QUANTRET: QUANTIZATION-ROBUST LANGUAGE MODEL PRETRAINING VIA GEOMETRIC RETRACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit post-training quantization (PTQ) can fail when a small number of coordinates set the quantization range or when high-gain directions amplify local errors. We introduce QuantRet, a pretraining method for residual Transformer decoders. QuantRet retracts selected activations to fixed-radius spheres and constrains the norms of hidden weight matrices throughout pretraining. These operations fix the radial scale of selected activations and constrain how hidden linear maps rescale their inputs without changing the quantizer. Our analysis connects matrix geometry to directional perturbations and bounds their injection through spherical residual mixing and subsequent propagation under explicit local sensitivity conditions. On Qwen3-like models, QuantRet lowers language-modeling loss under round-to-nearest quantization of weights, activations, and the key–value cache, and its strongest variants achieve higher average zero-shot task scores under the same quantization settings. Mechanism measurements reveal differences in local error amplification and depthwise propagation that complement local reconstruction statistics. These results support geometric constraints during pretraining as a route to improved low-bit PTQ quality.

1 INTRODUCTION

Post-training quantization (PTQ) converts a pretrained model to lower precision without repeating pretraining. At 4 bits, a few extreme values can determine the quantization range and leave little resolution for other values, making joint quantization of weights, activations, and the key–value cache challenging (Xiao et al., 2023; Ashkboos et al., 2024). The resulting prediction error also depends on how the network amplifies and propagates local perturbations (Vlassis et al., 2026). Pretraining shapes both the tensors presented to the quantizer and the network's sensitivity to their quantization errors. This motivates designing pretraining methods that improve the quality of the eventual low-bit model.

PTQ methods mitigate these difficulties by retaining exceptional activation channels at higher precision, redistributing scale between activations and weights, reconstructing quantized weights using approximate second-order information, or rotating hidden coordinates (Dettmers et al., 2022; Xiao et al., 2023; Frantar et al., 2023; Ashkboos et al., 2024; Liu et al., 2025). A complementary approach changes pretraining itself. Outlier-Safe Pre-Training (OSP) combines optimizer, normalization, and embedding modifications to suppress outlier formation (Park et al., 2025). Work on normalized architectures shows that constraining weights and representations can improve robustness to low-precision computation (Fishman et al., 2026). Manifold-constrained optimization further provides tools for controlling weight norms and activation scales during pretraining (An et al., 2026). These results motivate coordinating activation normalization and matrix constraints for low-bit PTQ.

For a residual Transformer, the design question is which geometric quantities to constrain and where to constrain them. Activation norms determine overall scale, while matrix geometry determines how input perturbations affect branch outputs. The roles of these matrices differ: reader projections map the residual stream into branch features, and writer projections return those features to the residual stream. Their errors then interact with normalization and residual mixing before reaching later layers. We therefore consider activation radii, matrix constraints, and residual updates together when designing for quantization robustness.

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.

We introduce **QuantRet**, a pretraining framework that combines matrix constraints with spherical activation retractions and residual updates for low-bit PTQ. QuantRet normalizes selected activations to fixed radii and retracts hidden weight matrices after each optimizer update. Its three variants constrain total matrix energy (Frob), maximum directional gain (Spec), or reader-row and writer-column norms (Obl). The variants share the same forward graph and optimizer, allowing us to compare these matrix constraints within the same framework. Learnable scales control attention scores, MLP gates, residual mixing, and vocabulary logits. Quantization is applied only after pretraining, allowing the resulting checkpoint to be used with different PTQ methods.

Our analysis examines how matrix geometry affects normalized branch directions, how quantization errors enter the residual stream, and how they propagate. We characterize the effects of matrix constraints on perturbation energy, relate a writer's output error to the change in its normalized direction, and bound the error introduced by spherical residual mixing. A conditional propagation bound then relates these local errors to final-state error and language-modeling loss. This analysis identifies how mixing coefficients, normalization-input norms, and the sensitivity of later sublayers govern error amplification or attenuation. Controlled perturbations and full-model quantization measurements examine these effects empirically.

Contributions. Our main contributions are:

- **Pretraining framework.** We introduce QuantRet, which combines spherical activation and residual retractions with three alternative hidden-matrix constraints. The variants control total matrix energy, maximum directional gain, or reader-row and writer-column norms, with quantization applied only after pretraining.
- **Error analysis.** We characterize finite errors after normalization and derive conditional bounds linking normalized branch errors, spherical residual mixing, and subsequent propagation to CE changes and output KL divergence. Controlled perturbations and quantization measurements examine the corresponding sensitivity and prediction changes.
- **Empirical results.** Across 0.3B, 0.6B, and 1.7B Qwen3-like models, all three variants achieve lower CE and PPL than AdamW and OSP under 4-bit RTN. At each scale, QuantRet also achieves the highest observed eight-task zero-shot average. An additional 0.3B Llama 3.1 architecture study reproduces the language-modeling improvements over both baselines.

2 QuantRet: Geometric Retractions for Quantization-Robust Pretraining

QuantRet coordinates activation radii, matrix geometry, and vocabulary maps in residual Transformer decoders. Selected activations are retracted to fixed-radius spheres, and hidden matrices are retracted to norm-constrained sets after each optimizer step. Learned orthogonal maps connect the vocabulary tables to the residual stream. A stop-gradient rule separates forward logit scaling from its local backward multiplier. Throughout, "retraction" denotes explicit normalization to a fixed-norm equality set. Activation retractions and learnable scales remain in the inference graph. Parameter retractions and the modified backward rule act during training.

2.1 Hyperspherical Transformer Block

In this subsection, we define the attention and feed-forward branches and their shared spherical residual update.

Let $d = D_{\text{model}}$ be the residual width, D_{head} the attention-head width, D_{ff} the feed-forward width, and $L = N_{\text{block}}$ the number of blocks. For a sequence of N_{tok} tokens, $\mathbf{H}_\ell \in \mathbb{R}^{N_{\text{tok}} \times d}$ denotes the state entering block $\ell \in \{0, \dots, L-1\}$, and $\mathbf{Z}_\ell$ denotes its post-attention state. Each row of $\mathbf{H}_0$ has norm $\sqrt{d}$, which the residual updates preserve. We use the batched linear convention

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^{\mathsf{T}}, \qquad \mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}, \tag{1}$$

so rows of $\mathbf{W}$ index output features and columns index input coordinates. Optional bias terms are omitted from the notation; matrix constraints apply only to $\mathbf{W}$. For a nonzero vector $\mathbf{x} \in \mathbb{R}^D$, define

the spherical retraction

$$\mathcal{S}_r(\mathbf{x}) = r\frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \qquad \mathrm{Sph}_D(\mathbf{x}) = \mathcal{S}_{\sqrt{D}}(\mathbf{x}), \tag{2}$$

and apply Sph_D row-wise to token matrices or along a tensor's feature axis. This fixes component RMS to one without a learned channel-wise affine transformation and replaces separate LayerNorm or RMSNorm operations on the residual branches. A fixed radius alone does not guarantee low coordinate concentration or outlier suppression.

Both branches use the same spherical residual update. For a residual state $\mathbf{H}$ and a branch output $\mathbf{F}$ of the same shape, define

$$\mathcal{R}_\lambda(\mathbf{H}, \mathbf{F}) = \mathrm{Sph}_d((1-\lambda)\mathbf{H} + \lambda\,\mathrm{Sph}_d(\mathbf{F})), \qquad 0 \le \lambda \le 1. \tag{3}$$

The learned, channel-shared coefficients α_ℓ and β_ℓ control interpolation toward the normalized attention and feed-forward branch directions, respectively.

Causal self-attention. Let N_{qh} and N_{kvh} denote the numbers of query and key–value heads, with $D_q = N_{\mathrm{qh}}D_{\mathrm{head}}$ and $D_{kv} = N_{\mathrm{kvh}}D_{\mathrm{head}}$. Compute $\mathbf{Q}_\ell = \mathbf{H}_\ell\mathbf{W}_{Q,\ell}^\mathsf{T} \in \mathbb{R}^{N_{\mathrm{tok}}\times D_q}$, $\mathbf{K}_\ell = \mathbf{H}_\ell\mathbf{W}_{K,\ell}^\mathsf{T} \in \mathbb{R}^{N_{\mathrm{tok}}\times D_{kv}}$, and $\mathbf{V}_\ell = \mathbf{H}_\ell\mathbf{W}_{V,\ell}^\mathsf{T} \in \mathbb{R}^{N_{\mathrm{tok}}\times D_{kv}}$. After reshaping into heads, apply the backbone's query/key positional transformations $\mathcal{P}_{\ell,h}^Q$ and $\mathcal{P}_{\ell,j}^K$, such as RoPE (Su et al., 2024). Each resulting query and key vector is retracted along its head-feature axis (Henry et al., 2020):

$$\bar{\mathbf{Q}}_\ell^{(h)} = \mathrm{Sph}_{D_{\mathrm{head}}}\left(\mathcal{P}_{\ell,h}^Q(\mathbf{Q}_\ell^{(h)})\right), \qquad \bar{\mathbf{K}}_\ell^{(j)} = \mathrm{Sph}_{D_{\mathrm{head}}}\left(\mathcal{P}_{\ell,j}^K(\mathbf{K}_\ell^{(j)})\right). \tag{4}$$

Let $g_\ell : [N_{\mathrm{qh}}] \to [N_{\mathrm{kvh}}]$ encode the backbone's head sharing. A positive learnable score scale $\tau_{\ell,h}$ controls attention sharpness for each query head. Causal scaled dot-product attention (Vaswani et al., 2017) is

$$\mathbf{A}_\ell = \mathrm{Concat}_{h=1}^{N_{\mathrm{qh}}}\left[\mathrm{softmax}\left(\frac{\tau_{\ell,h}}{\sqrt{D_{\mathrm{head}}}}\bar{\mathbf{Q}}_\ell^{(h)}\bar{\mathbf{K}}_\ell^{(g_\ell(h))\mathsf{T}} + \mathcal{M}_{\mathrm{causal}}\right)\mathbf{V}_\ell^{(g_\ell(h))}\right]\mathbf{W}_{O,\ell}^\mathsf{T}. \tag{5}$$

Here $\mathcal{M}_{\mathrm{causal}}$ is the causal mask. Values remain unnormalized during attention aggregation. The attention residual update is

$$\mathbf{Z}_\ell = \mathcal{R}_{\alpha_\ell}(\mathbf{H}_\ell, \mathbf{A}_\ell), \qquad 0 \le \alpha_\ell \le 1. \tag{6}$$

Feed-forward branch. For an ungated or gated feed-forward branch, define the intermediate features as

$$\Psi_\ell(\mathbf{Z}_\ell) = \begin{cases} \phi_\ell(\mathbf{Z}_\ell\mathbf{W}_{1,\ell}^\mathsf{T}), & \text{ungated MLP,} \\ \phi_{g,\ell}(\gamma_{g,\ell}\mathbf{Z}_\ell\mathbf{W}_{G,\ell}^\mathsf{T}) \odot \phi_{u,\ell}(\mathbf{Z}_\ell\mathbf{W}_{U,\ell}^\mathsf{T}), & \text{gated MLP.} \end{cases} \tag{7}$$

The backbone supplies ϕ_ℓ, $\phi_{g,\ell}$, and $\phi_{u,\ell}$; for SwiGLU, $\phi_{g,\ell} = \mathrm{SiLU}$ and $\phi_{u,\ell} = \mathrm{id}$. The learnable gate scale $\gamma_{g,\ell}$ is a positive channel-shared scalar. The intermediate features pass directly through the down projection before the spherical residual update:

$$\mathbf{F}_\ell = \Psi_\ell(\mathbf{Z}_\ell)\mathbf{W}_{D,\ell}^\mathsf{T}, \qquad \mathbf{H}_{\ell+1} = \mathcal{R}_{\beta_\ell}(\mathbf{Z}_\ell, \mathbf{F}_\ell), \qquad 0 \le \beta_\ell \le 1. \tag{8}$$

The retractions in $\mathcal{R}_{\beta_\ell}$ normalize the branch output and residual state.

The normalized final state $\mathbf{H}_L$ enters the vocabulary map defined below. Appendix B specifies architecture-dependent implementation details, including the additional pre-down normalization retained in evaluated checkpoints (Appendix B.1).

2.2 Orthogonal Vocabulary Maps and Logit Scaling

We learn orthogonal maps at the input embedding and LM head and decouple forward logit scaling from its local backward multiplier. Let $\mathbf{E}, \mathbf{M} \in \mathbb{R}^{N_{\mathrm{vocab}}\times d}$ be vocabulary tables with unit row norms, and let $\mathbf{R}_{\mathrm{in}}, \mathbf{R}_{\mathrm{out}} \in \mathrm{O}(d)$, where $\mathrm{O}(d) = \{\mathbf{R} : \mathbf{R}^\mathsf{T}\mathbf{R} = \mathbf{I}\}$. The effective tables are

$$\mathbf{E}^\star = \sqrt{d}\,\mathbf{E}\mathbf{R}_{\mathrm{in}}^\mathsf{T}, \qquad \mathbf{M}^\star = \sqrt{d}\,\mathbf{M}\mathbf{R}_{\mathrm{out}}^\mathsf{T}. \tag{9}$$

Each effective row has norm $\sqrt{d}$. The rotations learn how each vocabulary table is oriented relative to the residual coordinates while preserving its row norms and pairwise angles. For token indices $\mathbf{X}$, the input state is $\mathbf{H}_0 = \mathbf{E}^\star[\mathbf{X}]$, and the raw vocabulary scores are $\mathbf{L} = \mathbf{H}_L(\mathbf{M}^\star)^\mathsf{T}$. At deployment, the rotations are absorbed into these tables.

Logit scaling with gradient decoupling. We learn an unconstrained vector $\mathbf{s} \in \mathbb{R}^{N_{\text{vocab}}}$ to scale the logits for each vocabulary entry. We initialize $s_v(0) = c_{\text{bp}} = 1/\sqrt{d}$ using variance matching for independent isotropic vectors of radius $\sqrt{d}$. The logits are computed as

$$z_{t,v} = \text{sg}(L_{t,v})s_v + c_{\text{bp}}L_{t,v} - \text{sg}(c_{\text{bp}}L_{t,v}), \qquad c_{\text{bp}} = \frac{1}{\sqrt{d}}, \tag{10}$$

where sg denotes the stop-gradient operator, which is the identity in the forward pass and has zero derivative. The forward logits are $z_{t,v} = s_v L_{t,v}$, and the local derivatives are

$$\frac{\partial z_{t,v}}{\partial L_{t,v}} = \frac{1}{\sqrt{d}}, \qquad \frac{\partial z_{t,v}}{\partial s_v} = L_{t,v}. \tag{11}$$

The learned vector $\mathbf{s}$ controls logit scaling, and c_{bp} provides a shared, fixed local backward multiplier for the raw scores. Appendix B.3 specifies parameter sharing and the numerical implementation.

2.3 Matrix-Retraction Families

In this subsection, we define three matrix constraints that control total energy, maximum directional gain, or reader–writer norm balance. *Readers* map the residual stream to branch features (query, key, value, and feed-forward input projections). *Writers* map branch features back to the residual stream (attention and feed-forward output projections). For $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the three QuantRet variants share the forward graph and optimizer but use different post-step retractions:

$$\text{Retr}_{\text{Frob}}(\mathbf{W}) = \sqrt{d_{\text{out}}}\frac{\mathbf{W}}{\|\mathbf{W}\|_F}, \qquad \|\text{Retr}_{\text{Frob}}(\mathbf{W})\|_F = \sqrt{d_{\text{out}}}, \tag{12}$$

$$\text{Retr}_{\text{Spec}}(\mathbf{W}) = \sqrt{d_{\text{out}}/d_{\text{in}}}\frac{\mathbf{W}}{\|\mathbf{W}\|_2}, \qquad \|\text{Retr}_{\text{Spec}}(\mathbf{W})\|_2 = \sqrt{d_{\text{out}}/d_{\text{in}}}, \tag{13}$$

$$[\text{Retr}^{\text{read}}_{\text{Obl}}(\mathbf{W})]_{i:} = \frac{\mathbf{W}_{i:}}{\|\mathbf{W}_{i:}\|_2}, \qquad [\text{Retr}^{\text{write}}_{\text{Obl}}(\mathbf{W})]_{:j} = \sqrt{d_{\text{out}}/d_{\text{in}}}\frac{\mathbf{W}_{:j}}{\|\mathbf{W}_{:j}\|_2}. \tag{14}$$

These define *QuantRet-Frob*, *QuantRet-Spec*, and *QuantRet-Obl*: Frob fixes total matrix energy, Spec fixes the largest directional gain, and Obl equalizes reader-row and writer-column norms within each matrix. Frob and Obl both satisfy $\|\mathbf{W}\|_F^2 = d_{\text{out}}$. Spec imposes an equality constraint. Under its ideal constraint, an input of norm at most $\sqrt{d_{\text{in}}}$ maps to an output of norm at most $\sqrt{d_{\text{out}}}$.

2.4 Optimization with Hard Retractions

Muon's Newton–Schulz matrix-sign direction approximately equalizes update magnitudes across singular directions (Jordan et al., 2024). This motivates its use to limit anisotropy in hidden-matrix updates; the component ablation compares the FP and quantized quality of the selected Muon and projected AdamW training recipes.

