RC-VQ: RECOVERING WHAT CALIBRATION DISCARDS WITH VECTOR-QUANTIZED RESIDUAL CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibration data decides what training-free LLM compression keeps. In activation-aware factorization and post-training quantization alike, a small calibration set defines an importance measure over the weights, and whatever it does not certify is discarded as a residual. We argue that failing the calibration proxy does not make weights unimportant, and that the discarded residual should be recovered. The standard recovery tool, a low-rank adapter, is structurally mismatched to this task. On LLaMA-2 7B the residuals are dense and spectrally flat. Quantization residuals are statistically indistinguishable from a Marchenko–Pastur noise bulk, and truncation residuals form a gapless spectral continuum without isolated spikes, so a rank-r corrector recovers a vanishing fraction of either. Across four model families, the energy a low-rank corrector recovers grows only linearly in its rank. We propose RC-VQ, a full-rank vector-quantized corrector fitted to the residual without gradients or recovery data, which composes with SVD factors, INT4 weights, and their stacked combination. On structural compression of LLaMA-2 7B, LLaMA-3 8B, and Qwen3 8B, RC-VQ reduces perplexity degradation by 3–13$\times$ against the strongest reported structured baseline at matched checkpoint size. On the INT4-AWQ quantization gap of Qwen3.5 4B, it recovers 96% of the distance to FP16 at three bits per parameter, a budget at which an identically sized low-rank corrector recovers under 40% and a scalar residual branch needs 37% more bits to match. Sweeping codebook objectives across structural, quantization, and stacked residuals, we show that each pure objective has a failure regime, while the hierarchical direction-then-refinement objective, PolarVQ, never collapses, making it the robust default when residual geometry is unknown.

1 INTRODUCTION

Large language models increasingly run on-device, where the memory available to hold their weights, not their accuracy, is the binding constraint. Fitting a model into that budget means deciding what to keep, and the dominant training-free compression routes, activation-aware low-rank factorization and weight quantization, delegate this decision to a small calibration set. The calibration data defines an importance measure over the weights, the compressed representation keeps what that measure certifies, and everything else is discarded as a residual. This delegation equates two different things, unimportant under the calibration proxy and unimportant in fact. Weight directions that stay quiet under a few hundred calibration sequences may still support capabilities the proxy never probed, so the discarded residual is not noise but information the pipeline failed to certify.

The standard remedies, gradient-based fine-tuning or a LoRA-style low-rank adapter, reintroduce training and data curation and, more subtly, a structural assumption about the error they absorb. That assumption is wrong for the residuals that strong modern bases leave behind. Across both compression families, we find that the discarded residual is dense and spectrally flat. For activation-aligned SVD truncation this is almost definitional, since truncation removes precisely the high-energy singular directions, and we measure the spectral flatness of the remaining residual to be uniformly higher than that of the original weights, with complete separation between the two distributions at moderate compression. A random-matrix analysis sharpens the picture. The truncation residual is a gapless spectral continuum with no isolated spikes, while the quantization residual is statistically indistinguishable from the Marchenko–Pastur noise bulk itself (Marchenko & Pastur, 1967), so in

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.

either case a low-rank corrector, whatever its rank budget and whether trained or not, captures a vanishing fraction of the residual. In both regimes, the default corrector, a low-rank adapter, is structurally mismatched to the error it is asked to fix. The residual is informative yet spectrally featureless because its information is diffuse rather than concentrated in a few recoverable directions.

We propose the opposite tool, **RC-VQ**, a full-rank vector-quantized corrector fitted directly to the residual. The corrector partitions the residual into short sub-vectors, clusters them into a small per-projection codebook, and stores only the codebook and a packed index stream. Magnitudes are folded into the codewords, so no side-channel scales are required. It is fitted from the weights alone, with no gradients, no labels, and no recovery data, so the corrector recovers residual information without re-entering the calibration loop whose blind spots motivated it. It composes with either base representation, SVD factors (SVD-VQ) or INT4 weights (Quant-VQ), and with their stacked combination. Viewed as an adapter, it removes both of LoRA's defining constraints, the rank bottleneck and the need for training.

Our second finding is that the geometry of the residual decides which codebook objective is safe. We sweep three objectives, Euclidean (MSE) clustering, directional (spherical) clustering, and the hierarchical cosine→MSE composite we call **PolarVQ**, across three residual families with thirty seeds per configuration, namely whitened SVD-truncation residuals, AWQ quantization residuals of five models, and composite residuals of factorized-then-quantized weights of three models. The first two each have a failure regime. Directional clustering collapses on composite residuals (+0.5 to +5.4 perplexity) and trails by 0.2–0.4 on whitened-SVD residuals, while Euclidean clustering never wins on quantization residuals. PolarVQ collapses nowhere and is within 0.05 perplexity of the per-regime winner in nine of ten cells, within one standard error in the tenth. We therefore claim a robust default, the property that matters for a training-free pipeline applied to residuals whose geometry is not known in advance. Unlike polar-coordinate weight quantizers such as PCDVQ, which apply polar decoupling to the original weights and advocate it unconditionally, we target the residual with a single magnitude-folded codebook and measure when directional structure helps, ties, and fails.

Empirically, the corrector closes most of the gap that compression opens. On structural compression of LLaMA-2 7B, LLaMA-3 8B, and Qwen3 8B, spending a fixed storage budget on the VQ corrector instead of additional SVD rank reduces perplexity degradation by 3–13$\times$ (geometric mean 7$\times$) against the strongest reported structured baseline at matched checkpoint size. On the AWQ quantization gap of Qwen3.5 4B, a vision–language model whose language decoder we compress, the corrector recovers 96% of the gap to FP16 at three bits per parameter, where a low-rank corrector at the same budget recovers 39% and a scalar residual branch at equal rate recovers 77%. The compressed models trade decode throughput for decode-time memory on the current unfused VQ path, and we outline the LUT-fused kernel that removes this penalty (§ 5.7).

Contributions.

