THEORETICALLY SCALED QUANTIZATION: A DATA-FREE ANALYTICAL FRAMEWORK FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) require efficient post-training quantization (PTQ) to reduce memory and computational costs while preserving model quality. Existing PTQ methods commonly rely on calibration data, activation statistics, Hessian information, or iterative optimization, introducing additional preprocessing overhead and scale storage. We propose Theoretically Scaled Quantization (TSQ), a fully data-free analytical PTQ framework that derives quantization parameters directly from pretrained weight statistics. We formulate the TSQ Distortion Objective, $R_N(k)$, an exact distribution-independent geometric distortion objective for finite multi-region quantization, and define the optimal geometric scaling factor through one-dimensional minimization over k. We further derive exact region-wise absolute-error bounds that decrease geometrically with region index and a region-independent relative-error bound of $k/(2q_{\max})$ for all non-innermost regions. Under the empirically observed near-Gaussian structure of pretrained LLM weights, we derive the TSQ Governing Optimality Equation, $\Phi_N(k, r) = 0$, which analytically characterizes the optimal scaling factor using the normalized dynamic range $r = M/\sigma$, eliminating calibration data, activation statistics, Hessian computation, and iterative optimization. Building on the regional distortion analysis, we derive a generalized TSQ distortion law that decomposes normalized distortion into the weight dynamic range, optimized geometric distortion, and quantization resolution. TSQ also employs a scale-efficient representation requiring only one per-row maximum and a single global scaling parameter, from which all thresholds and quantization step sizes are analytically derived, reducing scale storage from $O(mn/g)$ to $O(m)$. Experiments across OPT, Llama, and Qwen model families show that TSQ achieves competitive or lower perplexity than AWQ across the evaluated settings while requiring no calibration data, substantially less scale storage, and lower full-model quantization time. These results demonstrate that analytical geometric scaling can provide a data-free and scalable approach to LLM post-training quantization.

1 INTRODUCTION

The remarkable success of Large Language Models (LLMs) has been accompanied by dramatic growth in model size, making memory capacity and bandwidth the primary bottlenecks for deployment. Post-training quantization (PTQ) compresses pretrained models without retraining, making it a fundamental component of efficient LLM deployment across cloud, edge, and resource-constrained platforms.

Despite their success, existing PTQ methods remain largely empirical: they determine quantization parameters through calibration data, activation statistics, or Hessian-guided optimization. These requirements increase quantization time, introduce per-group scale storage, and complicate deployment in privacy-sensitive or offline settings. This leaves open a basic question: can near-optimal quantization parameters be determined directly from pretrained weight statistics alone, without calibration data?

Pretrained LLM weights offer a clue: across architectures, they are concentrated near zero, nearly symmetric, and heavy-tailed relative to a Gaussian fit – a regime where a single global scale poorly

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.

summarizes the distribution. This suggests quantization parameters can be derived directly from the weights themselves.

We propose Theoretically Scaled Quantization (TSQ), a data-free analytical PTQ framework. TSQ partitions weight magnitudes into N geometric regions and selects the partition factor k by minimizing a distortion objective $R_N(k)$, with no distributional assumption; under a Gaussian approximation this reduces to a scalar root-find, $\Phi_N(k, r) = 0$. TSQ needs no calibration or per-group scale search, providing a scalable, scale-storage-efficient foundation for LLM quantization.

The rest of this paper derives $R_N(k)$ and $\Phi_N(k, r) = 0$, analyzes scale storage, and validates TSQ across multiple LLM families.

2 Related Work

Post-training quantization for LLMs. GPTQ (Frantar et al., 2022) uses Hessian-based reconstruction with calibration data, while AWQ (Lin et al., 2024) searches a grid of 20 per-channel scaling factors $\alpha \in [0, 1]$ plus a clipping search, also using calibration activations. Both depend on external optimization and calibration data; TSQ instead derives its parameters purely from pretrained weight statistics.

Classical quantization theory. Bennett's high-resolution theory (Bennett, 1948) models region-wise distortion as uniform noise weighted by probability mass, the basis for our distortion objective. The Lloyd–Max quantizer (Max, 1960) derives optimal thresholds for a known density, but for $N \geq 3$ requires an iterative, multi-dimensional search with no closed-form solution; Max also observed that the quantizer's output entropy falls below its nominal $\log_2 N$ rate, an early precedent for our entropy-coding gains. TSQ instead reduces the partition problem to a single scalar k, obtained either by a 1-D search over region probabilities (Prop. 3.1) or, under a Gaussian approximation, by a closed-form root-find $\Phi_N(k, r) = 0$.

Entropy coding for LLM weights. ANS (Duda, 2014) underlies production compressors such as Zstandard. Concurrent work (Tan et al., 2026) shows that AWQ4 retains 1.1–1.3$\times$ entropy redundancy and proposes a tile-level ANS decompression kernel for GEMM, motivating our choice to entropy-code the RTN and AWQ baselines alongside TSQ for bit-matched comparisons.

3 TSQ

3.1 General N-Region Partition

For each row of W, let $M = \max|W|$ and $q_{\max} = 2^{b-1} - 1$. Let $k > 1$ denote the geometric threshold factor. The $N - 1$ thresholds are

$$T_i = M/k^i, \qquad i = 1, \ldots, N-1, \tag{1}$$

which partition the weight magnitudes into $R_1 = \{|w| > T_1\}$, $R_i = \{T_i < |w| \leq T_{i-1}\}$ for $i = 2, \ldots, N-1$, and $R_N = \{|w| \leq T_{N-1}\}$, with $P_i = \int_{R_i} p(w)\, dw$ and $\sum_{i=1}^{N} P_i = 1$.

3.2 Region-Wise Quantization

The representative magnitude of region i is $T_i = \dfrac{M}{k^i}$, $i = 1, 2, 3, \ldots, N-1$. The corresponding step size is

$$s_i = \frac{T_i}{q_{\max}} = \frac{M}{k^{i-1} q_{\max}}, \qquad i = 1, 2, 3, \ldots, N, \tag{2}$$

with $s_1 > s_2 > \cdots > s_N$. For $w \in R_i$, the piecewise quantizer is $\hat{w} = s_i \,\mathrm{round}(w/s_i)$.

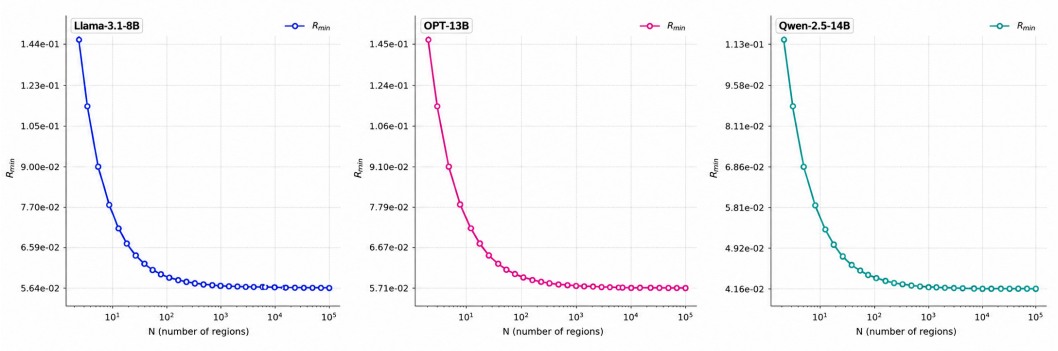

Figure 1: Normalized minimum distortion $R_{\min}$ versus the number of regions N for Llama-3.1-8B, OPT-13B, and Qwen-2.5-14B. Increasing the number of regions reduces distortion with diminishing improvement, approaching a finite saturation value, consistent with Proposition 3.3.

3.3 TSQ Distortion Objective

Under the high-resolution quantization model (Bennett, 1948), region i contributes $E_i = s_i^2 P_i/12$. Thus,

$$E_{\text{TSQ}} = \frac{M^2}{12q_{\max}^2} \sum_{i=1}^{N} \frac{P_i(k)}{k^{2(i-1)}}. \tag{3}$$

Since $E_{\text{uniform}} = M^2/(12q_{\max}^2)$, the normalized distortion is

$$R_N(k) = \frac{E_{\text{TSQ}}}{E_{\text{uniform}}} = \sum_{i=1}^{N} \frac{P_i(k)}{k^{2(i-1)}}. \tag{4}$$

Proposition 3.1. *The optimal threshold factor is $k_N^* = \arg\min_{k>1} R_N(k)$. The resulting optimal factor k_N^* is obtained without assuming any particular underlying weight distribution and is determined directly from the region probabilities induced by the geometric partition.*

Proposition 3.2. *For $k > 1$ and $N \geq 2$, $R_N(k) < 1$, implying that TSQ has lower normalized distortion than uniform quantization.*

Proposition 3.3 (Asymptotic Saturation)**.** *For fixed $k > 1$, $\{R_N(k)\}_{N\geq 2}$ is monotonically decreasing and bounded below, and hence*

$$\lim_{N\to\infty} R_N(k) = R_\infty, \qquad R_\infty \geq 0.$$

Moreover,

$$R_N(k) - R_{N+1}(k) = P_{N+1}\frac{k^2-1}{k^{2N}} \leq \frac{k^2-1}{k^{2N}} \to 0.$$

Thus, the distortion improvement from additional regions vanishes asymptotically, as confirmed empirically in Figure 1.

