CURVATURE IS A BUDGET: COORDINATES BEFORE RESOLUTION IN SECOND-ORDER QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A weight rotation preserves the full-precision model but changes the Hessian that measures quantization error. Multiple public repositories contain code paths that pair transformed weights with untransformed curvature: we reproduce this error in ModelTC/LightCompress and identify the corresponding metric-collection pattern in spcl/QuaRot's export source. Users can unknowingly trigger this silent failure by omitting one setting in LightCompress: the run completes without a frame warning while Llama-3.1-8B perplexity rises from 6.525 to 7.456, worse than rotated rounding at 7.283. In a controlled four-bit panel, stale dense curvature makes GPTQ worse than round-to-nearest on five of seven checkpoints, while aligned block-32 curvature beats stale dense curvature on all seven. At three bits, the median stale-to-aligned dense-metric perplexity ratio reaches 25.31. We derive the exact output error of a complete GPTQ sweep under a mismatched metric and characterize when block projection commutes with coordinate transport, enabling exact transport within compatible block representations using $O(db)$ stored coefficients. Curvature must be aligned before it is budgeted.

1 INTRODUCTION

Many low-bit LLM quantization pipelines run in two stages: transform the weights, then round them with a second-order quantizer. QuaRot, SpinQuant and QuIP# rotate (Ashkboos et al., 2024; Liu et al., 2025; Tseng et al., 2024), SmoothQuant and AWQ rescale channels (Xiao et al., 2023; Lin et al., 2024), and GPTQ (Frantar et al., 2023) or one of its descendants does the rounding. The transform is exact for the model, $W_T x_T = Wx$, and this invariance is why it is treated as free. It is not free for the quantizer. GPTQ minimizes a quadratic form in the input second moment H, and under $W \mapsto WT$ that form becomes $H_T = T^{-1} H T^{-\top}$. When the transform and Hessian live in different stages, their handoff must preserve this coordinate relation; the public-code paths studied here omit that check.

When that handoff fails, the value of curvature inverts. With weights, calibration data and quantizer fixed, GPTQ driven by a stale dense Hessian is *worse than round-to-nearest* on five of seven four-bit checkpoints, and an aligned block-32 Hessian, $1/128$ of the dense coefficients at width 4096, beats it on all seven (Figure 1). Intermediate metric blocks can fail catastrophically, and reducing precision amplifies the penalty. A stale block-128 Hessian drives Qwen2.5-3B to perplexity 2.8×10^6 at four bits and 3.2×10^8 at three, against 8.43 and 11.06 when aligned. At three bits the median stale/aligned perplexity ratio for dense Hessians across the panel is 25.31.

The failure is silent at the quantizer interface: the supplied-metric loss does not reveal whether the Hessian matches the weights. Under a Hadamard transform, a stale dense sweep's supplied-metric loss is a median $3.2\times$ smaller than the error the layer incurs. We find this mismatch in multiple public repositories: a measured failure in ModelTC/LightCompress (ModelTC Contributors, 2025) and a corresponding collection pattern in spcl/QuaRot's export source (spcl/QuaRot Contributors, 2024). In LightCompress, omitting the default-off rotation setting raises Llama-3.1-8B perplexity by 14%, from 6.525 to 7.456, past rotated round-to-nearest at 7.283. Users receive no frame-related warning when that run completes (§5.6).

QuIP, QuIP# and the lattice analysis of GPTQ analyze rounding with the Hessian matched to the weights (Chee et al., 2023; Tseng et al., 2024; Chen et al., 2026). We evaluate the completed sweep

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

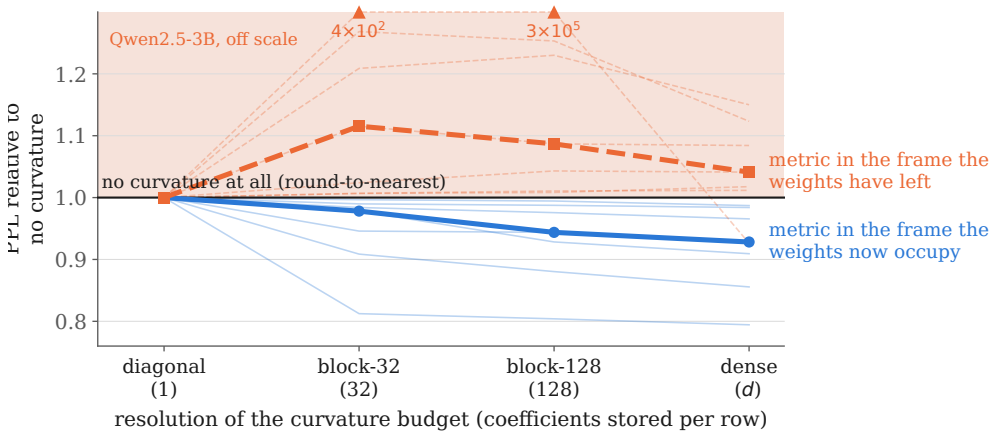

Figure 1: **More curvature makes quantization worse when it is in the wrong frame.** Seven checkpoints, symmetric W4A16 with group size 128 and a block-128 Hadamard transform. Faint lines show individual checkpoints; bold lines show medians. The horizontal axis is nominal metric resolution, from diagonal to dense. Perplexity is divided by round-to-nearest in the same rotated basis, so values above one are worse than using no curvature. Both diagonal endpoints equal that control because a diagonal metric makes no cross-column compensation. Aligned median: $1.000 \to 0.978 \to 0.944 \to 0.928$; stale median: $1.000 \to 1.115 \to 1.087 \to 1.041$. Table 5 gives every checkpoint's values.

when the supplied metric differs from the layer's metric, and characterize which compressed Hessians retain the information needed after a transform. To our knowledge, this is the first systematic mathematical and empirical study of coordinate mismatch together with curvature resolution in second-order LLM quantization.

Contributions.

1. **An exact account of GPTQ under the wrong metric** (§3). For any positive-definite supplied metric M, a completed sweep has supplied-metric loss $\|Z\|_F^2$ but incurs $\mathrm{tr}(ZK_MZ^\top)$, where Z are its realized rounding residuals and K_M is the supplied-to-physical distortion. The supplied-metric loss alone does not identify the mismatch. Without damping, a volume-normalized stale Hessian has $\det K_H = 1$ and $\mathrm{tr}\, K_H/d \geq 1$: any nontrivial distortion changes relative directions rather than applying a uniform scale.
2. **A complete characterization of which compressed Hessians can follow a transform** (§4). Block truncation commutes with transport if and only if the transform preserves the block subspace and its complement. A block-b rotation admits exact transport within a compatible $O(db)$ block representation: a block-128 cache uses 9.7 MB instead of 1.44 GB at width 18,944. For the mixing block rotations studied here, diagonal truncation and transport do not commute.
3. **The inversion, measured on fourteen checkpoints** (§5). A controlled seven-checkpoint panel with three calibration draws; an independent replication in which stale curvature is worse on all nine checkpoints; a median $25.31\times$ penalty at three bits; and a silent failure reproduced in unmodified LightCompress, alongside the QuaRot source finding. Under a Hadamard transform at four bits, aligned curvature wins on every one of the fourteen checkpoints.
4. **Three rules for compression pipelines** (§6). Tag every Hessian with its coordinate frame and reject mismatches before factorization; transport a truncated Hessian only when our criterion holds; and update curvature inside any curvature-scored rotation search.

2 Two Metrics for One Layer

A transform creates two metrics for one layer: the one GPTQ is given and the one that measures the deployed error. Every result in this paper follows from keeping them apart.

For a linear layer $y = Wx$ with input second moment $H = \mathbb{E}[xx^\top]$, quantizing W to A gives expected output error

$$\mathcal{L}_H(A; W) = \mathbb{E}\|(A - W)x\|_2^2 = \mathrm{tr}\big((A - W)H(A - W)^\top\big). \tag{1}$$

A full $d \times d$ matrix costs $O(d^2)$ storage. For example, one fp32 matrix at width $d = 18{,}944$ occupies 1.44 GB. Diagonal and block-diagonal matrices store less, at the cost of discarding correlations.

Let $W \in \mathbb{R}^{m\times d}$ and $T \in \mathbb{R}^{d\times d}$ be invertible. The pipeline transforms weights and inputs as

$$W_T = WT, \qquad x_T = T^{-1}x, \qquad W_T x_T = Wx. \tag{2}$$

The compressor returns weights Q in the new coordinates and deploys QT^{-1}. With $E = Q - W_T$, the deployed layer incurs

$$\mathbb{E}\|(QT^{-1} - W)x\|_2^2 = \mathrm{tr}(EH_TE^\top), \tag{3}$$

where the curvature in the working coordinates is

$$H_T = T^{-1}HT^{-\top}. \tag{4}$$

These identities hold for rotations, diagonal scales and invertible compositions. The full-precision function is unchanged; the quantization grid and the metric used to choose rounded weights depend on the coordinates.

Definition 2.1 (supplied and physical metric)**.** The *physical metric* is H_T, which measures the deployed layer's output error. The *supplied metric* $M \succ 0$ is the matrix GPTQ factors, including regularization. A supplied metric is *aligned* if it approximates H_T, and *stale* if it approximates H while GPTQ rounds W_T. Tables label the two arms "working coords" and "pre-transform coords".

