-MSE}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{L^2\boldsymbol{s} - \hat{\boldsymbol{s}}^2}{L^2\boldsymbol{s}^2} = \frac{\sum_{i=0}^{N-1}(s_i - \hat{s}_i)^2}{\sum_{i=0}^{N-1} s_i^2}. \tag{20}$$
Because standard mean squared error scales with the variance of each tensor, rel-MSE normalizes distortion by the total signal energy. Furthermore, the quadratic penalty makes rel-MSE highly sensitive to tail misfits, accurately punishing models that fail to track extreme outlier values.

4.2 CROSS-ARCHITECTURE UNIVERSALITY (EXPERIMENT 1)

Table 1 summarizes the reconstruction distortion across all three model architectures as the basis order K increases. Several critical patterns emerge:

First, the transition from $K = 2$ to $K = 5$ yields up to an eight-fold reduction in relative error across all evaluated architectures. At $K = 2$, where the formulation only calibrates location and scale, the average rel-MSE remains between 0.0035 and 0.0063. Incorporating the linear skewness term ($K = 3$) yields unchanged distortion ($0.0035 - 0.0063$), directly reflecting the pronounced bilateral symmetry of centered weight distributions around zero. The decisive performance gain occurs upon introducing higher-order terms ($K = 4, 5$): explicitly modeling excess kurtosis and heavy-tailed curvature drives the average relative error down to $0.0007 - 0.0019$.

Second, this convergence behavior is remarkably uniform across models. This proves that the near-perfect approximation achieved by the 5-term Metalog fit is not an artifact of a specific parameterization or training objective, but a universal property of transformer weight distributions.

Layer-Wise Robustness. To verify that this low distortion is not masking catastrophic failures in specific sub-modules, Figure 2 presents the layer-wise distribution of rel-MSE across functional projection groups in Qwen3-8B. Across attention projections ($\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v, \boldsymbol{W}_o$), MLP layers ($\boldsymbol{W}_{\text{gate}}, \boldsymbol{W}_{\text{up}}, \boldsymbol{W}_{\text{down}}$), and embedding heads, the error distributions remain tightly bounded with median rel-MSE < 0.001, demonstrating that the 5-term fit is robust across functionally distinct layer types.

4.3 BENCHMARK AGAINST ALTERNATIVE QUANTILE FAMILIES (EXPERIMENT 2)

To assess whether the Metalog formulation is uniquely suited for modeling weight spectra, we benchmark it against seven candidate probability distributions in Table 2. The baselines comprise classical parametric distributions (Gaussian, Lognormal, Exponential, Gumbel, Weibull) and flexible generalized quantile systems (Tukey Lambda, GLD).

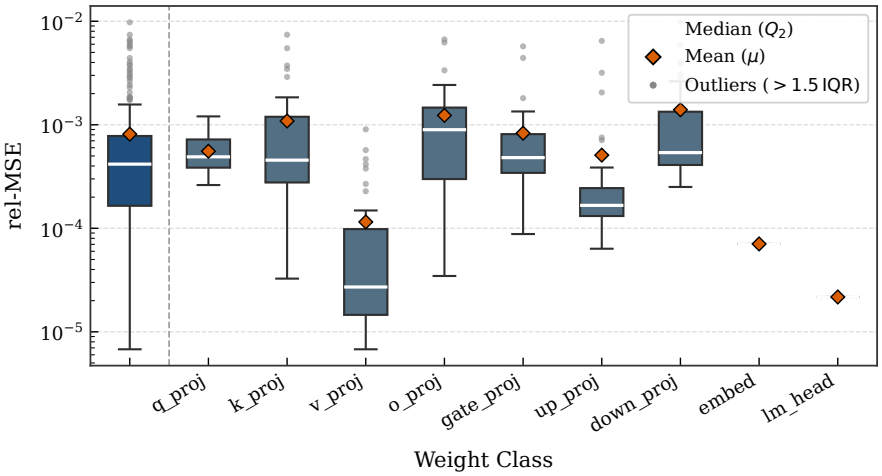


Figure 2: Distribution of layer-wise rel-MSE under a 5-term Metalog fit ($K = 5$) across all functional projection groups in Qwen3-8B. The fit remains consistently tight across self-attention, feed-forward projections, and embedding layers.

Table 2: Benchmark of distributional approximators fitted to empirical weight spectra across all layers of Meta-Llama-3.2-3B. Baselines include classical parametric families and generalized quantile distributions.

Distribution Family	Parameters (K)	MAE ($\times 10^{-4}$)	rel-MSE
Gaussian ($\mathcal{N}$)	2	8.176	0.00566
Lognormal (Shifted)	3	11.386	0.00797
Exponential (Shifted)	2	60.679	0.18038
Gumbel	2	33.076	0.05927
Weibull (Shifted)	3	12.091	0.01038
Tukey Lambda	3	1.597	0.00096
Generalized Lambda (GLD, FKML)	4	4.822	0.00342
Metalog (Ours)	**5**	**1.020**	**0.00068**

Failure of Classical Parametric Baselines. As reported in Table 2, standard parametric distributions fail to adequately capture empirical weight spectra. Strictly positive families (Exponential, Lognormal, Weibull) exhibit severe distortion (rel-MSE > 0.007, reaching 0.180 for Exponential) because their rigid hazard functions cannot accommodate symmetric parameter distributions around zero. Crucially, the Gaussian baseline—the core distributional prior in NF4 (Dettmers et al., 2023)—incurs an order of magnitude higher error (rel-MSE $= 0.00566$, MAE $= 8.176 \times 10^{-4}$) than Metalog, confirming that zero-mean Gaussian assumptions underfit the heavy tails and leptokurtic structure of LLM weights.