Proposition 3.4 (Asymptotic Behavior of the Geometric Factor)**.** *Let the geometric factor depend on the number of regions, $k = k(N) > 1$. As the number of regions increases, the first threshold approaches the maximum weight magnitude:*

$$\lim_{N\to\infty} T_1(N) = M.$$

Since $T_1(N) = M/k(N)$ and $M > 0$, this implies

$$\lim_{N\to\infty} k(N) = 1^+$$

Thus, the geometric factor approaches 1 from above as the number of regions increases, while $k(N) > 1$ for every finite N, as confirmed empirically in Figure 2.

All proofs are provided in Appendices A.1–A.8.

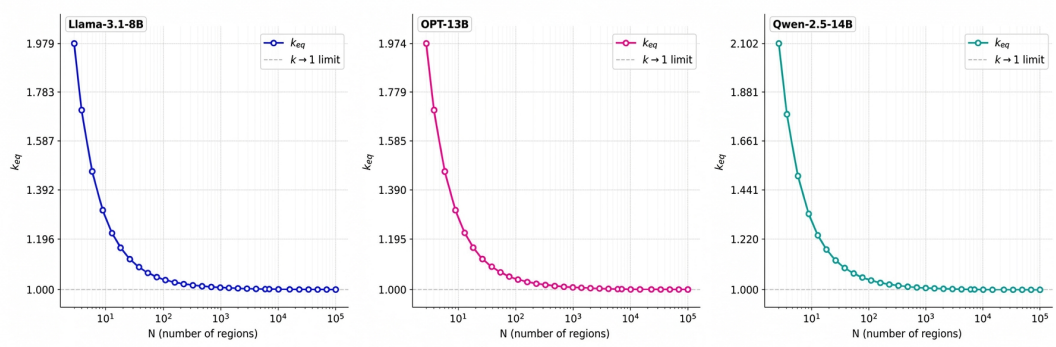

Figure 2: Evolution of the geometric factor $k(N)$ with the number of regions N for Llama-3.1-8B, OPT-13B, and Qwen-2.5-14B. Across all models, $k(N)$ decreases monotonically as N increases and approaches the limiting value $k = 1$ from above, while remaining strictly greater than 1 for every finite N. This empirical behavior is consistent with the asymptotic result $\lim_{N\to\infty} k(N) = 1$.

3.4 Region-Wise Absolute and Relative Error Bounds

The region-wise TSQ step size is $s_i = \frac{M}{k^{i-1}q_{\max}},\ i = 1, \ldots, N,\ k > 1$.

Proposition 3.5. *For $w \in R_i$, $i = 1, \ldots, N$,*

$$|w - \hat{w}| \le \frac{s_i}{2} = \frac{M}{2k^{i-1}q_{\max}}.$$

Hence, the absolute-error bound decreases geometrically by $1/k$.

Proposition 3.6. *For $w \in R_i$, $i = 1, \ldots, N-1$,*

$$\epsilon_{\text{rel}}(w) = \frac{|w - \hat{w}|}{|w|} < \frac{k}{2q_{\max}}.$$

Thus, TSQ provides a region-independent relative-error bound.

Proposition 3.7 (Innermost-Region Relative-Error Bound)**.** *For $w \in R_N$, $w \neq 0$,*

$$\epsilon_{\text{rel}}(w) \le \frac{T_{N-1}}{2q_{\max}|w|}, \qquad \epsilon_{\text{rel}}^{\text{env}} \ge \frac{1}{2q_{\max}}.$$

The minimum is attained at $|w| = T_{N-1}$

Proposition 3.8. *For uniform quantization with $s_{\text{unif}} = M/q_{\max}$,*

$$\frac{s_{\text{unif}}}{s_i^{\text{TSQ}}} = \frac{\epsilon_{\text{unif},i}^{\text{env}}}{\epsilon_{\text{TSQ},i}^{\text{env}}} = k^{i-1}.$$

3.5 Generalized TSQ Distortion Law

From the regional distortion analysis, the total distortion of an N-region TSQ quantizer is

$$E_{\text{TSQ}}^{(N)} = \frac{M^2}{12q_{\max}^2} R_N(k), \qquad R_N(k) = \sum_{i=1}^{N} \frac{P_i}{k^{2(i-1)}}.$$

Let $r = M/\sigma$ denote the normalized weight dynamic range, and let $R_N(k)$ be optimized over the scale factor k:

$$R_N^* = R_N(k^*), \qquad k^* = \arg\min_{k>1} R_N(k).$$

Proposition 3.9 (Generalized TSQ Distortion Law)**.** *Let $D_N = E_{\text{TSQ}}^{(N)}/\sigma^2$ denote the normalized distortion of an N-region TSQ quantizer operating at bit-width b on weights with dynamic range $r = M/\sigma$. Then D_N admits the closed-form decomposition*

$$D_N = \frac{r^2 R_N^*}{12\,(2^{b-1}-1)^2}. \tag{5}$$

Table 1: Theoretical versus measured MSE for TSQ quantization across model scales. The close agreement validates the generalized distortion law of Eq. equation 5.

Model	Theoretical MSE	Measured MSE	Error (%)
Llama-3.2-1B	1.64165×10^{-6}	1.7158×10^{-6}	4.52
Llama-3.2-3B	1.67744×10^{-6}	1.7069×10^{-6}	1.76
Llama-3.1-8B	8.62088×10^{-7}	8.6270×10^{-7}	0.071

Equation equation 5 separates the total distortion into three independent, interpretable factors: the normalized weight dynamic range r, governed by the source statistics; the quantization resolution term $(2^{b-1}-1)^2$, governed solely by the bit-width b; and the optimized geometric partition descriptor R_N^*, which fully captures the effect of the N-region threshold structure on distortion.

We validate Eq. equation 5 by comparing its predicted mean squared error (MSE) against the MSE measured after applying TSQ quantization and reconstructing the weights, as summarized in Table 1.

All proofs are provided in Appendices A.9

3.6 Gaussian Analytical Characterization

The distribution-independent objective $R_N(k)$ yields the optimal geometric factor for arbitrary region probabilities. For an analytical characterization, we specialize the region probabilities to a Gaussian weight distribution,

$$W \sim \mathcal{N}(0, \sigma^2), \qquad r = \frac{M}{\sigma}. \tag{6}$$

This assumption is empirically justified in Appendix A.10, where we verify near-zero mean and skewness across six models spanning two architecture families; it is used only to derive a closed-form estimate of k^* and does not affect the distribution-independent quantization mechanism itself.

The resulting Gaussian probabilities and their derivatives, substituted into $R'_N(k) = 0$, give the N-region optimality equation $\Phi_N(k, r) = 0$ (Proposition B.1, Appendix B):

$$\begin{aligned}\Phi_N(k,r) = \frac{2r}{k^2}\varphi\Big(\frac{r}{k}\Big) + \sum_{i=2}^{N-1}\Bigg[&- \frac{2(i-1)r}{k^{3i-2}}\varphi\Big(\frac{r}{k^{i-1}}\Big) + \frac{2ir}{k^{3i-1}}\varphi\Big(\frac{r}{k^i}\Big) \\ &+ \frac{2(i-1)}{k^{2i-1}}\left(\operatorname{erf}\left(\frac{r}{\sqrt{2}k^i}\right) - \operatorname{erf}\left(\frac{r}{\sqrt{2}k^{i-1}}\right)\right)\Bigg] \\ &- \frac{2(N-1)r}{k^{3N-2}}\varphi\Big(\frac{r}{k^{N-1}}\Big) - \frac{2(N-1)}{k^{2N-1}}\operatorname{erf}\left(\frac{r}{\sqrt{2}k^{N-1}}\right) = 0.\end{aligned} \tag{7}$$

Solving $\Phi_N(k, r) = 0$ gives a closed-form estimate of the Gaussian-optimal threshold k, replacing the numerical search over $R_N(k)$. Table 2 shows this estimate stays within 0.01–0.03 of the exact k^* while cutting runtime from 111–7270 s to 0.33–0.38 s, corresponding to a speedup exceeding $19{,}000\times$ at $N = 100$. The computational cost of the exact $R_N(k)$ search grows with both N and model size (Table 3), while $\Phi_N(k, r)$ only requires computing $r = M/\sigma$ once per row and a single closed-form evaluation, independent of this growth. Full derivation of $\Phi_N(k, r)$, including region-probability formulas and their derivatives, is given in Appendix B; monotonicity properties of the individual region probabilities are established in Proposition B.2 and validated empirically in Figures 4–5.

3.7 Region Probabilities and Monotonicity ($N = 3$)

For $N = 3$, the normalized distortion objective is

$$R_3(k) = \sum_{i=1}^{3} \frac{P_i(k)}{k^{2(i-1)}} = P_1(k) + \frac{P_2(k)}{k^2} + \frac{P_3(k)}{k^4}.$$

Table 2: Runtime and optimal scale factor k^*: exact $R_N(k)$ versus the closed-form Gaussian approximation $\Phi_N(k, r)$ on Llama-3.2-3B.

