UNDERSTANDING LLM QUANTIZATION THROUGH ACTIVATION-GUIDED COMPENSATION AND ORTHOGONAL RESIDUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training weight-activation quantization reduces the memory and inference costs of large language models, but aggressive W4A4 quantization remains difficult because activation outliers degrade effective quantization resolution. Although weight optimization, channel-wise scaling, and orthogonal rotation mitigate this problem, the error components they address and their relationship remain unclear. Using an exact decomposition of local weight-activation quantization error into an activation-guided weight compensation term and an orthogonal residual, we bound the residual using persistent channel-wise outlier and regular activation quantities. This decomposition clarifies which error components can be addressed by weight compensation and which require transformation design. We then use the residual bounds to derive practical guidelines for randomized Hadamard rotation, sign-based probability amplification, and channel scaling. In particular, the analysis explains how random signs suppress constructive interference among persistent outlier channels and how probability amplification through independent sign randomization tightens the high-probability bound on the minimum outlier term across sign configurations. It also shows how second-moment balancing leads to an L_2 scaling rule while a further relaxation recovers SmoothQuant-style L_∞ scaling. We evaluate these guidelines through backpropagation-free configurations across eight Llama and Mistral models, obtaining performance competitive with gradient-trained SpinQuant.

1 INTRODUCTION

Large language models (LLMs) achieve strong performance across a wide range of tasks, but their growing memory and inference costs hinder practical deployment. Post-training quantization (PTQ) reduces these costs using only a small calibration set, yet aggressive 4-bit weight and 4-bit activation quantization (W4A4) remains challenging because large activation values enlarge quantization ranges and reduce the effective resolution available for ordinary values (Frantar et al., 2023; Dettmers et al., 2022; Xiao et al., 2023). Existing methods address this difficulty in different ways: GPTQ-type methods optimize quantized weights to preserve full-precision layer outputs, scaling-based methods such as SmoothQuant redistribute quantization difficulty between activations and weights, and rotation-based methods such as QuaRot and SpinQuant transform activation coordinates to improve low-bit quantization (Frantar et al., 2023; Xiao et al., 2023; Ashkboos et al., 2024; Liu et al., 2025). Other hybrid and learned approaches combine or optimize these transformations in different ways (Czakó et al., 2025; Shao et al., 2024; Sun et al., 2025). Despite their effectiveness, it remains unclear which components of the quantization error these operations address and how their roles are related.

In this paper, we address this gap through a local error-decomposition framework for weight-activation PTQ. We exactly decompose the local quantization error into an activation-guided weight compensation (AGWC) term and an orthogonal residual. For a fixed transformation, weight optimization can reduce the AGWC term but cannot change the orthogonal residual. This decomposition therefore separates the part of the quantization error that can be addressed by weight compensation from the part that requires transformation design. By further analyzing the orthogonal residual, we clarify how

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.

rotation and scaling act on different error structures and provide theoretical guidance for how existing weight optimization, rotation, and scaling techniques should be combined.

Our analysis further connects several design choices that are otherwise treated separately in existing quantization methods. It explains the role of randomized rotation in controlling error associated with persistent activation outliers, and motivates our proposed sign-based probability amplification through a tighter high-probability bound on the minimum outlier term across independent sign configurations. It also characterizes channel scaling through the balance between activation-side and weight-side quantities, relating second-moment-based scaling to SmoothQuant-style scaling. In this way, randomized rotation, sign-based probability amplification, and scaling can be understood as complementary operations guided by the structure of the residual error rather than as independent heuristics.

To examine the practical implications of these theoretical guidelines, we evaluate backpropagation-free configurations that follow the analysis, including signed online Hadamard rotation, sign-based probability amplification, and L_2 channel scaling. Across eight Llama and Mistral models, these theoretically guided choices achieve performance competitive with gradient-trained SpinQuant without backpropagation. These results demonstrate that the proposed error-decomposition framework can provide practical guidance for configuring existing quantization techniques.

To summarize, our contributions are as follows:

- Using an exact decomposition of local weight-activation quantization error into an activation-guided weight compensation term and an orthogonal residual, we bound the residual through persistent channel-wise outlier and regular activation quantities, clarifying how transformation design complements weight compensation.
- We analyze the orthogonal residual to provide theoretical guidance for randomized rotation and channel scaling, placing these operations within a common framework.
- We derive a second-moment bound on the regular quantity and obtain an L_2 scaling rule from its Frobenius surrogate, relating it to SmoothQuant-style L_∞ scaling through a further relaxation.
- We translate our theoretical insights into practical, backpropagation-free quantization configurations and evaluate them across multiple Llama and Mistral models, achieving performance competitive with gradient-trained SpinQuant.