There are two correct ways to obtain H_T: collect activations after applying T^{-1}, or transport a sufficiently detailed cache of H with Equation (4). The equality is exact in arithmetic:

Lemma 2.2 (two constructions of the aligned metric)**.** $\mathbb{E}[(T^{-1}x)(T^{-1}x)^\top] = T^{-1}HT^{-\top}$.

On nine checkpoints, direct collection and dense-cache transport differ by at most 2.1×10^{-5} in relative Frobenius norm on any layer. The resulting median absolute perplexity difference is 0.003, with a maximum of 0.10423 (Lemma 2.2, Table 11). We use that maximum, 0.10423 perplexity, as the tie band throughout: it is the largest difference between two exact constructions of the same metric. Transporting a cache *after* truncating it is a different operation, exact within that format only when the transform and block partition satisfy Proposition 4.1. Passing the original H to GPTQ without transport is the stale arm studied below.

3 What GPTQ Optimizes When Handed the Wrong Metric

GPTQ compensates rounding errors using the supplied metric; the layer incurs error under its own metric. GPTQ rounds columns in order. After rounding column j, it uses the factored metric to change later columns. Those changes minimize loss according to the matrix supplied to GPTQ; their effect on the layer is measured by H_T. For any $G \succ 0$ write $\mathcal{L}_G(Q) = \mathrm{tr}((W_T - Q)G(W_T - Q)^\top)$. Let $C_{:j}$ be the working column just before it is rounded to $Q_{:j}$ and define $Z_{:j} = (C_{:j} - Q_{:j})/U_{jj}$, where U is the upper-triangular factor of the supplied metric defined below.

Proposition 3.1 (exact residual, supplied-metric and incurred loss)**.** *After a complete fixed-order GPTQ sweep with supplied metric M and upper-triangular U satisfying $M^{-1} = U^\top U$,*

$$W_T - Q = ZU, \qquad \underbrace{\mathcal{L}_{H_T}(Q) = \mathrm{tr}(ZK_MZ^\top)}_{\textit{incurred}}, \qquad \underbrace{\mathcal{L}_M(Q) = \|Z\|_F^2}_{\textit{supplied}}, \qquad K_M := UH_TU^\top, \tag{5}$$

and for $Z \neq 0$, $\lambda_{\min}(K_M) \le \mathcal{L}_{H_T}(Q)/\mathcal{L}_M(Q) \le \lambda_{\max}(K_M)$. The eigenvalues of K_M are the generalized eigenvalues of (H_T, M), and $K_M = I$ when $M = H_T$ undamped.

Our experiment tables call the supplied-metric objective the reported loss. The sweep measures loss as $\|Z\|_F^2$; the layer incurs $\mathrm{tr}(ZK_MZ^\top)$. With an aligned, undamped metric $K_M = I$ and the two coincide. Under rotation with stale curvature, the supplied-metric loss understates incurred layer error by a median factor above three in the controlled panel (§5.3). The interval bounds the ratio but does not fix it, because changing M also changes the rounding trajectory Z; in every measured case it straddles one, so the one-sided gap we observe is an empirical fact rather than a consequence of the bound (Appendix E). Appendix A proves the identity.

Corollary 3.2 (spectrum of a volume-normalized stale metric)**.** *Assume $H \succ 0$, no damping, and fix $|\det T| = 1$, covering orthogonal transforms and geometric-mean-normalized diagonal ones. For $M = H$, $\det K_H = 1$ and $S_0 := \frac{1}{d}\operatorname{tr}(K_H) = \frac{1}{d}\operatorname{tr}(H_T H^{-1}) \geq 1$, with equality iff $THT^\top = H$; for orthogonal T, $TH = HT$.*

An undamped stale Hessian is therefore never uniformly too large or too small: the eigenvalues of K_H have geometric mean one and arithmetic mean at least one, with equality only when the transform preserves H. A transform that spreads weight outliers can also change the input covariance. On the measured layers the median distortion $\operatorname{tr} K_M/d$ of a dense stale metric is 6.69, against 0.88 when aligned (Table 7).

4 Which Curvature Caches Can Follow a Transform

A dense Hessian is expensive. A block cache reduces storage by discarding cross-block correlations; its compatibility with a transform determines whether the retained information can be transported within that format.

Let B partition $\{1, \ldots, d\}$ into consecutive blocks, $\mathcal{P}_B \subset \operatorname{Sym}(d)$ the matrices supported on its diagonal blocks, Π_B the Frobenius projection onto $\mathcal{P}_B$, and $\mathcal{C}_T(H) = T^{-1}HT^{-\top}$. We count full block-array coefficients: $\Pi_B(H)$ represents $\sum_i |B_i|^2$ coefficients instead of d^2.

Proposition 4.1 (metric–transform compatibility)**.** $\Pi_B \circ \mathcal{C}_T = \mathcal{C}_T \circ \Pi_B$ *on* $\operatorname{Sym}(d)$ *if and only if $\mathcal{C}_T$ preserves both $\mathcal{P}_B$ and $\mathcal{P}_B^\perp$. It holds whenever T is block-diagonal with respect to B, hence for a uniform block-b transform whenever metric partition b_M is a multiple of b. Conversely, if T is block-b with every column of T^{-1} fully supported on its block (e.g., block Hadamard or Haar-random orthogonal), it fails whenever b_M is not a multiple of b.*

A block-128 Hadamard can be applied inside block-128 or block-256 metric partitions without moving information across their boundaries. Diagonal truncation does not commute with that rotation: only diagonal-times-permutation transforms preserve the entire class of diagonal matrices. The distinction is whether projection onto the stored blocks commutes with the coordinate transform, not whether the transform itself is inexpensive. Appendix A gives the proof.

Corollary 4.2 (what a curvature budget can be moved through)**.** *For every $H \in \operatorname{Sym}(d)$, project-after-transport and transport-after-project agree precisely when Π_B and $\mathcal{C}_T$ commute (Proposition 4.1). Outside that class, an already truncated cache generally lacks information to follow the transform. For uniform consecutive partitions and fully supported block transforms, a cached diagonal metric is not closed under nontrivial block rotations, a block-b_M metric is closed under block transforms whose size divides b_M, and the cheapest exact pairing for a block-b transform has nominal size $\Theta(db)$ rather than $\Theta(d^2)$.*

A compatible block-b transform therefore needs only $O(db)$ stored coefficients under the full-block convention: at $d = 18{,}944$, a block-128 cache occupies 9.7 MB in fp32 against 1.44 GB dense. Outside this compatibility class, transporting the truncated cache cannot reproduce project-after-transport for every input covariance. On all seven panel models the discrepancy is exactly zero for partitions that are multiples of 128 and of order one for every finer partition (Table 2).

Corollary 4.3 (within a matched budget the rotation is irrelevant)**.** *Let T be orthogonal and block-diagonal with respect to B. Then both the mass the budget discards, $\|H_T - \Pi_B(H_T)\|_F$, and the volume it keeps, $\det \Pi_B(H_T)$, are independent of T.*

For orthogonal block transforms, $(H_T)_{ij} = T_i^\top H_{ij} T_j$. The stored and discarded block norms are unchanged, although rounding in the rotated grid can still change model quality. The partition fixes what the curvature representation retains; the rotation changes which weights are rounded.

5 Experiments

Within each stale-versus-aligned comparison, transformed weights, calibration data and quantizer settings are fixed; the supplied curvature changes. The remaining controls vary the stated factor, including metric resolution, bit width, partition and calibration draw.

Table 1: **Matched W4 perplexity with and without transported curvature.** WikiText-2 perplexity, symmetric W4A16, group 128. RTN is round-to-nearest in the same basis: the no-curvature control. Under the Hadamard the same transformed weights are quantized with a dense metric in the coordinates the weights have left ("earlier") or the ones they occupy ("working"). Bold beats rotated RTN; †† marks a margin inside the 0.10423 PPL reproducibility band of §5. The last two columns are the earlier/working ratio at a dense metric, at four bits and at three; the three-bit ladder is in Appendix J. Random-orthogonal transforms behave the same and are in Appendix C.

		no transform		Hadamard, W4			ratio	
Model	BF16	RTN	GPTQ	RTN	earlier	working	W4	W3
Qwen2.5-0.5B	13.119	18.265	**14.610**	16.225	18.659	**13.882**	1.34	111.128
Qwen2.5-1.5B	9.687	12.162	**10.947**	11.097	12.466	**10.091**	1.24	25.310
Qwen2.5-3B	8.051	11.508	**8.520**	10.485	**9.711**	**8.329**	1.17	1.3×10^3
Qwen2.5-7B	7.769	8.216	8.231††	8.330	9.026	**8.044**	1.12	2.291
Llama-3.1-8B	6.225	7.000	10.429	7.105	7.398	**6.595**	1.12	29.236
Phi-4 (14B)	6.453	6.700	**6.593**	6.671	6.789	6.584††	1.03	1.515
OLMoE-1B-7B†	7.461	7.716	**7.581**	7.687	7.776††	**7.563**	1.03	1.363

† Attention projections only: 64 attention linears are quantized and the fused expert tensors remain BF16, so its perplexity is dominated by unquantized parameters and its margins are correspondingly small. Read it as an attention-only case study; it contributes 64 of the 1,259 pooled linears.