Method	N=2	N=3	N=4	N=5	N=10	N=100
$R_N(k)$						
Time (s)	111	182	255	327	694	7270
k^*	2.50	2.00	1.69	1.54	1.26	1.03
$\Phi_N(k, r)$						
Time (s)	0.33	0.33	0.34	0.34	0.34	0.38
k^*	2.52	1.99	1.69	1.54	1.26	1.03
Speedup (×)	336	552	750	962	2041	19132

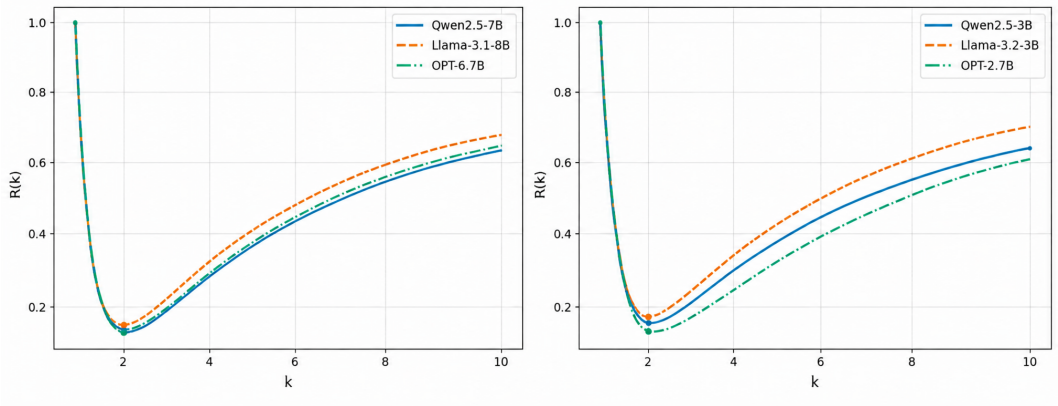

Figure 3: Empirical evaluation of $R_3(k)$ as a function of k at fixed $N = 3$ across three architectures. The left panel presents Qwen2.5-7B, Llama-3.1-8B, and OPT-6.7B, while the right panel presents Qwen2.5-3B, Llama-3.2-3B, and OPT-2.7B. Each model exhibits a unique minimum, yielding a well-defined optimal scale factor $k^* = \arg\min_{k>1} R_3(k)$. Across all three architectures, the resulting k^* remains close to 2, demonstrating the stability of the optimum across models.

We evaluate $R_3(k)$ empirically on the weight distributions of Qwen2.5-7B, Llama-3.1-8B, OPT-6.7B, Qwen2.5-3B, Llama-3.2-3B, and OPT-2.7B, plotted in Figure 3. Each curve exhibits a unique minimum, confirming the existence of an optimal scale factor $k^* = \arg\min_{k>1} R_3(k)$, with $k^* \approx 2.0$ across all models.

4 Results and Evaluation

4.1 Experimental Setup

We evaluate TSQ against AWQ on language modeling and zero-shot reasoning tasks. For language modeling, we measure perplexity (PPL) on the WikiText-2 test set (Merity et al., 2017) using a sequence length of 2048 and a stride of 512. Zero-shot performance is evaluated with `lm-evaluation-harness` (Gao et al., 2023) on PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2021), MMLU (Hendrycks et al., 2021), and GSM8K (Cobbe et al., 2021), reporting the average accuracy across the seven tasks. For the AWQ baseline, we use the official MIT Han Lab implementation (Lin et al., 2024) with 128 calibration samples from WikiText-2, sequence length 512, 4-bit quantization, and group size 128. TSQ is calibration-free and is evaluated using the proposed three-region configuration ($N = 3$) at both 3-bit and 4-bit precision. TSQ employs symmetric W4 quantization with levels $[-7, 7]$, yielding 15 representable signed levels and no zero-point. In contrast, AWQ uses asymmetric W4 quantization with 16 levels $[0, 15]$ and a zero-point, storing the scale

Table 3: Runtime (s) for computing the normalized dynamic range $r = M/\sigma$ and Gaussian optimality equation $\Phi_N(k, r)$ across model sizes and numbers of regions.

Model	N=2	N=3	N=4	N=5	N=10	N=100
Llama-3.2-3B	0.33	0.33	0.34	0.35	0.36	0.38
Llama-3.1-8B	0.73	0.73	0.74	0.76	0.78	0.85

Table 4: Full-model quantization runtime (min), AWQ vs. TSQ.

Method	L-3.2-1B	L-3.2-3B	L-3.1-8B	Opt-1.3B	Opt-2.7B	Opt-6.7B
AWQ	10	25	80	11	22	59
TSQ	7	13	31	8	12	25
Saving%	**30**	**48**	**62**	**27**	**46**	**58**

in FP16 and the zero-point in 4 bits. Thus, for group size G, its effective storage is $b + (16 + b)/G$ bits per weight.

4.2 Entropy Coding with rANS

Entropy Coding. We losslessly compress the quantized weight stream using asymmetric numeral systems (ANS) (Duda, 2014), a well-established entropy coder used in production systems (e.g., zstd, LZFSE). For TSQ, each weight is encoded as a single symbol $\rho = 16i + (q + Q_{\max})$, jointly representing the region index $i \in \{1, 2, \ldots, N\}$ and the quantization level $q \in [-Q_{\max}, Q_{\max}]$. For RTN and AWQ, ρ is simply the integer quantization code. In all cases, a per-group (or per-row, for TSQ) frequency table $\ell(\rho)$ is built from the empirical symbol distribution, with $L = \sum_\rho \ell(\rho)$ the total frequency and $c(\rho) = \sum_{\rho' < \rho} \ell(\rho')$ the cumulative frequency preceding ρ. Encoding and decoding follow the standard rANS recurrences,

$$\mathrm{Enc}(\rho, x) = L \left\lfloor \frac{x}{\ell(\rho)} \right\rfloor + c(\rho) + (x \bmod \ell(\rho)), \qquad \mathrm{Dec}(x) = \Big(\rho,\ \ell(\rho) \left\lfloor \frac{x}{L} \right\rfloor + (x \bmod L) - c(\rho) \Big),$$

where ρ in $\mathrm{Dec}(x)$ is identified from $x \bmod L \in [c(\rho), c(\rho) + \ell(\rho))$, and the state x is kept within a fixed interval $[L, L_{\max})$ by streaming excess bits to and from the output. Coding is exact under integer arithmetic, so every symbol is recovered losslessly; scales (and, for TSQ, per-row M) are stored separately for dequantization. We apply this identical entropy-coding procedure to TSQ, RTN, and AWQ alike, so that all reported bit-widths reflect the same lossless compression scheme applied to each method's native symbol stream.

4.3 Quantization Runtime

We measure end-to-end wall-clock time (loading, quantization, WikiText-2 perplexity on pseudo-quantized FP16 weights) for AWQ and TSQ under identical settings (AWQ calibration: 128 samples, seqlen 512) on two Tesla T4 GPUs. Both use pseudo-quantization; AWQ's INT4 kernel accelerates only inference, not the runtimes below. TSQ is consistently faster (Table 4): AWQ performs a grid search of size 20 over $\alpha \in [0, 1]$ to find the optimal activation-aware per-channel scale (Lin et al., 2024), plus a separate weight-clipping search to minimize quantization MSE, both requiring repeated forward passes over calibration data, whereas TSQ is data-free and computes its region thresholds analytically in a single pass.

4.4 Limitations and Future Work

TSQ's reported bit-widths reflect storage cost, not measured inference latency, because we do not provide a fused decode kernel. Reconstructing a TSQ weight requires two coupled steps: decoding the rANS-coded joint symbol $\rho = 16i + (q + Q_{\max})$ to recover the region index i and code q, then dequantizing with the scale s_i selected by i from the per-row value M. Existing low-bit GEMM kernels – AWQ, GPTQ, Marlin, and tile-level ANS decoders such as Tan et al. (2026) – assume a

Table 5: WikiText-2 PPL comparison for weight-only quantization. RTN, AWQ, and TSQ are entropy-coded with rANS. At W4 (g=32), the matched effective bit-width is 4.5–4.8 bits/weight; at W3, 3.42 bits/weight; at W2, 2.40 bits/weight (g=32), varying slightly across model architectures, at which AWQ-32 and TSQ are competitive in perplexity.AWQ-128 and AWQ-64 use fewer bits but incur substantially higher perplexity, and are included for reference only.

