CALIBRATION TIMING AND MODEL SENSITIVITY IN ROTATED KV-CACHE QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Does a rotated uniform KV-cache quantizer require locally fitted scales? A comparison can conflate scale granularity with calibration timing and metadata cost. We audit these factors with non-overlapping finite windows, a dense non-evicted reference, and one-write quantize–dequantize simulation. Following development experiments, a frozen 80-condition comparison crosses model, dataset, rotation seed, calibration timing, granularity, and symmetry, with eight-bit controls. On Llama-3.1-8B-Instruct, four-bit per-write global calibration yields post-prefill perplexities of 7.508–7.557 on WikiText-2 test (reference 7.404), whereas prefill-frozen global configurations yield 125.414–39,255.244. The timing effect persists on SQuAD contexts and for two new rotations. Grouped per-write scales reduce error further but incur more metadata, preventing an equal-total-budget claim. Qwen2.5-7B-Instruct behaves differently: all tested four-bit configurations fail severely across both datasets and rotations, while eight-bit controls remain close to their dense references. Unquantized cache and rotation-roundtrip controls also remain close on a diagnostic window. These results support calibration- and model-specific conclusions rather than universal necessity of local scales. They do not establish packed-kernel speed, cross-domain generalization, or a unique mechanism for the Qwen low-bit failure.

1 QUESTION AND SCOPE

Does a rotated uniform quantizer need locally fitted scales in a given evaluation setting? This question requires specifying both the quantizer and the reference. Rotation plus group-wise KV quantization is already part of prior systems such as QuaRot (Ashkboos et al., 2024); we do not present that combination as a new core quantizer. Nor does a comparison among uniform quantizers establish superiority over non-uniform codebooks, residual correction, or complete serving systems.

Our question is which apparent benefits of local scales survive when their calibration timing is controlled. An initial two-model comparison motivates this question, but confounds frozen global scales with per-write grouped scales. We then vary timing and granularity independently, equalize the number of quantization levels, and evaluate a fixed two-model, two-dataset, two-rotation matrix. The results separate a reproducible calibration effect on Llama from a severe four-bit sensitivity on Qwen. Eight-bit and unquantized controls bound the interpretation of the latter.

We contribute a controlled evaluation protocol, an explicit accounting of metadata, and complete measurements of the timing–granularity interaction. The mathematical observations clarify why rotation error envelopes alone cannot order realized model losses. The study concerns these specified uniform quantizers; it does not compare complete serving systems or propose a new rotation algorithm.

2 CONDITIONAL MATHEMATICAL MOTIVATION

2.1 ROTATION BOUNDS DO NOT IMPLY ACTUAL IMPROVEMENT

Let R be a normalized Hadamard transform after independent random signs, with a supported power-of-two dimension d. For a fixed unit vector x, each coordinate of Rx is a Rademacher sum.

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.

Proposition 1. *For $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\|Rx\|_\infty \le \sqrt{2\log(2d/\delta)/d}.$$

For a non-clipping symmetric nearest-rounding quantizer with $q_{\max} = 2^{b-1} - 1$, $b \ge 2$, and $s = \|Rx\|_\infty / q_{\max}$,

$$\frac{1}{d}\|x - R^\top Q(Rx)\|_2^2 \le \frac{\log(2d/\delta)}{2dq_{\max}^2}.$$

These are coordinate and error upper bounds, not a comparison to the realized unrotated error.

Proof. The coordinate tail is at most $2\exp(-dt^2/2)$. A union bound over d coordinates gives the first claim. Non-clipping rounding has coordinate error at most $s/2$. Orthogonality preserves its squared norm, and division by d yields the second bound. □

For example, $x = (1, 1/7, 0, 0)$ is exactly represented by unrotated symmetric 4-bit quantization, while a normalized Hadamard transform with all positive signs has nonzero rounding error. Normalizing x preserves the counterexample. A better upper bound cannot prove that every actual error improves. Peak magnitude is also not kurtosis.

2.2 Logit range is an upper-bound diagnostic

Let $a = \mathrm{softmax}(z)$ and $\widehat{a} = \mathrm{softmax}(z - \eta)$, and set $\Delta = \max_i \eta_i - \min_i \eta_i$.