Let $\mathcal{O}_t$ denote the optimizer state, η_t the step size, and $\mathbf{G}_t$ the matrix gradient supplied to the optimizer under the logit backward rule in Equation (10). QuantRet takes an ambient Muon step and then retracts the updated matrix:

$$(\widetilde{\mathbf{W}}_{t+1}, \mathcal{O}_{t+1}) = \texttt{MuonStep}(\mathbf{W}_t, \mathbf{G}_t, \mathcal{O}_t, \eta_t), \tag{15}$$

$$\mathbf{W}_{t+1} = \text{Retr}_v(\widetilde{\mathbf{W}}_{t+1}), \qquad v \in \{\text{Frob}, \text{Spec}, \text{Obl}\}. \tag{16}$$

The retraction acts on the matrix parameters, while gradients, momentum, and optimizer state remain in ambient coordinates. Vocabulary tables and learnable scales use auxiliary AdamW (Loshchilov & Hutter, 2019), with row retraction on the tables. The orthogonal maps use Cayley updates (Liu et al., 2025). QuantRet-Spec estimates the spectral norm numerically, so its exact-norm constraint and the analysis in Section 3 describe an idealized counterpart. Appendix B.4 gives the gradient preprocessing and numerical implementation.

3 Quantization Robustness of QuantRet

QuantRet passes normalized branch directions to the residual stream. We relate their quantization error to subsequent state and loss changes, following the distinction between local and propagated

error in Vlassis et al. (2026). We fix the checkpoint, use deterministic PTQ, and retain the same activation retractions and unquantized scales in both computations. Hats denote quantized counterparts. All normalization inputs are nonzero, and each residual token has radius $r = \sqrt{d}$.

Local error after normalization. At a linear map $\mathbf{u} = \mathbf{W}\mathbf{x}$, input error $\mathbf{e}$ and weight error $\Delta\mathbf{W}$ give $\boldsymbol{\delta} = \mathbf{W}\mathbf{e} + \Delta\mathbf{W}\mathbf{x} + \Delta\mathbf{W}\mathbf{e}$. A writer contributes $\mathcal{S}_r(\mathbf{u})$ to residual mixing, so the relevant local change is in its direction.

Lemma 1 (Finite error after normalization)**.** *Let $a = \|\mathbf{u}\|_2 > 0$, $b = \|\mathbf{u} + \boldsymbol{\delta}\|_2 > 0$, and $\delta_{\mathrm{dir}} = \|\mathcal{S}_r(\mathbf{u} + \boldsymbol{\delta}) - \mathcal{S}_r(\mathbf{u})\|_2 / r$. Then*

$$ab\,\delta_{\mathrm{dir}}^2 = \|\boldsymbol{\delta}\|_2^2 - (a-b)^2, \qquad \delta_{\mathrm{dir}} \le \frac{2\|\boldsymbol{\delta}\|_2}{a+b}. \tag{17}$$

The identity separates length changes from direction changes and applies to finite quantization errors. For small errors, the transverse component and output norm determine directional sensitivity. For the input-error term, write $\boldsymbol{\Sigma}_{\mathrm{e}} = \mathbb{E}[\mathbf{e}\mathbf{e}^{\mathsf{T}}]$ and $\mathcal{E}_{\mathrm{e}} = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{e}})$. Then

$$\mathbb{E}\|\mathbf{W}\mathbf{e}\|_2^2 = \mathrm{tr}(\mathbf{W}^{\mathsf{T}}\mathbf{W}\boldsymbol{\Sigma}_{\mathrm{e}}) \le \|\mathbf{W}\|_2^2 \mathcal{E}_{\mathrm{e}}. \tag{18}$$

Frob and Obl fix this energy to $\mathcal{E}_{\mathrm{e}} d_{\mathrm{out}}/d_{\mathrm{in}}$ for isotropic error. Ideal Spec bounds it by the same quantity for arbitrary error directions. At fixed Frobenius energy, Obl's reader rows minimize the largest output second moment under isotropic inputs, and its writer columns minimize worst-case error under grouped isotropy (Proposition 1). Directional error also depends on the endpoint norms in Lemma 1. Under homogeneous quantization, a positive rescaling of a writer preserves its normalized directions.

Propagation through spherical residuals. Index the $M = 2L$ sublayers by m, with $\mathbf{h}_{2\ell} = \mathrm{vec}(\mathbf{H}_\ell)$, $\mathbf{h}_{2\ell+1} = \mathrm{vec}(\mathbf{Z}_\ell)$, $\lambda_{2\ell} = \alpha_\ell$, and $\lambda_{2\ell+1} = \beta_\ell$. The normalized branch $\mathcal{U}_m$ acts on the entire sequence and has token radius r. Before residual retraction, $\mathbf{v}_m = (1-\lambda_m)\mathbf{h}_m + \lambda_m \mathcal{U}_m(\mathbf{h}_m)$, with the analogous quantized mixture $\widehat{\mathbf{v}}_m$. Set $r_{\mathrm{seq}} = \sqrt{N_{\mathrm{tok}}}\, r$ and define RMS state error and the endpoint factor by

$$\bar{\rho}_m = \frac{\|\widehat{\mathbf{h}}_m - \mathbf{h}_m\|_2}{r_{\mathrm{seq}}}, \qquad \mu_m = \min_t \frac{\|\mathbf{v}_{m,t}\|_2 + \|\widehat{\mathbf{v}}_{m,t}\|_2}{2r} > 0. \tag{19}$$

Theorem 1 (Propagation of normalized quantization error)**.** *Let $\lambda_m \in [0,1]$ and $\bar{\kappa}_m, \bar{\varepsilon}_m \ge 0$. Suppose the FP branch outputs at $\mathbf{h}_m$ and $\widehat{\mathbf{h}}_m$ differ by at most $\bar{\kappa}_m \|\widehat{\mathbf{h}}_m - \mathbf{h}_m\|_2$, and quantization at $\widehat{\mathbf{h}}_m$ changes the normalized branch output by at most $r_{\mathrm{seq}}\bar{\varepsilon}_m$. Then*

$$\bar{\rho}_{m+1} \le a_m \bar{\rho}_m + b_m \bar{\varepsilon}_m, \qquad a_m = \frac{1 - \lambda_m + \lambda_m \bar{\kappa}_m}{\mu_m}, \qquad b_m = \frac{\lambda_m}{\mu_m}, \tag{20}$$

$$\bar{\rho}_M \le \bar{\rho}_0 \prod_{j=0}^{M-1} a_j + \sum_{m=0}^{M-1} b_m \bar{\varepsilon}_m \prod_{j=m+1}^{M-1} a_j. \tag{21}$$

Moreover, $\mu_m \ge |1 - 2\lambda_m|$.

Local injection is scaled by b_m and propagated by subsequent a_j factors. Nearly opposing mixture directions can reduce μ_m and amplify error. RMS error measures direction changes; Appendix C.3 gives the proof and single-injection measurement relations.

From state error to predictions. The effective classifier is $\mathbf{B} = \mathrm{diag}(\mathbf{s})\mathbf{M}^\star$ from Section 2.2. Its maximum row norm is $\|\mathbf{B}\|_{2\to\infty} = \max_v \|\mathbf{B}_{v:}\|_2$.

Corollary 1 (Loss and prediction changes)**.** *For the same classifier and $1 \le K \le N_{\mathrm{tok}}$ scored token positions, let $\mathcal{L}$ be mean cross-entropy using the same reference tokens, and let $\overline{D}_{\mathrm{KL}}$ be the mean Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}(p_{\mathrm{FP},t} \| p_{Q,t})$ between the corresponding softmax output distributions. Then*

$$|\mathcal{L}_Q - \mathcal{L}_{\mathrm{FP}}| \le 2r\|\mathbf{B}\|_{2\to\infty} \sqrt{N_{\mathrm{tok}}/K}\, \bar{\rho}_M, \tag{22}$$

$$\overline{D}_{\mathrm{KL}} \le \frac{r^2}{2} \|\mathbf{B}\|_{2\to\infty}^2 \frac{N_{\mathrm{tok}}}{K}\, \bar{\rho}_M^2. \tag{23}$$

Table 1: **Language-modeling quality under RTN quantization.** We report FineWeb-Edu validation CE and WikiText-2 PPL. Lower values indicate better performance. Bold marks the lowest value in each column within each scale.

Scale	Method	FineWeb-Edu CE ↓				WikiText-2 PPL ↓			
		FP	W4A16	W4A4	W4A4KV4	FP	W4A16	W4A4	W4A4KV4
0.3B	AdamW	2.949	2.992	3.454	3.462	26.97	28.58	47.46	47.73
	OSP	**2.857**	3.017	3.447	3.470	24.58	28.87	46.70	47.86
	QuantRet-Frob	2.858	2.913	3.135	3.146	24.32	25.97	33.34	33.79
	QuantRet-Spec	2.975	3.012	3.157	3.164	27.86	29.65	34.98	35.27
	QuantRet-Obl	2.859	**2.910**	**3.123**	**3.132**	**23.89**	**25.57**	**32.08**	**32.44**
0.6B	AdamW	2.692	2.748	3.739	3.751	19.65	20.94	60.04	61.66
	OSP	**2.645**	2.797	3.219	3.238	**18.51**	21.28	32.15	32.76
	QuantRet-Frob	2.683	2.731	3.078	3.088	19.16	**20.31**	30.23	30.59
	QuantRet-Spec	2.840	2.868	3.064	3.070	22.41	23.25	**28.96**	**29.21**
	QuantRet-Obl	2.682	**2.722**	**3.049**	**3.056**	19.63	20.86	29.99	30.25
1.7B	AdamW	2.412	2.500	6.562	6.552	13.94	15.71	1728.91	1666.52
	OSP	**2.381**	2.674	3.498	3.521	**13.41**	17.41	40.31	41.45
	QuantRet-Frob	2.424	2.459	3.129	3.136	13.73	14.33	30.61	30.90
	QuantRet-Spec	2.614	2.637	**2.908**	**2.907**	16.86	17.40	**23.70**	**23.84**
	QuantRet-Obl	2.425	**2.456**	3.134	3.141	13.78	**14.29**	31.94	32.40

The state-to-prediction bounds also hold without geometric constraints (Appendix C.4). Compared with additive residuals, QuantRet has no larger upper bounds on quantization-induced CE changes and output KL divergence if $\lambda_m \leq 1/2$, $\mu_m \geq 1 - \lambda_m$, and its branch sensitivity, local error, and output gain are no larger (Appendix C.5). Absolute quantized quality also depends on FP loss.

4 Experiments

4.1 Experimental Setup

We compare AdamW, Outlier-Safe Pre-Training (OSP), and the three QuantRet variants on Qwen3-like decoders (Park et al., 2025; Yang et al., 2025). Within each scale, all methods share the architecture dimensions, tokenizer, FineWeb-Edu (Penedo et al., 2024) data order, and training budget. We use approximately 20 training tokens per parameter, following Chinchilla (Hoffmann et al., 2022). The 0.3B, 0.6B, and 1.7B experiments use approximately 5.89B, 11.94B, and 34.50B tokens, respectively. Frob, Spec, and Obl denote the matrix constraints in Section 2.3.

We report FineWeb-Edu validation cross-entropy (CE), WikiText-2 perplexity (PPL), and zero-shot task scores. W4 uses symmetric per-output-channel RTN, with clipping thresholds selected to minimize weight quantization error. A4 uses dynamic affine per-token quantization, with head grouping at the attention-output input. KV4 quantizes post-RoPE keys and projected values per token and KV head. Vocabulary weights remain unquantized. We use fake quantization to simulate low-bit inference. FP denotes the unquantized BF16 reference at all three scales. Quantized weights and activations are dequantized to BF16 for computation. Each model is pretrained once, with fixed random seeds for initialization and data ordering. Appendices B.1 and D.1 provide the model architectures, training configurations, and evaluation protocols.

4.2 Low-Bit Quality and Quantization Degradation

In this subsection, we compare language-modeling quality and zero-shot task scores before and after quantization to distinguish FP quality from quantization degradation.

Language modeling. Every QuantRet variant improves both metrics over AdamW and OSP under W4A4 and W4A4KV4 at all three scales (Table 1). At 1.7B, QuantRet-Spec achieves W4A4KV4 perplexity of 23.84, compared with 41.45 for OSP, despite its higher FP perplexity. Across all three scales, Spec has the smallest FP-to-W4A4/W4A4KV4 CE increase despite a higher FP CE than

Frob and Obl. Reporting FP and quantized quality separates pretraining quality from quantization degradation, which can respond differently to training choices (Catalan-Tatjer et al., 2026). The 0.3B Llama 3.1 architecture study reproduces the CE and PPL improvements over both baselines under W4A16, W4A4, and W4A4KV4 (Appendix D.8).

Table 2: **Zero-shot task scores (%, higher is better).** Avg. is the equal-weight average over eight tasks using their specified metrics (Appendix Table 7); MMLU is example-weighted accuracy over 57 subjects and is excluded from Avg. Bold marks the highest score in each column within each scale. Per-task low-bit scores are in Appendix D.2.

Scale	Method	FP		W4A4		W4A4KV4	
		Avg.	MMLU	Avg.	MMLU	Avg.	MMLU
0.3B	AdamW	43.50	23.32	41.81	23.10	41.75	22.91
	OSP	**45.55**	**23.68**	43.33	**23.58**	42.92	23.08
	QuantRet-Frob	45.39	23.19	43.41	23.46	43.75	**23.91**
	QuantRet-Spec	44.18	23.25	42.67	23.28	42.98	23.34
	QuantRet-Obl	44.60	23.59	**43.64**	23.29	**44.03**	23.19
0.6B	AdamW	47.66	24.42	43.39	23.12	43.43	23.14
	OSP	47.84	**25.25**	43.96	23.16	43.90	23.16
	QuantRet-Frob	48.49	24.58	**47.28**	23.89	47.11	23.68
	QuantRet-Spec	46.54	23.06	45.18	23.04	45.69	23.16
	QuantRet-Obl	**49.11**	23.99	47.06	**25.21**	**47.45**	**25.47**
1.7B	AdamW	54.26	25.26	40.19	23.39	40.16	23.52
	OSP	54.53	25.15	48.95	23.14	48.79	23.37
	QuantRet-Frob	**55.79**	25.10	**50.41**	25.07	**50.38**	24.69
	QuantRet-Spec	50.92	**25.99**	49.45	**25.64**	49.03	**25.25**
	QuantRet-Obl	54.71	25.17	49.83	24.88	49.76	25.10

Zero-shot evaluation. At each scale and under both W4A4 and W4A4KV4, a QuantRet variant achieves the highest observed eight-task average among the compared methods (Table 2). Under W4A4KV4, QuantRet-Obl scores 47.45 at 0.6B versus 43.90 for OSP, and QuantRet-Frob scores 50.38 at 1.7B versus 48.79 for OSP. FP quality and quantization degradation contribute differently to the zero-shot comparisons. At 0.3B, Frob has a slightly lower FP eight-task average than OSP but higher scores under both quantization settings. At 0.6B, Obl loses 2.05 percentage points from FP to W4A4, compared with 3.88 for OSP. At 1.7B, Frob combines a higher FP average than OSP with a smaller W4A4KV4 drop (5.41 versus 5.74 points). MMLU scores lie near the 25% random-choice baseline. Statistical limitations are discussed in Appendix D.1.

4.3 Quantization Sensitivity and Error Propagation

In this subsection, we examine how perturbation direction and error propagation affect predictions at 0.6B. We compare prediction sensitivity to controlled perturbations and measure cumulative errors and prediction changes under full W4A4 using the same RTN settings in Table 1. Appendix D.4.1 gives the diagnostic protocol; Appendix D.4.3 examines local operations and the effect of the quantized layer by quantizing one projection input at a time.

How sensitive are predictions to different perturbation directions? We probe prediction sensitivity with artificial perturbations, following prior analysis of LLM quantization (Gong et al., 2024). At attention-output and down-projection inputs in layers 0, 14, and 27, we use isotropic Gaussian directions, Gaussian directions with unequal coordinate variances, or estimated leading right singular vectors, scaling perturbations to 5% of each token's input norm before BF16 casting. All perturbations are evaluated on the same 32 FineWeb-Edu validation sequences, each containing 1024 tokens. Figure 1 reports Kullback–Leibler (KL) divergence and signed CE changes between FP and perturbed predictions. All three QuantRet variants have lower mean KL divergence than AdamW and OSP under the estimated leading-direction perturbations at both projection types. Under isotropic perturbations at the same amplitude, all three QuantRet variants have higher mean KL divergence than AdamW at down-projection inputs. Appendix D.4.2 specifies the perturbation construction.

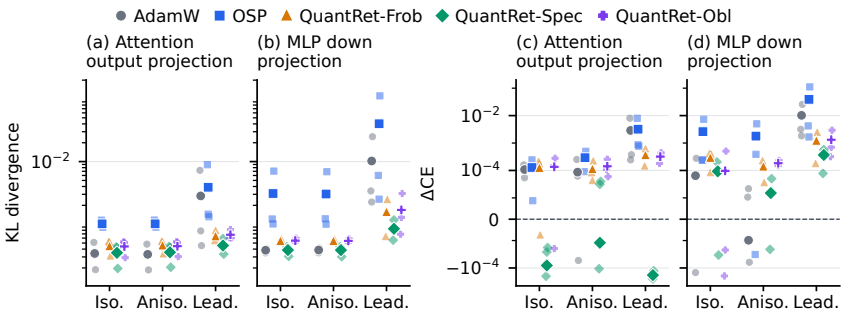

Figure 1: **Prediction sensitivity to directional perturbations.** Perturbations at 0.6B are scaled to 5% of each token's input norm before BF16 casting. The first two columns report output KL divergence after perturbing attention-output and down-projection inputs, respectively; the last two report the corresponding signed ΔCE (perturbed minus FP). Small points show layers 0, 14, and 27, and large markers their arithmetic means. Iso./Aniso./Lead. denote isotropic Gaussian directions, Gaussian directions with unequal coordinate variances, and estimated leading right singular vectors of the projection matrices. CE axes are symmetric-log, linear within $\pm 10^{-5}$.