- We identify spectral flatness of compression residuals, measured by SFM and confirmed by Marchenko–Pastur spike analysis, as the reason low-rank correctors fail and full-rank VQ succeeds across all three residual families.
- We introduce RC-VQ, a full-rank VQ corrector with magnitude-folded per-projection codebooks, fitted from the residual weights alone with no gradients and no recovery data, which composes with SVD factors, quantized bases, and their stacked combination, together with spectrum-driven rank and codebook allocation.
- We resolve the corrector's most consequential design choice, the codebook objective, by sweeping it across structural, quantization, and stacked residuals (four model families, thirty seeds per configuration). Each pure objective has a failure regime, whereas the hierarchical cosine→MSE objective (PolarVQ) never collapses and stays within 0.05 perplexity of the per-regime winner in nine of ten model–regime cells, making it a robust default when residual geometry is unknown.
- Across five models and three compression regimes, RC-VQ yields 3–13$\times$ lower perplexity degradation than the strongest reported structural baseline and recovers 96% of the INT4-AWQ gap at iso-storage, where low-rank correctors recover under 40% and scalar residual branches need 37% more bits, and we report the memory–throughput trade-off and the remaining kernel work.

2 Related Work

Training-free LLM compression has developed along a few well-established routes. The first replaces weight matrices with truncated factorizations. Plain SVD on weights is a poor proxy for functional error, and activation-aware variants such as FWSVD (Hsu et al., 2022), ASVD (Yuan et al., 2023), and SVD-LLM (Wang et al., 2025), which whitens weights with calibration statistics before truncation, largely close this gap, and MoDeGPT (Lin et al., 2025) extends the idea to modular decompositions of coupled matrix pairs, with later refinements (SAES-SVD, ERC-SVD; (Hu et al., 2026; Bai et al., 2025)). All of these methods decide what to keep and discard the remainder, falling back on recovery fine-tuning when quality at the target ratio is insufficient. We build on the strongest activation-aware SVD pipeline as our structural base and instead ask what to do with the remainder.

The second route maps weights to low-bit grids. GPTQ (Frantar et al., 2023) compensates rounding error with second-order information, AWQ (Lin et al., 2024) rescales salient channels from activation statistics, SmoothQuant (Xiao et al., 2023) migrates outlier energy into the weights, and QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2025) rotate outliers away before quantization. At 4-bit these are the deployment workhorse, and their residual error is simply accepted. Our corrector consumes only the residual $W - \hat{W}$, so it is complementary to this line and composes with any of these methods, of which we adopt AWQ as the strongest 4-bit representative.

A third line replaces the scalar grid with learned codebooks over sub-vectors for extreme 2-bit regimes. AQLM (Egiazarian et al., 2024) trains additive codebooks, QuIP# (Tseng et al., 2024) pairs Hadamard incoherence with E8 lattice codebooks, and VPTQ (Liu et al., 2024) and GPTVQ (van Baalen et al., 2024) refine clustering and update schedules. Closest in machinery is PCDVQ (Yue et al., 2025), which quantizes the original weights in polar coordinates with separate direction and magnitude codebooks. RC-VQ differs in role, since the model is carried by a cheap base of SVD factors or INT4 weights, and the VQ is fitted only to that operator's residual with a single magnitude-folded codebook and no codebook training. The same distinction separates us from residual VQ in audio codecs (SoundStream, EnCodec; (Zeghidour et al., 2022; Défossez et al., 2023)), which recursively quantizes its own previous stage's residual rather than the residual of an external operator.

When the base representation falls short, correction is typically low-rank. Trained variants attach LoRA-style adapters to a quantized base (QLoRA; (Dettmers et al., 2023)) or alternate quantization with low-rank fitting for fine-tuning initialization (LoftQ; (Li et al., 2024)). Training-free variants are the direct counterpart of our corrector. LoRC (Yao et al., 2023) applies truncated SVD to the quantization error $E = W - \hat{W}$ and stores the top-r factors, while LQER (Zhang et al., 2024) and CALDERA (Saha et al., 2024) develop joint low-rank-plus-low-precision decompositions in the same spirit. These works assume the compression error carries recoverable low-rank structure. Our spectral measurements (§ 4) challenge this premise for strong modern bases. The residual is a Marchenko–Pastur bulk after AWQ rounding and a gapless continuum after activation-aware truncation, so a rank-r corrector captures a vanishing fraction of it, and an iso-storage low-rank corrector closes under 40% of a gap that our full-rank corrector closes almost entirely.

Finally, quantizing magnitude and direction separately is classical in source coding, where polar quantizers suit circularly symmetric densities. The idea resurfaced for LLMs in two unrelated works both named PolarQuant, for KV-cache keys as radius–angle pairs (Wu et al., 2025) and with random preconditioning (Han et al., 2025), and in PCDVQ's direction/magnitude decoupling for weights. These works advocate polar structure unconditionally, whereas our finding is conditional. Directional objectives win or tie on quantization residuals, trail on whitened SVD residuals, and fail outright on heavy-tailed composite residuals. Residual geometry, not a universal preference, should select the codebook objective.

3 Method

3.1 Setting

Let C be a training-free compression operator producing $\hat{W} = C(W)$ from a projection weight $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, and let $R = W - \hat{W}$ be its residual. RC-VQ attaches a corrector $\hat{R} = Q_\theta(R)$ so the deployed weight is $\hat{W} + \hat{R}$, with θ (a codebook and an index stream) fitted from R alone by

weight-space k-means, without gradients or any recovery data. The base C carries the model, and the corrector carries only what C discarded. We instantiate C three ways.

SVD-VQ (structural base). Activation-aware whitened SVD: calibration activations define a whitening transform (Cholesky by default), the whitened weight is truncated at rank r, and the factors $A \in \mathbb{R}^{d_{\text{out}} \times r}$, $B \in \mathbb{R}^{r \times d_{\text{in}}}$ are stored in fp16. Ranks are allocated across projections by equal energy on the calibration spectra, and when r is not a power of two, a fixed-seed random orthogonal rotation replaces the Hadamard rotation.

Quant-VQ (quantization base). Activation-aware INT4, AWQ in our experiments: per-input-channel scales $s = a^{\alpha}$ with α grid-searched over 21 points against activation-weighted reconstruction error, followed by group-wise symmetric INT4 RTN (g=128). The base is unchanged from standard AWQ, but any scalar quantizer exposes the same interface, and the corrector is fitted to $R = W - \hat{W}$ regardless of how $\hat{W}$ was produced.