BPW (g=32)	Model	FP16	RTN-32	AWQ-128	AWQ-64	AWQ-32	TSQ
4.8	Llama-3.2-1B	8.60	9.40	9.61	9.21	9.13	**9.06**
	Llama-3.2-3B	6.91	7.25	7.31	7.20	7.11	**7.07**
	Llama-3.1-8B	5.53	5.82	5.89	5.81	5.76	**5.72**
4.7	Qwen2.5-3B	7.09	7.49	7.47	7.38	7.33	**7.29**
4.5	OPT-1.3B	12.64	12.94	13.06	12.98	12.88	**12.71**
	OPT-2.7B	10.66	11.09	10.88	10.85	10.81	**10.77**
	OPT-6.7B	9.30	9.52	9.40	9.38	9.35	**9.33**
W3A16 — Llama Models							
BPW(g=32)	Model	FP16	RTN-32	AWQ-128	AWQ-64	AWQ-32	TSQ
3.42	Llama-3.2-1B	8.60	15.02	14.88	12.76	11.77	**11.45**
	Llama-3.2-3B	6.91	9.35	9.06	8.46	8.19	**8.04**
	Llama-3.1-8B	5.53	7.25	7.22	6.83	6.49	**6.37**
W2A16 — Llama Models							
BPW(g=32)	Model	FP16	RTN-32	AWQ-128	AWQ-64	AWQ-32	TSQ
2.40	Llama-3.2-1B	8.60	2.6039e4	-	-	1.780e3	**172.03**
	Llama-3.2-3B	6.91	2.520e3	-	-	1.2200e4	**36.09**
	Llama-3.1-8B	5.53	6.877e3	-	-	1.5700e6	**19.42**

single fixed scale per element and are not designed for this region-conditional lookup after decode. Building such a kernel is not fundamentally hard: it requires extending a tile-aligned GPU rANS decoder with a small, per-region scale table indexed by the decoded region bits, kept resident in shared memory alongside the existing codebook. We view this as a natural extension of prior tile-level ANS decoding work rather than a new decoding primitive, and we leave its implementation and end-to-end throughput evaluation to future work.

5 Conclusion

We presented TSQ, a data-free post-training quantization framework that constructs geometric region partitions and their quantization scales analytically from pretrained weight statistics, without requiring calibration data, activation statistics, or second-order information. The proposed formulation reduces scale selection to a one-dimensional geometric optimization, yielding explicit region-wise error bounds and an analytical characterization of the optimal scaling factor. Across Llama, Qwen, and OPT models, TSQ achieves competitive or lower perplexity than AWQ and RTN at matched effective bit-widths, with the advantage becoming more pronounced at lower precision. Zero-shot evaluations further show that the resulting weight reconstruction maintains competitive downstream performance across multiple language and reasoning tasks.

Beyond accuracy, TSQ provides a substantially simpler representation: its quantization scales are determined from per-row maxima and a geometric scaling factor, giving $O(m)$ scale storage compared with the $O(mn/g)$ group-wise storage of AWQ and RTN. The analytical parameter selection also substantially reduces quantization overhead relative to search-based approaches. Overall, TSQ provides a theoretically grounded, data-free, and scale-efficient framework for practical low-bit post-training quantization of large language models.

LLM Usage. We used AI assistants (Claude and ChatGPT) to polish writing, verify appendix derivations, and cross-check citations against source papers.

Table 6: Zero-shot task accuracy for W4A16 quantization (rANS entropy-coded). AWQ-32 matches TSQ's bit-width; AWQ-128/64 are shown for reference.

Model	Task	FP16	AWQ-128	AWQ-64	AWQ-32	TSQ
Llama-3.2-1B	PIQA	75.14	74.16	**75.52**	74.48	75.03
	HellaSwag	48.18	46.87	46.57	46.92	**47.34**
	ARC-Easy	66.29	65.87	64.94	65.28	**66.67**
	ARC-Challenge	31.40	**31.31**	30.89	30.97	**31.31**
	WinoGrande	60.85	60.93	61.40	60.69	**61.96**
	MMLU	40.00	34.29	35.83	**37.36**	35.74
	GSM8K	2.96	2.35	3.34	2.27	**3.64**
	Average	46.40	45.11	45.50	45.42	**45.96**
Llama-3.2-3B	PIQA	76.39	76.01	76.33	**76.77**	75.95
	HellaSwag	55.83	54.93	55.07	**55.50**	55.49
	ARC-Easy	74.45	72.60	73.86	74.33	**75.59**
	ARC-Challenge	42.75	41.89	41.64	41.81	**43.26**
	WinoGrande	71.66	69.61	69.46	**70.17**	69.69
	MMLU	56.42	55.24	55.19	**56.53**	55.90
	GSM8K	10.24	7.35	9.25	9.17	**10.54**
	Average	55.39	53.95	54.40	54.90	**55.20**
Llama-3.1-8B	PIQA	79.16	79.11	79.05	78.89	**79.27**
	HellaSwag	60.70	59.67	59.83	60.27	**60.34**
	ARC-Easy	82.20	80.93	81.23	81.10	**81.40**
	ARC-Challenge	52.05	51.79	51.11	51.28	51.02
	WinoGrande	74.59	72.45	73.72	73.95	**74.03**
	MMLU	69.38	66.88	63.74	63.48	63.75
	GSM8K	25.09	21.08	23.58	24.17	**24.56**
	Average	63.31	61.70	61.75	61.88	**62.05**

Table 7: Zero-shot task accuracy for W3A16 quantization (rANS entropy-coded). AWQ-32 matches TSQ's bit-width; AWQ-128/64 are shown for reference.

Model	Task	FP16	AWQ-128	AWQ-64	AWQ-32	TSQ
Llama-3.2-1B	PIQA	75.14	69.64	70.18	71.60	**72.52**
	HellaSwag	48.18	41.30	42.45	43.71	**44.37**
	ARC-Easy	66.29	57.53	58.88	61.36	**63.93**
	ARC-Challenge	31.40	28.07	26.54	29.01	**30.29**
	WinoGrande	60.85	58.72	58.25	**59.98**	59.59
	MMLU	40.00	27.55	31.45	28.65	**32.53**
	GSM8K	2.96	2.27	2.12	2.27	**3.03**
	Average	46.40	40.73	41.41	42.37	**43.75**
Llama-3.2-3B	PIQA	76.39	74.16	**75.90**	75.73	75.14
	HellaSwag	55.83	50.49	51.86	52.39	**53.28**
	ARC-Easy	74.45	67.00	71.34	70.24	**71.55**
	ARC-Challenge	42.75	35.24	40.13	39.33	**40.53**
	WinoGrande	69.46	66.38	67.40	**69.38**	67.96
	MMLU	56.42	48.54	49.67	**51.16**	50.80
	GSM8K	10.24	5.23	4.62	5.91	**6.29**
	Average	55.08	49.58	51.56	52.02	**52.22**
Llama-3.1-8B	PIQA	79.16	77.69	77.91	77.97	**78.51**
	HellaSwag	60.70	55.42	57.03	57.66	**58.45**
	ARC-Easy	82.20	76.22	76.09	78.66	**80.13**
	ARC-Challenge	52.05	44.11	46.08	47.35	**49.74**
	WinoGrande	74.59	71.43	70.80	**72.53**	72.45
	MMLU	69.38	56.07	58.22	**60.55**	59.92
	GSM8K	25.09	9.10	11.68	14.71	**14.96**
	Average	63.31	55.72	56.83	58.49	**59.17**

REFERENCES

W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. In *Advances in Neural Information Processing Systems*, 2021.

Jarek Duda. Asymmetric numeral systems: Entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2014.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, 2024.

Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, 1960.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Hongshi Tan, Yao Chen, Gustavo Alonso, Weng-Fai Wong, and Bingsheng He. Approaching Shannon bound with lossless LLM weight compression. *arXiv preprint arXiv:2606.15789*, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

A APPENDIX

A.1 HIGH-RESOLUTION QUANTIZATION ERROR

Lemma A.1 (Bennett, 1948)**.** *Under the high-resolution quantization assumption (Bennett, 1948), the quantization error is modeled as*

$$e \sim \mathcal{U}\Big(-\frac{s}{2}, \frac{s}{2}\Big), \qquad \mathbb{E}[e^2] = \frac{s^2}{12}.$$

For a quantization region $\mathcal{R}_r$ with probability mass $P_r = \int_{\mathcal{R}_r} p(w)\,dw$ and step size s_r, the regional distortion is

$$E_r \approx \frac{s_r^2}{12} P_r.$$

Proof. Since the quantization error within $\mathcal{R}_r$ is uniformly distributed, $\mathbb{E}[e^2 \mid \mathcal{R}_r] = s_r^2/12$. Weighting this conditional error by the probability mass of the region gives

$$E_r = P_r\,\mathbb{E}[e^2 \mid \mathcal{R}_r] \approx \frac{s_r^2}{12} P_r.$$

□

A.2 General N-Region TSQ Partition

Let

$$M = \max|W|, \qquad q_{\max} = 2^{b-1} - 1$$

and choose a geometric threshold factor $k > 1$. The $N-1$ thresholds are defined as

$$T_i = \frac{M}{k^i}, \qquad i = 1, 2, \ldots, N-1.$$

These thresholds partition the weight space into N mutually exclusive regions:

$$\mathcal{R}_1 = \{w : |w| > T_1\}, \qquad \mathcal{R}_i = \{w : T_i < |w| \le T_{i-1}\},\ i = 2, \ldots, N-1, \qquad \mathcal{R}_N = \{w : |w| \le T_{N-1}\},$$

satisfying

$$\sum_{i=1}^{N} P_i = 1, \qquad P_i = \int_{\mathcal{R}_i} p(w)\,dw.$$

Region-wise step sizes. The representative magnitude of Region i is

$$T_i = \frac{M}{k^{i-1}}, \qquad i = 1, 2, \ldots, N.$$

The corresponding quantization step size is

$$s_i = \frac{T_i}{q_{\max}} = \frac{M}{k^{i-1} q_{\max}}, \qquad i = 1, 2, \ldots, N.$$

In particular,

$$s_1 = \frac{M}{q_{\max}}, \qquad s_N = \frac{M}{k^{N-1} q_{\max}}.$$

Since $k > 1$, it follows that $s_i > s_{i+1}$ for $i = 1, \ldots, N-1$, that is,

$$s_1 > s_2 > \cdots > s_N.$$

Therefore, larger-magnitude regions are quantized using progressively coarser step sizes, while smaller-magnitude regions receive progressively finer quantization.