Matched protocol. We cache each layer's activation second moment once, transform its weights, and run the same GPTQ sweep with either aligned or stale curvature. Both arms quantize the same transformed weights and use the same inverse map for evaluation, so they differ only in the matrix supplied to GPTQ. The main panel uses symmetric W4A16, group size 128, and 128 WikiText-2 (Merity et al., 2017) calibration windows of 2,048 tokens; perplexity uses 128 non-overlapping test windows. Unless stated otherwise, every sweep factors the supplied metric with relative ridge 10^{-2}, and Proposition 3.1 applies to that damped matrix (ridge controls in Table 9). Calibration seed 42 supplies the displayed values; seeds 43 and 44 test sensitivity (Appendix L). We test block-128 Hadamard and random orthogonal transforms, diagonal scaling, and scale-then-Hadamard. The seven checkpoints are four Qwen2.5 sizes, Llama-3.1-8B, Phi-4 and OLMoE-1B-7B; on OLMoE we quantize the 64 attention projections and leave expert tensors in BF16. The sweep uses one-pass cached layer statistics, symmetric grids and no activation reordering. We write Φ for stale/aligned perplexity. The checkpoint, not a layer or calibration draw, is the unit for model-level comparisons, and differences below 0.10423 perplexity are ties.

5.1 A Stale Dense Hessian Loses to Rounding

On five of seven four-bit checkpoints, GPTQ with a stale dense Hessian produces a worse model than round-to-nearest (RTN), which uses no curvature at all (Table 1). Under a Hadamard transform, aligned dense GPTQ beats RTN on 6 of 7 checkpoints; stale dense GPTQ loses to RTN on 5 of 7. Under random orthogonal transforms those counts are 7 of 7 and 5 of 7 (Table 4). RTN makes no compensation; stale GPTQ makes confident compensation toward the wrong objective. Across three calibration draws, $\Phi > 1$ in 27 of 28 transform/checkpoint cells (Table 19). The Llama identity-basis dense run is an unrelated outlier, finite but poor at 10.429 perplexity against 7.000 for RTN; it does not enter the frame comparison, which always uses the same transformed weights.

5.2 More Curvature, Worse Model

Curvature resolution pays off only in the right frame. Figure 1 and Table 5 sweep four metric sizes under the block-128 Hadamard: diagonal, block-32, block-128 and dense. An aligned block-32 metric beats the stale dense metric on 7 of 7 checkpoints, and aligned block-128 does so on 7 of 7. On Qwen2.5-1.5B, for example, aligned block-32 reaches 10.855 perplexity against 12.466 for stale dense. The aligned block-32 metric is formed by projecting curvature already transported to the working coordinates. These comparisons count nominal stored coefficients; collection time and peak device memory depend on the implementation.

Aligned ladders improve at every step on 7 of 7 checkpoints. No stale ladder does: four peak at an intermediate budget, one ties there, and on Phi-4 and OLMoE every additional coefficient makes the model worse. The diagonal endpoints coincide with RTN because a diagonal metric makes no cross-column update (Figure 3 resolves the ladders per checkpoint).

The stale Qwen2.5-3B ladder collapses: 10.485 at the diagonal anchor, 4.3×10^3 at block-32, 2.8×10^6 at block-128, then 9.711 at dense, while the aligned ladder falls from 10.485 to 8.329. The collapse recurs at all three calibration draws (Table 8). Its worst layer has physical reconstruction error $5.7{\times}10^4$ at block-128, and the top 1% of layers carry 48% of that arm's total error. In the layer diagnostics, normalized update size falls as resolution grows while physical error rises (Table 7). The location of the collapse depends on the group and transform widths, whereas the aligned ladder stays between 8.263 and 8.597 across the whole grid (Table 17).

5.3 GPTQ's Own Loss Hides the Error

The supplied-metric loss hides the coordinate mismatch. The reported and incurred losses of Proposition 3.1 separate on every real layer we measure under rotation. For the stale dense metric, the median incurred/reported ratio over the panel's 1259 linears is $3.21\times$ under Hadamard and $3.49\times$ under random orthogonal transforms; the composed scale-then-Hadamard class has $2.11\times$ and diagonal scaling $1.02\times$ (Table 3). Under rotation, the median layer incurs more than three times the loss measured in the supplied metric. Two further runs measure the same ratio: the budget sweep (median 3.094 over 1,260 layer records, Table 6) and the cached-layer ladder (3.188 at ridge 10^{-2}, Table 9). In both, the ratio rises monotonically with stale metric resolution, while the aligned ratio stays at or below one at every budget and ridge.

5.4 Three Bits Amplifies the Penalty

At three bits the stale penalty becomes a multiple, not a percentage. Removing one bit raises the median dense-metric Φ from 1.12 at W4 to 25.31 at W3, and raises it on 7 of 7 checkpoints (Table 1). At W3, stale block-128 curvature drives Qwen2.5-3B to $3.2{\times}10^8$ perplexity, and its stale dense value is 1.3×10^4. Aligned perplexity improves from diagonal to dense on every checkpoint, with one small block-32 to block-128 reversal on Llama (Table 16). No arm at three bits took a round-to-nearest fallback, so these values are the sweep's own output. Coarser grids leave larger rounding residuals Z, and Proposition 3.1 shows that the stale metric weights each of them by K_M.

5.5 Beyond the Panel: Nine Checkpoints and an 80B Model

The W4 penalty also appears in an independent replication that quantizes the same rotated weights with stale and with aligned curvature on nine checkpoints, seven outside the panel (Gemma-3-12B, Gemma-4-31B, GLM-4-9B, Llama-3.1-8B, Ministral-8B, Mistral-7B, Phi-4, Qwen3-8B and Qwen3.5-9B). Stale curvature raises perplexity on 9 of 9, with Φ from 1.034 to 1.430; Gemma-4-31B moves from 6.51 to 8.01. The same runs construct the aligned metric two exact ways, and the stale penalty exceeds the difference between them by $68\times$ to $1541\times$, median $171\times$ (Table 11).

On Qwen3-Next-80B-A3B-Instruct, two-scale NVFP4 uses E2M1 values, E4M3 scales for blocks of 16, and one FP32 scale per weight matrix. Both metric choices remain finite across five calibration draws; aligned curvature lowers WikiText-2 perplexity by 0.0045–0.0276 (Table 20). These small differences fall inside the panel's operational tie band. At seed 42, unrotated RTN reaches 5.840115, against 5.846894 for aligned GPTQ. Appendix M gives the expert metric, adaptive factorization and separate historical arithmetic diagnostic.

5.6 The Stage Boundary in Public Code

In public code, the failure is one configuration line away.

ModelTC/LightCompress: measured failure. The toolkit's documented QuaRot-to-GPTQ workflow passes rotation state between two stages (ModelTC Contributors, 2025). We fix the transformed Llama-3.1-8B checkpoint, calibration data and quantizer settings (commit `86f564d`, asymmetric W4A16, group 128, 141 evaluation windows). Omitting `online_rotate` from the GPTQ stage raises perplexity from 6.525 to 7.456, which is 2.4% worse than rotated RTN (7.283). The setting

defaults to off, and the run completes without a frame-related warning. The official combined example enables the setting; the defect is that an inconsistent handoff is accepted rather than rejected.

spcl/QuaRot: source finding. Its export source fuses inverse rotations into weights but proposes collecting GPTQ inputs on the bare linear, yielding an incompatible weight–metric pair (spcl/QuaRot Contributors, 2024). The separate fake-quant path collects after online rotation and is aligned. Appendix P documents the export path's execution boundary and pinned source lines, and the NVIDIA Model-Optimizer audit.

6 Consequences for Pipeline Design

The results translate into three rules for any pipeline that composes a transform with a second-order quantizer. First, a curvature cache must carry its coordinate frame, and GPTQ should reject a mismatch before factorization; the LightCompress failure occurs precisely at such a stage boundary. Second, exact transport within a retained block format for every covariance requires Proposition 4.1; otherwise collect in the new frame, or transport a richer cache before truncating it. Third, a rotation search scored with curvature must update that curvature as the rotation changes. In a local Cayley-parameterized search on 16 real layer pairs with three seeds each, recomputing the metric beats freezing it on 15 pairs and 44 of 48 runs (Appendix O); this measures that local objective, not the full training procedure of any published rotation method.

Retaining more covariance mass does not by itself decide GPTQ's output. Spectral and greedy channel groupings raise the captured squared-Frobenius mass from 18% to 51% and 68% at the same nominal size, yet physical layer error changes by less than 1.5% (Table 22), because the supplied metric also changes the rounding trajectory in Equation (5). Nor is a stale frame equivalent to a stale calibration draw. On two Qwen2.5 checkpoints, a fresh draw in the correct frame raises incurred layer error by 6.5% and 8.8%, whereas the original draw in the stale frame raises it $5.5\times$ and $7.7\times$ (Table 18).

Diagonal scaling marks the boundary of the effect. It produces the smallest mismatch in the panel, yet the aligned arm is still better by more than the 0.10423 tie band on four of its seven checkpoints; the three below-band cells of the panel's 28 transform/checkpoint cells are all diagonal. Across all 28 cells, 25 clear the band, and the diagonal effect persists across SmoothQuant exponents (Appendix H). The frame effect also survives activation reordering, and it grows monotonically with the number of quantized blocks (Appendix G). Pruning follows a different selection rule and is reported separately (Appendix I).

7 Related Work

Optimal Brain Damage (LeCun et al., 1989), Optimal Brain Surgeon (Hassibi & Stork, 1993), Optimal Brain Compression (Frantar et al., 2022), GPTQ (Frantar et al., 2023) and SparseGPT (Frantar & Alistarh, 2023) use curvature to reduce local compression error. Executed back-to-front, GPTQ's greedy sweep is equivalent to Babai's nearest-plane algorithm on a lattice defined by the layer Hessian (Chen et al., 2026). Our factorization distinguishes the metric that drives that sweep from the metric that measures its deployed output. ADMM-Q (Lucas et al., 2026) and KronQ (Lee et al., 2026) use other optimization or curvature structures; the coordinate frame still determines which error their matrices describe.