How does full W4A4 quantization affect predictions? We quantize linear-layer weights and inputs in all 28 Transformer blocks to 4 bits, keep the KV cache unquantized, and compare predictions and hidden states with those of the same model in FP on the same 32 FineWeb-Edu sequences of length 1024. Figure 2 tracks absolute error $E(\mathbf{h}_m) = \mathrm{RMS}(\widehat{\mathbf{h}}_m - \mathbf{h}_m)$ and relative error $\rho(\mathbf{h}_m) = E(\mathbf{h}_m)/\mathrm{RMS}(\mathbf{h}_m)$ after each attention and MLP sublayer. These cumulative differences include both upstream errors and new errors introduced by quantization. QuantRet has smaller absolute state errors throughout this trajectory, while OSP has the smallest final relative state error. All QuantRet variants nevertheless have lower output KL divergence and smaller CE increases (Table 3). Activation tails and RTN granularity provide complementary local evidence in Appendix D.3.

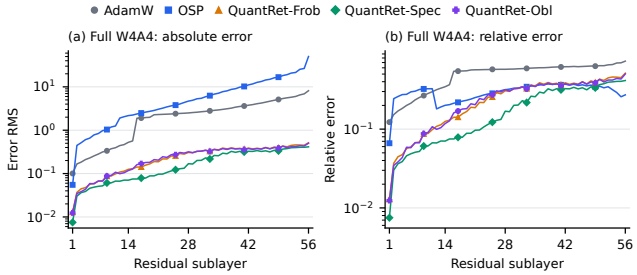

Figure 2: **Cumulative activation errors under full W4A4.** We compare the 0.6B quantized and FP computations at each of 56 attention/MLP sublayers within each checkpoint. **(a):** absolute error RMS. **(b):** error RMS divided by the corresponding FP activation RMS.

Table 3: **Prediction changes under full W4A4.** Results at 0.6B are relative to each checkpoint's FP computation. ΔCE is W4A4 minus FP; lower is better for both metrics.

	AdamW	OSP	QuantRet-Frob	QuantRet-Spec	QuantRet-Obl
KL divergence	1.046	0.613	0.396	**0.224**	0.378
ΔCE	1.067	0.607	0.403	**0.231**	0.377

4.4 Component Ablations

To assess the contribution of individual framework components, we compare three 0.3B controls with the main QuantRet-Obl reference (Table 4): disabling post-step hidden-matrix retraction, re-

placing Muon with projected AdamW at learning rate 3×10^{-4}, and imposing row constraints on all hidden matrices. Each control shares the reference's data order, seed, and training budget.

Table 4: **Component ablations at 0.3B.** All ablations use the main QuantRet-Obl reference. Entries are FineWeb-Edu CE and WikiText-2 PPL; lower is better.

	FineWeb-Edu CE ↓				WikiText-2 PPL ↓			
Configuration	FP	W4A16	W4A4	W4A4KV4	FP	W4A16	W4A4	W4A4KV4
QuantRet-Obl	2.859	2.910	3.123	3.132	23.89	25.57	32.08	32.44
No post-step retraction	2.888	2.973	3.239	3.255	26.21	28.93	38.22	38.82
Projected AdamW	3.030	3.073	3.265	3.277	30.04	31.73	38.45	38.94
All-row constraints	2.860	2.952	3.158	3.167	24.56	27.37	34.97	35.24

The reader–writer assignment improves W4A4 quality without a corresponding improvement in FP CE: QuantRet-Obl and the all-row control obtain FP CE of 2.859 and 2.860, respectively, while their W4A4 CE is 3.123 and 3.158 and PPL is 32.08 and 34.97. Removing post-step retraction also increases both W4A4 metrics. The Muon-based QuantRet-Obl reference achieves lower CE and PPL than the selected projected-AdamW recipe in both FP and W4A4.

4.5 Compatibility with PTQ Methods

To test compatibility with PTQ methods beyond RTN, we evaluate SmoothQuant, GPTQ, and QuaRot under W4A4KV4 on the same five pretrained 0.6B models in Table 5. The quantization settings and transformations for each method are specified in Appendix D.1. With QuaRot, QuantRet-Frob achieves CE/PPL of 2.783/21.83, compared with 2.864/24.09 for AdamW. The benefits nevertheless depend on the PTQ method: AdamW gives the lowest CE and PPL under the tested SmoothQuant configuration.

Table 5: **W4A4KV4 quality with additional PTQ methods.** We evaluate SmoothQuant, GPTQ, and QuaRot at 0.6B. Entries report FineWeb-Edu CE and WikiText-2 PPL; lower is better. For each metric, bold marks the lowest displayed value across all PTQ methods and pretraining methods in the table.

Pretraining method	SmoothQuant		GPTQ		QuaRot	
	CE ↓	PPL ↓	CE ↓	PPL ↓	CE ↓	PPL ↓
AdamW	3.059	29.49	3.712	60.01	2.864	24.09
OSP	3.404	37.54	3.089	28.35	2.978	25.25
QuantRet-Frob	3.156	33.04	3.056	29.53	**2.783**	**21.83**
QuantRet-Spec	3.103	30.42	3.046	28.26	2.913	24.66
QuantRet-Obl	3.110	32.50	3.031	29.31	2.785	22.21

5 Conclusion

We introduced QuantRet, a pretraining framework that combines spherical activation retractions with Frobenius, spectral, or reader–writer matrix constraints to improve post-training quantization. Our analysis connects local directional errors to residual mixing and subsequent propagation under explicit sensitivity conditions. The experiments show why activation ranges and local reconstruction error alone are insufficient to characterize quantized model quality: the direction of an error and its effect on final predictions also matter.

Across 0.3B, 0.6B, and 1.7B Qwen3-like models, all three variants improve CE and perplexity over AdamW and OSP under the common W4A4 and W4A4KV4 RTN protocol, while the strongest configurations also achieve higher average zero-shot task scores. The variants exhibit different trade-offs between FP quality and quantization degradation. Together, these findings support controlling model geometry during pretraining as a way to improve low-bit quality without introducing quantization into training.

AI USE STATEMENT

We used generative AI tools to assist with literature search, experiment planning and implementation, and improving readability of the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kang An, Jiaxiang Li, Donald Goldfarb, and Shiqian Ma. Demystifying manifold constraints in LLM pre-training. *arXiv preprint arXiv:2605.04418*, 2026.

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. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Albert Catalan-Tatjer, Niccolò Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6937a7c60361d05f5b6cfa04d2c27a5b-Abstract-Conference.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332, 2022.

Maxim Fishman, Brian Chmiel, Ron Banner, Daniel Soudry, and Boris Ginsburg. Normalized architectures are natively 4-bit. *arXiv preprint arXiv:2605.06067*, 2026.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhuocheng Gong, Jiahao Liu, Jingang Wang, Xunliang Cai, Dongyan Zhao, and Rui Yan. What makes quantization for large language model hard? an empirical study from the lens of perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18082–18089, 2024. doi: 10.1609/aaai.v38i16.29765. URL https://ojs.aaai.org/index.php/AAAI/article/view/29765.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yufei Gu and Zeke Xie. Mano: Restriking manifold optimization for LLM training. *arXiv preprint arXiv:2601.23000*, 2026.

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

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Yiping Ji, Mahalakshmi Sabanayagam, Peyman Moghadam, Hemanth Saratchandran, and Simon Lucey. The quantization benefits of residual-free transformers, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Yibang Li, Bihari Lal Pandey, Ravi Sah, Andi Han, Cyrus Mostajeran, Pratik Jawanpuria, and Bamdev Mishra. Intrinsic muon: Spectral optimization on riemannian matrix manifolds. *arXiv preprint arXiv:2605.09238*, 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. *Proceedings of machine learning and systems*, 6:87–100, 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*, volume 2025, pp. 92009–92032, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, volume 2025, pp. 74014–74038, 2025.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Jungwoo Park, Taewhoo Lee, Chanwoong Yoon, Hyeon Hwang, and Jaewoo Kang. Outlier-safe pre-training for robust 4-bit quantization of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12582–12600, 2025.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. SocialIQA: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Georgios Vlassis, Saleh Ashkboos, Alexandra Volkova, Torsten Hoefler, and Dan Alistarh. Beyond outliers: A study of optimizers under quantization. In *International Conference on Learning Representations*, volume 2026, pp. 95294–95313, 2026.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them ii: Hyperball optimization. *arXiv preprint arXiv:2606.16899*, 2026.

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*, pp. 38087–38099. PMLR, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

APPENDIX CONTENTS

A Related Work **14**

B Architecture Instantiation of QuantRet **14**

B.1 Evaluated Qwen3-Like Decoder Adapter 14

B.2 Learnable Scales and Stored Coordinates 15

B.3 Tied Vocabulary Boundary 15

B.4 Optimizer Details for the Reported Runs 16

B.5 Adaptation to Other Decoder Architectures 17

C Proofs and Geometric Error Analysis **17**

C.1 Error after Normalization 17

C.2 Matrix Geometry and Local Error 17

C.3 RMS Propagation through Residual Updates 18

C.4 From Final States to Loss and Prediction Changes 19

C.5 Comparison with Unconstrained Residual Models 20

C.6 Quantization Groups and Scale Symmetries 21

D Additional Experimental Protocols and Results **22**

D.1 Detailed Experimental Setup and Evaluation Protocol 22

D.2 Additional Zero-Shot Results 24

D.3 Local Reconstruction and Quantizer Dependence 24

D.4 Prediction Sensitivity and Error Propagation 29

D.5 Learnable-Scale Parameterization 31

D.6 Compatibility with PTQ Methods 35

D.7 NVFP4 Weight and Activation Quantization 35

D.8 Transfer to the Llama 3.1 Architecture 35

A Related Work

Outlier-aware quantization. Activation outliers are a central obstacle for low-bit LLM inference. LLM.int8() isolates outlier channels in higher precision (Dettmers et al., 2022). GPTQ and AWQ optimize weight quantization using calibration data (Frantar et al., 2023; Lin et al., 2024). SmoothQuant migrates activation difficulty into weights (Xiao et al., 2023). QuaRot and SpinQuant reduce outliers by rotating hidden representations before quantization (Ashkboos et al., 2024; Liu et al., 2025). QuantRet changes pretraining rather than the subsequent quantizer. We evaluate it with several PTQ methods and observe method-dependent benefits.

Quantization-friendly pretraining and normalized architectures. Bondarenko et al. (2023) modify attention during pretraining to reduce activation outliers and improve INT8 quantization. OSP addresses outlier formation through optimizer, normalization, and embedding modifications (Park et al., 2025). nGPT combines normalized hidden states, residual interpolation, and weight normalization along embedding dimensions (Loshchilov et al., 2025). Under our matrix convention, these weight-normalization axes correspond to reader rows and writer columns. Fishman et al. (2026) study normalized architectures for end-to-end NVFP4 training and analyze their dot-product signal-to-noise ratios. QuantRet builds on these geometric insights, using channel-shared residual mixing and comparing Frobenius, spectral, and reader–writer constraints within a shared framework for PTQ.

Quantization sensitivity and error propagation. Gong et al. (2024) study LLM quantization through perturbations to weights and activations. Vlassis et al. (2026) show that outlier statistics alone do not predict PTQ performance across optimizers and distinguish local errors from their accumulation and propagation. Catalan-Tatjer et al. (2026) examine how training dynamics affect FP quality and quantization robustness. Ji et al. (2026) investigate the quantization benefits of removing residual connections. QuantRet instead retains residual connections and analyzes finite directional errors after normalization, their injection through spherical mixing, and subsequent propagation under explicit sensitivity conditions.

Matrix geometry and Muon. Muon uses a matrix-sign update for hidden matrices, which can be interpreted as a spectral steepest-descent direction (Jordan et al., 2024). An et al. (2026) study Frobenius, spectral, and oblique constraints for LLM pretraining, while Hyperball combines normalized optimizer updates with fixed-radius weight normalization (Wen et al., 2026). Related methods also impose geometry on optimization itself: Mano shapes momentum using oblique geometry without constraining forward weights (Gu & Xie, 2026), whereas Intrinsic Muon formulates updates in Riemannian tangent spaces (Li et al., 2026). QuantRet applies ambient Muon updates followed by hidden-matrix retraction, leaving gradients, momentum, and optimizer states in ambient coordinates.

B Architecture Instantiation of QuantRet

An architecture adapter maps QuantRet's model-independent rules to a specific decoder. This appendix gives the Qwen3-like adapter used in the main experiments and the changes required for the Llama 3.1 transfer study. Appendix D.1 reports the exact pretraining and evaluation hyperparameters.

B.1 Evaluated Qwen3-Like Decoder Adapter

The evaluated models use a Qwen3-like decoder with grouped-query causal attention, RoPE, and SwiGLU feed-forward branches (Yang et al., 2025). They have no linear biases or dropout, and the embedding and output classifier share one parameter table. RoPE uses $\theta = 10^6$, the context length is 2048, and the frozen tokenizer vocabulary is padded to 151936 entries. Table 6 gives the widths, depths, head counts, token budgets, and parameter counts. These adapter choices are not requirements of QuantRet.

For each attention branch, the query, key, and value projections are residual readers. The attention output projection is a residual writer. Query and key vectors are retracted to the $\sqrt{D_{\text{head}}}$ sphere

after RoPE and before their dot product, while value vectors remain unnormalized. A positive head-wise score scale adjusts attention sharpness without changing the normalized query/key radius. The backbone determines the grouping from query heads to key–value heads.

For each SwiGLU branch, the gate and up projections are residual readers, and the down projection is a residual writer. A positive learnable scalar rescales the gate path as specified below. The attention and feed-forward outputs enter the spherical residual update in Equation (3).

The reader and writer matrix sets are

$$\mathcal{W}_{\mathrm{read}} = \{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_G, \mathbf{W}_U\}, \qquad \mathcal{W}_{\mathrm{write}} = \{\mathbf{W}_O, \mathbf{W}_D\}. \tag{24}$$

QuantRet-Frob and QuantRet-Spec apply their matrix-wide retractions to every matrix in these sets. QuantRet-Obl applies row retraction to $\mathcal{W}_{\mathrm{read}}$ and column retraction to $\mathcal{W}_{\mathrm{write}}$.

Forward graph of the reported checkpoints. The QuantRet checkpoints in Tables 1 and 17 additionally apply $\mathrm{Sph}_{D_{\mathrm{ff}}}$ to $\Psi_\ell(\mathbf{Z}_\ell)$ immediately before the down projection. This operation is not required by the framework in Section 2: for nonzero intermediate and branch-output vectors and a bias-free down projection, it only rescales the branch output by a positive token-wise factor, which cancels under the subsequent Sph_d in exact arithmetic. The evaluations retain the normalization present in each checkpoint, including its placement relative to activation quantization. We do not assume that this cancellation makes finite-precision or quantized implementations interchangeable. The separate learnable-scale study in Appendix D.5 feeds $\Psi_\ell(\mathbf{Z}_\ell)$ directly to the down projection in all eight configurations.

B.2 Learnable Scales and Stored Coordinates

We parameterize the positive attention-score and MLP gate scales in log coordinates,

$$\tau_{\ell,h} = \exp(\theta^{\tau}_{\ell,h}), \qquad \gamma_{g,\ell} = \exp(\theta^{g}_{\ell}), \tag{25}$$

with both effective scales initialized to one. For the SwiGLU adapter, we express the normalized branch using the feature map

$$\Psi_\ell(\mathbf{Z}_\ell) = \mathrm{SiLU}\big(\gamma_{g,\ell}\mathbf{Z}_\ell\mathbf{W}_{G,\ell}^{\mathsf{T}}\big) \odot \big(\mathbf{Z}_\ell\mathbf{W}_{U,\ell}^{\mathsf{T}}\big)\,. \tag{26}$$

For a bias-free branch at nonzero normalization inputs, a positive scalar on the linear up path cancels under branch normalization in exact arithmetic, with or without pre-down retraction.

The residual coefficients use the stored-coordinate device adopted from nGPT (Loshchilov et al., 2025). Let $s_0 = D_{\mathrm{model}}^{-1/2}$ and store one scalar a_ℓ for attention and b_ℓ for the MLP. The reported runs use

$$a_\ell(0) = b_\ell(0) = s_0, \qquad \alpha_\ell = \mathrm{clip}\left(|a_\ell|\frac{0.05}{s_0}, 0, 1\right), \qquad \beta_\ell = \mathrm{clip}\left(|b_\ell|\frac{0.05}{s_0}, 0, 1\right). \tag{27}$$

The forward coefficients therefore start at 0.05 at both model depths, while the stored parameters start at $1/\sqrt{D_{\mathrm{model}}}$. The multiplier changes the effective learning rate of the stored coordinate without changing its forward initialization. Clipping bounds each coefficient in $[0, 1]$ but does not constrain their sum across depth. This storage rule is inherited from nGPT and is separate from the QuantRet matrix constraints and reader/writer assignment. Section 2.2 defines the vocabulary scale and its backward decoupling.

B.3 Tied Vocabulary Boundary

The Qwen3 adapter shares the two vocabulary tables and rotations in Equation (9): $\mathbf{M} = \mathbf{E}$ and $\mathbf{R}_{\mathrm{in}} = \mathbf{R}_{\mathrm{out}} = \mathbf{R}$. The shared table has unit row norms, and $\mathbf{R} \in \mathrm{O}(D_{\mathrm{model}})$ is trainable. In the ideal orthogonal parameterization, the effective table at both boundaries is

$$\mathbf{E}^\star = \sqrt{D_{\mathrm{model}}}\,\mathbf{E}\mathbf{R}^{\mathsf{T}}, \qquad \|\mathbf{E}^\star_{v:}\|_2 = \sqrt{D_{\mathrm{model}}}. \tag{28}$$

For token indices $\mathbf{X}$, the input states and raw vocabulary scores are

$$\mathbf{H}_0 = \mathbf{E}^\star[\mathbf{X}], \qquad \mathbf{L} = \mathbf{H}_L(\mathbf{E}^\star)^{\mathsf{T}}. \tag{29}$$

Under this idealization, input tokens start on the residual sphere, and the output classifier uses the same effective table.