Generalized Quantile Systems vs. Metalog. Generalized quantile systems such as Tukey Lambda (Tukey, 1962) and GLD (Ramberg & Schmeiser, 1974) perform substantially better than rigid parametric models. Because they operate directly in quantile space, they adapt flexibly to empirical tails, with GLD achieving rel-MSE $= 0.00342$ and Tukey Lambda reaching 0.00096.

However, the Metalog distribution strictly outperforms all baselines, achieving the lowest distortion across both metrics (rel-MSE $= 0.00068$, MAE $= 1.020 \times 10^{-4}$). Beyond empirical accuracy, Metalog holds a decisive algorithmic advantage: while GLD requires iterative non-linear numerical optimization or method-of-moments heuristics that can suffer from non-convergence, Metalog parameter estimation is strictly convex and computed analytically via closed-form linear least squares (Equation 18).

4.4 Information-Theoretic Implication: The Permutation Bottleneck

Together, Experiments 1 and 2 reveal a structural truth regarding LLM parameter representations. The value spectrum of an LLM weight tensor—when stripped of coordinate positions—possesses an extremely low intrinsic dimensionality.

This result formalizes a profound informational asymmetry:

$$\text{Description Length}(\boldsymbol{W}) = \underbrace{\mathbb{O}(1)}_{\text{Values } \boldsymbol{s} \approx Q_5(u)} + \underbrace{\mathbb{O}(N \log N)}_{\text{Permutation } \boldsymbol{P}_\pi}. \tag{21}$$

Our findings prove that the value spectrum is an ultra-low-dimensional continuum. The true combinatorial bottleneck governing neural network storage and representation resides entirely within the spatial permutation $\boldsymbol{P}_\pi$.

5 Discussion and Future Directions

The finding that the value spectra of Large Language Models universally collapse onto an ultra-low-dimensional Metalog manifold carries fundamental implications for neural network representation, practical compression, and the theoretical understanding of overparameterized deep models.

5.1 Perspectives on Permutation Compression

Post-training quantization (PTQ) frameworks have achieved remarkable practical efficiency by refining the discretization of real-valued weight coordinates, employing techniques such as sensitivity-weighted codebooks, non-uniform clustering, and dedicated outlier preservation (Dettmers et al., 2023). In the context of our orthogonal decomposition, these methods optimize the coordinate-level representation of parameters. Our findings suggest that the underlying continuous value spectrum itself possesses a remarkably compact representation across layers.

At the same time, this decomposition highlights that the primary information-theoretic capacity remains associated with the spatial permutation operator $\boldsymbol{P}_\pi$. In the unconstrained setting, an arbitrary permutation over N elements requires $\lceil \log_2(N!) \rceil \approx N \log_2 N - 1.4427N$ bits, an $\mathbb{O}(N \log N)$ complexity that quickly surpasses standard floating-point representation sizes. However, parameter matrices in trained models are not uniformly distributed random permutations; neural weights inherently reflect functional structure, exhibiting channel-level correlations, rank alignments, and localized magnitude patterns.

Rather than viewing the permutation cost purely as a barrier, this observation suggests a promising and complementary direction for future compression research: exploring the structured, low-entropy characteristics of parameter ordering.

6 Conclusion

In this paper, we presented an empirical discovery concerning the structural composition of Large Language Model parameters. By formally decoupling weight matrices into spatial permutations and coordinate-free empirical quantile spectra, we demonstrated that the value distribution across trained transformer layers is strictly low-dimensional. Across open-weight models spanning multiple families and parameter scales, a 5-term Metalog distribution models the entire continuous weight spectrum with near-zero error ($\text{MAE} < 0.0002$, $\text{rel-MSE} < 0.002$).

This result exposes an informational asymmetry: while coordinate-free values collapse to an $\mathbb{O}(1)$ description length, spatial assignments retain an $\mathbb{O}(N \log N)$ combinatorial permutation entropy.

AI use statement

In this work, we used generative AI tools for editing and language polishing of draft sections. All theoretical formulations, empirical simulations, and data analyses were designed and executed by the authors. We have reviewed all text and artifacts and take full responsibility for the contents of this submission.

Ethics statement

This work studies the empirical properties and statistical distributions of weights in publicly available language models. It does not introduce novel ethical concerns, involve human subjects, or release sensitive data.

Reproducibility Statement

All model weights evaluated in this work are publicly accessible open-weight artifacts. The complete chunked least-squares fitting procedure, including basis evaluation routines and regularized normal equations, is detailed in Section 3. To ensure full reproducibility, our complete codebase, fitting scripts, and precomputed quantile parameters will be made publicly available upon publication.

References

Melih Barsbey, Milad Sefidgaran, Murat A. Erdogdu, Gaël Richard, and Umut Şimşekli. Heavy tails in sgd and compressibility of overparametrized neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29364–29378, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Emil J. Gumbel. Les valeurs extrêmes des distributions statistiques. *Annales de l'Institut Henri Poincaré*, 5(2):115–158, 1935.

Thomas W. Keelin. The metalog distributions. *Decision Analysis*, 13(4):243–277, 2016.

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

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

John S. Ramberg and Bruce W. Schmeiser. An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17(2):78–82, 1974.

John W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning (ICML)*, pp. 6458–6467. PMLR, 2019.

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

A Appendix