Proposition 2. *For finite logits, $\|a - \widehat{a}\|_{\mathrm{TV}} \le \min\{1, \Delta e^{\Delta}\}$. Token k overtakes token j exactly when $\eta_j - \eta_k > z_j - z_k = \log(a_j/a_k)$.*

Proof. Write $\widehat{a}_i/a_i = e^{u_i}$, where $u_i = -\eta_i + c$ and c is common. Since $\sum_i a_i e^{u_i} = 1$, $\min_i u_i \le 0 \le \max_i u_i$, and $\max_i u_i \le \Delta$. Summing $a_i(e^{u_i} - 1)$ over its nonnegative terms bounds TV by Δe^{Δ}. TV is also at most one. The overtaking statement is the direct comparison of the two perturbed logits. □

A larger total range does not guarantee misranking of any particular pair. The bound gives no positive lower bound on TV and is uninformative when its right-hand side reaches one.

For a unit query q and independent centered coordinate errors bounded by $s/2$, the projected logit error $q^\top \varepsilon/\sqrt{d}$ has a sub-Gaussian variance proxy at most $s^2/(4d)$. A maximal inequality then gives expected range $O(s\sqrt{\log n}/\sqrt{d})$ over n such logits. Independence and centering are assumptions, not consequences of deterministic rounding. A matching $\sqrt{\log n}$ order under an iid Gaussian model is an idealization, not an exact law for bounded rounding errors.

For coordinate-specific steps the analogous proxy is $\sum_j q_j^2 s_j^2/(4d)$. Group maxima cannot exceed a global maximum fitted on the same tensor with the same symmetric rule. This compares non-clipping envelopes only; it neither orders realized errors nor applies automatically against a frozen prefill scale on future tokens. The experiments below do not measure or validate the independence assumptions, and no theorem asserts a quality separation between their four configurations.

3 Four Quantizer Configurations

For each layer, generate the same signed orthogonal Hadamard matrix with seed $42 + \ell$ in all conditions. Rotate each FP16 key/value vector in FP32, quantize in the rotated basis, invert the rotation, and cast back to FP16 for ordinary model attention. Keys and values have separate calibration parameters. This QDQ route is not an implementation that directly attends to packed rotated vectors.

Symmetric. Use $q \in \{-7, \dots, 7\}$, with scale $s = \max\{\max|x|, 10^{-8}\}/7$, zero offset, nearest rounding, and clipping. It has 15 integer levels in a nominal 4-bit payload.

Asymmetric. Use $q \in \{0, \dots, 15\}$, offset $o = \min x$, and $s = \max\{\max x - \min x, 10^{-8}\}/15$, reconstructing $o + sq$. It has 16 levels. Symmetry therefore also changes the number and placement of representable levels; it is not solely a zero-point toggle.

Global. Fit one scale (and, when asymmetric, one offset) per layer and tensor kind across all prefill heads, tokens, and channels. Freeze those parameters for the remainder of that window. Later vectors may clip; no claim excludes this mechanism.

Grouped. Fit parameters for each token, KV head, and contiguous 32-channel group when written. Prefill vectors and subsequently generated KV vectors are each quantized once. Previously written entries are reused without re-quantization or eviction within a window.

Modeled versus actual storage. If E denotes the retained key/value element count and L the number of layers, global metadata costs $8L$ bytes for symmetric scales and $16L$ bytes for asymmetric scale–offset pairs. Grouped metadata costs $4E/32$ or $8E/32$ bytes. With the common payload $E/2$ bytes, the grouped effective bit-rates are 5 and 6 bits, whereas global rates approach 4 bits as E increases. These counts cover payload and scale/offset metadata, not every possible runtime allocation or packing-layout cost. Rotation matrices and floating-point tensors are recorded separately. Actual retained QDQ cache tensors use $2E$ bytes; no packed GPU allocation or kernel speedup is measured.

4 Windowed Evaluation

Models and evaluation text. We use Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3, each with its own tokenizer, on the same frozen WikiText test text joined with double newlines. We partition each token stream into non-overlapping windows of at most 4,096 tokens. Each window resets positions to zero, cache state, and global calibration. Different tokenizers need not yield the same number of tokens or windows.

Reference and targets. A dense FP16 causal forward pass with no cache supplies reference logits within each window. We prefill the first 512 tokens, or all but the last token for a shorter tail. We score only target positions after this prefix. Each such target is included exactly once; prefill targets are excluded. The first post-prefill prediction uses the prefill's unquantized last logits in the incremental path, so it is a reference-matching boundary prediction rather than an isolated test of QDQ. Subsequent predictions use the persistent compressed cache. No cache crosses a window boundary.