The vocabulary scale follows Equation (10), implemented with a fused autograd operation that returns $s_v L_{t,v}$ and applies the derivatives in Equation (11). The table and scale vector use auxiliary AdamW, with zero weight decay on the scale vector. The table is row-retracted after its optimizer step. The rotation uses the five-iteration Cayley update specified in Appendix B.4. Exact orthogonality describes the ideal map, and the truncated iteration introduces numerical error. At inference, $\sqrt{D_{\mathrm{model}}}\,\mathbf{E}\mathbf{R}^{\mathsf{T}}$ is fused once. The fused graph retains parameter tying and adds no rotation operator at runtime.

An untied vocabulary interface uses separate input and output tables in Equation (9). Both effective row radii remain $\sqrt{d}$.

B.4 Optimizer Details for the Reported Runs

The reported $\mathrm{MuonStep}$ in Equation (15) uses $\mathbf{G}_t$, the matrix component of the gradients computed with the logit backward rule in Equation (10) and then jointly clipped to a global norm threshold of 1.0. It forms a momentum buffer and Nesterov source as

$$\mathbf{B}_t = \mu\mathbf{B}_{t-1} + (1-\mu)\mathbf{G}_t, \qquad \mathbf{N}_t = \mu\mathbf{B}_t + (1-\mu)\mathbf{G}_t. \tag{30}$$

If $d_{\mathrm{out}} > d_{\mathrm{in}}$, $\mathbf{N}_t$ is transposed during orthogonalization and the result is transposed back. For the resulting wide matrix, the five-step Newton–Schulz approximation starts from $\mathbf{Z}_0 = \mathbf{N}_t/(\|\mathbf{N}_t\|_F + 10^{-7})$ and iterates

$$\mathbf{Z}_{k+1} = a_{\mathrm{NS}}\mathbf{Z}_k + b_{\mathrm{NS}}(\mathbf{Z}_k\mathbf{Z}_k^{\mathsf{T}})\mathbf{Z}_k + c_{\mathrm{NS}}(\mathbf{Z}_k\mathbf{Z}_k^{\mathsf{T}})^2\mathbf{Z}_k, \qquad (a_{\mathrm{NS}}, b_{\mathrm{NS}}, c_{\mathrm{NS}}) = (3.4445, -4.7750, 2.0315). \tag{31}$$

The ambient update uses

$$\widetilde{\mathbf{W}}_{t+1} = \mathbf{W}_t - \eta_t\sqrt{\max\left\{1, \frac{d_{\mathrm{out}}}{d_{\mathrm{in}}}\right\}}\mathbf{Z}_5, \tag{32}$$

Equation (16) retracts the parameter after this ambient update. The reported experiments use momentum 0.95 and the learning rates reported in Appendix D.1. The momentum buffer remains in ambient coordinates and is not changed by the parameter retraction.

QuantRet-Frob and QuantRet-Obl compute their normalization norms directly in FP32. QuantRet-Spec uses warm-started power iteration (Miyato et al., 2018) with 16 iterations at initialization and two per subsequent update. It rescales the matrix to the target using this estimate, rather than clipping only above the target. The estimated norm need not equal the exact spectral norm; the equality constraint in Equation (13) and the exact-norm analysis describe the idealized operation.

Vocabulary-rotation update. The stored rotation $\mathbf{R}_t$ starts orthogonally, with momentum $\mathbf{M}_0 = \mathbf{0}$. Let $\mathbf{G}_t^R$ be its gradient after global clipping, and let $\mathbf{U}_t$ divide each row $\mathbf{R}_{t,i:}$ by $\max\{\|\mathbf{R}_{t,i:}\|_2, \epsilon\}$, with $\epsilon = 10^{-8}$. The FP32 update forms

$$\mathbf{X}_t = \mathbf{U}_t^{\mathsf{T}}, \qquad \mathbf{V}_t = 0.95\mathbf{M}_t - (\mathbf{G}_t^R)^{\mathsf{T}}, \tag{33}$$

$$\mathbf{A}_t = \mathbf{V}_t\mathbf{U}_t - \tfrac{1}{2}\mathbf{U}_t^{\mathsf{T}}\big(\mathbf{U}_t(\mathbf{V}_t\mathbf{U}_t)\big), \qquad \mathbf{K}_t = \mathbf{A}_t - \mathbf{A}_t^{\mathsf{T}}. \tag{34}$$

Using the Muon learning-rate schedule η_t, set $h_t = \min\{\eta_t, (\|\mathbf{K}_t\|_1 + \epsilon)^{-1}\}$, where $\|\cdot\|_1$ is the maximum absolute column sum. The five fixed-point iterations are

$$\mathbf{Y}_0 = \mathbf{X}_t + h_t\mathbf{V}_t, \qquad \mathbf{Y}_{j+1} = \mathbf{X}_t + \tfrac{h_t}{2}\mathbf{K}_t(\mathbf{X}_t + \mathbf{Y}_j), \quad j = 0, \ldots, 4. \tag{35}$$

The stored updates are $\mathbf{R}_{t+1} = \mathbf{Y}_5^{\mathsf{T}}$ and $\mathbf{M}_{t+1} = \mathbf{K}_t\mathbf{X}_t$. The finite iteration approximates the Cayley map and does not enforce exact orthogonality.

Computational cost. For an $r \times c$ hidden matrix, five-step Muon costs $O(5rc\min\{r,c\})$ arithmetic and $O(rc + \min\{r,c\}^2)$ working memory, shared with OSP. Frob and Obl retraction cost $O(rc)$; Spec's fixed-count power iteration also costs $O(rc)$ per update. Vocabulary-table and activation retractions are linear in the number of elements normalized. A dense $d \times d$ Cayley update costs $O(d^3)$ arithmetic and $O(d^2)$ storage; applying boundary rotations to n training tokens costs $O(nd^2)$. These operation counts do not establish wall-clock overhead. The boundary rotations are fused into the vocabulary table for inference.

B.5 Adaptation to Other Decoder Architectures

To port QuantRet, the adapter retains the target decoder's tokenizer, positional encoding, attention grouping, activation, and vocabulary interface. It registers the attention and feed-forward outputs for spherical residual interpolation, identifies direction-only activation sites, assigns hidden matrices by reader/writer dataflow, and applies one selected matrix retraction after each ambient optimizer step. The decoder's attention and feed-forward equations otherwise remain unchanged.

We evaluate a Llama 3.1 instantiation with its corresponding tokenizer (Grattafiori et al., 2024). It retains the model family's untied 128256-entry vocabulary boundary, scaled RoPE, grouped-query attention, and SwiGLU structure. The QuantRet models use unit-norm rows in both stored vocabulary tables and independent fixed $\sqrt{d}$ forward multipliers. This instantiation fixes $\mathbf{R}_{\mathrm{in}} = \mathbf{R}_{\mathrm{out}} = \mathbf{I}$. The effective boundary rows have radius $\sqrt{d}$, and $s_v(0) = c_{\mathrm{bp}} = 1/\sqrt{d}$. Appendix D.8 reports the results. This evidence covers one approximately 0.3B scale, one training realization, and language-modeling fake-quantization metrics; broader cross-family and scaling claims remain untested.

C Proofs and Geometric Error Analysis

All results concern a fixed checkpoint, deterministic quantization, and exact arithmetic. We use column vectors for individual linear maps and vectorize full token sequences for the RMS analysis. This Euclidean sequence norm equals the Frobenius norm of the corresponding token matrix. Unquantized scales have the same values in both computations. Every normalization input, including those in the auxiliary branch comparisons, is nonzero.

C.1 Error after Normalization

Proof of Lemma 1. Write $\mathbf{u} = a\mathbf{p}$ and $\mathbf{u} + \boldsymbol{\delta} = b\mathbf{q}$, where $a, b > 0$ and $\|\mathbf{p}\|_2 = \|\mathbf{q}\|_2 = 1$. Then $\delta_{\mathrm{dir}} = \|\mathbf{q} - \mathbf{p}\|_2 \le 2$ and

$$\|\boldsymbol{\delta}\|_2^2 = (a-b)^2 + ab\|\mathbf{q} - \mathbf{p}\|_2^2.$$

This proves the identity. The finite bound follows from

$$4\|\boldsymbol{\delta}\|_2^2 - (a+b)^2\delta_{\mathrm{dir}}^2 = (a-b)^2(4 - \delta_{\mathrm{dir}}^2) \ge 0.$$

Equality in the bound holds when $a = b$ or $\mathbf{q} = -\mathbf{p}$. □

Local directional sensitivity. For $\mathbf{P}_{\mathbf{u}}^{\perp} = \mathbf{I} - \mathbf{u}\mathbf{u}^{\mathsf{T}}/\|\mathbf{u}\|_2^2$, differentiation gives

$$D\mathcal{S}_r(\mathbf{u}) = \frac{r}{\|\mathbf{u}\|_2}\mathbf{P}_{\mathbf{u}}^{\perp}, \qquad \mathcal{S}_r(\mathbf{u} + \boldsymbol{\delta}) - \mathcal{S}_r(\mathbf{u}) = \frac{r}{\|\mathbf{u}\|_2}\mathbf{P}_{\mathbf{u}}^{\perp}\boldsymbol{\delta} + O\left(\frac{r\|\boldsymbol{\delta}\|_2^2}{\|\mathbf{u}\|_2^2}\right)$$

as $\|\boldsymbol{\delta}\|_2/\|\mathbf{u}\|_2 \to 0$. Small output norms can increase directional sensitivity even at a fixed spectral norm. For example, $\mathbf{W} = \mathrm{diag}(1, \eta)$ with $0 < \eta \le 1$ has spectral norm one. At $\mathbf{x} = (0, 1)^{\mathsf{T}}$, an input perturbation $(t, 0)^{\mathsf{T}}$ gives an output angle θ with $\tan\theta = |t|/\eta$. For numerical normalization $r\mathbf{u}/\max\{\|\mathbf{u}\|_2, \epsilon\}$, the sphere identities apply when each evaluated endpoint is above the threshold.

C.2 Matrix Geometry and Local Error

Weight and activation errors. Expanding $\widehat{\mathbf{W}} = \mathbf{W} + \Delta\mathbf{W}$ and $\widehat{\mathbf{x}} = \mathbf{x} + \mathbf{e}$ gives

$$\widehat{\mathbf{W}}\widehat{\mathbf{x}} - \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{e} + \Delta\mathbf{W}\mathbf{x} + \Delta\mathbf{W}\mathbf{e}. \tag{36}$$

For fixed $\mathbf{W}, \Delta\mathbf{W}$ and finite second moments $\boldsymbol{\Sigma}_{\mathrm{e}} = \mathbb{E}[\mathbf{e}\mathbf{e}^{\mathsf{T}}]$ and $\boldsymbol{\Sigma}_{\mathrm{x}} = \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}]$,

$$\mathbb{E}\|\mathbf{W}\mathbf{e}\|_2^2 = \mathrm{tr}(\mathbf{W}^{\mathsf{T}}\mathbf{W}\boldsymbol{\Sigma}_{\mathrm{e}}), \qquad \mathbb{E}\|\Delta\mathbf{W}\mathbf{x}\|_2^2 = \mathrm{tr}(\Delta\mathbf{W}^{\mathsf{T}}\Delta\mathbf{W}\boldsymbol{\Sigma}_{\mathrm{x}}). \tag{37}$$

These follow by writing each squared norm as a trace and taking expectations. They hold when $\mathbf{e}$ depends on $\mathbf{x}$. The full squared error also contains cross terms among the three channels. For a quantized state $\widehat{\mathbf{x}}$, the perturbation entering a quantized linear map is $Q_A(\widehat{\mathbf{x}}) - \mathbf{x} = (\widehat{\mathbf{x}} - \mathbf{x}) + [Q_A(\widehat{\mathbf{x}}) - \widehat{\mathbf{x}}]$. This includes accumulated state error and fresh activation rounding.

Energy and directional constraints. For isotropic error $\mathbf{\Sigma}_{\mathrm{e}} = (\mathcal{E}_{\mathrm{e}}/d_{\mathrm{in}})\mathbf{I}$, Equation (37) gives $\mathbb{E}\|\mathbf{We}\|_2^2 = (\mathcal{E}_{\mathrm{e}}/d_{\mathrm{in}})\|\mathbf{W}\|_F^2$. Frob and Obl fix $\|\mathbf{W}\|_F^2 = d_{\mathrm{out}}$. For general error, $\|\mathbf{We}\|_2 \le \|\mathbf{W}\|_2\|\mathbf{e}\|_2$ proves Equation (18). Ideal Spec sets its multiplier to $d_{\mathrm{out}}/d_{\mathrm{in}}$. If retraction uses a positive estimate $\widehat{\sigma}$ of the spectral norm σ before retraction, the multiplier becomes $(d_{\mathrm{out}}/d_{\mathrm{in}})(\sigma/\widehat{\sigma})^2$.

Under the exact Spec constraint, the sum of squared singular values satisfies

$$\|\mathbf{W}\|_F^2 \le \min(d_{\mathrm{out}}, d_{\mathrm{in}})\frac{d_{\mathrm{out}}}{d_{\mathrm{in}}}.$$

For wide writers, this upper bound is smaller than the Frob and Obl energy d_{out}. Raw output-error measurements therefore reflect both total matrix energy and its distribution across directions. A bound on normalized error also requires output norms. For example, if $\|\mathbf{Wx}\|_2 + \|\mathbf{W}(\mathbf{x}+\mathbf{e})\|_2 \ge 2r\mu_0 > 0$ for every sample, Lemma 1 gives $\mathbb{E}\delta_{\mathrm{dir}}^2 \le \mathbb{E}\|\mathbf{We}\|_2^2/(r^2\mu_0^2)$ for this channel.

Proposition 1 (Minimax properties of mixed oblique retractions)**.** *Let $\mathbf{W} \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$ and constrain $\|\mathbf{W}\|_F^2 = d_{\mathrm{out}}$.*

(i) *For $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}] = \sigma_{\mathrm{x}}^2\mathbf{I}$ with $\sigma_{\mathrm{x}}^2 > 0$,*

$$\min_{\|\mathbf{W}\|_F^2 = d_{\mathrm{out}}} \max_i \mathbb{E}[(\mathbf{W}_{i:}\mathbf{x})^2] = \sigma_{\mathrm{x}}^2,$$

and equality holds exactly when every reader row has norm one.

(ii) *Let $\mathcal{P} = \{\mathcal{G}_g\}$ partition the input coordinates into nonempty groups and let $\mathbf{\Sigma}_{\mathrm{e}} = \mathrm{blockdiag}_g((v_g/|\mathcal{G}_g|)\mathbf{I})$, where $v_g \ge 0$ and $\sum_g v_g = \mathcal{E}_{\mathrm{e}} > 0$. Then*

$$\min_{\|\mathbf{W}\|_F^2 = d_{\mathrm{out}}} \sup_{\{v_g\}} \mathbb{E}\|\mathbf{We}\|_2^2 = \mathcal{E}_{\mathrm{e}}\frac{d_{\mathrm{out}}}{d_{\mathrm{in}}}.$$

Equal column norms of writers, $\|\mathbf{W}_{:j}\|_2^2 = d_{\mathrm{out}}/d_{\mathrm{in}}$, attain this minimum simultaneously for every partition. They are necessary for the singleton partition.

Proof. For part (i), isotropy gives $\mathbb{E}[(\mathbf{W}_{i:}\mathbf{x})^2] = \sigma_{\mathrm{x}}^2\|\mathbf{W}_{i:}\|_2^2$. The squared row norms sum to d_{out}, so their maximum is at least one, with equality exactly when all equal one. For part (ii), define $c_g = \|\mathbf{W}_{:\mathcal{G}_g}\|_F^2/|\mathcal{G}_g|$. The objective is $\sum_g v_g c_g$, whose supremum is $\mathcal{E}_{\mathrm{e}}\max_g c_g$. Since $\sum_g |\mathcal{G}_g| c_g = d_{\mathrm{out}}$, the largest c_g is at least $d_{\mathrm{out}}/d_{\mathrm{in}}$. Equal column norms attain equality for every partition. A single fixed partition requires only equal group averages. Singleton groups require every column norm to agree. Positive budgets give the stated necessity results; at zero budget every feasible matrix attains zero. □

Meaning of the diagonal prediction. Define the diagonal matrix $\mathbf{D}_{\mathrm{e}} = \mathrm{diag}((\mathbf{\Sigma}_{\mathrm{e}})_{11}, \ldots, (\mathbf{\Sigma}_{\mathrm{e}})_{d_{\mathrm{in}}d_{\mathrm{in}}})$. Then

$$E_{\mathrm{diag}} = \mathrm{tr}(\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{D}_{\mathrm{e}}), \qquad \mathbb{E}\|\mathbf{We}\|_2^2 - E_{\mathrm{diag}} = \mathrm{tr}\big[\mathbf{W}^{\mathsf{T}}\mathbf{W}(\mathbf{\Sigma}_{\mathrm{e}} - \mathbf{D}_{\mathrm{e}})\big]. \tag{38}$$

For $E_{\mathrm{diag}} > 0$, the actual-to-diagonal ratio minus one is this net remainder divided by E_{diag}. The observer in Appendix D.3.2 uses $\mathbf{e} = Q_A(\mathbf{x}) - \mathbf{x}$ and FP weights, matching this formula. Its table averages ratios over matrices, so agreement of the mean permits variation among individual matrices and cancellation among relative remainders. Reader isotropy concerns the separate matrix $\mathbf{\Sigma}_{\mathrm{x}}$.