QuIP and QuIP# explicitly transform weights and Hessians together and analyze adaptive rounding under the transformed quadratic objective (Chee et al., 2023; Tseng et al., 2024). QuIP shows that its LDLQ procedure is equivalent to GPTQ and factorizes the rounding residual under the matched Hessian; Proposition 3.1 evaluates the same factorization when the supplied metric differs from the layer's. HeRo-Q also supplies transformed curvature while optimizing its conditioning (Zhang et al., 2026). These constructions establish the setting with matched coordinates. We study a sweep driven by a different metric from the one measuring output error, and characterize when a compressed block cache can be transported exactly. QuIP#'s block rounding groups quantization decisions; our block criterion concerns the correlations retained in the curvature representation.

QuaRot, SpinQuant and OptRot use transforms to improve quantization (Ashkboos et al., 2024; Liu et al., 2025; Gadhikar et al., 2025); SmoothQuant and AWQ use channel scaling (Xiao et al., 2023;

Lin et al., 2024). Whenever such a transform is followed by GPTQ, as in the QuaRot and SpinQuant weight pipelines, the quantizer must receive H_T, and our results specify what happens when it does not. Our controlled comparisons hold the transformed weights fixed and change only the supplied curvature, which separates the cost of metric mismatch from the benefit of the transform itself.

8 Conclusion

A transform that leaves a network unchanged still changes what its quantizer must minimize. We derive the exact loss of a GPTQ sweep under a mismatched metric and characterize when block truncation commutes with transport. The measured consequence is an inversion of the value of curvature: stale dense GPTQ loses to rounding on five of seven checkpoints and to aligned block-32 GPTQ on all seven, while the median three-bit penalty reaches $25.31\times$. The mismatch also appears in public code, and an omitted LightCompress setting causes a silent quality loss. Its supplied-metric objective gives the quantizer no check that the coordinate handoff was correct. A valid handoff preserves the inference format: identify the Hessian's frame, align it with the weights, and retain a compatible block representation. Curvature must be aligned before it is budgeted.

Limitations. The theory concerns layer output error; perplexity and numerical stability are measured separately, and we do not report downstream task accuracy for the controlled comparison. Storage counts are nominal coefficients, not peak memory, and collecting or transporting curvature costs time and working memory. The controlled panel calibrates and evaluates on WikiText-2. The two-scale NVFP4 study reports small perplexity differences across five calibration draws; the historical arithmetic failures are analyzed separately. Efficient updates of compact metrics under learned rotations remain open.

AI use statement

Generative AI tools were used to assist with manuscript restructuring, language editing, and literature discovery. The authors reviewed the AI-assisted text, checked the mathematical arguments and source attributions, and verified numerical claims against the cited artifacts. The authors take responsibility for the final content of the work, including all text, claims, and artifacts produced with AI assistance.

Ethics statement

This work uses public model checkpoints and benchmark text to study a model-compression workflow. It uses no human-subject data. The named LightCompress experiment tests a reachable configuration mismatch; the documented combined example enables the correct setting, and we describe the issue as a missing validation rather than a defect in that recipe.

Reproducibility statement

Appendix Q specifies the calibration and evaluation protocol, arm definitions, transform conventions, and artifact map. The controlled-panel, budget, mechanism, and partition tables are generated from per-run JSON records by a build script, with a cell manifest linking values to their sources. The commutation criterion of Proposition 4.1 is checked numerically on every panel model (Appendix B). For the third-party workflow (LightCompress commit `86f564ddb1d6548b228c67a10509a4ed7264345c`), the BF16 and rotated round-to-nearest values are re-parsed from that toolchain's own evaluation output with file, line, and matched text recorded; the two GPTQ arms are measured by our evaluator on the checkpoints it wrote, and agree with the toolchain's own evaluator to 0.019% where both were run. Complete proofs are in Appendix A. The anonymous code supplement includes the raw records, generators, recorded commands and a README mapping each result to its source.

References

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*, 2024.

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*, 2023.

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The geometry of LLM quantization: GPTQ as Babai's nearest plane algorithm. In *International Conference on Learning Representations*, 2026.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, 2023.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, 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*, 2023.

Advait Gadhikar, Riccardo Grazzi, and James Hensman. OptRot: Mitigating weight outliers via data-free rotations for post-training quantization. *arXiv preprint arXiv:2512.24124*, 2025.

Babak Hassibi and David G Stork. Second order derivatives for network pruning: Optimal Brain Surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pp. 164–171, 1993.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pp. 598–605, 1989.

Donghyun Lee, Yuhang Li, Ruokai Yin, and Priyadarshini Panda. KronQ: LLM quantization via Kronecker-factored hessian. *arXiv preprint arXiv:2607.07964*, 2026.

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 *Proceedings of Machine Learning and Systems*, 2024.

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*, 2025.

Ryan Lucas, Mehdi Makni, Xiang Meng, Adam Deng, and Rahul Mazumder. ADMM-Q: An improved hessian-based weight quantizer for post-training quantization of large language models. *arXiv preprint arXiv:2605.11222*, 2026.

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

ModelTC Contributors. LightCompress: A versatile toolkit for compressing large models. `https://github.com/ModelTC/LightCompress/tree/86f564ddb1d6548b228c67a10509a4ed7264345c`, 2025. Commit 86f564ddb1d6548b228c67a10509a4ed7264345c.

NVIDIA Corporation. NVIDIA Model-Optimizer: A state-of-the-art library for model compression. `https://github.com/NVIDIA/Model-Optimizer`, 2025. Modular PTQ, FP4, and rotation-based calibration toolchain.

spcl/QuaRot Contributors. QuaRot source code, pinned export and fake-quant paths. `https://github.com/spcl/QuaRot/tree/5008669b08c1f11f9b64d52d16fddd47ca754c5a`, 2024. Source audit at commit 5008669b08c1f11f9b64d52d16fddd47ca754c5a.

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*, 2024.

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*, 2023.

Jinhao Zhang, Yunquan Zhang, Zicheng Yan, Boyang Zhang, Jun Sun, and Daning Cheng. HeRo-Q: A general framework for stable low bit quantization via hessian conditioning. *arXiv preprint arXiv:2601.21626*, 2026.

A PROOFS

Proof of Proposition 3.1. Let U be the upper-triangular Cholesky factor of M^{-1} used by GPTQ. GPTQ initializes its working matrix as $C = W_T$. Immediately before column j is rounded, the accumulated updates from earlier columns give $C_{:j} = W_{T,:j} - \sum_{i<j} Z_{:i}U_{ij}$. By definition of Z, $C_{:j} - Q_{:j} = Z_{:j}U_{jj}$, so $W_{T,:j} - Q_{:j} = \sum_{i\le j} Z_{:i}U_{ij}$ for every column, hence $W_T - Q = ZU$ and $\mathcal{L}_{H_T}(Q) = \mathrm{tr}(ZUH_TU^\top Z^\top) = \mathrm{tr}(ZK_MZ^\top)$. Since $M^{-1} = U^\top U$ we have $M = U^{-1}U^{-\top}$ and $UMU^\top = I$, so $\mathcal{L}_M(Q) = \mathrm{tr}(ZUMU^\top Z^\top) = \|Z\|_F^2$. Applying the Rayleigh bounds for K_M to each row of Z and summing gives the interval. When $M = H_T$, $K_M = UH_TU^\top = I$. □

Proof of Corollary 3.2. Because $H_T = T^{-1}HT^{-\top}$, $\det H_T = \det H/(\det T)^2$. For $|\det T| = 1$ and $M = H$, $\det K_H = \det(U)^2 \det(H_T) = \det(H^{-1})\det(H_T) = 1$, and cyclicity gives $\mathrm{tr}\, K_H = \mathrm{tr}(H_TH^{-1})$. The arithmetic–geometric mean inequality on the positive eigenvalues of K_H yields $S_0 \ge 1$, with equality exactly when all eigenvalues are 1, i.e. $K_H = I$ and $H_T = H$. Multiplying $T^{-1}HT^{-\top} = H$ on the left by T and the right by $T^\top$ gives $THT^\top = H$, which for orthogonal T is $TH = HT$. □