We report

$$\mathrm{PPL}_{\mathrm{post}} = \exp\{\mathrm{NLL}_{\mathrm{post}}/N_{\mathrm{post}}\},$$

with the exact input, window, and target counts. This is a finite-window post-prefill metric, not standard full-token WikiText perplexity. We also report mean $\mathrm{KL}(p_{\mathrm{reference}} \| p_{\mathrm{compressed}})$ and top-1 agreement on the same targets. Models use FP16, SDPA attention, and one H100 per condition; metric calculations use FP32 logits.

Reference numerical check. Before full evaluation, we compare dense and unquantized incremental outputs on the first 128 tokens, with a 32-token prefill and 96 compared targets. Prespecified numerical tolerances are maximum absolute logit difference at most 0.25 and mean absolute target-NLL difference at most 0.02. Passing does not prove bitwise equivalence or certify every later window. Partition and storage arithmetic are checked separately from model-quality measurements.

Analysis and pairing. For each model, all four conditions use the same token sequence, window boundaries, input counts, and reference NLL. We aggregate complete windows before comparing NLL per target. Let $y_{gs}, y_{ga}, y_{ps}, y_{pa}$ be compressed-minus-reference mean NLL for global/grouped and symmetric/asymmetric conditions. Descriptive contrasts are

$$D_{\mathrm{group}} = ((y_{ps} - y_{gs}) + (y_{pa} - y_{ga}))/2, \tag{1}$$

$$D_{\mathrm{asym}} = ((y_{ga} - y_{gs}) + (y_{pa} - y_{ps}))/2, \tag{2}$$

$$I = (y_{pa} - y_{ps}) - (y_{ga} - y_{gs}). \tag{3}$$

They describe this fixed corpus and implementation. Windows are not asserted to be iid samples, and there are no multi-seed confidence intervals or causal mechanism estimates.

Scale	Symmetry	Modeled bits	$\mathrm{PPL}_{\mathrm{post}}$	Mean KL	Top-1 (%)
Llama-3.1-8B-Instruct; reference 6.4473					
Global	Symmetric	$\simeq 4$	19.5195	1.11051	56.39
Global	Asymmetric	$\simeq 4$	144.2616	3.09987	34.37
Grouped	Symmetric	5	6.5198	0.01075	95.37
Grouped	Asymmetric	6	6.4986	0.00710	96.28
Mistral-7B-Instruct-v0.3; reference 4.8772					
Global	Symmetric	$\simeq 4$	6.3925	0.30844	78.08
Global	Asymmetric	$\simeq 4$	50.6098	2.40121	49.78
Grouped	Symmetric	5	4.8990	0.00624	96.69
Grouped	Asymmetric	6	4.8913	0.00408	97.34

Table 1: Complete two-model finite-window comparison. KL is reference-to-compressed; top-1 is agreement with reference predictions, not task accuracy. All configurations have 4 payload bits, but modeled bits include unequal scale/offset costs. Actual retained caches remain FP16 QDQ tensors. Only the independent-window protocol is reported.

4.1 Initial comparison with coupled calibration

The initial comparison includes eight conditions. Each Llama-3.1-8B-Instruct condition consumes 289,077 input tokens in 71 windows and scores 252,725 post-prefill targets; each Mistral-7B-Instruct-v0.3 condition consumes 334,661 tokens in 82 windows and scores 292,677 targets. Within each model, all four token hashes, window boundaries, and per-window dense-reference NLL sums agree. Reference $\mathrm{PPL}_{\mathrm{post}}$ is 6.4473 for Llama and 4.8772 for Mistral. Table 1 includes every condition, including both severely degraded global/asymmetric results.

On this fixed corpus, the descriptive mean-NLL contrasts $(D_{\mathrm{group}}, D_{\mathrm{asym}}, I)$ are $(-2.09831, 0.99848, -2.00348)$ for Llama and $(-1.30139, 1.03372, -2.07061)$ for Mistral, in nats per target. In both models, the positive averaged asymmetry contrast masks opposite directions: asymmetric quantization is much worse globally but slightly better within groups. Neither grouped configuration improves on its dense reference. The observations do not isolate clipping, scale locality, or calibration timing, and the grouped conditions use more modeled bits. They therefore do not establish a matched-total-budget advantage or a universal need for grouped scales.