C.3 RMS Propagation through Residual Updates

Proof of Theorem 1. Add and subtract $\mathcal{U}_m(\widehat{\mathbf{h}}_m)$ in the difference between the actual branch outputs. The theorem's assumptions give

$$\|\widehat{\mathbf{v}}_m - \mathbf{v}_m\|_2 \le (1 - \lambda_m + \lambda_m\bar{\kappa}_m)r_{\mathrm{seq}}\bar{\rho}_m + \lambda_m r_{\mathrm{seq}}\bar{\varepsilon}_m.$$

Apply Lemma 1 to each pair of mixture tokens. Their norm sum is at least $2r\mu_m$, so

$$\|\widehat{\mathbf{h}}_{m+1,t} - \mathbf{h}_{m+1,t}\|_2 \le \mu_m^{-1}\|\widehat{\mathbf{v}}_{m,t} - \mathbf{v}_{m,t}\|_2.$$

Square and sum over tokens, take square roots, and divide by r_{seq}. This proves Equation (20). Repeated substitution gives Equation (21), since $a_m, b_m \ge 0$. Empty products equal one.

For tokens $\mathbf{h}, \mathbf{u}$ of norm r, direct expansion gives

$$\|(1-\lambda)\mathbf{h}+\lambda\mathbf{u}\|_2^2 = r^2\left[1-2\lambda(1-\lambda)\left(1-\frac{\langle \mathbf{h},\mathbf{u}\rangle}{r^2}\right)\right] \geq r^2(1-2\lambda)^2. \tag{39}$$

Applying this to both actual mixtures proves $\mu_m \geq |1-2\lambda_m|$. At $\lambda_m = 1/2$, the endpoint definition still applies to nonzero mixtures. □

Let $\widehat{\mathcal{U}}_m$ denote the quantized normalized branch, including all causal keys and values. For distinct input states, the smallest admissible factors are

$$\begin{aligned} \bar{\kappa}_m &= \frac{\|\mathcal{U}_m(\widehat{\mathbf{h}}_m) - \mathcal{U}_m(\mathbf{h}_m)\|_2}{\|\widehat{\mathbf{h}}_m - \mathbf{h}_m\|_2}, \\ \bar{\varepsilon}_m &= \frac{\|\widehat{\mathcal{U}}_m(\widehat{\mathbf{h}}_m) - \mathcal{U}_m(\widehat{\mathbf{h}}_m)\|_2}{r_{\mathrm{seq}}}. \end{aligned} \tag{40}$$

At coincident states, set $\bar{\kappa}_m = 0$ and retain the same definition of $\bar{\varepsilon}_m$. These finite differences do not require the quantizer to be differentiable.

Single injections and measured RMS. If the computations agree through the input of sublayer k and only that sublayer is perturbed, $\bar{\rho}_k = 0$ and we can take $\bar{\varepsilon}_j = 0$ for $j \neq k$. Therefore

$$\bar{\rho}_{k+1} \leq b_k\bar{\varepsilon}_k, \qquad \bar{\rho}_M \leq b_k\bar{\varepsilon}_k \prod_{j=k+1}^{M-1} a_j. \tag{41}$$

For the local A4 measurements in Appendix D.4.3, let E_{branch}, E_{mix}, and E_{residual} denote component RMS errors after branch normalization, residual mixing, and residual retraction, respectively, over the same tokens. With the residual input unchanged, scalar mixing and Lemma 1 give, in exact arithmetic and for nonzero denominators,

$$\frac{E_{\mathrm{mix}}}{E_{\mathrm{branch}}} = \lambda_k, \qquad \frac{E_{\mathrm{residual}}}{E_{\mathrm{mix}}} \leq \mu_k^{-1}. \tag{42}$$

The choice in Equation (40) gives $\bar{\varepsilon}_k = E_{\mathrm{branch}}$ because $r = \sqrt{d}$. Figure 5 reports the corresponding BF16 measurements. Evaluating the full propagation bound also requires the mixture norms and auxiliary branch comparisons.

For radius-r token states, RMS error has the exact angular interpretation

$$\bar{\rho}_m^2 = \frac{2}{N_{\mathrm{tok}}}\sum_t (1-\cos\theta_{m,t}), \tag{43}$$

where $\theta_{m,t}$ is the angle between the paired token states. Thus $\bar{\rho}_m \leq 2$. A maximum-token version replaces sequence Euclidean norms by the maximum token norm and r_{seq} by r throughout the error and sensitivity definitions. The endpoint factor is unchanged. The maximum-token distance lies between $\bar{\rho}_m$ and $\sqrt{N_{\mathrm{tok}}}\,\bar{\rho}_m$.

C.4 From Final States to Loss and Prediction Changes

Proof of Corollary 1. For a target y and logits $\mathbf{z}$, let $\ell_y(\mathbf{z}) = \log\sum_v e^{z_v} - z_y$. Its gradient is $\mathrm{softmax}(\mathbf{z}) - \mathbf{e}_y$, with ℓ_1 norm at most two. Integrating along the line segment between two logit vectors gives $|\ell_y(\widehat{\mathbf{z}}) - \ell_y(\mathbf{z})| \leq 2\|\widehat{\mathbf{z}} - \mathbf{z}\|_\infty$. For a common classifier, the logit difference is bounded by $\|\mathbf{B}\|_{2\to\infty}\|\widehat{\mathbf{h}}_{M,t} - \mathbf{h}_{M,t}\|_2$. If $\mathcal{T}$ contains the K scored positions, Cauchy–Schwarz gives

$$\frac{1}{K}\sum_{t\in\mathcal{T}}\|\widehat{\mathbf{h}}_{M,t} - \mathbf{h}_{M,t}\|_2 \leq \frac{\|\widehat{\mathbf{h}}_M - \mathbf{h}_M\|_2}{\sqrt{K}} = r\sqrt{N_{\mathrm{tok}}/K}\,\bar{\rho}_M.$$

Combining the two inequalities proves Equation (22).

For logits $\widehat{\mathbf{z}} = \mathbf{z}+\Delta\mathbf{z}$, set $p = \mathrm{softmax}(\mathbf{z})$ and $\widehat{p} = \mathrm{softmax}(\widehat{\mathbf{z}})$. The bounded-variable exponential-moment inequality (Hoeffding, 1963) gives

$$\begin{aligned} D_{\mathrm{KL}}(p\|\widehat{p}) &= \log\left(\sum_v p_v e^{\Delta z_v}\right) - \sum_v p_v \Delta z_v \\ &\le \frac{(\max_v \Delta z_v - \min_v \Delta z_v)^2}{8} \le \frac{1}{2}\|\Delta\mathbf{z}\|_\infty^2. \end{aligned} \tag{44}$$

All logarithms are natural. Averaging over $\mathcal{T}$ and applying the common-classifier bound yields

$$\overline{D}_{\mathrm{KL}} \le \frac{\|\mathbf{B}\|_{2\to\infty}^2}{2K}\sum_{t\in\mathcal{T}}\|\widehat{\mathbf{h}}_{M,t} - \mathbf{h}_{M,t}\|_2^2 \le \frac{r^2}{2}\|\mathbf{B}\|_{2\to\infty}^2 \frac{N_{\mathrm{tok}}}{K}\bar{\rho}_M^2.$$

This proves Equation (23). □

For a reference token y, the same logit expansion gives

$$\ell_y(\widehat{\mathbf{z}}) - \ell_y(\mathbf{z}) = D_{\mathrm{KL}}(p\|\widehat{p}) + \sum_v p_v\Delta z_v - \Delta z_y. \tag{45}$$

KL divergence measures output-distribution change, while the reference-token term permits signed CE changes in either direction. These are the two prediction metrics in Section 4.3.

When every position is scored, the factor $\sqrt{N_{\mathrm{tok}}/K}$ equals one. Dataset CE and KL divergence use token-weighted averages of the corresponding sequence bounds; the KL bound averages squared state errors. With $\|\mathbf{M}^\star_{v:}\|_2 = r$ for every vocabulary row, $\|\mathbf{B}\|_{2\to\infty} = r\|\mathbf{s}\|_\infty$. Measured row norms can replace this ideal equality. The stop-gradient rule in Equation (10) leaves the forward logits unchanged, so the bound uses the learned forward scales. These bounds describe prediction changes under PTQ at a fixed checkpoint. Comparing absolute quantized losses across checkpoints also requires their FP losses.

C.5 Comparison with Unconstrained Residual Models

We compare QuantRet (C) with an additive-residual model with unconstrained branch weights (U), both with M sublayers of width d. Each model is compared with its own FP checkpoint on the same token sequence and K scored positions. Embeddings, residual coefficients, and output maps remain unquantized, so the initial state error is zero within each model. The additive update is $\mathbf{h}^{\mathrm{U}}_{m+1} = \mathbf{h}^{\mathrm{U}}_m + F^{\mathrm{U}}_m(\mathbf{h}^{\mathrm{U}}_m)$, where F^{U}_m includes the model's existing branch normalizations. For QuantRet, set $F^{\mathrm{C}}_m = \mathcal{U}_m$ and use the spherical update in Theorem 1.

For $j \in \{\mathrm{U},\mathrm{C}\}$, let $e^j_m = \|\widehat{\mathbf{h}}^j_m - \mathbf{h}^j_m\|_2$ be the absolute sequence error and define

$$\begin{aligned} \kappa^j_m &= \frac{\|F^j_m(\widehat{\mathbf{h}}^j_m) - F^j_m(\mathbf{h}^j_m)\|_2}{e^j_m}, \\ q^j_m &= \|\widehat{F}^j_m(\widehat{\mathbf{h}}^j_m) - F^j_m(\widehat{\mathbf{h}}^j_m)\|_2. \end{aligned} \tag{46}$$

Set $\kappa^j_m = 0$ when $e^j_m = 0$. Thus q^j_m isolates quantization of the branch at a fixed input. For QuantRet, these definitions give $\kappa^{\mathrm{C}}_m = \bar{\kappa}_m$ and $q^{\mathrm{C}}_m = r_{\mathrm{seq}}\bar{\varepsilon}_m$ with the choices in Equation (40). Adding and subtracting $F^{\mathrm{U}}_m(\widehat{\mathbf{h}}^{\mathrm{U}}_m)$ in the additive update and applying the triangle inequality gives the first recurrence below; Theorem 1 gives the second:

$$\begin{aligned} e^{\mathrm{U}}_{m+1} &\le (1+\kappa^{\mathrm{U}}_m)e^{\mathrm{U}}_m + q^{\mathrm{U}}_m, \\ e^{\mathrm{C}}_{m+1} &\le \frac{1-\lambda_m+\lambda_m\kappa^{\mathrm{C}}_m}{\mu_m}e^{\mathrm{C}}_m + \frac{\lambda_m}{\mu_m}q^{\mathrm{C}}_m. \end{aligned} \tag{47}$$

Write these as $e^j_{m+1} \le a^j_m e^j_m + b^j_m q^j_m$. Since $e^j_0 = 0$, repeated substitution yields

$$e^j_M \le R_j := \sum_{m=0}^{M-1} b^j_m q^j_m \prod_{\ell=m+1}^{M-1} a^j_\ell. \tag{48}$$

Let g_j be the fixed map from the final residual state to logits, and choose $c_j \geq 0$ such that

$$\|g_j(\widehat{\mathbf{h}}^j_{M,t}) - g_j(\mathbf{h}^j_{M,t})\|_\infty \leq c_j \|\widehat{\mathbf{h}}^j_{M,t} - \mathbf{h}^j_{M,t}\|_2 \tag{49}$$

at all scored positions. For QuantRet, $c_{\mathrm{C}} = \|\mathbf{B}_{\mathrm{C}}\|_{2\to\infty}$ is admissible. For the additive model, c_{U} includes any final normalization as well as the classifier. The proof of Corollary 1 does not require fixed state norms. With $\Delta\mathcal{L}_j = \mathcal{L}_{Q,j} - \mathcal{L}_{\mathrm{FP},j}$, it gives

$$\begin{aligned} |\Delta\mathcal{L}_j| &\leq \frac{2c_j e^j_M}{\sqrt{K}} \leq \mathcal{B}^j_{\mathrm{CE}} := \frac{2c_j R_j}{\sqrt{K}}, \\ \overline{D}_{\mathrm{KL},j} &\leq \frac{c_j^2 (e^j_M)^2}{2K} \leq \mathcal{B}^j_{\mathrm{KL}} := \frac{c_j^2 R_j^2}{2K}. \end{aligned} \tag{50}$$

Proposition 2 (Conditional comparison of loss-error bounds)**.** *Suppose $c_{\mathrm{C}} \leq c_{\mathrm{U}}$ and, at every sublayer,*

$$\kappa^{\mathrm{C}}_m \leq \kappa^{\mathrm{U}}_m, \qquad q^{\mathrm{C}}_m \leq q^{\mathrm{U}}_m, \qquad 0 \leq \lambda_m \leq \tfrac{1}{2}, \qquad \mu_m \geq 1 - \lambda_m. \tag{51}$$

Then $R_{\mathrm{C}} \leq R_{\mathrm{U}}$, $\mathcal{B}^{\mathrm{C}}_{\mathrm{CE}} \leq \mathcal{B}^{\mathrm{U}}_{\mathrm{CE}}$, and $\mathcal{B}^{\mathrm{C}}_{\mathrm{KL}} \leq \mathcal{B}^{\mathrm{U}}_{\mathrm{KL}}$. Both prediction-bound inequalities are strict if $c_{\mathrm{U}} > 0$ and at least one sublayer has $q^{\mathrm{U}}_m > 0$ and $\lambda_m < 1/2$.

Proof. The assumptions imply

$$\begin{aligned} a^{\mathrm{C}}_m &\leq 1 + \frac{\lambda_m}{1-\lambda_m}\kappa^{\mathrm{U}}_m \leq 1 + \kappa^{\mathrm{U}}_m = a^{\mathrm{U}}_m, \\ b^{\mathrm{C}}_m q^{\mathrm{C}}_m &\leq \frac{\lambda_m}{1-\lambda_m} q^{\mathrm{U}}_m \leq q^{\mathrm{U}}_m = b^{\mathrm{U}}_m q^{\mathrm{U}}_m. \end{aligned}$$

Every term in Equation (48) is nonnegative, so comparison term by term gives $R_{\mathrm{C}} \leq R_{\mathrm{U}}$. Equation (50) and $c_{\mathrm{C}} \leq c_{\mathrm{U}}$ give the claimed loss bounds. If $q^{\mathrm{U}}_m > 0$ and $\lambda_m < 1/2$, the local-injection inequality is strict. Since every $a^{\mathrm{U}}_\ell \geq 1$, that contribution gives $R_{\mathrm{C}} < R_{\mathrm{U}}$; $c_{\mathrm{U}} > 0$ then makes both prediction-bound inequalities strict. □

The condition $\mu_m \geq 1 - \lambda_m$ has a geometric sufficient condition: residual and branch vectors have nonnegative inner products at every token in both FP and quantized computations. Equation (39) then gives mixture norms at least $(1 - \lambda_m)r$. In this regime, scalar mixing reduces local error injection without increasing the propagation coefficient above its additive counterpart under Equation (51).

The comparison conditions concern complete branches and output maps, including their learned scales. The matrix results in Appendix C.2 control individual linear error channels; they do not by themselves establish these conditions for whole branches. Proposition 2 orders the stated upper bounds, while realized CE changes also depend on logit direction through Equation (45). If only matrix constraints are removed while spherical residuals are retained, both models instead obey Theorem 1, with their own branch errors, sensitivities, mixture factors, and output gains.

C.6 Quantization Groups and Scale Symmetries

For a nonzero quantization group $\mathbf{t}_g \in \mathbb{R}^{m_g}$, unclipped symmetric RTN at $b \geq 2$ bits has $q_b = 2^{b-1} - 1$ and step $\Delta_g = \|\mathbf{t}_g\|_\infty / q_b$. Nearest rounding gives

$$\|Q^{\mathrm{sym}}_b(\mathbf{t}_g) - \mathbf{t}_g\|_2^2 \leq \frac{m_g \|\mathbf{t}_g\|_\infty^2}{4q_b^2}.$$

For an affine grid with represented endpoints $[\ell_g, u_g]$ covering the group, the bound is $m_g(u_g - \ell_g)^2/[4(2^b - 1)^2]$. An exactly represented constant group has zero error. With exactly represented saturation endpoints, clipping and rounding residuals have disjoint support and their squared errors add. Other endpoint conventions require a separate bound. Refining a partition cannot increase the minimum reconstruction error when each finer group can independently choose the coarse scale, zero point, and clipping threshold. The selected quantizer determines the attained error.

Positive scaling and normalized outputs. For $c > 0$, $\mathcal{S}_r(c\mathbf{u}) = \mathcal{S}_r(\mathbf{u})$. If $Q_W(c\mathbf{W}) = cQ_W(\mathbf{W})$, as for ideal MaxAbs RTN with fixed groups, a bias-free writer also obeys

$$\mathcal{S}_r\big(Q_W(c\mathbf{W})Q_A(\mathbf{x})\big) = \mathcal{S}_r\big(Q_W(\mathbf{W})Q_A(\mathbf{x})\big).$$

Matrix radii can affect normalized paths through the training trajectory, other operations, or departures from quantizer homogeneity. Spherical retraction itself preserves coordinate concentration: $\sqrt{d}\|\mathcal{S}_r(\mathbf{u})\|_\infty/\|\mathcal{S}_r(\mathbf{u})\|_2 = \sqrt{d}\|\mathbf{u}\|_\infty/\|\mathbf{u}\|_2$.

For intermediate MLP features ψ and any radius $r_{\text{mid}} > 0$, linearity gives $\mathcal{S}_r(\mathbf{W}_D\mathcal{S}_{r_{\text{mid}}}(\psi)) = \mathcal{S}_r(\mathbf{W}_D\psi)$ at nonzero normalization inputs. The identity extends to an activation quantizer satisfying $Q_A(c\psi) = cQ_A(\psi)$ for positive token scales. A positive scalar up scale in bias-free SwiGLU similarly factors through the down projection and cancels at branch normalization. Coordinate scales can change feature direction, and the gate scale acts inside SiLU. Appendix B.1 records the evaluated graphs and their numerical conditions.