Stacked base (SVD ∘ INT4). For deployments that factorize and quantize, we quantize the SVD factors themselves with a per-factor AWQ scheme: the scale for A is searched on the intermediate activations $H = XB^{\top}$ and absorbed exactly into B (the fp product AB is unchanged, so the fold has zero runtime cost); the scale for B is searched on the raw inputs and, since it cannot be absorbed upstream, is stored as a per-channel input scale applied at runtime. The corrector then targets the composite residual $R = W - \mathrm{dq}(A_q)\,\mathrm{dq}(B_q)\,\mathrm{diag}(1/s_B)$, which absorbs the truncation error and both factor quantization errors in a single stage. The searched α is near zero for most late layers, but the scheme guarantees the stacked base is never worse than plain RTN on the factors.

The pipeline's calibration boundary is explicit, since whitening, rank allocation, and AWQ scales consume calibration data, whereas the corrector does not, being a function of the residual matrix only.

3.2 PolarVQ corrector

R is flattened row-major and cut into sub-vectors of length D, with $D \mid d_{\text{in}}$, so that each sub-vector covers D consecutive input channels of a single output channel. This locality is what the fused-kernel design (§ 5.7) relies on. Sub-vectors are clustered into a per-projection codebook of K codewords under one of three objectives: **MSE**, standard k-means in $\mathbb{R}^D$; **directional**, spherical k-means on normalized sub-vectors with magnitudes folded into codewords at update time; or **PolarVQ (cos→MSE)**, a hierarchical composite that forms G direction groups by spherical assignment and then refines K/G codewords within each group by MSE. All three store exactly the same bytes: K fp16 codewords of length D plus one $\lceil \log_2 K \rceil$-bit index per sub-vector. Magnitudes are folded into codewords, so no per-group scales or norms are stored, and the mode choice is free at deployment. The corrector's rate is $\rho = \lceil \log_2 K \rceil / D$ bits per parameter plus a codebook term $16KD/(d_{\text{out}} d_{\text{in}})$ that is negligible in practice (0.006% of the LLaMA-2 checkpoint). K and D are two knobs on the same rate, and quality tracks ρ (§ 5.3).

3.3 Allocation

Rank across projections follows equal energy on calibration spectra. Codebook capacity across layers follows a tail-entropy signal under a fixed total budget, and entropy predicts codebook demand no worse than tail energy, and slightly ahead of it (§ 5.3). In the configurations of this paper the allocator converges to uniform per-layer K (the normalized signal is too flat to cross power-of-two thresholds), which we report as such.

4 Why Full-Rank Correction: The Geometry of Compression Residuals

The discarded residual is functionally valuable, yet every measurement in this section finds it spectrally featureless. The two are compatible because the residual's value lies in how its information is distributed rather than in how much structure its spectrum shows. The residual is informative but diffuse, its content spread across all directions rather than concentrated in the few that the calibration proxy certified and the base kept.

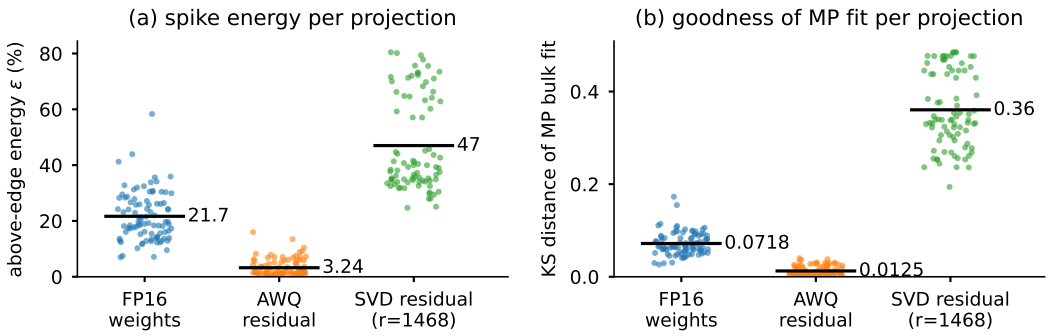


Figure 1: Marchenko–Pastur fit statistics per MLP projection of LLaMA-2 7B (90 projections for weights and AWQ residual, the 84 compressed ones for the SVD residual; bars mark means). (a) Energy above the fitted bulk edge: FP16 weights carry isolated spikes, the AWQ residual almost none, and the SVD residual a large but spread-out excess. (b) KS distance of the bulk fit: the AWQ residual is indistinguishable from a noise bulk, whereas the SVD-truncation residual is not a bulk at all.

Flat spectra. We measure the spectral flatness (SFM) of the compressed MLP projections of LLaMA-2 7B and of their SVD-truncation residuals. SFM, also known as Wiener entropy, is the ratio of the geometric to the arithmetic mean of the squared singular values, ranging from 0 for a spectrum dominated by a few directions to 1 for one that is perfectly flat. Original weights average 0.722 [0.615, 0.772], while residuals at 20% compression average 0.851 [0.831, 0.867]. The two distributions are completely separated, since the flattest residual (0.831) exceeds the least-flat weight (0.772). Truncation removes exactly the structure a low-rank object can represent, and what remains is closer to white noise than the weights ever were.

No recoverable spikes. We fit the empirical spectral density of each MLP projection of LLaMA-2 7B to the Marchenko–Pastur law (Marchenko & Pastur, 1967), with bulk edge $\lambda_+ = \hat{\sigma}^2(1+\sqrt{q})^2$, and report per projection the energy above the edge and the Kolmogorov–Smirnov distance (Kolmogorov, 1933; Smirnov, 1948), KS hereafter, between the empirical bulk and its fit (Figure 1). The three residual families separate cleanly. The FP16 weights follow a spiked model, a well-matched bulk at mean KS 0.072 plus isolated spikes carrying 21.7% of the spectral energy on average, which is precisely the structure that activation-aware factorization extracts. The AWQ residual collapses to the bulk alone, with 3.2% above-edge energy at KS 0.013. It is statistically indistinguishable from iid noise, and by the BBP argument (Baik et al., 2005) the energy available to any rank-r corrector, trained or not, is bounded by this near-zero spike fraction. The SVD-truncation residual at r=1468 fails the fit in the opposite direction, at KS 0.36, because its spectrum is not a noise bulk but a gapless, decaying continuum, the truncated tail itself. It carries 47% of its energy above the fitted edge, yet spread across the whole tail rather than concentrated in isolated spikes, leaving nothing for a low-rank corrector to seize there either. The functional counterpart appears in § 5.4, where the residual energy a rank-r corrector recovers grows linearly in r, by roughly 0.05% per rank on Qwen3.5 4B and 0.03% on 9B, the signature of a spectrum with no privileged directions, so low-rank correction buys noise at a flat per-bit price and leaves most of the quantization gap open even beyond the VQ corrector's budget.