Proof of Proposition 4.1. $\mathrm{Sym}(d) = \mathcal{P}_B \oplus \mathcal{P}_B^\perp$ and Π_B is the projection along that decomposition, so Π_B commutes with a linear map $\mathcal{C}$ iff $\mathcal{C}$ preserves both summands: for $H \in \mathcal{P}_B$, $\Pi_B\mathcal{C}H = \mathcal{C}\Pi_BH = \mathcal{C}H$ forces $\mathcal{C}H \in \mathcal{P}_B$; for $H \in \mathcal{P}_B^\perp$, $\Pi_B\mathcal{C}H = \mathcal{C}\Pi_BH = 0$ forces $\mathcal{C}H \in \mathcal{P}_B^\perp$; the converse is immediate. If T is block-diagonal with respect to B with blocks T_i, then $(\mathcal{C}_TH)_{ij} = T_i^{-1}H_{ij}T_j^{-\top}$ blockwise, so a matrix supported on the diagonal blocks stays there and one with zero diagonal blocks keeps them zero. A T block-diagonal with uniform block b is also block-diagonal with respect to any coarser uniform partition of size a multiple of b, and with respect to the trivial partition when $b_M \ge d$. For the converse, let T be block-diagonal with respect to a uniform partition B' of size b that does not refine B, and suppose every column of T^{-1} is fully supported on its B' block. Some B' block then meets two B blocks, giving indices $i \in B_p$ and $j \in B_q$ with $p \ne q$ in one B' block. Applying $\mathcal{C}_T$ to the B-block-diagonal matrix $e_ie_i^\top$ gives $(T^{-1})_{:,i}(T^{-1})_{:,i}^\top$, whose (i,j) entry is $(T^{-1})_{ii}(T^{-1})_{ji} \ne 0$ by the support assumption; this lies outside $\mathcal{P}_B$, so $\mathcal{P}_B$ is not preserved. The support assumption is what rules out the counterexamples that make the general converse false—a block permutation matrix has one nonzero per column and does commute—and it holds for the block Hadamard and block Haar transforms used here. Finally, when $\mathcal{P}_B$ is the diagonal matrices, preserving $\mathcal{P}_B$ requires $T^{-1}e_ie_i^\top T^{-\top} = (T^{-1})_{:,i}(T^{-1})_{:,i}^\top$ to be diagonal for every i, i.e. every *column* of T^{-1} to have a single nonzero, so T is diagonal times a permutation. □

B COMPATIBILITY CHECK

For each panel model we take the first quantized linear with $d \ge 512$, form a block-128 Hadamard, and compare $\Pi_B(\mathcal{C}_T(H))$ against $\mathcal{C}_T(\Pi_B(H))$ at several metric partitions. Agreement in the predicted cases is exact because both sides reduce to the same masking; the informative entries are the partitions finer than the transform, where no such reduction exists and the discrepancy is $O(1)$ relative. This is a check that the implementation matches Proposition 4.1, not evidence for it.

C FULL CONTROLLED PANEL

Within each transform, the stale and aligned arms share one cached activation second moment per layer and differ in the metric supplied to GPTQ. The budget ladder is a separate run of this protocol.

Table 2: Relative Frobenius discrepancy between truncating after the congruence and truncating before it, $\|\Pi_B(\mathcal{C}_T H) - \mathcal{C}_T(\Pi_B H)\|/\|\Pi_B(\mathcal{C}_T H)\|$, for a block-128 Hadamard T and several metric partitions B. Proposition 4.1 predicts agreement whenever B is a multiple of 128 and, within this block-structured family, failure otherwise; the bold entries are that prediction. The point is not the magnitude of the zeros—the operations are literally the same masking—but that no partition finer than the transform admits it.

Model / metric partition B	1	32	64	128	256	dense
Qwen2.5-0.5B	4.13	1.53	0.99	**0**	**0**	**0**
Qwen2.5-1.5B	8.34	1.70	1.00	**0**	**0**	**0**
Qwen2.5-3B	2.25	1.27	0.95	**0**	**0**	**0**
Qwen2.5-7B	3.20	1.47	0.94	**0**	**0**	**0**
Llama-3.1-8B	7.47	1.68	1.00	**0**	**0**	**0**
Phi-4 (14B)	2.86	1.37	0.95	**0**	**0**	**0**
OLMoE-1B-7B†	7.98	1.70	1.00	**0**	**0**	**0**

Table 3: **What the sweep reports against what the layer incurs** when curvature is held in the pre-transform coordinates. Each entry is the median over that checkpoint's quantized linears of incurred/reported loss (Proposition 3.1); > 1 means the sweep's own objective understates the error it caused. Checkpoints, not layers, are the unit of replication: layers within a checkpoint share one calibration draw. The final rows pool all layers to report the fraction on which the sign holds.

Checkpoint	linears	Hadamard	random orth.	diagonal	composed
Qwen2.5-0.5B	168	4.126	4.189	1.055	2.226
Qwen2.5-1.5B	196	3.752	3.992	1.021	2.004
Qwen2.5-3B	252	3.577	3.907	1.011	1.963
Qwen2.5-7B	195	3.271	3.512	0.976	2.007
Llama-3.1-8B	224	2.867	2.995	1.030	2.250
Phi-4 (14B)	160	2.551	2.731	1.019	2.420
OLMoE-1B-7B†	64	2.307	2.291	0.994	1.724
all layers pooled, median	1259	**3.210**	**3.486**	**1.017**	**2.112**
fraction > 1	1259	1.0000	1.0000	0.5727	0.9865
maximum	1259	233	218	20	33

On Qwen2.5-7B one fallback linear changes the dense perplexities by less than 0.006 between runs: 9.026 versus 9.031 with stale curvature and 8.044 versus 8.043 with aligned curvature. It remains in model perplexity, but is excluded from the 1,259 paired linears; the budget diagnostic includes 1,260 layer records per cell. OLMoE expert tensors remain BF16.

At seed 42, stale/aligned perplexity exceeds one in all 28 matched transform/checkpoint cells; 25 exceed the 0.10423 perplexity band. Across all three calibration draws the ratio exceeds one in 27 of 28 cells. The smallest ratio, 1.0016, belongs to a diagonal-scale cell within the tie band. The Qwen2.5-7B diagonal cell also falls inside the band: its unrounded difference is 0.10415, although the rounded entries in Table 4 differ by 0.105. All 5036 paired layer comparisons have incurred-error ratio above one, with minimum 1.008 and median 2.809. The layer comparisons share checkpoints and calibration data; Table 3 gives checkpoint-level results.

D CURVATURE BUDGET: LAYER-LEVEL LADDER AND DAMPING

The layer-level ladder reuses cached (W, H) pairs. Arms vary the supplied metric and ridge without recollecting statistics. It records incurred error from Equation (3) and supplied-metric loss $\|Z\|_F^2$ from each sweep. Across the tested ridge values, stale ratios remain elevated while aligned ratios stay near one.

Layer-level error and end-to-end perplexity do not track perfectly, which is why the model-level and layer-level tables are reported separately rather than pooled. The clearest instance is Qwen2.5-0.5B in the identity basis, where the block-32 metric has lower aggregate layer error than the diagonal one but slightly higher perplexity.

Table 4: Full controlled panel. Each transform cell is pre-transform → working-coordinate perplexity. All arms are evaluated after the same deployment-consistent inverse transform.

Model	BF16	RTN	GPTQ	Hadamard	random orthogonal	diagonal	diagonal∘Hadamard
Qwen2.5-0.5B	13.119	18.265	14.610	18.659→13.882	18.138→13.917	15.584→14.625	15.752→13.897
Qwen2.5-1.5B	9.687	12.162	10.947	12.466→10.091	12.203→10.092	11.136→10.456	10.855→10.009
Qwen2.5-3B	8.051	11.508	8.520	9.711→8.329	10.239→8.339	8.835→8.538	8.773→8.324
Qwen2.5-7B	7.769	8.216	8.231	9.026→8.044	8.705→7.990	8.170→8.065	8.613→7.957
Llama-3.1-8B	6.225	7.000	10.429	7.398→6.595	7.434→6.598	7.039→6.705	6.932→6.615
Phi-4 (14B)	6.453	6.700	6.593	6.789→6.584	6.778→6.569	6.632→6.621	6.833→6.572
OLMoE-1B-7B†	7.461	7.716	7.581	7.776→7.563	7.747→7.548	7.628→7.606	7.683→7.567

Table 5: **A curvature budget spent in the working coordinates or in the coordinates the weights have left.** WikiText-2 perplexity, symmetric W4A16 group 128, block-128 Hadamard basis. Columns are the supplied metric: its diagonal, its block-32 and block-128 truncations, and the dense matrix. "no transform" is the same ladder in the identity basis. Under the full-block convention, nominal coefficients represented per row are 1, 32, 128 and d, so at d=4096 the dense column costs 32× the block-128 column and 4096× the diagonal one. This is representation accounting, not measured peak memory. Round-to-nearest in each basis coincides exactly with the diagonal column, because a diagonal metric makes no cross-column update.

Model	curvature in	diagonal	block-32	block-128	dense
Qwen2.5-0.5B	no transform	18.265	18.629	16.914	14.610
	working coords	16.225	14.743	14.284	13.882
	pre-transform coords	16.225	19.613	19.957	18.659
Qwen2.5-1.5B	no transform	12.162	11.363	11.841	10.947
	working coords	11.097	10.855	10.302	10.091
	pre-transform coords	11.097	14.075	13.907	12.466
Qwen2.5-3B	no transform	11.508	11.141	9.218	8.520
	working coords	10.485	8.518	8.430	8.329
	pre-transform coords	10.485	4.3×10^3	2.8×10^6	9.711
Qwen2.5-7B	no transform	8.216	8.252	8.213	8.231
	working coords	8.330	8.196	8.127	8.043
	pre-transform coords	8.330	9.291	9.053	9.031
Llama-3.1-8B	no transform	7.000	6.981	6.908	10.429
	working coords	7.105	6.721	6.706	6.595
	pre-transform coords	7.105	7.276	7.411	7.398
Phi-4 (14B)	no transform	6.700	6.691	6.671	6.593
	working coords	6.671	6.648	6.634	6.584
	pre-transform coords	6.671	6.715	6.726	6.789
OLMoE-1B-7B†	no transform	7.716	7.694	7.661	7.581
	working coords	7.687	7.609	7.591	7.563
	pre-transform coords	7.687	7.740	7.769	7.776

E MISMATCH SPECTRUM: HOW INFORMATIVE IS THE INTERVAL

The eigenvalue interval of Proposition 3.1 can straddle one, allowing either direction of loss mismatch. The one-sided result in §5.3 is measured, not forced by that bound (Table 10, Figure 2). A pre-quantization spectral score ranks realized damage with Spearman correlation 0.95 across layer–transform pairs and 0.88 across model–transform arms. Its numerical scale is unreliable under damping: on the worst-conditioned tested layer it is 40 times below the undamped statistic, and 23 of 5036 scores fall below the undamped bound's floor. We use the score to describe ordering in this appendix, not to predict a layer's exact error.