Scalar residual coefficients and optimization. Scalar mixing and spherical retraction commute with orthogonal changes of coordinates: $\mathcal{S}_r((1-\lambda)\mathbf{Q}\mathbf{h} + \lambda\mathbf{Q}\mathbf{u}) = \mathbf{Q}\mathcal{S}_r((1-\lambda)\mathbf{h} + \lambda\mathbf{u})$. A diagonal coefficient matrix commutes with every orthogonal matrix exactly when its entries are equal, as follows by considering coordinate swaps. This property concerns the mixing operation. The propagation theorem assumes scalar coefficients. Its bounds concern checkpoint forward computations; the ambient Muon updates, hard parameter retractions, and logit backward rule specify how those checkpoints are trained.

D Additional Experimental Protocols and Results

D.1 Detailed Experimental Setup and Evaluation Protocol

Models and baselines. The Qwen3-like decoder configurations use grouped-query attention, RoPE with $\theta = 10^6$, SwiGLU blocks, no linear biases or dropout, a context length of 2048, and shared token-embedding and output-classifier weights. We use a fixed version of the Qwen3 tokenizer (Yang et al., 2025) and a vocabulary of 151936 entries. Table 6 lists the configurations at each scale. Scale labels are approximate; normalization and rotation parameters cause small differences in parameter count across methods.

Table 6: **Main pretraining configurations.** Parameter counts and token-to-parameter ratios are listed in AdamW/OSP/QuantRet order. The global batch contains 128 sequences of length 2048.

Configuration	0.3B	0.6B	1.7B
Transformer blocks	20	28	28
Hidden width d_{model}	768	1024	2048
FFN width d_{ff}	2304	3072	6144
Query/KV heads	12/6	16/8	16/8
Head dimension	128	128	128
Parameters (millions)	293.671/294.224/294.376	596.050/597.033/597.185	1720.575/1724.645/1724.798
Optimizer steps	22460	45562	131592
Warmup steps	1123	2278	6580
Training tokens	5887754240	11943804928	34496053248
Training tokens/parameter	20.049/20.011/20.001	20.038/20.005/20.000	20.049/20.002/20.000

AdamW uses a pre-norm Transformer with learned per-channel RMSNorm (Zhang & Sennrich, 2019). Our matched OSP adaptation uses single-scale normalization (SSNorm) and embedding projection (EmbProj) (Park et al., 2025). An orthogonally initialized shared projection preserves the shared input and output vocabulary weights. OSP shares the architecture dimensions, tokenizer, data stream, and token budget of the other methods. QuantRet-Frob, QuantRet-Spec, and QuantRet-Obl apply the constraints in Section 2.3 with the same architecture and optimization settings.

Pretraining and optimization. We train on a fixed, tokenized FineWeb-Edu training sequence (Penedo et al., 2024). Following the Chinchilla token budget (Hoffmann et al., 2022), the

token budget at each scale is 20 times the largest trainable parameter count (QuantRet), rounded up to a complete optimizer step. Matched AdamW and OSP models use the same data order and budget. The resulting ratios span 20.000–20.049 tokens per parameter (Table 6). Model initialization uses seed 1234 and data ordering uses seed 42. AdamW uses peak learning rate 3×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and matrix weight decay 0.1. OSP and QuantRet optimize hidden matrices with Muon at learning rate 10^{-2}, momentum 0.95, Nesterov momentum, and five Newton–Schulz iterations. Their auxiliary parameters use AdamW at learning rate 10^{-3} with zero weight decay. QuantRet applies hard retraction after each ambient Muon update (Section 2.4). Gradients, momentum, and optimizer states remain in ambient coordinates. The warmups in Table 6 are followed by cosine decay to 10% of the peak learning rate. Training uses BF16 mixed precision and global gradient clipping at 1.0.

Language-modeling and quantization protocol. FineWeb-Edu CE is averaged over 10485760 next-token predictions from a fixed validation prefix, using batch size 4 and context length 1024. The next token after each input sequence is included when computing CE. All precision settings use the same validation text. WikiText-2 uses nonoverlapping blocks of 2048 tokens from the Qwen3-tokenized test stream, evaluating next-token probabilities at 2047 positions per block. The 146 complete blocks contribute 298862 evaluated token positions; the remaining 70 tokens are discarded. Batch size is 4, with the final partial batch retained.

FP denotes the unquantized BF16 reference. All three model scales, the Llama transfer, and all diagnostics use BF16 for model computation. W4 quantizes the seven linear weight matrices in each block; vocabulary weights, normalization operations, and learned scales remain unquantized. A4 quantizes the inputs to these linear layers, and KV4 acts on post-RoPE keys and projected values before attention. W4A16 denotes weight-only quantization with unquantized activations in BF16. Quantizer fitting and range calculations use FP32. Normalization and evaluation statistics use FP32/FP64 accumulation. We use fake quantization: quantized weights and activations are dequantized to BF16 for computation. The main RTN evaluation and the additional W4A4KV4 methods at 0.6B use the numerical protocols below. Calibration and selection details appear in Section D.6.

Main RTN. W4 uses one scale per output row and integers $[-8, 7]$. For each row, we search clipping radii $r(1 - i/100)$ for $i = 0, \ldots, 79$, where $r = \max |W|$, minimizing $\sum |Q(W) - W|^{2.4}$. A4 uses affine $[0, 15]$ quantization per token, with per-head groups at the attention-output input. K/V4 uses affine quantization per token and KV head. Affine ranges include zero.

SmoothQuant. W4 uses signed-symmetric, unclipped MaxAbs quantization per output row. A4 and KV4 use the dynamic affine quantizers of the main RTN evaluation. Q/K/V share each smoothing vector, and gate/up share another. We divide activations by the smoothing vectors and multiply the corresponding weight columns by the same vectors before quantization. The attention-output and down projections are quantized without smoothing.

GPTQ. W4 uses affine MinMax quantization in groups of 128, with ranges that include zero. The reconstruction block size is 128 and damping is 0.01. Activation ordering and static groups are disabled, and `true_sequential=false`. A4 and KV4 use the dynamic affine quantizers of the main RTN evaluation.

QuaRot. W4 uses symmetric per-output-channel quantization with the clipping search of main RTN. A4 uses symmetric quantization per token, with clipping ratio 0.9 and scale $0.9 \max |x|/7$. K/V4 uses affine quantization per token and KV head, with clipping ratio 0.95, no forced zero in the range, and an unclamped rounded zero point. We apply random orthogonal rotations to the residual stream and Hadamard transforms within attention and MLP layers.

Zero-shot evaluation protocol. We evaluate FP, W4A4, and W4A4KV4 at all three scales with `lm-eval` (Biderman et al., 2024) v0.4.12, zero-shot prompting, batch size 4, and maximum sequence length 2048. The FP reference uses the same model with quantization disabled. Evaluation uses the same Qwen3 tokenizer without a BOS token, with `<|endoftext|>` (ID 151643) for padding and EOS ID 151645. The eight-task evaluation suite contains ARC-Easy, ARC-Challenge (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), and WinoGrande (Sakaguchi et al., 2021), totaling 22419 examples. We report the equally weighted average of these eight task scores. Table 7 lists the evaluation splits and metrics. The BoolQ prompt contains

the passage, question, and answer prefix, with "no" and "yes" as the candidate answers. The `acc` metric selects answers by log-likelihood, while `acc_norm` uses length-normalized log-likelihood. MMLU uses example-weighted accuracy across 57 subjects and 14042 examples (Hendrycks et al., 2021). No decontamination filter is applied.

Table 7: **Zero-shot evaluation splits and primary metrics.** The eight task scores contribute equally to their reported average; MMLU is reported separately.

Task	Split	Metric	Examples
ARC-Easy	Test	`acc_norm`	2376
ARC-Challenge	Test	`acc_norm`	1172
BoolQ	Validation	`acc`	3270
HellaSwag	Validation	`acc_norm`	10042
OpenBookQA	Test	`acc_norm`	500
PIQA	Validation	`acc_norm`	1838
SocialIQA	Validation	`acc`	1954
WinoGrande	Validation	`acc`	1267
MMLU	Test	`acc`	14042

Scope and limitations. The main comparisons cover 0.3B, 0.6B, and 1.7B models; mechanism analyses focus on 0.6B, and component and scale ablations on 0.3B. Each configuration is pretrained once with fixed initialization and data-order seeds. All result comparisons are descriptive; we do not perform significance tests or report confidence intervals for the eight-task average, and no paired-example resampling is performed. Variation across layers and data blocks does not measure variability across independent training runs. Additional pretraining seeds, 0.6B component ablations, full cross-scale mechanism replication, and larger-scale cross-architecture transfer remain unmeasured. Llama 3.1 evaluation covers language modeling only. Quantization is simulated to evaluate model quality. We do not measure the training overhead of retraction or the latency, throughput, and memory use of native low-bit inference.

D.2 Additional Zero-Shot Results

We break down the low-bit scores from Table 2 by task in Tables 8 and 9, using the metrics in Table 7. Bold identifies the highest observed score. For Frob, adding KV4 changes the eight-task average by $+0.34$, -0.17, and -0.03 percentage points at 0.3B, 0.6B, and 1.7B, respectively. On BoolQ, OSP has the highest score at 0.3B and 1.7B and AdamW at 0.6B under both quantization settings.

D.3 Local Reconstruction and Quantizer Dependence

In this subsection, we examine activation distributions, local linear-output errors, and the effects of quantizer granularity and range to complement the prediction-level measurements in Section 4.3. We study prediction sensitivity and error propagation separately in Appendix D.4.

D.3.1 Activation Tails and Local Reconstruction Error

Figure 3 reports survival curves of token-level statistics at attention and down-projection inputs. Each curve pools histogram counts over the same 32768 tokens at each of the 28 layers of a 0.6B model. In all activation-tail figures, curves follow histogram bin boundaries and show only positive survival probabilities on logarithmic axes. Alongside MaxAbs/RMS, we report a fourth-moment statistic motivated by OSP's analysis of activation tails (Park et al., 2025). For a token vector $x_t \in \mathbb{R}^D$, define

$$K_t = \left[\frac{D^{-1} \sum_j x_{t,j}^4}{\max\{\hat{\sigma}_t, 10^{-8}\}^4} - 3 \right]_+, \qquad \hat{\sigma}_t^2 = \frac{1}{D-1} \sum_j (x_{t,j} - \bar{x}_t)^2,$$

where $[a]_+ = \max\{a, 0\}$ and $\bar{x}_t$ is the channel mean. The numerator is the uncentered fourth moment; the denominator uses the sample standard deviation, with variance computed using the

Table 8: **Zero-shot task scores (%) under W4A4.** The eight-task average and MMLU are reported in Table 2. Bold marks the highest task scores within each scale.

Model scale	Method	Evaluation Tasks							
		ARC-E	ARC-C	BoolQ	HellaSwag	OBQA	PIQA	SIQA	WinoGrande
0.3B	AdamW	41.84	25.26	60.52	31.81	29.00	59.09	35.93	51.07
	OSP	46.89	26.79	**60.64**	33.98	28.00	61.53	37.15	**51.62**
	QuantRet-Frob	**50.08**	**27.82**	53.52	**36.34**	30.20	**62.51**	**37.92**	48.86
	QuantRet-Spec	43.73	25.94	59.57	33.82	29.20	61.53	37.77	49.80
	QuantRet-Obl	49.12	27.65	56.33	36.21	**30.80**	62.35	37.36	49.33
0.6B	AdamW	48.06	26.02	**62.14**	34.55	27.80	61.48	36.34	50.75
	OSP	51.47	27.65	46.61	40.06	30.00	64.47	37.31	**54.14**
	QuantRet-Frob	**53.07**	**31.91**	58.99	43.16	**34.40**	**65.67**	37.97	53.04
	QuantRet-Spec	51.52	26.62	57.52	38.61	32.20	63.76	**38.69**	52.49
	QuantRet-Obl	51.56	30.80	60.73	**44.12**	34.00	65.02	38.33	51.93
1.7B	AdamW	41.54	23.98	60.09	29.49	27.20	57.78	34.49	46.96
	OSP	59.85	34.04	**60.80**	46.41	32.80	64.96	39.20	53.51
	QuantRet-Frob	**60.23**	**36.09**	54.80	51.59	**39.20**	**68.23**	**40.23**	52.88
	QuantRet-Spec	57.66	33.62	57.89	47.62	36.60	67.14	40.17	**54.93**
	QuantRet-Obl	58.63	**36.09**	53.55	**51.65**	38.00	67.46	39.15	54.14

Table 9: **Zero-shot task scores (%) under W4A4KV4.** The eight-task average and MMLU are reported in Table 2. Bold marks the highest task scores within each scale.

Model scale	Method	Evaluation Tasks							
		ARC-E	ARC-C	BoolQ	HellaSwag	OBQA	PIQA	SIQA	WinoGrande
0.3B	AdamW	41.54	24.91	60.80	32.04	29.00	59.85	35.41	50.43
	OSP	46.25	25.77	**61.28**	33.73	27.60	60.94	36.54	51.22
	QuantRet-Frob	**50.59**	**27.73**	53.79	35.86	**32.00**	61.92	37.82	50.28
	QuantRet-Spec	45.37	26.71	60.18	33.42	30.80	60.55	37.46	49.33
	QuantRet-Obl	48.36	27.22	55.87	**36.16**	30.60	**62.68**	**37.97**	**53.35**
0.6B	AdamW	48.27	24.49	**61.59**	34.79	30.00	62.68	36.08	49.57
	OSP	51.22	27.82	45.96	39.70	31.60	64.20	37.92	**52.80**
	QuantRet-Frob	**52.82**	**32.85**	60.28	43.11	34.20	64.64	37.77	51.22
	QuantRet-Spec	51.94	28.92	58.01	38.65	33.00	63.87	38.59	52.57
	QuantRet-Obl	52.48	30.97	60.21	**43.55**	**35.40**	**64.96**	**39.41**	52.64
1.7B	AdamW	39.23	23.21	59.91	28.91	27.00	58.00	33.98	51.07
	OSP	57.91	33.62	**60.89**	46.04	31.20	65.67	39.56	**55.41**
	QuantRet-Frob	59.18	**35.58**	54.04	**51.73**	**40.40**	68.06	**41.30**	52.72
	QuantRet-Spec	57.41	33.19	58.17	47.91	35.80	65.72	40.48	53.59
	QuantRet-Obl	**59.34**	34.30	51.80	51.15	39.60	**68.28**	40.23	53.35

divisor $D - 1$. At the attention input, the mean of layer-wise maximum MaxAbs/RMS values is 28.12 for AdamW, 5.57 for OSP, and 5.34–5.48 for QuantRet. The K_t distributions show the same separation.

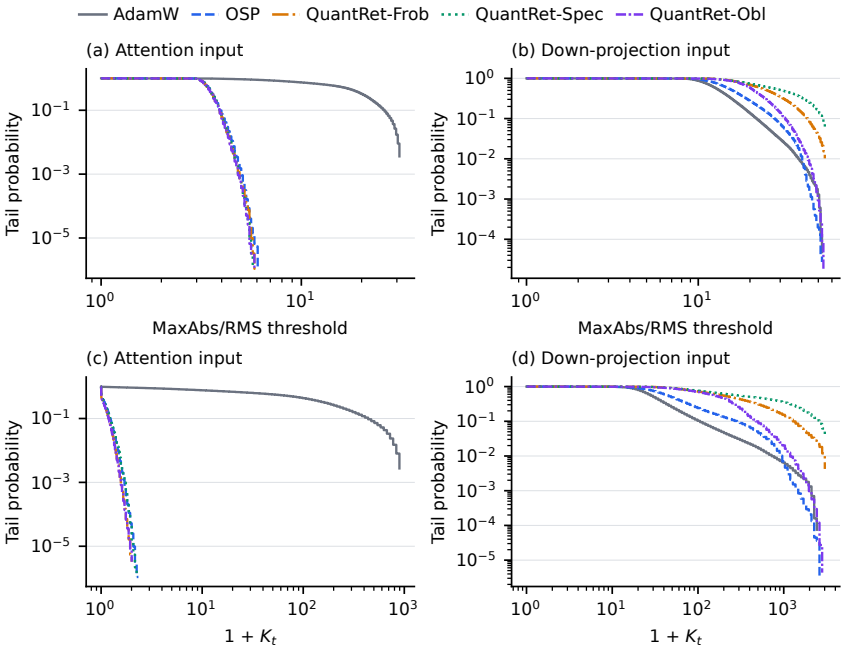

Figure 3: **Token-level activation statistics at 0.6B.** Top: MaxAbs/RMS at attention and down-projection inputs. Bottom: $1 + K_t$ at the same projection inputs. Survival curves pool histogram counts over 28 layers with 32768 tokens per layer.

At the down-projection input, the mean of layer-wise maximum MaxAbs/RMS values is 49.25/49.17 for AdamW/OSP and 53.40/53.64/52.00 for QuantRet-Frob/Spec/Obl. QuantRet-Spec has the lowest mean dynamic-A4 normalized mean squared error (NMSE) at the down-projection input: 0.107, compared with 0.132 for AdamW and 0.142 for OSP. The ordering by MaxAbs/RMS differs from the ordering by local quantization error.

D.3.2 Local Output Error and Norm Balance

Reader/writer norm balance. Using activations from the FP model, we measure norm balance and the local output error $\mathbf{W}[Q_A(\mathbf{x}) - \mathbf{x}]$ induced by activation quantization with unquantized weights. The measurements cover 32768 input tokens, 140 reading matrices, and 56 writing matrices. We report the coefficient of variation (CV) of row squared norms for readers and column squared norms for writers (Table 10).

Table 10: **Reader/writer norm balance and local output errors.** Measurements use the 0.6B models. CV columns average row-squared-norm coefficients of variation for readers and column-squared-norm coefficients of variation for writers. Actual/diag. averages the measured local output-error energy divided by its diagonal second-moment prediction.