Rate–distortion of flat residuals. A flat, dense residual is precisely the regime where vector quantization dominates scalar quantization through dimension-wise dependency and packing gains, and § 5.4 confirms it, since at equal side-channel rate (3.1 bpw), the VQ corrector recovers 96% of the quantization gap where a scalar residual branch recovers 77%.

Energy versus performance. Recovered residual energy and recovered quality are not the same quantity. At the same 3.01 bpw on Qwen3.5 4B, the purely directional codebook recovers 67.6% of the residual energy against 84.2% for Euclidean clustering, 17 pp less, yet matches or beats it in perplexity on every quantization regime we test (§ 5.6), so directional fidelity of the correction matters more than the energy it restores. The scalar comparison makes the same point from the other side. PolarVQ and the RTN-3 residual branch recover nearly equal energy at nearly equal rate, 84.4%

against 82.9%, yet differ by 19 pp of the quality gap (§ 5.4), so how the remaining error is distributed matters as much as how large it is. As a methodological corollary, the hierarchical objective is also a better optimizer of the Euclidean criterion itself, recovering 84.43% of the residual energy against 84.24% for flat k-means with four restarts, a small margin that is stable across seeds. Direction-first partitioning escapes local optima that flat k-means settles into.

Directional structure and its limits. Three residual regimes behave differently. On whitened SVD residuals Euclidean and hierarchical clustering tie exactly, while purely directional clustering, blind to magnitude, trails by 0.2–0.4 PPL. Quantization residuals retain directional structure that Euclidean clustering consistently fails to exploit. Composite residuals of factorized-then-quantized weights behave differently again, because equal-energy rank allocation compresses some projections to double-digit ranks. Down-projections of late layers reach r=64–96, where the residual retains 97–98% of the weight norm, against 65–73% for their gate and up siblings, so sub-vector magnitudes across the layer become heavy-tailed. Magnitude-blind directional assignment then places small-magnitude sub-vectors into groups anchored by large codewords, and collapses (+0.5 to +5.4 PPL, § 5.6). The hierarchical objective retains the directional stage but refines within groups by MSE, and survives every regime we test.

All three probes point the same way. The residual carries the information the calibration proxy failed to certify, but diffusely, with no privileged directions for a rank-limited corrector to exploit, so recovering it requires capacity spread across every direction at once, which is what the full-rank corrector of § 3 provides.

5 Experiments

5.1 Setup

Models. LLaMA-2 7B, LLaMA-3 8B, Qwen3 8B (structural + stacked regimes); Qwen3.5 4B and 9B (quantization regime), vision–language models whose language decoder we compress and evaluate. Compression targets MLP projections, which hold roughly two thirds of the parameters in these models, while embeddings, lm_head, and attention are untouched, so the checkpoint-size ratios below are bounded by the MLP share (§ 6). Structural experiments preserve boundary layers (LLaMA: 0,1,30,31; Qwen3: 0–3,34,35), whereas the stacked-regime sweep preserves none and compresses all layers, which we exploit in § 5.6.

Protocols. Two WikiText-2 perplexity protocols, never mixed: **P1** (sliding window, stride 512, window 1124) matches MoDeGPT's reported dense perplexity and anchors the structural and stacked tables; **P2** (non-overlapping 2048-token segments, the GPTQ convention) anchors the quantization Pareto. Calibration: WikiText-2, 128 samples. k-means seeds 0–29 for every VQ configuration.

Operating points. D=4 throughout; K=1024 (LLaMA, stacked) or 2048 (Qwen3) for the structural tracks and K=4096 (3.0 bpw) for the quantization Pareto, with the full settings in Appendix H. Storage is byte-identical across codebook objectives. **Iso-storage LoRA ranks.** A rank-r corrector on a $d_{\text{out}} \times d_{\text{in}}$ projection stores fp16 factors at $16r(d_{\text{out}}+d_{\text{in}})/(d_{\text{out}}d_{\text{in}})$ bits per parameter; equating this to the VQ corrector's rate $\rho = \lceil \log_2 K \rceil / D + 16KD/(d_{\text{out}}d_{\text{in}})$ gives $r_{\text{iso}} = \rho\, d_{\text{out}}d_{\text{in}}/\big(16(d_{\text{out}}+d_{\text{in}})\big)$. With K=4096, D=4 this yields r=377 for Qwen3.5 4B (MLP 2560×9216, $\rho = 3.011$) and r=577 for 9B (4096×12288, $\rho = 3.005$); the measured extra storage of these LoRA runs (3.011 / 3.005 bpw) matches the VQ corrector's to three decimals.

5.2 Structural compression

At matched checkpoint sizes (measured, GiB: LLaMA-2 11.36/10.07/8.87/7.58/6.34 against 12.55 FP16), allocating budget to the corrector instead of additional rank sharply reduces WikiText-2 degradation. On LLaMA-2, where MoDeGPT reports perplexity under the same protocol (dense 5.12), its degradation at 10/20/30/40% compression is +0.36/+1.04/+2.39/+3.29 PPL against our +0.117/+0.178/+0.185/+0.388, a 3.1–12.9× reduction (geometric mean 6.7×); at 50% both methods leave the graceful regime. MoDeGPT reports no WikiText-2 perplexity for LLaMA-3 8B or Qwen3 8B, so there the iso-size comparison is against our activation-aware rank-only baseline (Appendix E),

Table 1: RC-VQ (SVD-VQ instance) across models, sizes, and domains: P1 protocol, mean ± sd over 30 seeds, WikiText-2 calibration only.