F ALGEBRAIC TRANSPORT VERSUS DIRECT RECOLLECTION

The direct path applies the online Hadamard to live activations before accumulating its metric; the algebraic path accumulates in the original coordinates and applies (4); the stale path supplies the untransformed H. Remaining transported–recollected differences arise because low-order floating-

Table 6: **The sweep's self-assessment degrades monotonically in the resolution of a metric held in the wrong coordinates.** Median ratio of incurred to reported loss over 1,260 layer records per budget cell. The paired panel diagnostic excludes one fallback linear and has 1,259 distinct linears. A diagonal metric makes no cross-column compensation, so the two coordinate choices coincide there.

Transform	curvature in	diagonal	block-32	block-128	dense
Hadamard	working coords	0.990 (1260)	0.984 (1260)	0.978 (1260)	0.958 (1260)
	pre-transform coords	0.990 (1260)	1.241 (1260)	1.567 (1260)	3.094 (1260)
random orth.	working coords	0.990 (1260)	–	0.978 (1260)	0.958 (1260)
	pre-transform coords	0.990 (1260)	–	1.579 (1260)	3.368 (1260)

Table 7: Aggregate mechanism measurements on cached (W, H) pairs at ridge 10^{-2}; medians use the same layer–transform pairs in every cell. Increasing resolution adds compensation and coincides with rising median distortion in the pre-transform frame, shrinking innovations, and rising median physical error. These separate medians do not resolve the alignment in $\mathrm{tr}(ZK_MZ^\top)$ and do not explain the catastrophic tail. The log-volume row is measured after ridge regularization; $\det K_M = 1$ applies only to the undamped dense pre-transform setting of Corollary 3.2.

quantity	curvature in	diagonal	block-32	block-128	dense
compensation, $\|U_{\mathrm{off}}\|_F/\|U_{\mathrm{diag}}\|_F$	working coords	0.000	0.438	0.522	0.734
	pre-transform coords	0.000	0.213	0.325	0.612
distortion, $\mathrm{tr}\, K_M/d$	working coords	0.98	0.97	0.95	0.88
	pre-transform coords	1.88	2.52	3.35	6.69
innovation size, $\|Z\|_F$	working coords	1.85	1.74	1.52	1.15
	pre-transform coords	1.85	1.84	1.64	1.22
log-volume, $\log_{10} \det K_M$	working coords	-1208	-786	-581	-102
	pre-transform coords	-825	-720	-581	-102
incurred error	working coords	3.42	3.00	2.26	1.25
	pre-transform coords	3.42	3.47	3.56	3.85

point differences move weights across rounding boundaries. Two checkpoints, Llama-3.1-8B and Phi-4, also appear in the controlled panel; these are separate runs.

G Robustness controls

On two models and one fixed draw each, activation-order reordering retains the frame effect (Table 12). A separate depth intervention on GLM-4-9B quantizes 8, 16, 24, 32 or 40 of its blocks at fixed format; Φ increases monotonically across all five depths (Table 13), and a cumulative spectral score accounts for $R^2 = 0.419$ of its variation.

H Diagonal-scale sensitivity

The diagonal class has the smallest effect in the panel and is therefore the class most exposed to the objection that the effect is a hyperparameter artifact. Varying the SmoothQuant exponent (Table 14) over $\{0.25, 0.5, 0.75\}$ on Qwen2.5-1.5B leaves the ratio above one throughout for both the pure diagonal and composed classes, and every one of the six cells clears the 0.10423 band, the smallest margin being 0.68 perplexity.

I Pruning controls

We test compensated OBS pruning in the style of SparseGPT (Frantar & Alistarh, 2023) at 50% unstructured and 2:4 sparsity (Table 15). At fixed transformed weights, aligned curvature gives lower WikiText-2 and C4 perplexity in all 16 model/transform/mask comparisons. The aligned Hadamard arm remains worse than the unrotated arm in all eight model/mask comparisons. The transform

Table 8: Calibration-draw replication of the intermediate-budget failure on Qwen2.5-3B under a block-128 Hadamard. The working-coordinate ladder is stable across draws to within 0.03 perplexity; the pre-transform ladder collapses at every draw, at the same two budgets, reaching perplexities of 2.6×10^3 to 1.8×10^7.

seed	curvature in	diagonal	block-32	block-128	dense
42	working coords	10.485	8.518	8.430	8.329
	pre-transform coords	10.485	4.3×10^3	2.8×10^6	9.711
43	working coords	10.485	8.516	8.410	8.329
	pre-transform coords	10.485	2.6×10^3	1.8×10^7	9.810
44	working coords	10.485	8.529	8.436	8.332
	pre-transform coords	10.485	5.7×10^3	2.7×10^6	9.750

Table 9: Layer-level ladder on cached (W, H) pairs, replicated across three ridge levels so that the resolution effect is not confounded with damping. Values are the median incurred/reported ratio over all layer–transform pairs at that budget, rotation classes only. The divergence in the pre-transform rows survives every ridge; the working rows stay at or below one throughout.

ridge	curvature in	diagonal	block-32	block-128	dense
10^{-1}	working coords	0.908	0.867	0.821	0.693
10^{-1}	pre-transform coords	0.908	1.151	1.369	2.096
10^{-2}	working coords	0.989	0.984	0.978	0.943
10^{-2}	pre-transform coords	0.990	1.364	1.711	3.188
10^{-3}	working coords	0.998	0.998	0.998	0.993
10^{-3}	pre-transform coords	0.999	1.431	2.063	3.473

changes which weights the mask removes, while its coordinate frame determines the metric used for compensation. Undoing the Hadamard transform generally destroys sparsity in the deployed weights. These separate controls use 128 calibration windows of 2,048 tokens, seed 42 and relative ridge 0.01; evaluation averages cross-entropy over predicted target tokens. The supplement provides both dataset results and the corrected compensation implementation.

J THE BUDGET LADDER AT THREE BITS

Table 16 uses the four-bit panel's protocol with three-bit symmetric grids at group 128. No arm in any run took a round-to-nearest fallback, so the large stale values are the sweep's own output rather than a numerical failure.

K GROUP AND TRANSFORM WIDTH CONTROLS

Table 17 varies the group and the transform block around the catastrophic cell, where both equal 128, with everything else fixed. No arm in any configuration took a round-to-nearest fallback.

L EVERY PANEL CELL AT THREE CALIBRATION DRAWS

Each of seeds 42, 43 and 44 selects 128 WikiText-2 calibration windows; weights, evaluation inputs and quantizer settings remain fixed. The stale/aligned ratio exceeds one on all three draws in 27 of 28 cells and exceeds each cell's draw-to-draw spread in 26 of 28. Draws from the same checkpoint share its weights and evaluation data.

M FRONTIER MOE MULTI-SEED EVALUATION

Table 20 reports five calibration draws (seeds 7, 11, 23, 42, 137) on the 48-layer, 512-expert `Qwen/Qwen3-Next-80B-A3B-Instruct` checkpoint. Both arms use the same block-16 Hadamard-

Table 10: Width and informativeness of the Rayleigh interval of Proposition 3.1 for the rotation classes with curvature in the earlier coordinates, over the layer-level ladder. "straddles 1" is the fraction of cases with $\lambda_{\min}(K_M) < 1 < \lambda_{\max}(K_M)$, where the proposition permits either sign of the incurred/reported gap.

metric budget	cases	median $\lambda_{\min}$	median $\lambda_{\max}$	median $\operatorname{tr} K_M/d$	straddles 1
diagonal	54	0.000602	884	1.61	1.00
block-32	54	0.000575	1.36e+03	2.15	1.00
block-128	54	0.000447	1.67e+03	2.8	1.00
block-1536	30	0.000221	2.43e+03	4.46	1.00
dense	24	0.000254	2.27e+03	3.83	1.00

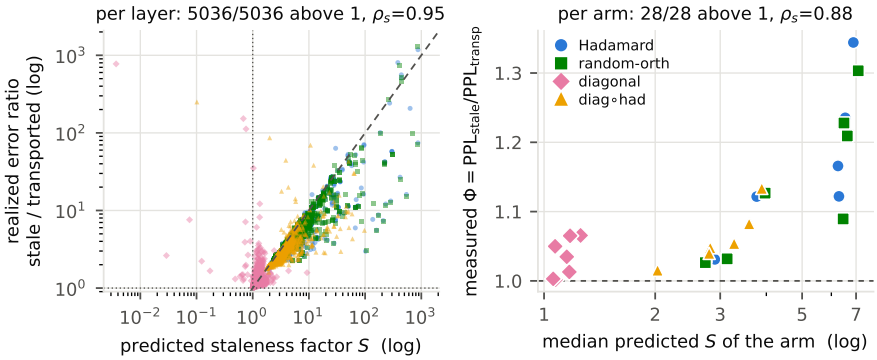

Figure 2: Damped spectral score versus incurred/reported layer loss. Across the measured layer–transform pairs, the score ranks the loss ratio with Spearman $\rho = 0.95$. Its magnitude is not a calibrated estimate of that ratio.

transformed weights and a two-scale NVFP4 reference: E2M1 values, E4M3 block-16 scales, and one FP32 global scale per weight matrix. The reference saturates finite scales and handles zero blocks explicitly. Calibration uses 128 WikiText-2 training windows of length 2,048; evaluation uses 128 test windows and averages cross-entropy over predicted targets. All ten paired arms have finite weights and forward losses.