	Squared-norm CV		Actual/diag.	
Method	Reader row	Writer column	Reader	Writer
AdamW	0.1812	0.1182	1.0423	1.1590
OSP	0.3038	0.2670	0.9999	1.0330
QuantRet-Frob	0.3138	0.3048	1.0001	1.0765
QuantRet-Spec	0.8856	1.0091	1.0005	1.0381
QuantRet-Obl	0.0002	0.0002	1.0000	1.1063

QuantRet-Obl retains nearly equal reader-row and writer-column norms after BF16 casting. Mean actual-to-diagonal ratios differ from one by less than 16% for every method and role, and less than 11% for both QuantRet-Obl roles. These matrix averages summarize agreement with diagonal predictions for the measured activation errors. Proposition 1 retains its reader and grouped-error isotropy assumptions.

Joint W4A4 output error. To include both weight and activation rounding, we measure the signal-to-noise ratio (SNR) (Fishman et al., 2026) of local W4A4 matrix outputs on 32768 input tokens from FineWeb-Edu. The diagnostic covers seven linear maps at layers 0, 14, and 27: 15 readers and six writers. Weight quantization uses symmetric per-output-channel RTN, selecting clipping thresholds to minimize weight quantization error. Activations use dynamic per-token affine RTN, with head grouping at attention-output inputs. We quantize copies of the recorded weights and activations to measure local reconstruction error without propagating it to subsequent layers. For a set of matrices $\mathcal{M}$, with $\mathbf{x}_t$ denoting each map's recorded input, the pooled statistic is

$$\mathrm{SNR}_{\mathcal{M}} = 10 \log_{10} \frac{\sum_{\mathbf{W} \in \mathcal{M}, t} \|\mathbf{W}\mathbf{x}_t\|_2^2}{\sum_{\mathbf{W} \in \mathcal{M}, t} \|Q_4(\mathbf{W})Q_4(\mathbf{x}_t) - \mathbf{W}\mathbf{x}_t\|_2^2}.$$

Signal and error energies are pooled before conversion to decibels. Figure 4 reports reader, writer, and combined values. All three QuantRet variants have higher reader and combined SNR than AdamW and OSP. Writer SNR is higher than both baselines for QuantRet-Obl and lower than both baselines for QuantRet-Frob and QuantRet-Spec. This statistic measures error in the linear output after joint weight and activation rounding; input A4 NMSE measures activation reconstruction alone.

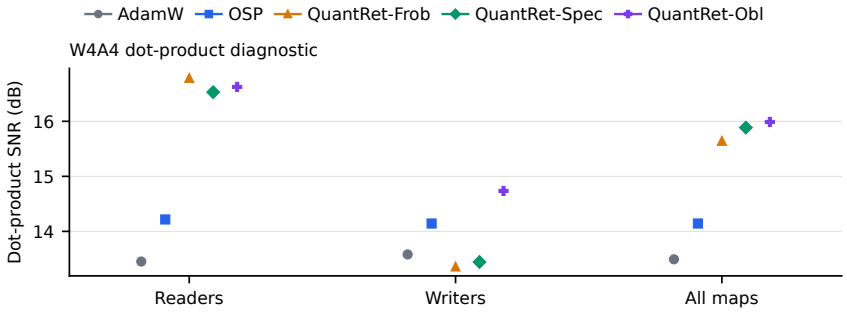

Figure 4: **Local W4A4 dot-product SNR at 0.6B.** Each point is computed from pooled signal and error energies for 15 reading matrices, six writing matrices, or all 21 matrices at layers 0, 14, and 27. All matrices are evaluated on the same 32768 input tokens. Higher values indicate a larger ratio of FP output energy to local quantization-error energy.

D.3.3 Quantizer Granularity and Range

Weight granularity. We compare quantization granularities under uniform RTN on the five 0.6B checkpoints. Signed-symmetric W4 with unclipped MaxAbs ranges is applied to the seven linear maps in each Transformer block. Activations, the KV cache, token embeddings, and the language-model head remain in full precision. Function-preserving rotations are fused before quantization. Evaluation follows the FineWeb-Edu and WikiText-2 protocols in Appendix D.1. Table 11 defines the seven partitions.

At 0.6B, the per-tensor partition produces the largest separation: all QuantRet variants remain below CE 3.0, and QuantRet-Obl gives the lowest CE and perplexity.

Channel and group partitions improve resolution. Using the unrounded results, QuantRet-Obl has the lowest CE in six of seven partitions, while QuantRet-Frob has the lowest perplexity in five. The benefit of reader/writer partitioning varies across methods and metrics. Group schemes use 3440640 scales, compared with 196 for per-tensor, 344064 for output-channel, 286720 for input-channel, and 430080 for reader/writer channel quantization. Finer groups improve quantization resolution but require more quantization scales.

Table 11: **Weight RTN granularity.** Signed-symmetric W4 uses unclipped MaxAbs ranges; activations, the KV cache, embeddings, and the language-model head remain in full precision. CE is evaluated on FineWeb-Edu and PPL on WikiText-2. Lower CE and PPL are better. Out./In. denote output/input channels (rows/columns). Role denotes reader/writer alignment: rows for Q, K, V, gate, and up; columns for O and down. G128 splits each aligned vector into groups of 128 values. The 0.3B replication tests only the per-tensor partition.

Method	Tensor	Out. channel	In. channel	Out. G128	In. G128	Role channel	Role G128
0.6B: FineWeb-Edu CE							
AdamW	9.260	2.768	2.799	2.736	2.733	2.809	2.739
OSP	4.590	2.921	2.908	2.800	2.805	2.884	2.802
QuantRet-Frob	2.933	2.759	2.758	2.727	2.730	2.751	2.728
QuantRet-Spec	2.999	2.897	2.896	2.871	2.873	2.893	2.872
QuantRet-Obl	2.855	2.756	2.755	2.727	2.727	2.752	2.726
0.6B: WikiText-2 PPL							
AdamW	2.16×10^4	21.34	22.79	20.76	20.61	23.13	20.77
OSP	247.85	23.77	23.88	21.56	21.80	23.03	21.67
QuantRet-Frob	27.25	21.35	20.78	20.32	20.26	21.03	20.29
QuantRet-Spec	28.27	24.14	24.08	23.36	23.28	24.10	23.34
QuantRet-Obl	24.96	21.73	21.50	21.11	20.79	21.53	20.99

0.3B replication: per-tensor W4		
Method	FineWeb-Edu CE	WikiText-2 PPL
AdamW	3.449	49.55
OSP	3.966	78.70
QuantRet-Frob	3.181	36.49
QuantRet-Spec	3.179	37.28
QuantRet-Obl	3.071	31.62

The 0.3B replication in Table 11 uses the same signed-symmetric, unclipped per-tensor W4 rule. At both scales, QuantRet-Obl gives the lowest CE and perplexity, and every QuantRet variant outperforms both baselines under this coarse partition.

Range and clipping robustness. We vary the per-tensor W4 range rule while keeping the other weight-only settings and evaluation streams fixed. One rule uses a signed-symmetric range with a clipping threshold selected to minimize weight quantization error. The other uses an unclipped affine MinMax range that includes zero.

Table 12: **Range selection for per-tensor W4.** We compare symmetric clipped ranges with affine MinMax ranges. Activations, the KV cache, embeddings, and the language-model head remain unquantized in BF16. Values are FineWeb-Edu CE and WikiText-2 perplexity. Bold marks the lowest displayed value within each scale and column.

		Symmetric clipped		Affine MinMax	
Scale	Method	CE	PPL	CE	PPL
0.3B	AdamW	3.086	30.91	3.313	41.91
	OSP	3.035	29.71	3.578	51.34
	QuantRet-Frob	2.951	27.12	3.101	32.94
	QuantRet-Spec	3.023	30.01	3.124	34.67
	QuantRet-Obl	**2.912**	**25.63**	**3.027**	**29.60**
0.6B	AdamW	3.135	31.93	6.920	1.79×10^{3}
	OSP	2.829	22.14	3.440	42.50
	QuantRet-Frob	2.764	20.95	2.883	24.45
	QuantRet-Spec	2.879	23.42	2.968	26.66
	QuantRet-Obl	**2.722**	**20.76**	**2.823**	**24.03**

QuantRet-Obl gives the lowest CE and PPL at both scales under both range rules, and QuantRet-Frob improves both metrics over AdamW and OSP in every comparison (Table 12). QuantRet-Spec improves both metrics over both baselines with affine MinMax. Under symmetric clipping, it improves CE at 0.3B but has higher PPL than OSP; at 0.6B, it has higher CE and PPL than OSP.

Activation granularity. We keep weights and the KV cache in full precision and apply dynamic A4 RTN to the inputs of the 196 linear layers within the Transformer blocks. We compare per-tensor, per-token, per-channel, and group-wise quantization (128 values per group), each with symmetric and affine ranges and without clipping or calibration. All conditions use the FineWeb-Edu and WikiText-2 streams from the weight-only experiment.

Activation RTN produces a different ordering from weight RTN (Table 13). QuantRet-Spec gives the lowest CE and perplexity with either per-tensor range and with per-token symmetric ranges, consistent with the spectral bound on worst-direction gain. With affine groups of 128, QuantRet-Frob gives the lowest absolute CE and perplexity. QuantRet-Spec has the smallest increase from its FP reference and a higher FP CE. Affine ranges outperform symmetric ranges in every matched method–partition pair.

D.4 Prediction Sensitivity and Error Propagation

We detail the protocols and supplementary results for Section 4.3, covering controlled perturbations, quantization of one projection input at a time, and full W4A4. Local reconstruction and norm-balance measurements are in Appendix D.3.2.

D.4.1 Diagnostic Protocol

We use the five pretrained 0.6B models with their original architectures, including QuantRet's gated-feature normalization before the down projection (Appendix B.1). The forward computation uses BF16 with TF32 disabled, FP32/FP64 accumulation for the diagnostic statistics, and context length 1024. Unless stated otherwise, batch size is one and all conditions use the same sequential FineWeb-Edu validation prefix. Each perturbation is applied independently and compared with the same model's unperturbed predictions on identical tokens.

Table 13: **Activation RTN granularity at 0.6B.** Dynamic A4 uses symmetric (Sym.) or affine ranges without clipping or calibration; weights and the KV cache remain in full precision. CE is evaluated on FineWeb-Edu and PPL on WikiText-2. Lower CE and PPL are better. Tensor, Token, and Chan. denote per-tensor, per-token, and per-channel partitions, respectively; per-channel scales are shared across batch and sequence positions. G128 assigns one scale per group of 128 values.

Method	Tensor Sym.	Tensor Affine	Token Sym.	Token Affine	Chan. Sym.	Chan. Affine	G128 Sym.	G128 Affine
0.6B: FineWeb-Edu CE								
AdamW	10.330	9.218	5.718	3.597	3.042	2.869	2.835	2.753
OSP	8.972	8.440	3.348	3.050	3.047	2.833	2.786	2.735
QuantRet-Frob	8.994	7.349	3.284	3.039	2.983	2.808	2.748	2.721
QuantRet-Spec	5.548	4.928	3.172	3.038	3.021	2.927	2.885	2.868
QuantRet-Obl	7.718	6.581	3.259	3.015	2.984	2.808	2.750	2.722
0.6B: WikiText-2 PPL								
AdamW	3.76×10^4	1.37×10^4	455.93	53.35	29.80	24.02	24.81	21.81
OSP	1.12×10^4	6.35×10^3	37.57	27.23	29.53	22.77	21.27	20.28
QuantRet-Frob	1.36×10^4	3.33×10^3	38.49	28.85	27.21	22.29	20.62	20.01
QuantRet-Spec	404.53	207.30	32.44	28.05	27.92	24.98	23.57	23.17
QuantRet-Obl	5.42×10^3	1.46×10^3	37.89	28.54	28.02	22.94	21.19	20.58

Quantization and samples. In 56 conditions, we quantize one attention-output or down-projection input to A4 at a time across all 28 layers, keeping all weights and other activations in FP. Attention-output projection inputs use affine quantization within each head; down-projection inputs use affine per-token quantization. Each condition at layers 0, 14, and 27 is evaluated on 32 sequences of 1024 tokens; the other layers use the first eight sequences. The 18 relative-perturbation conditions reported here combine these six layer/projection choices with three directions at 5% of each token's input norm, each evaluated on all 32 sequences. Full W4A4 quantization is evaluated on the same 32 sequences using the main RTN settings, including weight clipping and attention head grouping. The KV cache, vocabulary weights, and classifier input remain unquantized.

Error measurement and aggregation. The projection output, normalized branch, residual mixture, and retracted residual are defined in Appendix D.4.3. For attention at layer ℓ, $m = 2\ell$ and $\mathbf{h}_{m+1} = \mathrm{vec}(\mathbf{Z}_\ell)$; for the MLP, $m = 2\ell + 1$ and $\mathbf{h}_{m+1} = \mathrm{vec}(\mathbf{H}_{\ell+1})$ (Equations (6) and (8)). The attention observation therefore precedes the same block's MLP. The final residual is measured before terminal normalization, and classifier-input error is measured after it.

For $E(\mathbf{z}) = \mathrm{RMS}(\widehat{\mathbf{z}} - \mathbf{z})$, squared differences and element counts are pooled over matching tokens before taking the square root. Relative error divides $E(\mathbf{z})$ by the FP reference RMS. Each local-operation ratio in Figure 5(a),(d) is calculated from the same first eight sequences before averaging the 28 layer-specific ratios. Prediction changes use token-mean $D_{\mathrm{KL}}(p_{\mathrm{FP}} \| p_{\mathrm{perturbed}})$ and signed $\Delta\mathrm{CE} = \mathrm{CE}_{\mathrm{perturbed}} - \mathrm{CE}_{\mathrm{FP}}$, computed from the model logits with FP64 accumulation.

D.4.2 Relative-Perturbation Protocol

We add noise to one attention-output or down-projection input in layers 0, 14, and 27, keeping weights and all other computations in FP. Isotropic noise uses a Gaussian direction. Anisotropic noise uses Gaussian coordinates with a shared, deterministically permuted log-spaced variance profile and a 64:1 variance ratio. Leading-direction noise uses the projection matrix's estimated leading right singular vector with independent token signs. Seed 1234 fixes the random draws across models; the singular vector is specific to each model and projection.

Each noise vector is normalized in FP32 to $\|\mathbf{e}_t\|_2 = \eta\|\mathbf{x}_t\|_2$, where $\eta = 0.05$, before the perturbed input is cast to BF16. The singular-vector estimate uses 32 power iterations. Across the 30 model/projection/layer combinations evaluated at relative amplitude 0.05, the relative eigenpair residuals $\|\mathbf{W}^\top\mathbf{W}\hat{\mathbf{v}} - \hat{\sigma}^2\hat{\mathbf{v}}\|_2 / \|\mathbf{W}^\top\mathbf{W}\hat{\mathbf{v}}\|_2$, with estimated singular value $\hat{\sigma}$ and right singular vector $\hat{\mathbf{v}}$, range from 1.25×10^{-7} to 0.0297, with eight above 0.01; these estimates do not certify the exact worst perturbation direction.

Figure 1 reports output KL divergence and signed CE changes at 5% relative amplitude, for all three directions and all three perturbed layers.

D.4.3 Quantizing One Projection Input

We quantize one attention-output or down-projection input to A4 at a time across all 28 layers, keeping all remaining computation in FP. Figure 5 uses the same eight 1024-token sequences for every layer and compares each quantized computation with its own checkpoint's FP computation. The two rows correspond to the two projection types.

How do QuantRet's local operations change quantization error? For the three QuantRet variants, we measure how each operation changes the activation error. With the notation of Section 3, $\mathbf{u}_m$ concatenates the token-wise projection outputs $\mathbf{u} = \mathbf{W}\mathbf{x}$. Branch normalization gives $\mathcal{U}_m(\mathbf{h}_m) = \mathcal{S}_r(\mathbf{u}_m)$, residual mixing gives $\mathbf{v}_m = (1-\lambda_m)\mathbf{h}_m + \lambda_m \mathcal{U}_m(\mathbf{h}_m)$, and spherical retraction gives $\mathbf{h}_{m+1} = \mathcal{S}_r(\mathbf{v}_m)$, with $\mathcal{S}_r$ applied token-wise. Figure 5(a),(d) plots $E(\text{after})/E(\text{before})$ for these three operations, where $E(\mathbf{z}) = \mathrm{RMS}(\widehat{\mathbf{z}} - \mathbf{z})$ compares quantized and FP activations from the same checkpoint. Values below one indicate smaller absolute error. Branch normalization increases Spec's error, residual mixing reduces error, and retraction can increase it. These measurements describe local error scaling; they do not by themselves determine the final prediction change.

How does the impact of local quantization error vary across depth? Figure 5 shows the output effects of local A4 errors across network depth: relative classifier-input error in (b),(e) and output KL divergence in (c),(f). The horizontal axis identifies the quantized layer: each point comes from a separate computation in which only that projection input is quantized. Full-W4A4 errors accumulated during one computation are shown in Figure 2.

QuantRet has lower KL divergence at early layers, with reversals against AdamW at some later layers. When evaluated on 32 sequences of 1024 tokens at layer 14's down-projection input, KL divergence is 0.00293 for AdamW, 0.00431 for OSP, and 0.00644, 0.00384, and 0.00655 for Frob, Spec, and Obl. The comparisons use each checkpoint's actual rounding error.

D.5 Learnable-Scale Parameterization

In this subsection, we examine how the granularity of attention, gate/up, and residual scales affects FP quality, quantization degradation, and activation distributions.