Model	Size	Rank	WikiText-2	C4	PTB	LAMBADA
LLaMA-2 7B	FP16	—	5.12	6.92	22.18	11.65
	90%	2016	5.237 ±.007	7.074 ±.004	22.775 ±.159	11.845 ±.016
	80%	1468	5.298 ±.007	7.170 ±.005	23.113 ±.167	11.942 ±.016
	70%	912	5.305 ±.007	7.181 ±.004	23.216 ±.166	11.934 ±.020
	60%	414	5.508 ±.009	7.531 ±.007	24.701 ±.236	12.339 ±.035
	50%	32	8.719 ±.115	12.076 ±.115	48.859 ±1.55	19.891 ±.528
LLaMA-3 8B	FP16	—	5.75	8.94	9.18	17.09
	90%	2096	6.029 ±.007	9.445 ±.012	9.402 ±.016	17.613 ±.039
	80%	1600	6.142 ±.008	9.660 ±.013	9.493 ±.022	17.797 ±.046
	70%	1056	6.296 ±.013	9.965 ±.014	9.633 ±.026	18.046 ±.056
	60%	544	6.488 ±.021	10.353 ±.021	9.840 ±.033	18.381 ±.053
Qwen3 8B	FP16	—	9.16	14.72	14.45	25.05
	90%	2000	9.806 ±.241	15.766 ±.243	15.539 ±.324	26.767 ±.440
	80%	1472	9.992 ±.193	16.067 ±.242	15.848 ±.277	27.235 ±.402
	70%	912	10.058 ±.291	16.173 ±.440	15.855 ±.516	27.321 ±.733
	60%	352	10.222 ±.202	16.464 ±.261	16.173 ±.335	27.747 ±.496

which collapses outright: 23.1/78.9/310.7/1495.8 PPL on LLaMA-3 and 12.4/16.9/29.8/60.8 on Qwen3 at 90–60% size, from 5× to three orders of magnitude more degradation than RC-VQ at identical storage. Within the LLaMA family the contrast is sharp, +2.2 PPL for LLaMA-2 against +17.3 for LLaMA-3 at 90% size, consistent with § 4, since the more heavily trained the model, the less low-rank slack its weights retain. Although the entire pipeline is calibrated on WikiText-2 alone, degradation relative to FP16 stays in the same range on the three held-out domains as on the calibration-matched one (LLaMA-2 at 60%: +7.6% WikiText-2, +8.8% C4, +11.4% PTB, +5.9% LAMBADA), so the corrector's benefit is not an artifact of evaluating inside the calibration domain. Zero-shot commonsense (5-task average, acc_norm): LLaMA-2 declines 0.5–1.6 pp from dense 68.85 down to 60% size and collapses at 50% (−8.3 pp, the practical limit of MLP-only compression); LLaMA-3 declines 0.65–1.44 pp from 72.77 and Qwen3 1.9–3.3 pp from 71.47. Against MoDeGPT's zero-shot at equal size the margin grows with compression, +3.5/+7.2/+15.3 pp at 80/70/60% on LLaMA-2 and +10.9 pp on LLaMA-3 at 70% even against its Alpaca-calibrated variant. Seed sensitivity is model-dominated: LLaMA $\sigma \leq 0.021$ PPL from 90–60%, Qwen3 0.19–0.29 at every size.

5.3 Rate allocation

Where the bits go dominates quality. Splitting the 90%-size budget three ways (rank-only / D=8 / D=4), perplexity falls monotonically as budget moves from SVD factors to the corrector (7.24 → 5.58 → 5.24), and the ordering holds at all five sizes. This budget-split series uses a uniform-rank preset, whereas the equal-energy iso-size rank-only point of Appendix A9 evaluates at 7.28. K and D are two knobs on one rate ρ, and configurations at equal ρ perform alike, so we tune ρ. For layer-wise codebook capacity, tail entropy predicts demand no worse than tail energy (5.237 vs 5.251), and in these models the normalized signal is flat and allocation converges to uniform.

5.4 Quantization gap recovery

On Qwen3.5 4B, FP16 evaluates at 9.585 and INT4-AWQ at 9.735, a gap of +0.150; on 9B, 8.649 versus 8.814, a gap of +0.165. RTN-8 on the residual restores FP16 on both models, confirming the ceiling. The 4B Pareto:

Three points follow. (i) Iso-storage versus low-rank. At exactly matched budgets (3.01 bpw), PolarVQ recovers 96% of the gap where LoRA recovers 39%, and the LoRA curve stays nearly flat across 2.5–4.5 bpw (39→42%) because the energy it recovers grows only linearly in rank (§ 4). (ii) Iso-rate

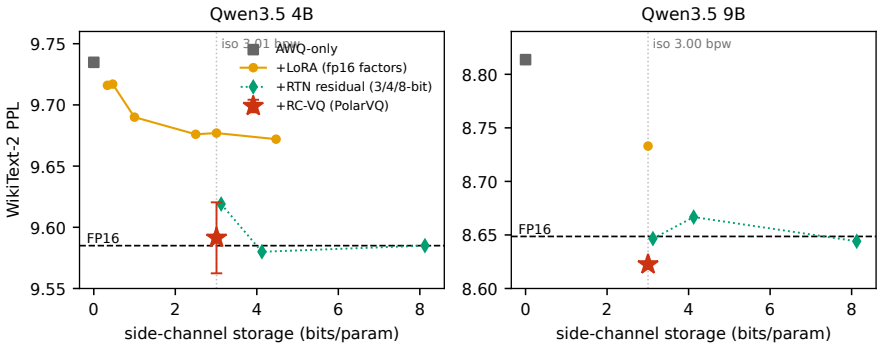


Figure 2: Quantization-gap Pareto on Qwen3.5 4B (left) and 9B (right): perplexity versus corrector storage for LoRA, scalar residual branches, and PolarVQ.

versus scalar. At 3.1 bpw the vector codebook recovers 19 pp more of the gap than the scalar grid, which needs 4.13 bpw, 37% more, to reach the ceiling. (iii) Scale. The 9B gap is larger in absolute terms and low-rank correction fares worse at small budgets, where LoRA at $r \leq 125$ lands behind the uncorrected base, turns helpful only from $r \approx 313$, and settles into a flat 49–51% band around the matched budget (49% at r=577, 3.01 bpw). At the same 3.01 bpw, PolarVQ on 9B reaches 8.623 $\pm$ .040, closing the entire WikiText-2 gap, and the purely directional variant reaches 8.581, slightly below the FP16 reference, while the scalar RTN-3 branch sits at the ceiling (8.647) with 4% more bits. The sub-FP16 WikiText-2 perplexity does not transfer to held-out domains, since re-evaluated under the same protocol the directional corrector lands 0.034 above FP16 on C4 and at FP16 on PTB (Appendix D), so we read it as a calibration-domain effect rather than a regularization gain. What does transfer is the recovery itself, which closes 86–101% of the AWQ gap on 9B across C4 and PTB. The pipeline transfers to a VLM language backbone without modification on both models.