Regional distortion. Using the high-resolution quantization result from Lemma A.1 (Appendix A.1), the distortion contributed by Region i is

$$E_i \approx \frac{s_i^2 P_i}{12}, \qquad i = 1, \ldots, N.$$

The total quantization distortion is therefore

$$E_{\mathrm{TSQ}} = \sum_{i=1}^{N} E_i.$$

Piecewise quantizer. For a weight $w \in \mathcal{R}_i$, the quantized value is

$$\hat{w} = s_i \,\mathrm{round}\left(\frac{w}{s_i}\right), \qquad i = 1, \ldots, N.$$

Remark A.2. *The proposed N-region TSQ requires only two parameters: the per-row maximum magnitude M and the global threshold factor k. All thresholds, region-wise step sizes, and quantization parameters are determined analytically from these quantities, requiring no calibration data or activation statistics.*

A.3 Geometric Interpretation of the Threshold Factor

The parameter $k > 1$ is the geometric partition factor of the proposed TSQ framework. It determines the locations of all threshold boundaries and therefore controls the trade-off between allocating finer quantization resolution to high-magnitude and low-magnitude weights.

The thresholds satisfy

$$T_i = \frac{M}{k^i}, \qquad i = 1, \ldots, N-1,$$

which form a geometric progression with common ratio

$$\frac{T_{i+1}}{T_i} = \frac{1}{k}.$$

Consequently, k is the single parameter governing the geometric spacing of the thresholds, and its optimal value is obtained by minimizing the normalized distortion objective

$$k_N^* = \arg\min_{k>1} R_N(k).$$

Larger values of k produce a more rapidly decaying sequence of thresholds, resulting in a sparser partition of the weight range. Conversely, as k approaches unity, the threshold spacing becomes progressively denser, yielding a finer geometric partition of the weight interval.

As established in Proposition 3.4 (Appendix A.7),

$$\lim_{N\to\infty} k_N^* = 1,$$

implying that the geometric partition becomes arbitrarily dense as the number of regions increases.

A.4 General N-Region TSQ Distortion

Using the high-resolution quantization result from Lemma A.1 (Appendix A.1), the distortion contributed by Region i is

$$E_i = \frac{s_i^2}{12} P_i, \qquad i = 1, \ldots, N.$$

From Appendix A.2,

$$s_i = \frac{M}{k^{i-1} q_{\max}}.$$

Therefore,

$$E_i = \frac{1}{12}\left(\frac{M}{k^{i-1} q_{\max}}\right)^2 P_i = \frac{M^2}{12 q_{\max}^2} \frac{P_i}{k^{2(i-1)}}.$$

Summing over all N regions gives the total TSQ distortion,

$$E_{\mathrm{TSQ}} = \sum_{i=1}^{N} E_i = \frac{M^2}{12 q_{\max}^2} \sum_{i=1}^{N} \frac{P_i}{k^{2(i-1)}}.$$

Equivalently,

$$E_{\mathrm{TSQ}} = \frac{M^2}{12 q_{\max}^2}\left(P_1 + \sum_{i=2}^{N-1} \frac{P_i}{k^{2(i-1)}} + \frac{P_N}{k^{2(N-1)}}\right).$$

The corresponding uniform quantization distortion is

$$E_{\text{uniform}} = \frac{M^2}{12q_{\max}^2}.$$

Hence, the normalized relative distortion is

$$R_N(k) = \frac{E_{\text{TSQ}}}{E_{\text{uniform}}} = \sum_{i=1}^{N} \frac{P_i(k)}{k^{2(i-1)}}.$$

Equivalently,

$$R_N(k) = P_1(k) + \sum_{i=2}^{N-1} \frac{P_i(k)}{k^{2(i-1)}} + \frac{P_N(k)}{k^{2(N-1)}}.$$

The optimal threshold factor is obtained by minimizing the normalized distortion,

$$k_N^* = \arg\min_{k>1} R_N(k).$$

A.5 Distribution-Independent Upper Bound on TSQ Distortion

The following theorem establishes that the normalized distortion of the proposed N-region TSQ is always strictly smaller than that of uniform quantization. Importantly, this result is completely distribution-independent and holds for any weight distribution.

Theorem A.3 (Distribution-Independent Dominance)**.** *Let*

$$R_N(k) = P_1 + \sum_{i=2}^{N} \frac{P_i}{k^{2(i-1)}},$$

where $P_i \geq 0$, $\sum_{i=1}^{N} P_i = 1$, *and* $k > 1$. *Then, for every* $N \geq 2$,

$$R_N(k) < 1.$$

Consequently, $E_{\text{TSQ}} < E_{\text{uniform}}$, *and this result is independent of the underlying weight distribution.*

Proof. Since $k > 1$, we have $k^{2(i-1)} > 1$ for $i = 2, \ldots, N$, which implies

$$\frac{1}{k^{2(i-1)}} < 1, \qquad i = 2, \ldots, N.$$

Because $P_i \geq 0$, multiplying both sides by P_i gives

$$\frac{P_i}{k^{2(i-1)}} < P_i, \qquad i = 2, \ldots, N.$$

Substituting into the normalized distortion,

$$R_N(k) = P_1 + \sum_{i=2}^{N} \frac{P_i}{k^{2(i-1)}} < P_1 + \sum_{i=2}^{N} P_i = \sum_{i=1}^{N} P_i = 1.$$

Therefore,

$$R_N(k) < 1, \qquad \forall\, k > 1,\ N \geq 2.$$

Since $E_{\text{TSQ}} = E_{\text{uniform}}\, R_N(k)$, it follows that

$$E_{\text{TSQ}} < E_{\text{uniform}}.$$

The proof relies only on the facts that the regional probabilities are non-negative and sum to one. No assumption is made on the form of the weight distribution. Hence, the result holds for arbitrary weight distributions. □

A.6 Monotonic Improvement with Increasing Regions

A fundamental property of the proposed N-region TSQ framework is that introducing one additional region can never increase the normalized distortion.

Theorem A.4 (Monotonic Improvement)**.** *For any fixed threshold factor $k > 1$,*

$$R_{N+1}(k) \leq R_N(k),$$

with equality if and only if $P_{N+1} = 0$.

Proof. The normalized distortion of the N-region quantizer is

$$R_N(k) = P_1 + \sum_{i=2}^{N-1} \frac{P_i}{k^{2(i-1)}} + \frac{P_N}{k^{2(N-1)}}.$$

When one additional region is introduced, the last region is partitioned into $P_N = P'_N + P_{N+1}$, where P'_N denotes the remaining probability mass of the original last region. Therefore,

$$R_{N+1}(k) = P_1 + \sum_{i=2}^{N-1} \frac{P_i}{k^{2(i-1)}} + \frac{P'_N}{k^{2(N-1)}} + \frac{P_{N+1}}{k^{2N}}.$$

Subtracting,

$$R_N(k) - R_{N+1}(k) = \frac{P_N}{k^{2(N-1)}} - \frac{P'_N}{k^{2(N-1)}} - \frac{P_{N+1}}{k^{2N}} = P_{N+1}\left(\frac{1}{k^{2(N-1)}} - \frac{1}{k^{2N}}\right) = \frac{P_{N+1}(k^2-1)}{k^{2N}}.$$

Since $P_{N+1} \geq 0$ and $k > 1$, we have $k^2 - 1 > 0$, hence

$$R_N(k) - R_{N+1}(k) \geq 0, \qquad \text{i.e.,} \qquad R_{N+1}(k) \leq R_N(k).$$

Equality holds if and only if $P_{N+1} = 0$, that is, when the newly introduced region is empty. □

A.7 Asymptotic Behavior of the Geometric Thresholds

The following theorem establishes the limiting behavior of the first threshold as the number of geometric regions increases.

Theorem A.5 (Asymptotic Convergence of the First Threshold)**.** *Let the absolute weight range satisfy $0 \leq |W| \leq M$, $M = \max|W|$. Suppose the interval $[0, M]$ is partitioned into N ordered geometric regions $\mathcal{R}_1, \ldots, \mathcal{R}_N$ with thresholds $M > T_1 > T_2 > \cdots > T_{N-1} > 0$. If the partition becomes arbitrarily fine as $N \to \infty$, then*

$$\lim_{N\to\infty} T_1 = M.$$

Proof. The interval $[0, M]$ has finite length M. The first region is $\mathcal{R}_1 = (T_1, M]$, with width $L_1 = M - T_1$. As the number of regions tends to infinity, the finite interval $[0, M]$ is partitioned into infinitely many progressively finer regions, so every region width approaches zero; in particular, $\lim_{N\to\infty} L_1 = 0$. Since $L_1 = M - T_1$, we obtain $\lim_{N\to\infty}(M - T_1) = 0$, which gives

$$\lim_{N\to\infty} T_1 = M.$$

□

A.8 Asymptotic Saturation of the N-Region TSQ Objective

A natural question is whether the normalized distortion can be reduced indefinitely by continuously increasing the number of quantization regions. The following theorem shows that this is impossible. Although each additional region decreases the distortion, the amount of improvement becomes progressively smaller, causing the objective to converge to a finite saturation value.

