

Q-ROAR: Outlier-Aware Rescaling for RoPE Position Interpolation in Quantized Long-Context LLMs

Ye Qiao, Sitaoh Huang

University of California, Irvine, Irvine, California, USA
 {yeq6, sitaoh}@uci.edu

Abstract

Extending LLM context windows is crucial for long range tasks. RoPE-based position interpolation (PI) methods like linear and frequency-aware scaling extend input lengths without retraining, while post-training quantization (PTQ) enables practical deployment. We show that combining PI with PTQ degrades accuracy due to coupled effects long context aliasing, dynamic range dilation, axis grid anisotropy, and outlier shifting that induce position-dependent logit noise. We provide the first systematic analysis of PI plus PTQ and introduce two diagnostics: *Interpolation Pressure* (per-band phase scaling sensitivity) and *Tail Inflation Ratios* (outlier shift from short to long contexts). To address this, we propose **Q-ROAR**, a RoPE-aware, weight-only stabilization that groups RoPE dimensions into a few frequency bands and performs a small search over per-band scales for W_Q, W_K , with an optional symmetric variant to preserve logit scale. The diagnostics guided search uses a tiny long-context dev set and requires no fine-tuning, kernel, or architecture changes. Empirically, Q-ROAR recovers up to 0.7% accuracy on standard tasks and reduces GovReport perplexity by more than 10%, while preserving short-context performance and compatibility with existing inference stacks.

Motivation and Problem

LLMs(Touvron 2023) increasingly rely on long contexts for summarization, RAG, code, chain-of-thought, and argentic workflows. Inference-time RoPE scaling (PI) (Peng et al. 2023) extends the window is effective even without finetuning, but it is typically studied in full precision. Meanwhile, Post Training Quantization(PTQ)(Frantar et al. 2022; Lin et al. 2024) is essential for practical serving.

We observe that naively applying position interpolation (PI) to post-training quantized (PTQ) LLM models degrades accuracy both within and beyond the pretraining window (Figure 1). We compare quantized LLaMA-2-7B models with and without YaRN interpolation (Peng et al. 2023). In the non-interpolated control (Figure 1a), quantized models behave as expected, showing only modest degradation. In contrast, under YaRN, all quantized variants deteriorate and most sharply for the generic RTN (round-to-nearest) configuration. AWQ performs better out of the box than RTN,

suggesting that explicit activation outlier handling is implicated in the robustness gap. Motivated by this, we conduct a principled analysis and attribute the failures to *coupling* between RoPE scaling and quantization: (i) *aliasing* as high-frequency phases wrap; (ii) *dynamic-range dilation* that inflates pre-activation tails; (iii) *anisotropy* when axis-aligned quantizers operate on RoPE-rotated pairs; and (iv) *outlier shift/amplification*. Together, these effects induce position-dependent logit noise.

Interpolation Pressure and Tail Inflation

Most RoPE scaling(interpolation) methods share a unified form:

$$\phi_i^{\text{scaled}}(m) = \omega_i \frac{f(m)}{s_i}, \quad s_i > 0, \quad (1)$$

where $f(\cdot)$ warps positions and s_i rescales per-dimension frequency. Let the training regime support $|m - n| \leq L_0$, target displacement D , and deviation $\varepsilon_i(D) = \omega_i \left(\frac{f(D)}{s_i} - D_0 \right)$. We define the sensitivity

$$\text{IP}_i = \left| \frac{\partial \varepsilon_i(D)}{\partial s_i} \right| = \omega_i \frac{f(D)}{s_i^2}, \quad (2)$$

which grows with ω_i and D , identifying high-frequency bands as fragile.