5.5 Recovering visual behavior (VisualWebBench)

Text-calibrated compression can discard capability that the calibration proxy never probes. Testing this requires a task outside the calibration modality, so we evaluate both Qwen3.5 models on VisualWebBench across seven subtasks, comparing FP16, AWQ-only, AWQ+RC-VQ (one representative seed per model, chosen by WikiText-2 PPL before this benchmark), and AWQ+LoRA at iso-storage (r=377 / r=577), all under greedy decoding with fp16 lm_head.

Raw scores (appendix) show that text-calibrated quantization perturbs visual tasks in both directions, up to +16.4 pp on one subtask, so benchmark score alone is the wrong axis, and what the corrector should restore is the FP16 model's behavior. Measuring behavioral fidelity as the mean absolute per-subtask deviation from FP16 (Table 3, appendix), RC-VQ lowers drift from 1.72 pp (AWQ-only) and 2.09 pp (LoRA) to 1.21 pp on 4B, and from 3.97 and 4.62 to 0.36 pp on 9B. On 9B the corrector thus restores the FP16 model almost exactly across seven heterogeneous subtasks, with two subtasks (webqa, action_prediction) matching FP16 to the reported precision, which under greedy decoding is evidence of answer-level behavioral restoration rather than aggregate coincidence. The iso-storage low-rank corrector increases drift on both models (and on 4B leaves one subtask's outputs bit-identical to uncorrected AWQ), consistent with § 4, since the perturbation it would need to undo is diffuse, and a rank-limited correction cannot reach it. This is the opening claim made concrete. The residual carries behavior that a text-only calibration set never certified, and a corrector fitted from the weights alone, without touching any data, returns it.

5.6 Codebook objective vs residual geometry

Across five models and three residual families (30 seeds per cell), purely directional clustering collapses on every composite regime (+0.49 LLaMA-2, +4.60 Qwen3, +5.35 LLaMA-3, ordered by residual tail weight) and trails by 0.20–0.37 on whitened SVD residuals, where Euclidean and hierarchical clustering tie to within 0.0001 PPL; purely Euclidean clustering wins no quantization regime (behind in four models, tied in LLaMA-3). PolarVQ never collapses, and it is within 0.05 PPL

Table 2: Codebook objective vs residual geometry. WikiText-2 PPL, mean ± sd over 30 seeds (structural-row cos: 25 seeds); best per row in bold; PolarVQ = cos→MSE.

Regime	Model	MSE	Directional (cos)	PolarVQ	PolarVQ − best
Composite (SVD∘INT4, P1)	LLaMA-2 7B	**5.874 ± .061**	6.365 ± .049	5.893 ± .055	+0.019
	LLaMA-3 8B	**7.829 ± .146**	13.179 ± .473	7.871 ± .166	+0.043
	Qwen3 8B	**10.418 ± .340**	15.015 ± .630	10.526 ± .352	+0.108 (< 1 s.e.)
Quantization (INT4-AWQ, P1)	LLaMA-2 7B	5.149 ± .022	**5.128 ± .003**	5.146 ± .020	+0.018
	LLaMA-3 8B	5.791 ± .010	5.786 ± .002	**5.786 ± .010**	0
	Qwen3 8B	9.192 ± .079	**9.138 ± .037**	9.178 ± .089	+0.040
Quantization (INT4-AWQ, P2)	Qwen3.5 4B	9.644 ± .070	9.636 ± .019	**9.591 ± .029**	0
	Qwen3.5 9B	8.645 ± .061	**8.581 ± .015**	8.623 ± .040	+0.041
Structural (whitened SVD, P1)	LLaMA-2 7B (90%)	5.237 ± .005	5.432 ± .006	**5.237 ± .007**	0

of the per-regime winner in nine of ten cells (at most 0.043, LLaMA-3 composite) and, on the seed-noisiest Qwen3 composite regime (σ=0.34), trails MSE by 0.11 PPL, within one standard error. On quantization residuals, directional objectives are also the most seed-stable (e.g., σ=0.0028 vs 0.0222 for MSE on LLaMA-2), whereas on composite residuals the collapsed directional objective has the largest variance instead. The usage rule is therefore PolarVQ by default, and the pure directional objective only when the residual is known to be magnitude-homogeneous.

5.7 Inference cost

On the current unfused PyTorch path (LLaMA-2 7B, prompt 512 / generate 128) compression buys memory, not speed, as the hybrid schedule cuts decode memory from 33.7 to 23.1 GiB (31%) at 24.4 versus 74.5 tok/s (Appendix G). The sub-vector layout (§ 3.2) is designed for the LUT-fused kernel that removes this penalty, and re-quantizing the VQ branch onto an INT grid loses recovery to double quantization, so the fp16-codebook LUT path is the right deployment shape. Selective application per projection is a complementary lever we are exploring.

6 Conclusion

Calibration decides what training-free compression keeps, and whatever it cannot certify as important is discarded as a residual. We show that this residual, while spectrally featureless, is functionally valuable and recoverable. Because its spectrum carries no recoverable spikes, a Marchenko–Pastur noise bulk for quantization error and a gapless tail continuum for truncation error, the standard low-rank corrector is the wrong tool. It recovers residual energy only linearly in rank and closes under half of the quantization gap even beyond the VQ corrector's budget. A full-rank vector-quantized corrector, fitted from the residual alone without gradients or data, is the right one, since it recovers 96% of the same gap at equal storage, cuts structural-compression degradation by up to 13×, and composes unchanged across SVD factors, INT4 weights, and their stacked combination.