Theorem A.6 (Asymptotic Saturation)**.** *For any fixed threshold factor $k > 1$, the normalized distortion sequence $\{R_N(k)\}_{N=2}^{\infty}$ is monotonically decreasing and bounded below. Consequently,*

$$\lim_{N\to\infty} R_N(k) = R_\infty,$$

where R_∞ is a finite constant.

Proof. By Theorem A.4,

$$R_N(k) - R_{N+1}(k) = \frac{P_{N+1}(k^2-1)}{k^{2N}} \geq 0,$$

with equality only if $P_{N+1} = 0$, so $R_2(k) \geq R_3(k) \geq \cdots \geq R_N(k) \geq \cdots$. Since $R_N(k) = P_1 + \sum_{i=2}^{N} P_i/k^{2(i-1)}$ and $P_i \geq 0$, every term is non-negative, so $R_N(k) \geq 0$ for all N. Thus $\{R_N(k)\}$ is monotone decreasing and bounded below by zero; by the Monotone Convergence Theorem,

$$\lim_{N\to\infty} R_N(k) = R_\infty, \qquad R_\infty \geq 0.$$

□

Remark A.7 (Diminishing Improvement)**.** *Since $0 \leq P_{N+1} \leq 1$,*

$$0 \leq R_N(k) - R_{N+1}(k) \leq \frac{k^2-1}{k^{2N}}.$$

Because $k > 1$, the denominator grows exponentially with N, so

$$\lim_{N\to\infty} \big(R_N(k) - R_{N+1}(k)\big) = 0.$$

Hence, although each additional region always reduces the normalized distortion, the magnitude of this reduction vanishes asymptotically, and the distortion approaches the finite saturation value R_∞.

For sufficiently large N, introducing additional geometric regions yields only marginal improvement in normalized distortion while increasing the computational complexity of the quantizer.

A.9 Generalized TSQ Distortion Law

Let $r = M/\sigma$ denote the normalized weight dynamic range, where σ is the standard deviation of the weights, and let $D_N = E_{\mathrm{TSQ}}^{(N)}/\sigma^2$ denote the normalized mean squared distortion of an N-region TSQ quantizer, with $R_N^* = \min_{k>1} \sum_{i=1}^{N} P_i(k)/k^{2(i-1)}$ the optimized geometric distortion factor.

Proposition A.8 (Generalized TSQ Distortion Law)**.**

$$D_N = \frac{r^2 R_N^*}{12\,(2^{b-1}-1)^2}.$$

Proof. From Appendix A.4, $E_{\mathrm{TSQ}}^{(N)} = \dfrac{M^2}{12q_{\max}^2} R_N(k)$. At the optimal threshold factor k_N^*,

$$E_{\mathrm{TSQ}}^{(N)} = \frac{M^2}{12q_{\max}^2} R_N^*.$$

Dividing by σ^2 and using $r = M/\sigma$,

$$D_N = \frac{M^2}{12\sigma^2 q_{\max}^2} R_N^* = \frac{r^2 R_N^*}{12q_{\max}^2}.$$

Substituting $q_{\max} = 2^{b-1} - 1$ for b-bit signed magnitude quantization yields

$$D_N = \frac{r^2 R_N^*}{12\,(2^{b-1}-1)^2}.$$

□

Table 8: Weight distribution statistics for quantizable layers.

Model	Mean	Std	Skewness	Excess Kurtosis
Llama-3.2-1B	-1.50×10^{-5}	0.01968	-0.0031	3.9609
Llama-3.2-3B	-4.71×10^{-6}	0.01963	-0.0001	2.2459
Llama-3.1-8B	-4.95×10^{-6}	0.01376	-0.0010	2.4968
OPT-1.3B	2.97×10^{-6}	0.02093	0.0042	2.2936
OPT-2.7B	-5.17×10^{-6}	0.01895	-0.0039	1.9778
OPT-6.7B	-8.43×10^{-6}	0.01655	0.0007	0.8899

Remark A.9. *This decomposition separates the normalized distortion into three interpretable factors:*

$$\underbrace{r^2}_{\text{weight dynamic range}} \qquad \underbrace{R_N^*}_{\text{optimized geometric partition}} \qquad \underbrace{\frac{1}{12(2^{b-1}-1)^2}}_{\text{quantization resolution}}.$$

Thus, the weight statistics enter through the normalized dynamic range r, the N-region TSQ structure enters through the optimized factor R_N^*, and the bit width enters only through the quantization-resolution term.

A.10 Empirical Weight Distribution Structure

TSQ's region partition (Eq. 1) relies on weight magnitudes being symmetric about zero and heavy-tailed enough that a single global scale is a poor summary. We verify both properties across Llama (3.2-1B, 3.2-3B, 3.1-8B) and OPT (1.3B, 2.7B, 6.7B), restricted to the layers TSQ quantizes (`q_proj`, `k_proj`, `v_proj`, `o_proj`/`out_proj`, `gate_proj`/`fc1`, `up_proj`/`fc1`, `down_proj`/`fc2`)
. Mean is negligible ($|\mu| \leq 1.5 \times 10^{-5}$), confirming the zero-mean assumption $W \sim \mathcal{N}(0, \sigma^2)$ in Section 3.6. Skewness is negligible ($|\gamma_1| \leq 0.0042$), confirming symmetry, while excess kurtosis is strictly positive (0.89–3.96), confirming heavy tails – the regime TSQ's multi-scale hierarchy targets.

Remark A.10 (Gaussian Assumption Selects k; It Does Not Define the Quantizer)**.** *TSQ's mechanism is distribution-independent (Theorem A.3 guarantees $R_N(k) < 1$ for any distribution). The Gaussian assumption only yields a closed-form k^* (Proposition B.1) in place of a per-row numerical search, and does not affect the quantizer once k^* is chosen. Table 2 confirms it stays within 0.01–0.03 of the exact optimum despite the heavy tails above.*

B Gaussian Region Probabilities and the Optimality Equation

B.1 Assumptions and Definitions

Let the weights follow

$$W \sim \mathcal{N}(0, \sigma^2), \qquad p(w) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{w^2}{2\sigma^2}\right).$$

Let

$$M = \max|w|, \qquad r = \frac{M}{\sigma},$$

and let the thresholds be

$$T_i = \frac{M}{k^i}, \qquad i = 1, \ldots, N-1, \qquad k > 1,$$

inducing regions

$$R_1 = \{|w| > T_1\}, \quad R_i = \{T_i < |w| \leq T_{i-1}\}, \ (i = 2, \ldots, N-1), \quad R_N = \{|w| \leq T_{N-1}\}.$$

Define the standard Gaussian tail function

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

B.2 Region 1

By definition,

$$P_1(k) = \Pr(|w| > M/k).$$

By symmetry of $p(w)$,

$$P_1(k) = \Pr(w > M/k) + \Pr(w < -M/k) = 2\Pr(w > M/k).$$

Writing the probability as an integral of the Gaussian density,

$$P_1(k) = 2\int_{M/k}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{w^2}{2\sigma^2}\right) dw.$$

Substitute $x = w/\sigma$, so $w = \sigma x$, $dw = \sigma\, dx$:

$$\frac{w^2}{2\sigma^2} = \frac{x^2}{2}, \qquad \frac{1}{\sigma\sqrt{2\pi}}\, dw = \frac{1}{\sqrt{2\pi}}\, dx.$$

The lower limit transforms as

$$x_{\text{lower}} = \frac{M/k}{\sigma} = \frac{M}{k\sigma} = \frac{r}{k}.$$

Hence

$$P_1(k) = 2\int_{r/k}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 2\,Q\Big(\frac{r}{k}\Big).$$

B.3 Intermediate Regions ($i = 2, \ldots, N-1$)

By definition,

$$R_i = \left\{ \frac{M}{k^i} < |w| \le \frac{M}{k^{i-1}} \right\}.$$

The region probability is the difference of two tail probabilities:

$$P_i(k) = \Pr(|w| > M/k^i) - \Pr(|w| > M/k^{i-1}),$$

since $\Pr(|w| > M/k^i)$ includes R_i together with all outer regions, and subtracting $\Pr(|w| > M/k^{i-1})$ removes exactly the outer part. By the same derivation as Region 1, with threshold M/k^j in place of M/k,

$$\Pr(|w| > M/k^j) = 2\,Q\Big(\frac{r}{k^j}\Big), \qquad j = 1, \ldots, N-1.$$

Applying this with $j = i$ and $j = i-1$,

$$P_i(k) = 2\,Q\Big(\frac{r}{k^i}\Big) - 2\,Q\Big(\frac{r}{k^{i-1}}\Big) = 2\left[Q\Big(\frac{r}{k^i}\Big) - Q\Big(\frac{r}{k^{i-1}}\Big)\right].$$

Since $k > 1 \Rightarrow r/k^i < r/k^{i-1}$, and Q is monotonically decreasing,

$$Q\Big(\frac{r}{k^i}\Big) > Q\Big(\frac{r}{k^{i-1}}\Big) \quad \Longrightarrow \quad P_i(k) > 0.$$