Factorial design. We pretrain eight 0.3B QuantRet-Obl models, one for each combination of three binary factors (a full 2^3 factorial design). The factors are attention-score granularity (one scalar per query head, H, or one value per query-head coordinate, C), gate/up parameterization (a learnable scalar gate with an equivalent unit up factor, S, or learnable channel vectors on both paths, V), and joint attention/MLP residual granularity (S or V). The vocabulary scale is a vector in every configuration. Table 14 defines the eight configurations by their attention, gate/up, and residual settings. Coordinate-wise attention scaling multiplies each normalized query coordinate by its positive learned scale before the dot product. The gate scale is learned in every configuration and acts inside SiLU; configurations with vector gate/up scales also learn a channel-wise up scale. Vector residual coefficients apply the interpolation coordinate-wise before the final spherical normalization. Positive attention, gate, and vector up scales initialize to one, and effective residual coefficients initialize to 0.05 using the same parameter rescaling and upper bound as the scalar parameterization (Appendix B.2).

All eight runs use the 20-layer, width-768 architecture, tied Qwen3 vocabulary table, tokenizer, and optimizer described in Appendix D.1, with initialization seed 1234 and data-order seed 42. Each run processes 5887754240 tokens in 22460 optimizer steps. Gated features pass directly to the down projection, with spherical normalization at the branch output and residual mixture. The models in the main comparison use the normalization before the down projection described in Appendix B.1; C0 is the separately pretrained reference for this factorial study. Quantizer settings and evaluation streams match Table 1. Table 14 reports results for all eight configurations at four precision settings. The gate and up settings are varied jointly, as are the two residual branches.

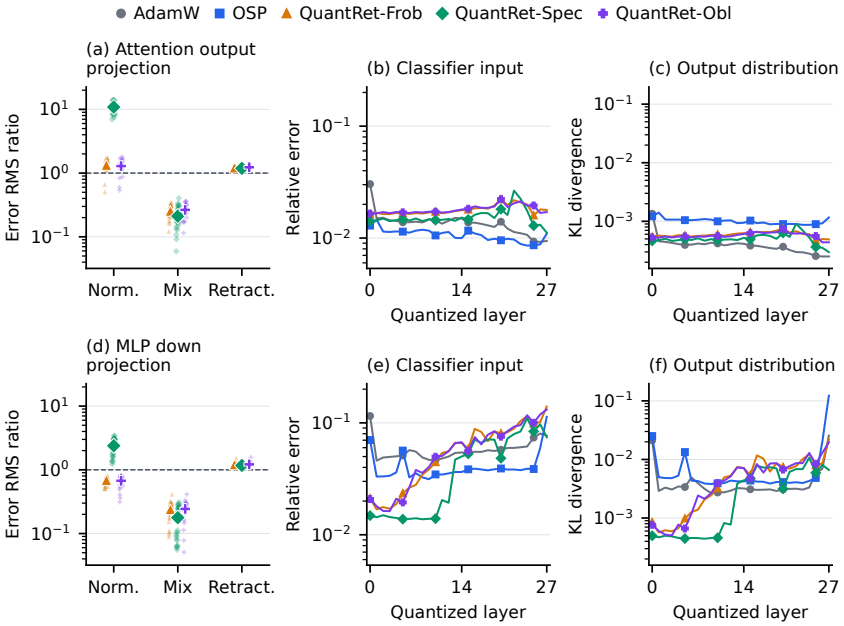

Figure 5: **Local error scaling and prediction changes under single-input A4.** We evaluate the 0.6B models on eight 1024-token sequences. Rows identify attention-output and down-projection inputs. **(a),(d):** absolute error RMS ratios across branch normalization, residual mixing, and spherical retraction for the three QuantRet variants; small points show layers, large markers their means, and dashed lines mark one. **(b),(e):** relative classifier-input error. **(c),(f):** output KL divergence. The last two columns include all five methods; lines connect separate quantized-layer experiments.

Table 14: **Learnable-scale ablations at 0.3B.** Each adjacent pair varies only residual granularity. The configuration column lists attention, gate/up, and residual settings: H denotes one scale per query head, C one per head coordinate, S a scalar, and V a vector. *For Gate/up, S denotes a learnable scalar gate with the up factor represented equivalently by the constant 1; V denotes learnable vectors on both paths. Vocabulary scales are vectors, and gated features pass directly to the down projection. Small subscripts give ΔCE (quantized minus FP), computed before rounding. Lower absolute CE and PPL are better.

		FineWeb-Edu CE ↓				WikiText-2 PPL ↓			
Config.	Attention-Gate/up*-Residual	FP	W4A16	W4A4	W4A4KV4	FP	W4A16	W4A4	W4A4KV4
C0	H-S-S	2.860	$2.907_{+0.048}$	$3.118_{+0.258}$	$3.126_{+0.266}$	25.04	26.66	33.52	33.87
C4	H-S-V	2.843	$2.911_{+0.068}$	$3.132_{+0.289}$	$3.142_{+0.300}$	24.18	26.50	33.27	33.66
C1	C-S-S	2.859	$2.909_{+0.051}$	$3.116_{+0.257}$	$3.124_{+0.265}$	25.03	26.82	33.65	33.92
C5	C-S-V	2.843	$2.904_{+0.061}$	$3.115_{+0.272}$	$3.124_{+0.281}$	23.81	25.68	32.08	32.39
C2	H-V-S	2.857	$2.908_{+0.050}$	$3.168_{+0.311}$	$3.176_{+0.319}$	24.68	26.46	36.01	36.33
C6	H-V-V	2.841	$2.907_{+0.066}$	$3.163_{+0.321}$	$3.173_{+0.332}$	23.83	25.82	34.05	34.49
C3	C-V-S	2.857	$2.907_{+0.050}$	$3.158_{+0.301}$	$3.165_{+0.308}$	24.58	26.06	34.86	35.28
C7	C-V-V	2.841	$2.902_{+0.061}$	$3.144_{+0.303}$	$3.152_{+0.311}$	24.12	26.07	33.41	33.89

Quantized quality. Across all four matched residual pairs, vector scales improve FP CE. Under W4A4 and W4A4KV4, they improve PPL but increase the CE gap from FP, with mixed changes in absolute CE (Table 14). Joint vector gate/up scales worsen CE and PPL under both quantization settings in every matched pair.

For the attention-scale pairs C0/C1, C2/C3, C4/C5, and C6/C7, coordinate scales lower W4A4 CE in each pair. W4A4 PPL improves in three pairs and increases slightly for C0/C1 (Table 14).

Activation diagnostics. We measure activations in BF16 on the same 131072 FineWeb-Edu validation input positions for each model, comprising four disjoint blocks of 32768 positions with context length 1024. Measurements cover all 20 layers without modifying the model outputs. For a token vector x_t, define $y_{t,j} = x_{t,j}/\max\{\mathrm{RMS}(x_t), 10^{-8}\}$. Residual channel concentration is the fraction of total $\sum_{t,j} y_{t,j}^2$ carried by the eight channels with largest $\sum_t y_{t,j}^2$, corresponding to the top $\lceil 0.01d \rceil$ channels at $d = 768$. We estimate the 99th percentile across tokens of the maximum normalized activation magnitude, $\max_j |y_{t,j}|$, from activation histograms. Local A4 normalized mean squared error (NMSE) is $\sum_t \|Q_4(x_t) - x_t\|_2^2 / \sum_t \|x_t\|_2^2$. We apply the main protocol's dynamic affine per-token quantizer to the recorded activations; the resulting errors are not propagated through the model.

Table 15: **Activation statistics across learnable-scale configurations.** Results for the eight configurations are averaged equally over 20 layers on the shared 131072-token prefix. P99 is the histogram estimate of the per-layer 99th percentile of token MaxAbs/RMS. Energy is the share of normalized residual energy in the top 1% of channels. NMSE columns give local A4 error multiplied by 100. MLP input denotes the post-attention residual.

	Block-output residual		A4 NMSE (%)		
Config.	P99	Energy (%)	Attention input	MLP input	Down input
C0	4.360	3.63	1.509	1.514	13.288
C4	7.025	11.29	2.885	2.865	13.201
C1	4.318	3.22	1.498	1.497	13.318
C5	7.336	10.60	2.929	2.868	13.055
C2	4.327	3.63	1.512	1.514	12.730
C6	7.210	11.70	3.018	3.101	13.024
C3	4.343	3.34	1.502	1.501	12.506
C7	6.985	10.33	2.705	2.715	12.900

Activation concentration and quantization error. Across all four matched comparisons of scalar and vector residual coefficients, mean residual P99 increases by 60.8–69.9%, and the energy share of the top 1% of channels increases from 3.22–3.63% to 10.33–11.70% (Table 15). Attention-input A4 NMSE increases by 80.1–99.7% and MLP-input NMSE by 81.0–104.9%. The concentration and local-error differences are positive at all 20 layers in each pair and after averaging within each data block. Figure 6 shows larger separations in shallow and middle layers, with smaller differences near the final residual output. Figure 7 separates element-level and token-maximum tails: the scalar/vector residual contrast is pronounced in block-output residuals and MLP inputs, while down-input curves largely overlap.

Joint vector gate/up scales increase the mean energy share of the top 1% of channels at the down-projection input by 7.4–17.0% and reduce mean down-input A4 NMSE by 1.2–6.1% across the four pairs. Their W4A4 CE and perplexity increase despite the lower local error. These comparisons show associations between activation statistics and quantized model quality, without establishing that one causes the other.

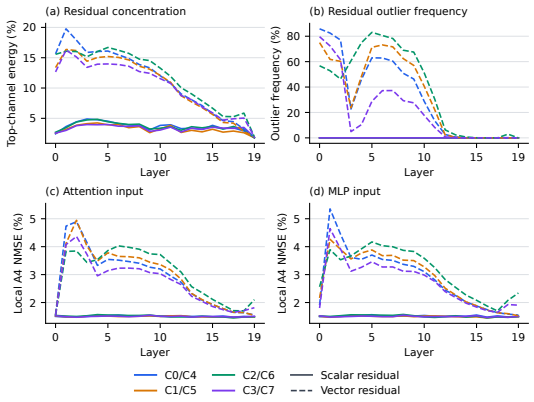

Figure 6: **Scale-ablation diagnostics by layer.** Colors identify matched pairs; solid and dashed lines denote scalar and vector residual scales. Pair labels refer to the configurations in Table 14. Top left: the share of normalized residual energy in the eight highest-energy channels. Top right: the maximum, across channels, of the fraction of tokens satisfying $|x_{t,j}|/\,\mathrm{RMS}(x_t) > 6$. Bottom: local A4 NMSE at attention and MLP inputs. Every layer uses the same 131072 input positions.

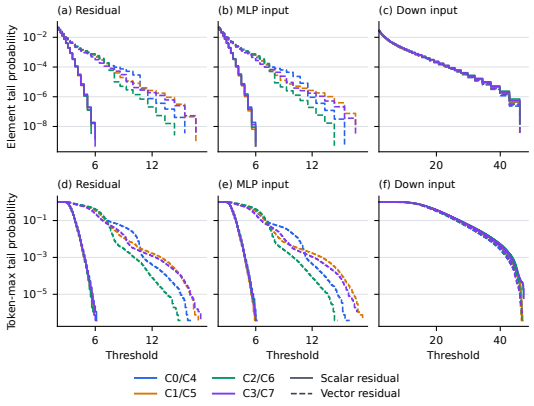

Figure 7: **Normalized activation tails for the eight scale configurations.** Columns show block-output residuals, MLP inputs, and down-projection inputs. Top: survival probabilities of individual $|x_{t,j}|/\,\mathrm{RMS}(x_t)$ values, normalized by the number of elements. Bottom: survival probabilities of per-token maxima, normalized by the number of tokens. Curves pool histogram counts across 20 layers with 131072 positions per layer. Colors and line styles match Figure 6.

D.6 Compatibility with PTQ Methods

We compare SmoothQuant, GPTQ, and QuaRot on the same five pretrained 0.6B models and evaluation samples (Table 5). All three methods quantize weights and linear-layer inputs within Transformer blocks, post-RoPE keys, and projected values to 4 bits. Embeddings and the language-model head remain unquantized. Appendix D.1 specifies the quantizers and transformations used for each method.

SmoothQuant (Xiao et al., 2023) estimates activation maxima from 128 shared calibration sequences of length 2048. We sample nonoverlapping training sequences of 2049 tokens without replacement using seed 42, yielding 2048 input–target pairs per sequence. A disjoint development set of 64 sequences of length 2048 was evaluated in FP32 to select $\alpha \in \{0.5, 0.7, 0.8\}$ by minimizing the five-model mean CE increase from each model's FP reference, with smaller α breaking ties. The selected global value, 0.8, is retained for BF16 without retuning. GPTQ (Frantar et al., 2023) uses the same 128 calibration sequences and no model-specific tuning. QuaRot (Ashkboos et al., 2024) uses randomized residual rotations and online attention/MLP Hadamard transforms. Its activation and cache clipping ratios are fixed across checkpoints.

Relative to the main clipped RTN baseline (Table 1), SmoothQuant improves AdamW and gives higher CE and PPL for the other four checkpoints. This comparison changes both smoothing and the weight-range rule. QuaRot gives the lowest CE and PPL across the tested integer W4A4KV4 methods for each pretrained model.

D.7 NVFP4 Weight and Activation Quantization

We evaluate NVFP4 W4A4 on the five pretrained models at each of 0.3B, 0.6B, and 1.7B, using the same language-modeling streams as Table 1. Weights and inputs of the seven linear layers in each Transformer block use E2M1 values with an E4M3FN scale per 16 contiguous input coordinates and one FP32 global scale per weight tensor or linear input. The global scale is $a_{\max}/(6 \cdot 448)$; weight maxima come from each tensor, and activation maxima are measured with the unquantized BF16 model on 128 calibration sequences of 2048 tokens each from a shared FineWeb-Edu training prefix and then held fixed. Block scales are dynamic, saturate at 448, and may round to zero; both scale and value rounding use nearest-even rules. Reconstructed values are cast to BF16. The KV cache and vocabulary weights remain unquantized. We evaluate quantization through simulation, without storing tensors in packed FP4 format or executing native FP4 arithmetic.

Table 16: **Language-modeling quality under NVFP4 W4A4.** The KV cache remains unquantized in BF16. Lower CE and PPL are better; bold marks the lowest value in each column.

	0.3B		0.6B		1.7B	
Method	CE	PPL	CE	PPL	CE	PPL
AdamW	3.005	28.98	2.741	20.90	**2.452**	14.60
OSP	3.038	30.10	2.810	21.90	2.682	18.02
QuantRet-Frob	**2.926**	26.60	**2.737**	**20.51**	2.467	**14.40**
QuantRet-Spec	3.023	29.70	2.880	23.56	2.644	17.61
QuantRet-Obl	2.929	**26.22**	2.740	20.94	2.469	14.59

At 0.3B, Frob and Obl improve both metrics over AdamW and OSP; Frob gives the lowest CE and Obl the lowest PPL. At 0.6B, Frob gives the lowest CE and PPL, with smaller gains over AdamW. At 1.7B, all three QuantRet variants improve both metrics over OSP, but AdamW gives the lowest CE and Frob the lowest PPL. Spec has higher CE and PPL than AdamW at every scale. The finer floating-point groups and unquantized KV cache distinguish this setting from the integer W4A4KV4 comparisons.

D.8 Transfer to the Llama 3.1 Architecture

We test transfer to an approximately 0.3B Llama 3.1 decoder (Grattafiori et al., 2024). The model has 12 blocks, width 768, SwiGLU width 2816, six query heads, two key–value heads, head dimension 128, untied input/output tables with 128256 vocabulary entries, and the Llama 3.1 scaled RoPE

rule. Training and evaluation use the official Llama 3.1 tokenizer at revision `d04e592`. AdamW, OSP, and the three QuantRet variants share context length 2048, 22460 optimizer steps, global batch size 128, data-order seed 42, and 5887754240 training tokens. Their trainable parameter counts are 293.752M, 294.912M, and 293.861M, respectively. QuantRet uses unit-norm latent rows for both untied vocabulary tables, with independent fixed $\sqrt{d}$ forward multipliers and identity vocabulary rotations. The effective rows have radius $\sqrt{d}$. The initial per-vocabulary scale and fixed raw-logit backward coefficient are $s_v(0) = c_{\mathrm{bp}} = 1/\sqrt{d}$.

Table 17 uses the FP, W4A16, W4A4, and W4A4KV4 definitions from the main experiments; embeddings and the language-model head remain in full precision. FineWeb-Edu CE is averaged over 10485760 next-token predictions. WikiText-2 perplexity is computed over 288627 token positions in complete sequences tokenized with the Llama 3.1 tokenizer. The tokenizer and vocabulary determine the units of these metrics, so comparisons are made within this table.

Table 17: **QuantRet transfer to the Llama 3.1 architecture.** We evaluate 0.3B models with the official tokenizer and common RTN protocol. Values are FineWeb-Edu CE and WikiText-2 perplexity. Bold marks the lowest displayed value in each column.

	FineWeb-Edu CE				WikiText-2 PPL			
Pretraining method	FP	W4A16	W4A4	W4A4KV4	FP	W4A16	W4A4	W4A4KV4
AdamW	3.131	3.214	5.833	5.863	35.79	40.04	313.53	323.45
OSP	3.025	3.294	3.812	3.855	31.38	40.01	69.34	72.73
QuantRet-Frob	3.019	3.124	3.437	3.457	**30.54**	34.72	48.76	50.08
QuantRet-Spec	3.138	3.200	**3.394**	**3.406**	35.03	38.34	**47.78**	**48.49**
QuantRet-Obl	**3.017**	**3.098**	3.415	3.431	30.90	**34.50**	48.31	49.22

Every QuantRet variant gives lower CE and perplexity than AdamW and OSP under W4A16, W4A4, and W4A4KV4. QuantRet-Spec leads both W4A4/W4A4KV4 metrics. These results support transfer of the observed low-bit improvements to the Llama 3.1 architecture at this scale.