The corrector's most consequential design choice is the codebook objective, and we resolve it empirically. Pure objectives each have a failure regime, while the hierarchical cosine→MSE objective never collapses across five models and three residual families, staying within 0.05 perplexity of the per-regime winner in nine of ten cells and within one standard error in the tenth. Two limitations remain. We compress only MLP projections, which bounds the achievable ratio, and all configurations degrade sharply near 50% size, but attention projections expose the same residual interface and are the immediate extension. The unfused VQ decode path trades throughput for memory, a penalty that a LUT-fused kernel folding codebook gathers into the matrix multiply would remove.

Ethics Statement

This work studies post-training compression of publicly released language models and involves no human subjects, no personal data, and no data collection. All models (LLaMA-2, LLaMA-3, Qwen3, Qwen3.5) and evaluation sets (WikiText-2, C4, PTB, LAMBADA, the zero-shot commonsense suite, VisualWebBench) are used under their respective licenses. Compression makes deployment of these models cheaper, which lowers the barrier to both beneficial and harmful uses of the underlying

models, but the corrector introduces no capability beyond restoring behavior the uncompressed model already has. We are not aware of any ethical concern specific to this work.

REPRODUCIBILITY STATEMENT

Section 5.1 specifies the models, the two perplexity protocols (P1 and P2), the calibration set, and the seed range used for every vector-quantized configuration. Appendix H lists the operating points (K, D, G), the representative-seed rule used for VisualWebBench, and the software and hardware environment. The iso-storage rank of every low-rank baseline follows the closed-form expression in Section 5.1, and all reported numbers are means over the stated number of seeds. Every fixed choice of the pipeline (whitening transform, rank and codebook allocation, AWQ search grid) is stated in Section 3. The compression and evaluation code and all sweep configurations are provided as supplementary material.

AI USE STATEMENT

We used large language model assistants in four ways during this work. First, for writing assistance, to polish the grammar, wording, and concision of author-written text. Second, for literature retrieval, to help locate and summarize related work, with every cited paper checked against its original source by the authors. Third, for drafting, where the assistant produced initial drafts of some sections that the authors then rewrote and verified. Fourth, for research support, including discussion of experimental design, writing and debugging of experiment and plotting scripts, and cross-checking of reported numbers against experiment logs. All experiments were designed and run by the authors, every number in the paper comes from the authors' own runs, and no synthetic data or mathematical proofs were generated by an assistant. The authors take full responsibility for the content of this paper.

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 (NeurIPS)*, 2024.

Haolei Bai, Siyong Jian, Tuo Liang, Yu Yin, and Huan Wang. ERC-SVD: Error-controlled SVD for large language model compression. *arXiv preprint arXiv:2505.20112*, 2025.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research (TMLR)*, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *International Conference on Machine Learning (ICML)*, 2024.

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.

Insu Han, Praneeth Kacham, Amin Karbasi, Vahab Mirrokni, and Amir Zandieh. PolarQuant: Quantizing KV caches with polar transformation. *arXiv preprint arXiv:2502.02617*, 2025.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations (ICLR)*, 2022.

Xing Hu, Dawei Yang, Yuan Cheng, Zhixuan Chen, and Zukang Xu. SAES-SVD: Self-adaptive suppression of accumulated and local errors for SVD-based LLM compression. In *International Conference on Learning Representations (ICLR)*, 2026.

Andrey Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Chi-Heng Lin, Shangqian Gao, James Seale Smith, Abhishek Patel, Shikhar Tuli, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. MoDeGPT: Modular decomposition for large language model compression. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (MLSys)*, 2024.

Yifei Liu, Jicheng Wen, Yang Wang, Shengyu Ye, Li Lyna Zhang, Ting Cao, Cheng Li, and Mao Yang. VPTQ: Extreme low-bit vector post-training quantization for large language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 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 (ICLR)*, 2025.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Rajarshi Saha, Naomi Sagan, Varun Srivastava, Andrea J. Goldsmith, and Mert Pilanci. Compressing large language models using low rank and low precision decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Nikolai Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2):279–281, 1948.

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.

Mart van Baalen, Andrey Kuzmin, Ivan Koryakovskiy, Markus Nagel, Peter Couperus, Cedric Bastoul, Eric Mahurin, Tijmen Blankevoort, and Paul Whatmough. GPTVQ: The blessing of dimensionality for LLM quantization. *arXiv preprint arXiv:2402.15319*, 2024.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations (ICLR)*, 2025.

Songhao Wu, Ang Lv, Xiao Feng, Yufei Zhang, Xun Zhang, Guojun Yin, Wei Lin, and Rui Yan. PolarQuant: Leveraging polar transformation for key cache quantization and decoding acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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.

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. ZeroQuant-V2: Exploring post-training quantization in LLMs from comprehensive study to low rank compensation. *arXiv preprint arXiv:2303.08302*, 2023.

Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. ASVD: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

Yuxuan Yue, Zukang Xu, Zhihang Yuan, Dawei Yang, Jianlong Wu, and Liqiang Nie. PCDVQ: Enhancing vector quantization for large language models via polar coordinate decoupling. *arXiv preprint arXiv:2506.05432*, 2025.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

Cheng Zhang, Jianyi Cheng, George A. Constantinides, and Yiren Zhao. LQER: Low-rank quantization error reconstruction for LLMs. In *International Conference on Machine Learning (ICML)*, 2024.

Table 3: Behavioral fidelity on VisualWebBench: mean absolute per-subtask deviation from FP16 (pp), lower is better.

mean $\lvert\Delta\rvert$ vs FP16 (pp)	AWQ-only	+LoRA (iso)	+RC-VQ
Qwen3.5 4B	1.72	2.09	**1.21**
Qwen3.5 9B	3.97	4.62	**0.36**

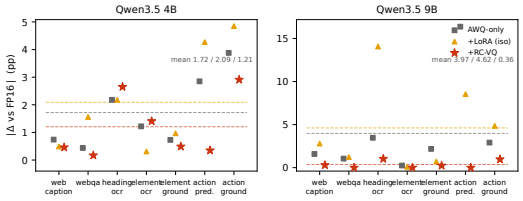


Figure 3: Per-subtask deviation from FP16 on VisualWebBench for AWQ-only, AWQ+LoRA (iso-storage), and AWQ+RC-VQ.