To capture PI-induced outlier shift, we use *Tail-Inflation Ratios* at a high quantile $1 - \varepsilon$:

$$\text{TIR}_i^W = \frac{Q_{|w_i^\top h|, \text{long}}(1 - \varepsilon)}{Q_{|w_i^\top h|, \text{short}}(1 - \varepsilon)}, \quad \text{TIR}_i^A(m) = \frac{Q_{\|R(\phi_i^{\text{scaled}}(m))u_i\|_\infty}(1 - \varepsilon)}{Q_{\|R(\phi_i(m))u_i\|_\infty}(1 - \varepsilon)}. \quad (3)$$

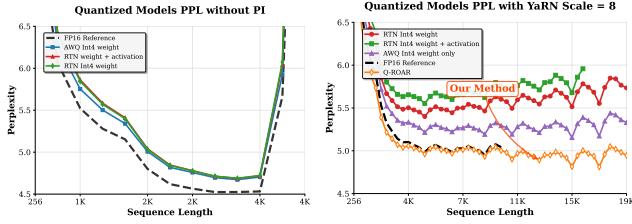
TIR^W reflects pre-activation tail growth; TIR^A reflects phase-axis amplitude inflation that increases activation clipping.

Geometric intuition. RoPE preserves ℓ_2 energy but rotates signal and quantization noise relative to axis-aligned bins; PI changes the angular trajectory over long spans. Thus, even uniform step sizes calibrated at short contexts become phase-misaligned at long contexts, increasing effective error along certain axes and perturbing attention logits.

Q-ROAR (Band-wise Weight Rescaling)

We partition RoPE pairs into B log-spaced frequency bands $\{\mathcal{B}_b\}_{b=1}^B$ and assign a single scale g_b :

$$W_Q^{(b)} \leftarrow g_b W_Q^{(b)}, \quad W_K^{(b)} \leftarrow \begin{cases} g_b W_K^{(b)} & (\text{shared mode}) \\ g_b^{-1} W_K^{(b)} & (\text{symmetric mode}) \end{cases} \quad (4)$$



(a) None Interpolated models (b) YaRN interpolation ($s = 8$)

Figure 1: Quantized Llama-2-7b vs. GovReport PPL

Algorithm 1: Q-ROAR search

- 1: Partition RoPE dims into $\{\mathcal{B}_b\}_{b=1}^B$
- 2: Estimate per-band IP_b and TIR_b^W from short vs. PI long-context caches
- 3: Derive band windows \mathcal{G}_b ; build log-spaced grids $G_b \subset \mathcal{G}_b$
- 4: **for** candidates $\{g_b \in G_b\}$ **do** evaluate ppl on tiny dev set
- 5: Select $\{g_b^*\}$ minimizing the objective; prefer symmetric mode; fallback to shared if unstable
- 6: Serialize $\{g_b^*\}$ and mode in model metadata

Symmetric mode approximately preserves logit magnitude. We set per-band windows $\mathcal{G}_b = [g_b^{\min}, g_b^{\max}]$ by combining IP (tight for high frequencies) with TIR^W (shrink inflated tails). A tiny long-context dev set (~ 10 docs) drives a lightweight grid search that minimizes a length-weighted perplexity objective.

We choose **Band count** $B \in \{6, 8\}$ with log-frequency grouping; **Grid** 5–9 log-spaced candidates per band with coordinate or small joint search and early stopping ($\text{gain} < \eta$); **Objective** $\sum_{L \in \mathcal{L}} w_L \text{ppl}(L; \{g_b\})$ with w_L emphasizing longer lengths and KV reuse; **Safety** bound g_b via $\gamma_b = 1 + \frac{\tau}{1 + \log(\omega_{b,\text{med}}/\omega_{\min})}$ and κ/TIR_b^W , with $\kappa \in [1.0, 1.3]$, $\tau \in [0.2, 0.5]$.

We focus on rescaling *weights* of key and query projection layers (W_Q, W_K) rather than adjusting activation quantization for three reasons. (i) Activation statistics drift with content/position under PI; changing activation clips/steps couples to kernels/runtimes. (ii) Weight rescaling is static, quantizer and kernel agnostic (AWQ/RTN), and portable. (iii) Symmetric scaling keeps logit scale stable, avoiding retuning.