B.4 Innermost Region

By definition,

$$R_N = \left\{ |w| \le \frac{M}{k^{N-1}} \right\}.$$

Since $\sum_{i=1}^{N} P_i = 1$,

$$P_N(k) = 1 - \Pr(|w| > M/k^{N-1}) = 1 - 2\,Q\Big(\frac{r}{k^{N-1}}\Big).$$

Using the standard identity relating the Gaussian tail to the error function, $1 - 2Q(x) = \operatorname{erf}(x/\sqrt{2})$ — which follows from $\Pr(-x < X < x) = 1 - 2Q(x)$ for $X \sim \mathcal{N}(0,1)$ and the definition $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du$ — we substitute $x = r/k^{N-1}$:

$$P_N(k) = \operatorname{erf}\left(\frac{r/k^{N-1}}{\sqrt{2}}\right) = \operatorname{erf}\left(\frac{r}{\sqrt{2}\,k^{N-1}}\right).$$

B.5 Consistency Check

Summing the intermediate-region result over $i = 2, \ldots, N-1$ telescopes:

$$\sum_{i=2}^{N-1} P_i(k) = 2 \sum_{i=2}^{N-1} \left[Q\Big(\frac{r}{k^i}\Big) - Q\Big(\frac{r}{k^{i-1}}\Big) \right] = 2 \left[Q\Big(\frac{r}{k^{N-1}}\Big) - Q\Big(\frac{r}{k}\Big) \right].$$

Adding $P_1(k) = 2Q(r/k)$ cancels the $Q(r/k)$ term:

$$P_1(k) + \sum_{i=2}^{N-1} P_i(k) = 2\, Q\Big(\frac{r}{k^{N-1}}\Big).$$

Adding $P_N(k) = 1 - 2Q(r/k^{N-1})$:

$$\sum_{i=1}^{N} P_i(k) = 2\, Q\Big(\frac{r}{k^{N-1}}\Big) + 1 - 2\, Q\Big(\frac{r}{k^{N-1}}\Big) = 1.$$

This confirms that the Gaussian region probabilities are internally consistent and correctly partition the full probability mass. These expressions for $P_i(k)$ are used in Section B.9 to construct $\Phi_N(k, r)$ via $R'_N(k) = 0$.

B.6 Derivative of $P_1(k)$

Recall

$$P_1(k) = 2\, Q(r/k), \qquad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad Q'(x) = -\varphi(x).$$

Let $x = r/k$. Then $dx/dk = -r/k^2$, so by the chain rule

$$P_1'(k) = 2\, Q'(r/k) \cdot \frac{dx}{dk} = 2\, [-\varphi(r/k)] \left[-\frac{r}{k^2} \right] = \frac{2r}{k^2}\, \varphi\Big(\frac{r}{k}\Big).$$

B.7 Derivative of $P_i(k)$, $i = 2, \ldots, N-1$

Recall

$$P_i(k) = 2 \left[Q(r/k^i) - Q(r/k^{i-1}) \right].$$

For a general threshold r/k^j, let $x = r/k^j = r\, k^{-j}$, so $dx/dk = -jr/k^{j+1}$, and

$$\frac{d}{dk} Q(r/k^j) = Q'(x) \cdot \frac{dx}{dk} = \left[-\varphi(r/k^j) \right] \left[-\frac{jr}{k^{j+1}} \right] = \frac{jr}{k^{j+1}}\, \varphi\Big(\frac{r}{k^j}\Big).$$

Applying this with $j = i$ (first term) and $j = i - 1$ (second term),

$$P_i'(k) = 2 \left[\frac{ir}{k^{i+1}} \varphi\Big(\frac{r}{k^i}\Big) - \frac{(i-1)r}{k^i} \varphi\Big(\frac{r}{k^{i-1}}\Big) \right] = \frac{2ir}{k^{i+1}}\, \varphi\Big(\frac{r}{k^i}\Big) - \frac{2(i-1)r}{k^i}\, \varphi\Big(\frac{r}{k^{i-1}}\Big).$$

B.8 Derivative of $P_N(k)$

Recall

$$P_N(k) = 1 - 2\, Q(r/k^{N-1}).$$

Applying the same result with $j = N - 1$,

$$\frac{d}{dk} Q(r/k^{N-1}) = \frac{(N-1)r}{k^N}\, \varphi\Big(\frac{r}{k^{N-1}}\Big).$$

Since the constant 1 vanishes under differentiation,

$$P_N'(k) = -\frac{2(N-1)r}{k^N}\, \varphi\Big(\frac{r}{k^{N-1}}\Big).$$

B.9 The Gaussian Optimality Equation

Recall

$$R_N(k) = P_1(k) + \sum_{i=2}^{N-1} \frac{P_i(k)}{k^{2(i-1)}} + \frac{P_N(k)}{k^{2(N-1)}}.$$

The $P_1(k)$ term carries no k-dependent prefactor, so it differentiates directly: $\frac{d}{dk}P_1(k) = P_1'(k)$. Each remaining term is a product $P_i(k)\,k^{-2(i-1)}$; by the product rule,

$$\frac{d}{dk}\left[P_i(k)\,k^{-2(i-1)}\right] = P_i'(k)\,k^{-2(i-1)} + P_i(k)\cdot(-2(i-1))\,k^{-2i+1} = \frac{P_i'(k)}{k^{2(i-1)}} - \frac{2(i-1)P_i(k)}{k^{2i-1}}.$$

The same rule applies to the final term with $i = N$. Summing all contributions,

$$R_N'(k) = P_1'(k) + \sum_{i=2}^{N-1}\left[\frac{P_i'(k)}{k^{2(i-1)}} - \frac{2(i-1)P_i(k)}{k^{2i-1}}\right] + \left[\frac{P_N'(k)}{k^{2(N-1)}} - \frac{2(N-1)P_N(k)}{k^{2N-1}}\right]. \tag{8}$$

We denote this general form $\Phi_N(k,r) := R_N'(k)$. Substituting $P_1'(k)$, $P_i'(k)$, $P_N'(k)$ (Appendices B.6–B.8) into Eq. equation 8, and converting each $Q(\cdot)$ term via $Q(x) = \frac{1}{2}\left[1 - \operatorname{erf}(x/\sqrt{2})\right]$ (so the constant $\frac{1}{2}$ contributions cancel identically across the sum), yields the closed-form Gaussian optimality equation stated below.

Proposition B.1 (Gaussian N-Region Optimality Equation)**.**

$$\begin{aligned}\Phi_N(k,r) = \frac{2r}{k^2}\,\varphi\Big(\frac{r}{k}\Big) &\\ + \sum_{i=2}^{N-1}\Bigg[&-\frac{2(i-1)r}{k^{3i-2}}\,\varphi\Big(\frac{r}{k^{i-1}}\Big) + \frac{2ir}{k^{3i-1}}\,\varphi\Big(\frac{r}{k^{i}}\Big)\\ &+ \frac{2(i-1)}{k^{2i-1}}\left(\operatorname{erf}\Big(\frac{r}{\sqrt{2}\,k^{i}}\Big) - \operatorname{erf}\Big(\frac{r}{\sqrt{2}\,k^{i-1}}\Big)\right)\Bigg]\\ &- \frac{2(N-1)r}{k^{3N-2}}\,\varphi\Big(\frac{r}{k^{N-1}}\Big) - \frac{2(N-1)}{k^{2N-1}}\,\operatorname{erf}\Big(\frac{r}{\sqrt{2}\,k^{N-1}}\Big),\end{aligned}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. The optimal Gaussian scale factor k^* satisfies $\Phi_N(k^*, r) = 0$, giving a closed-form condition for the optimal geometric threshold factor directly in terms of $r = M/\sigma$, without iterative minimization of $R_N(k)$.

B.10 Monotonicity of the Region Probabilities

Proposition B.2 (Sign Behaviour of Region-Probability Derivatives)**.** *Let $P_1(k)$, $P_i(k)$ $(i = 2,\dots,N-1)$, and $P_N(k)$ be the Gaussian region probabilities above, with $r = M/\sigma > 0$ fixed. Then for all $k > 1$:*

- *(i) $P_1(k)$ is strictly increasing: $P_1'(k) > 0$.*
- *(ii) Each intermediate $P_i(k)$, $i = 2,\dots,N-1$, is unimodal: $P_i'(k) > 0$ for small k, $P_i'(k) < 0$ for large k, with a unique interior maximum.*
- *(iii) $P_N(k)$ is strictly decreasing: $P_N'(k) < 0$.*

Proof. **(i)** From Appendix B.6, $P_1'(k) = \frac{2r}{k^2}\,\varphi(r/k)$. Since $r > 0$, $k > 0$, and $\varphi(x) > 0$ for all $x \in \mathbb{R}$, every factor is positive, so $P_1'(k) > 0$ for all $k > 1$. Geometrically, as k increases, $T_1 = M/k$ decreases, so $R_1 = \{|w| > T_1\}$ grows monotonically, consistent with $P_1(k)$ increasing.