A VisualWebBench raw scores

Per-subtask scores for FP16, AWQ-only, AWQ+LoRA, and AWQ+RC-VQ on both models are reported in the supplementary material.

B Zero-shot commonsense accuracy

Per-task accuracies (ARC-e, ARC-c, HellaSwag, PIQA, WinoGrande) across models and sizes are reported in the supplementary material; Section 5.2 gives the five-task averages.

C Quantization Pareto (numbers)

D Cross-domain re-evaluation

E Iso-size rank-only baseline

F Rate–distortion of flat residuals

G Inference cost

LLaMA-2 7B, prompt 512 / generate 128, unfused PyTorch. FP16: 74.5 tok/s at 33.7 GiB decode memory. Compressed model: 57.5 tok/s / 28.2 GiB with the corrector materialized; 24.4 tok/s / 23.1 GiB in the hybrid schedule (31% memory reduction); 16.3 tok/s / 21.4 GiB reconstructing on the fly.

H Reproduction details

Operating points. Structural: K_{ref}=1024 for LLaMA and 2048 for Qwen3, D=4. Stacked sweep: K=1024, D=4, G=4 (2.5 bpw). Quantization Pareto: K=4096, D=4, G=4 (3.0 bpw). Representative seeds for VisualWebBench: WikiText-2 PPL closest to the 30-seed mean, selected independently of the benchmark. Environment: PyTorch 2.7.0+cu126, one H100 80GB, lm-eval-harness 0.4.8.

Table 4: Quantization-gap recovery on Qwen3.5 4B (P2). Side bpw is the corrector's storage per parameter; gap recovered is the fraction of the AWQ-to-FP16 perplexity gap closed.

corrector	side bpw	energy recovered	PPL	gap recovered
— (AWQ only)	0	0%	9.735	0%
LoRA r=42 / 58 / 125	0.34 / 0.46 / 1.00	2.2 / 3.1 / 6.5%	9.716 / 9.717 / 9.690	13 / 12 / 30%
LoRA r=313	2.50	15.4%	9.676	39%
LoRA r=377 (iso-storage)	3.01	18.2%	9.677	39%
LoRA r=560	4.47	25.6%	9.672	42%
PolarVQ (K4096/D4)	**3.01**	84.4%	**9.591 ± .029**	**96%**
RTN-3 residual	3.13	82.9%	9.619	77%
RTN-4 residual	4.13	92.7%	9.580	ceiling

Table 5: Cross-domain re-evaluation of the quantization track (P2, non-overlapping 2048-token segments). All twelve checkpoints re-evaluated by a single implementation; WT2 columns agree with the P2 numbers of Section 5.4 to within 0.003 for FP16 and AWQ. Corrector rows use seed 0. C4: allenai/c4 validation shard 0, first 1100 documents; PTB: test split.

	FP16	AWQ-only	+PolarVQ cos	+PolarVQ cos_mse
Qwen3.5 4B WT2	9.5876	9.7376	9.6344	9.6201
C4	13.8011	14.1939	13.8554	13.8152
PTB	17.7292	17.8868	17.8441	17.8264
Qwen3.5 9B WT2	8.6485	8.8145	8.5898	8.6400
C4	12.0356	12.2754	12.0700	12.0419
PTB	13.5706	13.8871	13.5671	13.5954

Table 6: Iso-size rank-only baseline (P1, seed 0). rank-only rank = Table 1 rank $+\Delta r$ (Section 5.1). Checkpoint sizes are the packed deployable format, measured for both methods at all twelve points; at each point the two checkpoints match to within 0.01%, the residual being a constant per-model bookkeeping offset (codebook tables and index metadata, at most 1.4 MB). At the 70% point the iso rank is set by the measured bytes of the RC-VQ checkpoint, whose allocator stores 47 ranks above nominal at that budget. † MoDeGPT as reported (WikiText-2-calibrated, same protocol; not reported for LLaMA-3/Qwen3).

Model	Size	RC-VQ PPL (30-seed mean)	rank-only PPL	rank-only rank	RC-VQ ckpt (B, packed)	rank-only ckpt (B)	MoDeGPT†
LLaMA-2 7B	90%	5.237	7.28	2483	12,207,111,468	12,206,664,454	5.48
	80%	5.298	10.91	1935	10,816,576,749	10,816,129,798	6.16
	70%	5.305	22.75	1426	9,524,788,264	9,524,556,531	7.51
	60%	5.508	64.25	881	8,142,080,860	8,141,634,140	8.41
LLaMA-3 8B	90%	6.029	23.09	2594	14,246,751,141	14,245,381,762	—
	80%	6.142	78.92	2098	12,710,849,301	12,709,480,042	—
	70%	6.296	310.68	1554	11,026,311,811	11,024,942,672	—
	60%	6.488	1495.77	1042	9,440,864,543	9,439,495,716	—
Qwen3 8B	90%	9.806	12.43	2529	14,795,955,507	14,796,095,413	—
	80%	9.992	16.85	2001	13,238,820,000	13,238,960,044	—
	70%	10.058	29.80	1441	11,587,312,602	11,587,452,792	—
	60%	10.222	60.80	881	9,935,804,929	9,935,945,538	—

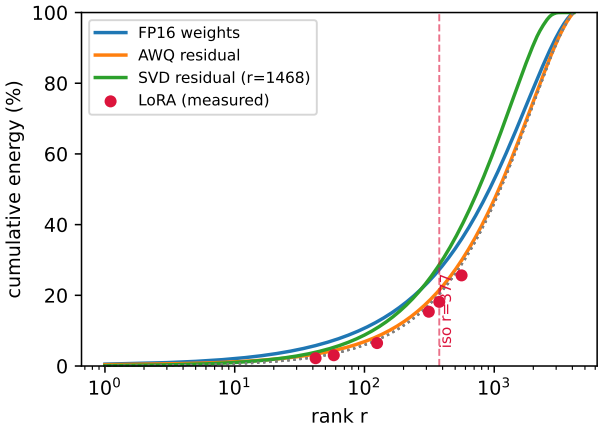


Figure 4: Cumulative residual energy versus rank with the measured LoRA points (Qwen3.5 4B): energy recovered grows linearly in rank, the signature of a spectrum with no privileged directions.