The actual retained floating-point cache reaches 536,739,840 bytes (511.875 MiB) in every condition; the modeled packed totals at that point are 134,185,216/134,185,472 bytes for global symmetric/asymmetric and 167,731,200/201,277,440 bytes for grouped symmetric/asymmetric. These are representation estimates, not measured packed allocations or peak GPU memory. The dense/incremental reference check yields maximum logit differences 0.04297/0.02734 and mean target-NLL differences 0.00260/0.00155 for Llama/Mistral, within the prespecified tolerances; it remains a short engineering check rather than a proof of numerical equivalence everywhere.

5 Equal-level factorial protocol

The factorial experiments use the same number of levels for both grids. Let x denote rotated coordinates in one calibration region and let b be payload precision. For an asymmetric grid, set $\ell = \min x$ and $h = \max x$; for a symmetric grid, set $h = \max |x|$ and $\ell = -h$. In both cases define

$$s = \frac{\max(h - \ell, 10^{-8})}{2^b - 1}, \qquad \widehat{x} = \ell + s \ \mathrm{clip}\left(\mathrm{round}\,\frac{x - \ell}{s}, 0, 2^b - 1\right).$$

The symmetric grid is mid-rise with an even number of levels. Except in degenerate cases it has no zero level; it is not the 15-level signed grid used in the initial comparison or the non-clipping quantizer in Proposition 1. Consequently, numerical differences between those studies are not pure replications with more data. The factorial design addresses the level-count confound explicitly and retains the original comparison as motivation.

Calibration axes. Write a KV tensor as $[B, H, T, G, C]$, where $C = 32$ and $G = d/32$. Per-write global calibration reduces over H, G, C while retaining T; per-write grouped calibration reduces

Table 2: Four-bit development results. All eight cells are reported; the other payload widths are retained with the experimental records. Effective bit-rates include scale/offset metadata.

Calibration	Grid	Global PPL	Group PPL	Bits: global/group
Prefill	Symmetric	568.139	41.997	4.000/4.002
Prefill	Asymmetric	56.183	5.889	4.000/4.004
Write	Symmetric	5.008	5.032	4.031/5.000
Write	Asymmetric	4.924	4.947	4.062/6.000

over C alone. During prefill this same rule is applied separately to every token, so the write rule does not use later prefill tokens to calibrate earlier ones. Frozen calibration additionally reduces over T and stores the resulting parameters for the whole window. After prefill, future entries use those stored parameters. Keys and values are calibrated separately, and each layer has its own signed rotation. No experiment pools scales between models, layers, windows, or key and value tensors.

Following a cache entry. An entry is rotated in FP32, quantized once, inverse-rotated in FP32, and stored in FP16. Subsequent decoding concatenates the new entry with the previously reconstructed entries. It does not rotate or quantize those old entries a second time. Dense-reference logits and incremental-path logits are compared at the same target position. The first scored target is a common unquantized prefill prediction, which contributes one of 1,536 targets per follow-up window. Reporting this boundary avoids attributing that reference-matching prediction to the compressed cache.

6 Separating calibration timing in a development pilot

We crossed calibration timing (prefill-frozen or per-write), scale granularity (global or grouped), symmetry, and three payload widths (3, 4, and 5 bits), for 24 conditions on Llama-3.1-8B-Instruct. Two 512-token windows from the WikiText-2 validation text each supply 448 prefill tokens and 64 scored targets. The first target uses the common dense prefill prediction; later targets use quantized cache states. These 128 targets form a development diagnostic, not an independent benchmark-level estimate. The dense-reference PPL is 4.956.

Both symmetries use 2^b uniformly spaced levels: the symmetric grid is mid-rise, unlike the earlier $2^b - 1$-level integer-symmetric grid. Within prefill-frozen conditions, both granularities fit scales once on prefill; within per-write conditions, each uses only the current token. Grouped scales retain head and channel-group distinctions; global scales pool heads and channels. FP32 scale and offset metadata are included in the modeled bit-rate. This is a factorial comparison at common payload widths, not equal total storage.