Expert metrics are normalized input second moments weighted by squared full-softmax router probabilities over all 512 experts. They are local calibration surrogates rather than top-k-masked routed-output Hessians. Factorization starts at relative ridge 0.01 and increases it on a failed Cholesky attempt, so effective damping can differ between the supplied matrices. The source records retain each retry and fallback count. The five aligned-minus-stale comparisons have small perplexity differences; seed 42 also includes the unrotated and rotated RTN baselines in the table.

Historical arithmetic diagnostic. The earlier instrument used E2M1 values with directly rounded E4M3 block scales, omitting the FP32 tensor scale. It also used different scale saturation, zero handling and midpoint rounding. Its stale arm produced non-finite forward loss on seeds 23 and 42; all aligned arms finished (Table 21). These are failures of that particular arithmetic and metric combination. The separate two-scale reference keeps both arms finite on all five draws. The supplement preserves the historical results and source alongside the two-scale runs, with distinct format labels.

N Choosing the metric's partition

Every partitioned arm represents 128 nominal coefficients per row and applies a block transform respecting its partition. The permutation changes both the metric blocks and the channels sharing a quantization scale, so Table 22 tests each complete grouping rule. Captured mass is computed on H before the transform, where Corollary 4.3 makes it transform-invariant.

Table 11: **Stale curvature against two exact constructions of the aligned metric** (block-128 Hadamard, symmetric W4A16, group 128, seed 42). All three arms quantize the same rotated weights. "Transported" applies Equation (4) to a cached H; "recollected" accumulates H_T from live post-transform activations. Φ is stale/transported perplexity; separation is the stale penalty divided by $|\Delta|$, the transported–recollected difference. [‡]One stale linear took a round-to-nearest fallback after a failed factorization. [§]Linears that receive no activations from text calibration (162 on Gemma-3-12B, 191 on Gemma-4-31B) keep BF16 weights in every arm. This experiment uses its own protocol; absolute values are not commensurable with the panel's.

Model	stale	transported	recollected	max rel. Fro.	$\lvert\Delta\rvert$	Φ	separation
Gemma-3-12B[§]	36.4405	25.4812	25.5854	8.53×10^{-6}	0.10423	1.430	105×
Gemma-4-31B[§]	8.0144	6.5095	6.5007	1.19×10^{-5}	0.00881	1.231	171×
GLM-4-9B	8.3657	7.7754	7.7725	5.62×10^{-6}	0.00285	1.076	207×
Llama-3.1-8B	7.3615	6.6039	6.6052	5.46×10^{-6}	0.00130	1.115	581×
Ministral-8B	7.7637	7.2244	7.2252	1.04×10^{-5}	0.00086	1.075	626×
Mistral-7B-v0.3	5.9276	5.6355	5.6353	6.21×10^{-6}	0.00019	1.052	1541×
Phi-4	6.8078	6.5818	6.5848	2.07×10^{-5}	0.00299	1.034	75×
Qwen3.5-9B	11.1510[‡]	8.7321	8.7048	1.46×10^{-5}	0.02721	1.277	89×
Qwen3-8B	11.1595	10.0002	10.0174	8.51×10^{-6}	0.01715	1.116	68×

Table 12: Activation reordering does not remove the effect. Values are earlier/working perplexity ratios; calibration-draw replication for the whole panel is in Table 19.

Control	Setting	ratio
Activation order	Gemma-4-31B, act-order enabled	1.305
Activation order	GLM-4-9B, act-order enabled	1.088

Table 13: Within-model depth intervention on GLM-4-9B. Quantizing more blocks monotonically increases Φ, with all other model and format choices held fixed.

Compressed blocks	8	16	24	32	40
Φ	1.0153	1.0190	1.0272	1.0339	1.0481

Table 14: Diagonal-scale sensitivity on Qwen2.5-1.5B. Φ is the pre-transform/working perplexity ratio; the $\alpha = 0.5$ column is the panel cell.

Transform class	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$
diagonal	1.0691	1.0651	1.0928
diagonal∘Hadamard	1.1169	1.0845	1.0979

Table 15: Pruning with compensated updates. WikiText-2 perplexity for 50% unstructured and 2:4 masks. Aligned and stale arms use identical transformed weights; the identity arm uses no transform. These separate controls use 128 calibration windows of 2,048 tokens, seed 42, relative ridge 0.01, and target-count-weighted evaluation on up to 128 test windows. Masks are sparse in the working coordinates; undoing a Hadamard transform generally gives dense deployed weights.

			Hadamard		diagonal scale	
Model	mask	identity	aligned	stale	aligned	stale
Qwen2.5-0.5B	50%	17.589	21.311	8596.370	17.925	23.188
Qwen2.5-0.5B	2:4	28.561	76.472	137287.331	28.915	44.703
Qwen2.5-1.5B	50%	12.589	13.836	270.990	12.571	12.895
Qwen2.5-1.5B	2:4	21.299	29.088	10220.502	20.032	22.880
Qwen2.5-3B	50%	9.789	10.262	7422.847	9.797	10.035
Qwen2.5-3B	2:4	13.071	15.305	107247.799	13.069	13.723
Llama-3.1-8B	50%	8.248	8.508	1575.678	8.281	8.350
Llama-3.1-8B	2:4	12.311	14.812	113761.179	12.385	12.619

Table 16: The budget ladder at three bits, symmetric, group 128. The working-coordinate dense result improves on the diagonal anchor for all seven checkpoints, with a block-32 to block-128 reversal on Llama; the earlier-coordinate maximum is at an intermediate budget on five of seven checkpoints. Ratios at a dense metric are in Table 1.

		working coordinates				earlier coordinates			
Model	BF16	diag	blk-32	blk-128	dense	diag	blk-32	blk-128	dense
Qwen2.5-0.5B	13.119	115.891	33.808	24.724	20.342	115.891	1.0×10^3	3.4×10^3	2.3×10^3
Qwen2.5-1.5B	9.687	46.437	18.719	15.343	13.778	46.437	278.739	513.385	348.718
Qwen2.5-3B	8.051	9.4×10^3	12.015	11.057	10.205	9.4×10^3	5.9×10^6	3.2×10^8	1.3×10^4
Qwen2.5-7B	7.769	13.630	10.310	9.867	9.302	13.630	68.065	63.265	21.316
Llama-3.1-8B	6.225	41.573	10.457	10.603	9.051	41.573	133.290	621.333	264.623
Phi-4 (14B)	6.453	8.523	8.020	7.936	7.403	8.523	9.299	9.804	11.215
OLMoE-1B-7B†	7.461	9.222	8.452	8.310	8.092	9.222	9.755	10.395	11.033

Table 17: Group and transform width controls on Qwen2.5-3B. Varying group size with transform width 128 keeps the largest earlier-coordinate error at block-128, while changing transform width moves its location in this grid. The working-coordinate results remain between 8.263 and 8.597 perplexity.

		earlier coordinates			working coordinates		
group	transform	blk-32	blk-128	dense	blk-32	blk-128	dense
128	64	7.2×10^5	6.9×10^3	10.153	8.466	8.444	8.323
64	128	1.8×10^3	3.4×10^4	9.944	8.441	8.356	8.263
128	128	4.3×10^3	2.8×10^6	9.711	8.518	8.430	8.329
256	128	1.4×10^4	1.2×10^7	10.218	8.597	8.506	8.362
128	256	129.019	3.6×10^4	10.923	8.561	8.448	8.341

Table 18: Calibration-draw reuse and coordinate mismatch on two Qwen2.5 checkpoints. All arms quantize the same transformed weights; only supplied curvature changes. The pre-quantization score $\mathrm{tr}\, K_M/d$ and incurred layer error rise in the same order. Perplexity for the reused draw on Qwen2.5-3B is slightly lower than the matched-draw value, within the stated tie band.

Model	supplied curvature	$\mathrm{tr}\, K_M/d$	incurred	PPL
Qwen2.5-1.5B	correct basis, correct data	0.943	2332	10.0908
	correct basis, another draw	0.981	2484	10.1640
	pre-transform basis, correct data	5.102	1.279e+04	12.4658
Qwen2.5-3B	correct basis, correct data	0.946	3734	8.3292
	correct basis, another draw	1.004	4061	8.3261
	pre-transform basis, correct data	4.887	2.879e+04	9.7106

Table 19: **Every panel cell at three calibration draws.** Each entry is the earlier/working perplexity ratio, mean over seeds 42/43/44 with the spread across draws in parentheses. Cells whose effect exceeds their own draw-to-draw spread are bold.