We sampled 10 long documents from Proof-pile (Zhangir Azerbayev 2022) dataset that each of them are longer than 60k tokens; cache short vs. PI distributions to compute TIR. **Complexity.** with running average sliding window size of 256. If each band has K candidates, evaluation cost is $\mathcal{O}(BK)$ runs (coordinate search) or $\mathcal{O}(K^B)$ (small B only) with token reuse via sliding windows. **Serialization.** Store $\{g_b\}$ alongside checkpoint; apply at load time to (W_Q, W_K) tensors only. The search was conducted using two NVIDIA RTX 4090 and roughly took 4 GPU hours.

Experiments

For long-context evaluation, we benchmarked GovReport, and for standard LLM performance we used WikiText2(Merity et al. 2016) and five zero-shot tasks. At the base 4K context, quantized models track FP16 with only modest degradation. Under YaRN interpolation (32K/64K), RTN and AWQ exhibit clear drops in accuracy and higher perplexity, confirming the destructive interaction between position interpolation and quantization. In contrast, Q-ROAR consistently improves robustness that recovering up to 0.7% accuracy on the standard suite (notably, even where PI-PTQ coupling is min-

Table 1: GovReport perplexity on LLaMA-2-7B across evaluation context sizes (lower is better).

Setting	Context	Evaluation Context Window Size				
		2048	4096	8192	16384	32768
<i>Extended Context with YaRN (64K tokens, s=16)</i>						
FP16	64K	4.437	4.359	4.329	4.175	6.069
RTN W4	64K	4.544	4.485	4.470	4.485	6.713
AWQ W4	64K	4.489	4.421	4.405	4.414	6.302
Q-ROAR W4	64K	4.444	4.393	4.321	4.181	5.833

Table 2: Performance of LLaMA-2-7B on standard LLM benchmarks under different quantization and PI settings.

Setting	Context	5-shot Avg.	WikiText2 PPL	
			Base Context (4096 context window)	
Baseline FP16	4k	70.83%	5.47	
RTN W4	4k	64.14%	5.73	
RTN W4-A4	4k	62.02%	5.92	
AWQ W4	4k	64.25%	5.61	
AWQ W4-A4	4k	62.31%	5.77	
<i>Extended Context with YaRN (32K context window, s=8)</i>				
Baseline FP16	32K	63.71%	6.09	
RTN W4	32K	63.32%	6.41	
RTN W4-A4	32K	62.84%	6.60	
AWQ W4	32K	63.52%	6.31	
AWQ W4-A4	32K	62.53%	6.49	
Q-ROAR W4	32K	63.96%	6.19	
Q-ROAR W4-A4	32K	63.21%	6.40	

imal) and yielding over 10% relative perplexity reductions on GovReport versus RTN while preserving short-context performance. Overall, Q-ROAR mitigates aliasing and outlier amplification, enabling stable long context inference in quantized LLMs without retraining or kernel changes.

Conclusion

We analyzed why PI harms PTQ LLMs and introduced IP and TIR to quantify the coupling between RoPE scaling and quantization. Building on these diagnostics, **Q-ROAR** provides a portable, bandwidth-rescaled rescaling of (W_Q, W_K) that delivers consistent long context gains with negligible overhead and no kernel changes. While Q-ROAR assumes RoPE and targets weight-quantized models (heavy activation/KV quantization may still benefit from modest clip expansion guided by TIR), it substantially mitigates aliasing and outlier amplification at extended lengths. Future work includes extending the approach to non-RoPE position encoding and integrating lightweight, context-aware activation calibration.

References

- Frantar, E.; Ashkboos, S.; Hoefer, T.; and Alistarh, D. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.
- Lin, J.; et al. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100.
- Merity, S.; et al. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.
- Touvron, H. o. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Zhangir Azerbayev, B. P., Edward Ayers. 2022. Proof-pile.