The large prefill-frozen gaps shrink when global scales are recalibrated at each write. At four bits, per-write global and grouped PPL differ by less than 0.03 under either symmetry. This small development sample does not establish equivalence, but it does not support attributing the earlier gap to scale granularity alone. Changes in calibration and the representation budget must be separated before asserting that local scales are necessary. No packed-memory or serving-speed benefit is measured.

6.1 Longer-window, two-model follow-up

The same eight four-bit conditions were evaluated on Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. For each tokenizer, windows 1–4 have length 2,048 and a 512-token prefill, giving 6,144 scored targets per condition. These windows exclude the initial short-pilot segment, but remain in the same validation corpus. They do not constitute cross-corpus confirmation. Per-write calibration again reduces the Llama error, but does not rescue Qwen: all eight Qwen configurations degrade severely. These results rule out a model-independent claim that per-write calibration suffices at four bits. They also prevent interpreting grouped scales as a general remedy. The mechanism of the Qwen degradation is unresolved by this factorial comparison; quantizer sensitivity and implementation-specific numerical behavior require separate controls. We therefore retain the full matrix and restrict the calibration finding to the measured Llama setting.

Table 3: All sixteen longer-window conditions. Entries are post-prefill perplexities. Dense references are 5.771 for Llama and 5.994 for Qwen. Fixed payload width does not match total metadata cost.

		Llama		Qwen	
Calibration	Grid	Global	Group	Global	Group
Prefill	Symmetric	14076.133	150.321	14608.471	16006.765
Prefill	Asymmetric	72.259	8.006	12012.552	7056.662
Write	Symmetric	5.889	5.844	4546.073	14097.966
Write	Asymmetric	5.874	5.818	11015.934	6354.512

Numerical controls on one Qwen window. On the first follow-up window (1,536 targets), the dense reference has PPL 5.48754. Rebuilding the cache without quantization gives 5.48803; applying only FP32 rotation and inverse rotation gives 5.48741. With per-write asymmetric eight-bit quantization, global and grouped PPLs are 5.49623 and 5.48409. These controls show no comparable degradation from cache rebuilding or the rotation roundtrip on this window and locate the observed sensitivity in the low-bit setting. They do not establish its mechanism, exclude every implementation issue, or replace broader low-bit validation.

7 Frozen follow-up across datasets and rotations

We fixed a follow-up matrix after the development experiments: two models, two data sources, two new rotation seeds (104729 and 130363), and eight four-bit configurations per combination. Two eight-bit, per-write asymmetric controls bring the total to 80 conditions. Each condition scores 6,144 targets in four independent 2,048-token windows with 512-token prefill. The data are the first four tokenized windows of WikiText-2 test and of the SQuAD validation contexts, deduplicated in first-occurrence order; questions and answer labels are excluded. WikiText test also appears in the initial comparison; the factorial follow-up is not a previously untouched test set across the entire study. Both datasets derive from Wikipedia, so this comparison does not establish cross-domain generalization. The same text windows are reused across rotations; seeds are not independent text samples.

On Llama, four-bit per-write global PPL ranges from 7.508 to 7.557 on WikiText-2 (reference 7.404) and from 5.490 to 5.516 on SQuAD contexts (reference 5.393). Grouped per-write values range from 7.450 to 7.484 and from 5.426 to 5.437, respectively. Frozen calibration instead produces much larger errors, with substantial rotation sensitivity. Thus, the importance of calibration timing persists beyond the original development windows. The smaller grouped error must still be weighed against its metadata cost; these are not equal-total-budget comparisons.

On Qwen, every four-bit configuration fails severely on both datasets and rotations: PPL ranges from 2,685.735 to 36,262.921, against dense references 5.178 and 6.229. In contrast, all eight-bit per-write asymmetric controls remain close to their references: 6.203–6.252 on WikiText-2 and 5.177–5.200 on SQuAD contexts. Together with the unquantized controls, this narrows the observed failure to the tested low-bit configurations; it does not identify a unique cause or invalidate other quantization systems. The supported conclusion is a calibration-by-model interaction in this uniform-QDQ protocol, not universal sufficiency of recalibration or grouped scales. Per-window losses and every configuration are retained for replication.