Model	Hadamard	random orth.	diagonal	diag.∘Had.
Qwen2.5-0.5B	**1.331** (0.028)	**1.327** (0.056)	**1.062** (0.006)	**1.128** (0.009)
Qwen2.5-1.5B	**1.210** (0.078)	**1.237** (0.047)	**1.080** (0.056)	**1.057** (0.049)
Qwen2.5-3B	**1.171** (0.012)	**1.210** (0.028)	**1.034** (0.007)	**1.051** (0.007)
Qwen2.5-7B	**1.093** (0.069)	**1.099** (0.024)	1.022 (0.023)	**1.060** (0.034)
Llama-3.1-8B	**1.124** (0.005)	**1.121** (0.014)	**1.050** (0.005)	**1.049** (0.005)
Phi-4 (14B)	**1.031** (0.001)	**1.030** (0.008)	**1.004** (0.003)	**1.036** (0.005)
OLMoE-1B-7B†	**1.028** (0.002)	**1.029** (0.003)	1.001 (0.003)	**1.015** (0.001)

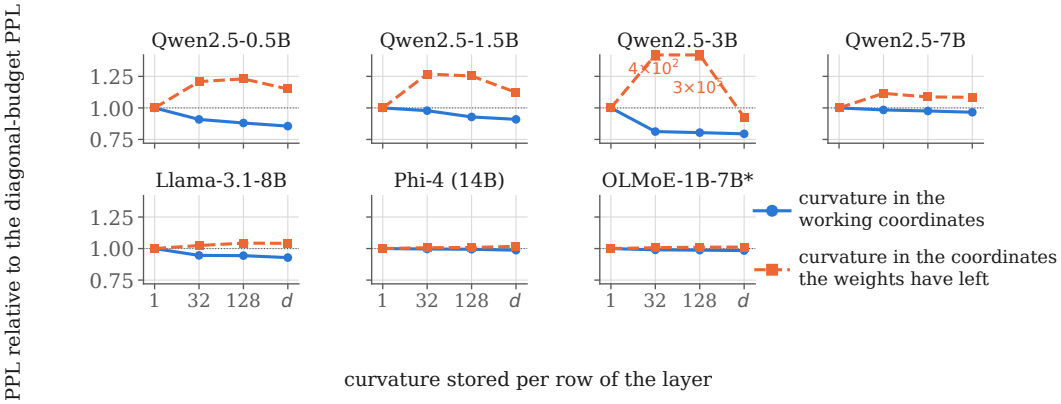

Figure 3: **Figure 1 resolved per checkpoint.** Each panel is one checkpoint under a block-128 Hadamard; x is the nominal number of full-block coefficients represented per row, not measured peak memory, and y is perplexity divided by that checkpoint's own diagonal-budget perplexity—the point where the two coordinate choices provably coincide, so both series start at exactly 1. Triangles mark points above the axis, labeled with their value. Series are distinguished by color, dash pattern and marker. OLMoE* is the attention-only case study. Raw perplexities are in Table 5.

Table 20: Two-scale NVFP4 Qwen3-Next-80B-A3B results across five calibration draws. Each pair fixes transformed weights and calibration; both arms use local expert-metric weighting and adaptive Cholesky factorization. At seed 42, unrotated and rotated RTN score 5.840115 and 5.845345, respectively. WikiText-2 perplexity uses this experiment's target-count-weighted evaluator; its absolute values are not directly comparable to the quantization panel.

Calibration seed	stale PPL	aligned PPL	decrease
7	5.862435	5.848137	0.014297
11	5.872295	5.844661	0.027634
23	5.852682	5.848189	0.004493
42	5.858655	5.846894	0.011761
137	5.849240	5.844357	0.004884

Table 21: Historical implementation diagnostic, not two-scale NVFP4. This earlier driver rounded E4M3 block scales without an FP32 tensor scale; its nonfinite observations cannot establish two-scale NVFP4 instability. Arms are matched within each row; cached references with different evaluation lengths are excluded.

Calibration seed	stale PPL	transported PPL
7	5.881801	5.863951
11	5.883024	5.867842
23	non-finite	5.871161
42	non-finite	5.864822
137	5.870477	5.870502

Table 22: **Captured curvature mass does not predict the best partition in this experiment.** Block size 128; every partitioned arm represents 128 nominal coefficients per row and transports exactly. "Captured" is the fraction of squared curvature mass inside the diagonal blocks; "error" is the incurred reconstruction error of Equation (3), relative to the consecutive-channel partition on the same layer. Higher capture yields no systematic error reduction.

metric partition	coeff./row	captured	error (rel.)	wins
consecutive channels	128	0.179	1.000	–
spectral ordering of $\lvert H \rvert$	128	0.514	1.001	9/18
greedy max-affinity grouping	128	0.682	1.012	9/18
dense reference	d	1.000	0.558	18/18

O ROTATION SEARCH UNDER A FROZEN METRIC

In a local Cayley-parameterized rotation search, following the parameterization used by SpinQuant (Liu et al., 2025), one loop updates the metric as T changes and the other keeps its initial metric. The updated loop has lower final physical error on 15 of 16 real layer pairs and 44 of 48 layer/seed runs. On 11 pairs, the frozen loop ends within 5% of the no-rotation error. These are local reconstruction objectives: the experiment does not run a full SpinQuant or OptRot training pipeline, and the updated loop does not beat the fixed Hadamard control here.

P SOFTWARE BOUNDARY TAXONOMY

Three implementation patterns exist. **Live collection after the transform:** the metric hook observes x_T and accumulates H_T; the budget is in the working coordinates. **Cached collection plus explicit transport:** the pipeline stores H before T is applied and computes H_T by (4); also working coordinates, and the pattern offline-Hessian workflows need. **Separated state with no coordinate handoff:** the pipeline transforms weights or installs an online transform after H is cached and passes the bare cached matrix to the compressor; the budget is in the earlier coordinates.

At commit `86f564ddb1d6548b228c67a10509a4ed7264345c`, the LightCompress run in §5.6 reaches the third pattern by omitting `online_rotate` from its GPTQ-stage configuration. Its official guide says to keep rotation settings consistent across stages; the combined example enables this field. Our two GPTQ arms differ at this setting in an otherwise matched unmodified-toolkit run. The release artifact contains the configurations, toolchain logs, evaluator source and parsed result so that the stage handoff can be checked.

The pinned spcl/QuaRot commit exposes a different issue. In `e2e/checkpoint_utils/rotation_utils.py`, the exporter fuses inverse Hadamard transforms into projection weights (lines 84 and 100–101). In `e2e/checkpoint_utils/gptq_utils.py`, the proposed GPTQ hook reads `inp[0].data` on the bare linear (lines 327 and 331). Before reaching that hook, however, the module list contains `self_attn.v_proj=` (line 299); a lookup on the tested Llama checkpoint raises `KeyError`. This branch therefore does not provide an end-to-end stale-metric result. The separate `fake_quant` implementation hooks `self_attn.v_proj.module` after online rotation. The artifact includes the pinned source audit and the real-checkpoint module-lookup record.

For NVIDIA Model-Optimizer at commit 9392dfe (NVIDIA Corporation, 2025), `model_calib.py` captures inputs in a linear pre-hook, while `tensor_quantizer.py` can rotate weights. Candidate evaluation can rotate the weights back, however, leaving the raw-input metric appropriate for that path. We do not classify the repository as a demonstrated stale-metric failure. These audits show why an interface should store a frame identifier with each metric, reject incompatible handoffs before factorization, and check Proposition 4.1 before transporting a block cache.

Q REPRODUCIBILITY AND ARTIFACT MAP

The underlying run records store script arguments, seed, Torch and CUDA versions, GPU type, and timestamp; per-layer files inherit provenance through their enclosing run. Controlled-panel values come from the per-model panel records and the paired per-layer files; the curvature-budget arms, the per-layer supplied-metric/physical logs, and the compatibility check come from the budget records; the layer-level ladder and its ridge replications come from the ladder record; the exact-identity gate is its own artifact. Algebraic-versus-live comparisons are nine separate records, the calibration-draw replication is one record per model and seed, and the third-party workflow values are parsed from that toolchain's stage and evaluation logs. A cell manifest records the source file and extraction rule for generated table values.

Data is WikiText-2 (Merity et al., 2017), configuration `wikitext-2-raw-v1`; calibration windows are drawn from the training split and perplexity is measured on non-overlapping test windows. Controlled-panel checkpoints are `Qwen/Qwen2.5-0.5B`,

Qwen/Qwen2.5-1.5B, Qwen/Qwen2.5-3B, Qwen/Qwen2.5-7B, meta-llama/Llama-3.1-8B, microsoft/phi-4, and allenai/OLMoE-1B-7B-0924. Live-recollection checkpoints add google/gemma-3-12b-it, google/gemma-4-31B, THUDM/GLM-4-9B, mistralai/Ministral-8B-Instruct-2410, mistralai/Mistral-7B-Instruct-v0.3, Qwen/Qwen3.5-9B, and Qwen/Qwen3-8B. The third-party workflow uses meta-llama/Llama-3.1-8B. Tensor-coverage exceptions, transform parameters, and evaluation counts are stated where the corresponding numbers appear. The anonymous code supplement contains the raw records, table and figure generators, per-cell manifest, and recorded experiment commands. Its README.md and docs/PAPER_RESULTS.md map each table to its inputs; docs/EXPERIMENT_COMMANDS.md gives setup and rerun instructions. The older panel records retain arguments, Torch/CUDA versions and GPU type but lack complete model, data and dependency pins; the supplied reconstruction environment is identified separately. A fresh ten-arm Qwen2.5-0.5B run reproduces every archived perplexity exactly. Corrected pruning and two-scale frontier runs include pinned inputs and token hashes, and are stored separately from the historical diagnostics.