(ii) At $k = 1$, the thresholds coincide, $r/k^i = r/k^{i-1} = r$, giving $P_i(1) = 2[Q(r) - Q(r)] = 0$. As $k \to \infty$, both $r/k^i \to 0$ and $r/k^{i-1} \to 0$, so $Q(r/k^i), Q(r/k^{i-1}) \to \frac{1}{2}$ and $P_i(k) \to 0$. Since $P_i(k) \geq 0$ for all k (it is a probability), $P_i(k)$ must rise from 0 at $k = 1$ and return to 0 as $k \to \infty$, and is therefore unimodal by continuity and Rolle's theorem: there exists $k_i > 1$ with $P_i'(k) > 0$ for $k < k_i$, $P_i'(k_i) = 0$, and $P_i'(k) < 0$ for $k > k_i$. Hence $P_i(k)$ is not globally monotonic.

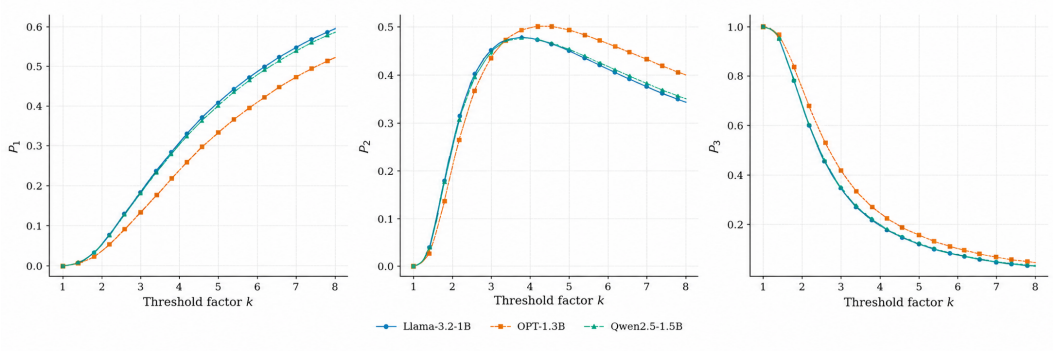

Figure 4: Empirical region probabilities $P_1(k)$, $P_2(k)$, $P_3(k)$ for $N = 3$ as functions of the threshold factor k, computed from actual pretrained weights of Llama-3.2-1B, OPT-1.3B, and Qwen2.5-1.5B. Across all three models, $P_1(k)$ is strictly increasing, $P_2(k)$ is unimodal with an interior maximum near $k \approx 4$, and $P_3(k)$ is strictly decreasing — confirming Proposition B.2 empirically.

(iii) From Appendix B.8, $P'_N(k) = -\frac{2(N-1)r}{k^N}\,\varphi(r/k^{N-1})$. Every factor is positive except the leading minus sign, so $P'_N(k) < 0$ for all $k > 1$. Geometrically, as k increases, $T_{N-1} = M/k^{N-1}$ decreases, so $R_N = \{|w| \le T_{N-1}\}$ shrinks monotonically, consistent with $P_N(k) = \operatorname{erf}\big(r/(\sqrt{2}\,k^{N-1})\big) \to 0$ as $k \to \infty$. □

Proposition B.2 is geometrically intuitive: as k increases, the outer region R_1 grows, the inner region R_N shrinks, and each intermediate region R_i is swept through in turn, gaining and then losing probability mass exactly once. This structure guarantees that $R_N(k) = \sum_i P_i(k)/k^{2(i-1)}$ admits a nontrivial interior root $k^* > 1$ of $\Phi_N(k, r) = R'_N(k) = 0$ (Proposition B.1), rather than a degenerate boundary optimum at $k = 1$ or $k \to \infty$. The zero-crossing of an individual $P'_i(k)$ marks only where that region's probability mass peaks, not the global optimum k^*: for $N = 3$, $P'_2(k)$ crosses zero near $k \approx 4$ (Figure 5), while the true optimum lies near $k^* \approx 2$ — the optimum balances all regions jointly through the weighted objective $R_N(k)$, not any single region's turning point. Figures 4 and 5 confirm this sign behaviour numerically.

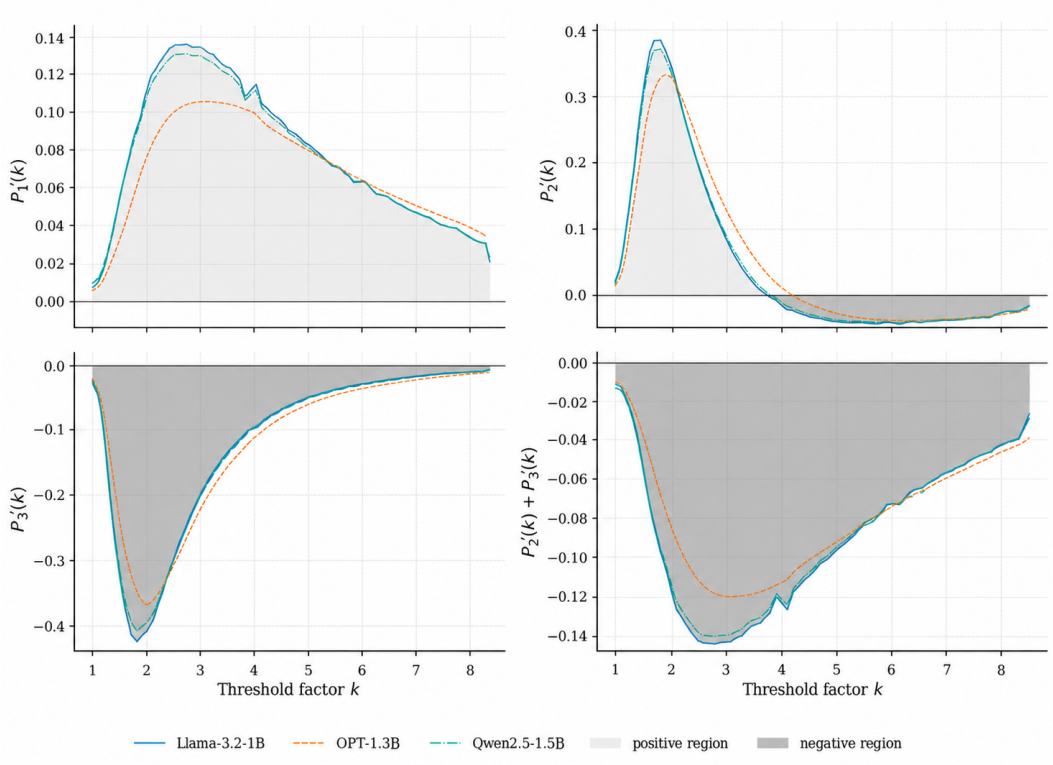

Figure 5: Empirical derivatives $P_1'(k)$, $P_2'(k)$, $P_3'(k)$ for $N = 3$, computed from real pretrained weights of Llama-3.2-1B, OPT-1.3B, and Qwen2.5-1.5B. Consistent with Proposition B.2, $P_1'(k) > 0$ and $P_3'(k) < 0$ throughout, while $P_2'(k)$ changes sign once. Bottom-right shows $P_2'(k) + P_3'(k) = -P_1'(k)$, an identity following from $\sum_{i=1}^{3} P_i(k) = 1$, verified empirically on real weight statistics across all three models.

Table 9: Notation used throughout the paper.

Symbol	Meaning
W, w	Weight matrix / individual weight value
$p(w)$	Probability density of weights
M	Per-row max weight magnitude, $M = \max \lvert W \rvert$
σ	Standard deviation of weights
r	Normalized dynamic range, $r = M/\sigma$
b	Bit-width
$q_{\max}$	Max quantization level, $q_{\max} = 2^{b-1} - 1$
N	Number of geometric regions
k	Geometric threshold factor
T_i	i-th threshold, $T_i = M/k^i$
$\mathcal{R}_i$	i-th quantization region
P_i, $P_i(k)$	Probability mass of region i
s_i	Step size of region i, $s_i = M/(k^{i-1} q_{\max})$
$\hat{w}$	Quantized weight
E_i	Distortion contributed by region i
E_{TSQ}, $E_{\mathrm{TSQ}}^{(N)}$	Total TSQ distortion
E_{uniform}	Uniform quantizer distortion
$R_N(k)$	Normalized TSQ distortion objective
k_N^*	Optimal threshold factor, $k_N^* = \arg\min_{k>1} R_N(k)$
R_N^*	$R_N(k)$ evaluated at k_N^*
R_∞	Saturation value of $R_N(k)$, $N \to \infty$
D_N	Normalized MSE distortion, $D_N = E_{\mathrm{TSQ}}^{(N)}/\sigma^2$
$Q(x)$	Standard Gaussian tail function
$\varphi(x)$	Standard Gaussian density
$\mathrm{erf}(z)$	Error function
$\Phi_N(k, r)$	Gaussian optimality function, $\Phi_N = R'_N(k)$
$\epsilon_{\mathrm{rel}}(w)$	Relative quantization error at w
$\epsilon_{\mathrm{rel}}^{\mathrm{env}}$	Envelope (worst-case) relative-error bound
G	AWQ calibration group size
i	Region index used in entropy coding
q	Quantization level, $q \in [-q_{\max}, q_{\max}]$
ρ	Packed rANS symbol, $\rho = 16i + (q + q_{\max})$
$\ell(\rho)$	Frequency of rANS symbol ρ
$c(\rho)$	Cumulative frequency preceding ρ
L	Total rANS frequency, $L = \sum_\rho \ell(\rho)$
x	rANS coding state
$L_{\max}$	Upper bound of the rANS renormalization interval