7.1 Aggregation and comparison units

Each reported perplexity is computed from the sum of token negative log-likelihoods, not the average of window perplexities. For window j, let T_j denote the number of scored targets and S_j their summed loss. The aggregate is

$$\mathrm{PPL} = \exp\left(\frac{\sum_j S_j}{\sum_j T_j}\right). \tag{4}$$

This weighting matters when the final corpus window is shorter. A comparison uses identical target positions for its compressed and reference paths. For any such pair, the log perplexity ratio equals the difference in mean token loss. Thus a large ratio represents an accumulated predictive penalty under

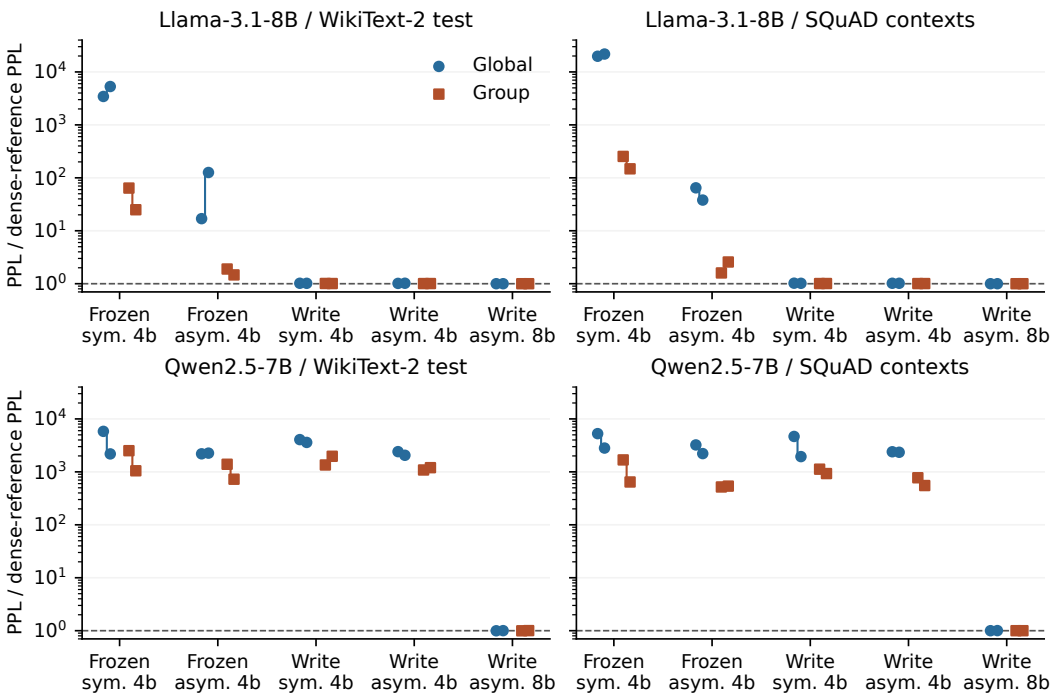


Figure 1: All 80 frozen follow-up conditions. Each point is one rotation seed; vertical segments connect the two seeds and are not confidence intervals. Perplexity ratios use each model and dataset's dense reference. The log scale retains the severe failures. Global and grouped configurations share payload width, but not total representation cost.

Table 4: Complete frozen Llama-3.1-8B-Instruct comparison. Each row reports one rotation and quantizer; all values are post-prefill PPL. Seeds a and b are 104729 and 130363. Global/grouped columns have equal payload but unequal metadata costs.

			WikiText-2 test		SQuAD contexts	
Bits / timing	Grid	Seed	Global	Group	Global	Group
4 / Frozen	Sym.	a	25480.902	474.125	106700.538	1368.159
4 / Frozen	Sym.	b	39255.244	184.857	117446.511	797.994
4 / Frozen	Asym.	a	125.414	14.035	348.973	8.619
4 / Frozen	Asym.	b	933.662	10.866	204.938	13.874
4 / Write	Sym.	a	7.552	7.484	5.516	5.432
4 / Write	Sym.	b	7.534	7.460	5.490	5.437
4 / Write	Asym.	a	7.508	7.450	5.497	5.426
4 / Write	Asym.	b	7.557	7.458	5.497	5.427
8 / Write	Asym.	a	7.407	7.405	5.392	5.392
8 / Write	Asym.	b	7.405	7.405	5.393	5.393

this scoring protocol; it is not a percentage change in downstream answer accuracy. Reporting both the reference and the ratio avoids confusing a harder text segment with a stronger quantization effect.

The frozen follow-up holds token windows fixed within each model and corpus, then varies rotation and quantizer settings. Model tokenizers can split the same source text differently, so matching window lengths does not guarantee identical character spans across models. We therefore interpret each model's quantized-versus-dense comparison internally, rather than comparing absolute perplexities between tokenizers. Likewise, repeating rotations on fixed text probes numerical sensitivity without creating new observations of language. The two corpora broaden the text selection, but their common Wikipedia origin limits the domain contrast.

Table 5: Complete frozen Qwen2.5-7B-Instruct comparison. Each row reports one rotation and quantizer; all values are post-prefill PPL. Seeds a and b are 104729 and 130363. Global/grouped columns have equal payload but unequal metadata costs.

			WikiText-2 test		SQuAD contexts	
Bits / timing	Grid	Seed	Global	Group	Global	Group
4 / Frozen	Sym.	a	36262.921	15648.829	27300.978	8711.918
4 / Frozen	Sym.	b	13602.725	6531.580	14575.838	3335.311
4 / Frozen	Asym.	a	13654.175	8668.692	16703.485	2685.735
4 / Frozen	Asym.	b	14064.992	4524.976	11458.965	2785.817
4 / Write	Sym.	a	25322.046	8404.180	24158.329	5817.157
4 / Write	Sym.	b	22410.598	12268.257	10052.773	4807.496
4 / Write	Asym.	a	15007.092	6760.992	12408.545	4008.070
4 / Write	Asym.	b	12809.624	7474.197	12140.107	2855.600
8 / Write	Asym.	a	6.206	6.203	5.178	5.178
8 / Write	Asym.	b	6.243	6.252	5.200	5.177

These units also determine what a future uncertainty analysis would need to resample. Tokens inside a window are dependent, and configurations sharing a window are paired. Treating every token, rotation, and configuration as an independent replicate would overstate effective sample size. We retain the window-level loss sums so that the reported aggregates can be recomputed, but four follow-up windows per cell are insufficient for a strong population-level claim. The complete configuration matrix is consequently reported as a finite comparison, including the severe low-bit failures.

8 Interpreting calibration, precision, and cost

Equal payload is not equal storage. The four-bit factorial cells have the same number of payload bits and levels, but grouping creates more scale parameters. With FP32 metadata, a symmetric per-write 32-channel group adds one bit per element; an asymmetric group adds two. For global per-write calibration, one parameter is shared across all KV heads and channels at the current token, so the overhead depends on the model's KV-head count and head dimension. This is a different sharing rule from both a single sequence-wide scale and a per-channel temporal scale. Reported bit-rates must therefore state the axis of parameter sharing. Our saved per-window records expose payload and metadata separately.

Two contrasts and their reference. Changing granularity changes spatial sharing and metadata; changing calibration timing changes temporal adaptation and parameter storage. Neither is an equal-total-storage comparison. Equal-budget superiority would require implementable configurations accounting for scales, offsets, and residual buffers. A dense causal reference and a token-by-token quantized path retain the same context within each window. Entries are quantized once when written, testing accumulated perturbations without requantization. Window resets delimit this comparison; they do not reproduce a continuous long stream.

What the numerical controls resolve. The unquantized and rotation-roundtrip controls address the cache reconstruction and orthogonal-transform paths on one diagnostic window. Eight-bit controls extend the precision comparison across the full follow-up matrix. Their near-reference scores make a failure common to every use of the reconstructed cache less plausible, but do not isolate key versus value sensitivity, layer-specific amplification, clipping, or rounding. Those mechanisms would require separate interventions. The observed Qwen degradation is thus a reproducible result for this implementation, not a statement that the model cannot be quantized to four bits by another method.

9 Related Work

Rotation and numerical representation. QuaRot uses fixed rotations (Ashkboos et al., 2024), while SpinQuant learns them (Liu et al., 2025). QuIP and QuIP# connect incoherence with low-bit weight quantization (Chee et al., 2023; Tseng et al., 2024); their guarantees do not establish our

KV-cache claim. GPTQ and AWQ adapt weight quantization to reconstruction or activation information (Frantar et al., 2023; Lin et al., 2024). SmoothQuant redistributes activation difficulty (Xiao et al., 2023), while mixed-precision outlier handling and dense–sparse quantization provide other numerical choices (Dettmers et al., 2022; Kim et al., 2024). We isolate a uniform cache quantizer after a fixed signed Hadamard transform, rather than reproduce these full systems.

Cache compression. KIVI and KVQuant make KV-specific choices about axes and numerical treatment (Liu et al., 2024b; Hooper et al., 2024). Palu instead reduces the hidden dimension with low-rank projections (Chang et al., 2025). Token retention is another axis: H_2O, Scissorhands, and SnapKV select cache entries using attention information (Zhang et al., 2023; Liu et al., 2023; Li et al., 2024), while StreamingLLM preserves attention sinks (Xiao et al., 2024). Our experiment retains every token in each window. Its losses therefore cannot rank eviction, projection, and quantization under a common memory budget.

Architecture and deployment. Transformer attention (Vaswani et al., 2017), grouped-query heads (Ainslie et al., 2023), and rotary positional embeddings (Su et al., 2024) determine the tensors being quantized. They should be distinguished from an additional quantization rotation. FlashAttention and FlashAttention-2 optimize exact attention execution (Dao et al., 2022; Dao, 2024); PagedAttention manages cache allocation (Kwon et al., 2023). QServe jointly designs quantization and serving kernels (Lin et al., 2025), and CacheGen addresses cache compression and streaming (Liu et al., 2024a). These systems motivate measuring actual memory and throughput: our dequantized FP16 tensors and modeled bit-rates do not supply such measurements.

Evaluation sources. WikiText is a language-modeling corpus (Merity et al., 2017). We also evaluate SQuAD's Wikipedia contexts (Rajpurkar et al., 2016), without its questions or answer labels. The resulting perplexity is not a question-answering score, and the shared Wikipedia origin limits domain breadth.

10 Limitations

The follow-up uses two instruction-tuned models, two rotations, and four windows per model–dataset cell. Dataset reuse and the shared Wikipedia origin restrict independent confirmation and domain breadth. Rotation seeds measure transform sensitivity on the same text and do not increase the number of independent text samples. We do not report population confidence intervals from these seeds. The initial Mistral comparison uses a different quantizer grid and is not an additional cell of the equal-level factorial study. All cache tensors remain FP16 after dequantization; modeled bit-rates are not measured packed allocations, and shared-device timings are not serving benchmarks. Calibration and grouping do not identify the cause of the Qwen low-bit failure. Our claims are restricted to the observed losses and contrasts under the declared protocol.

11 Conclusion

A comparison of rotated uniform KV quantizers needs both a valid reference and an explicit representation budget. In the complete eight-condition finite-window audit, grouped configurations stay closer to the dense reference for both models, while the direction of the symmetry contrast reverses between granularities. Extra metadata and calibration timing remain confounded with these configurations. Conditional error envelopes alone cannot establish an actual performance separation or universal necessity; the additional development pilot isolates calibration timing and finds that per-write global calibration sharply reduces the observed gap on Llama. The longer-window Qwen comparison instead fails across all four-bit conditions, limiting the scope of that finding. The frozen follow-up reproduces the calibration finding on Llama and the four-bit failure on Qwen across two Wikipedia-derived datasets and two new rotations. Matched-total-budget controls, non-Wikipedia validation, and packed-kernel deployment remain untested.

Reproducibility Statement

The supplement contains the quantizers, finite-window evaluator, complete per-window measurements, and scripts for regenerating the result tables and figure. Configurations specify dataset splits, context preprocessing, rotation seeds, model identifiers, and calibration axes. Model weights and third-party datasets are obtained separately from their original providers.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: Kv-cache compression with low-rank projection. In *International Conference on Learning Representations (ICLR)*, 2025.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. Squeezellm: Dense-and-sparse quantization. In *International Conference on Machine Learning (ICML)*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 on-device llm compression and acceleration. In *MLSys*, 2024.

Yujun Lin, Haotian Tang, Shang Yang, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, and Song Han. Qserve: W4a8kv4 quantization and system co-design for efficient llm serving. In *MLSys*, 2025.

Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. Cachegen: Kv cache compression and streaming for fast large language model serving. In *ACM SIGCOMM*, 2024a.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations (ICLR)*, 2025.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. In *International Conference on Machine Learning (ICML)*, 2024b.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. URL https://aclanthology.org/D16-1264/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning (ICML)*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.