2 Related Work

Post-training weight-activation quantization. PTQ reduces the memory and inference costs of large language models using only a small calibration set, without requiring full retraining (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023). Reconstruction-based methods such as GPTQ optimize quantized weights to preserve full-precision layer outputs (Frantar et al., 2023), while activation-aware methods such as AWQ use activation statistics to identify and protect salient weight channels (Lin et al., 2024). Compared with weight-only quantization, aggressive low-bit weight-activation quantization is more challenging because activation quantization introduces input-dependent error, and large activation values expand the quantization range and reduce the effective resolution available for ordinary values (Dettmers et al., 2022; Xiao et al., 2023). In particular, persistent channel-wise outliers, namely channels that remain consistently large across tokens, can repeatedly dominate per-token activation quantization scales (Xiao et al., 2023; Raman et al., 2025). A recent study further provides an explicit statistical criterion for detecting such channels and empirically analyzes their formation through normalization and weight-matrix multiplication (Raman et al., 2025). We therefore focus on persistent channel-wise outliers as a principal activation-side source of error in weight-activation quantization.

Output-preserving transformations. Outlier-aware quantization often applies transformations that preserve the full-precision output while modifying activations and weights before quantization. SmoothQuant transfers activation magnitudes to weights through output-preserving channel-wise scaling (Xiao et al., 2023). Rotation-based methods include QuIP, which uses structured random orthogonal transformations to spread large weight entries and locally sensitive directions more evenly across coordinates before quantization (Chee et al., 2023), and QuIP#, which efficiently

implements a related preprocessing strategy using randomized Hadamard transforms (Tseng et al., 2024). QuaRot applies Hadamard rotations throughout the model (Ashkboos et al., 2024), while SmoothRot combines Hadamard rotation with calibrated channel scaling (Czakó et al., 2025). DFRot empirically compares randomized orthogonal and Hadamard rotations under per-token activation quantization (Xiang & Zhang, 2025). SpinQuant learns quantization-aware rotations through gradient-based calibration (Liu et al., 2025). OmniQuant (Shao et al., 2024) and FlatQuant (Sun et al., 2025) learn more general equivalent transformations through gradient-based calibration. Our work develops a common analytical basis for these transformations through local quantization-error decomposition. The residual analysis explains how rotation and scaling complement weight compensation and guides their design without backpropagation.

3 Problem Setting

We first define the local reconstruction problem for a linear layer under an output-preserving transformation. We then introduce a simple decomposition of activations that separates persistent outlier structure from the remaining variation, which will be used to analyze rotation and scaling.

3.1 Local Weight-Activation Quantization

Consider a linear layer $\mathbf{Y} = \mathbf{X}\mathbf{W}^\top$, where $\mathbf{X} \in \mathbb{R}^{T\times d}$ and $\mathbf{W} \in \mathbb{R}^{d'\times d}$. For an invertible transformation $\mathbf{P} \in \mathbb{R}^{d\times d}$, define $\mathbf{Z} = \mathbf{X}\mathbf{P}$ and $\mathbf{V} = \mathbf{W}\mathbf{P}^{-\top}$, so that $\mathbf{X}\mathbf{W}^\top = \mathbf{Z}\mathbf{V}^\top$. Thus, $\mathbf{P}$ changes the representation seen by the quantizers without changing the full-precision output. To clarify the effects of rotation and scaling, we focus on quantization errors introduced at the current layer, excluding errors propagated from preceding layers.

For activation quantization, let b_A denote the quantized activation bit-width, and set $q_A = 2^{b_A-1} - 1$, the maximum value after quantization. For analytical tractability, we consider no-clipping dynamic per-token symmetric uniform quantization. For each token t, define the dynamic scale by $\Delta_t = \|(\mathbf{Z})_{t,:}\|_\infty / q_A$. Then the quantized activation $\widetilde{\mathbf{Z}} = \mathcal{Q}_A(\mathbf{Z}) = \mathbf{Z} + \mathbf{A}$ is given element-wise by

$$(\widetilde{\mathbf{Z}})_{t,j} = \mathcal{Q}_A((\mathbf{Z})_{t,j}) = \Delta_t \,\mathrm{round}\left(\frac{(\mathbf{Z})_{t,j}}{\Delta_t}\right), \qquad (\mathbf{A})_{t,j} = (\widetilde{\mathbf{Z}})_{t,j} - (\mathbf{Z})_{t,j}. \tag{1}$$

For weight quantization, we write $\widetilde{\mathbf{V}} = \mathcal{Q}_W(\mathbf{V})$. Unlike the activation quantizer above, $\mathcal{Q}_W$ is not restricted to a fixed rounding rule. It may itself be the result of an optimization procedure, such as GPTQ, or any other feasible weight quantization method.

We investigate the local reconstruction objective

$$\mathcal{J}(\mathbf{P}, \widetilde{\mathbf{V}}) = \frac{1}{T}\left\|\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}^\top - \mathbf{Z}\mathbf{V}^\top\right\|_F^2. \tag{2}$$

This objective measures the output reconstruction error of a single linear layer due to quantization of its input activations and weights.

3.2 Persistent Channel-Wise Outlier Decomposition

Under the per-token activation quantizer introduced above, the quantization scale for token t is determined by $\|(\mathbf{Z})_{t,:}\|_\infty$, so a small number of consistently large activation channels can repeatedly dominate the quantization range and reduce the effective resolution available for the remaining coordinates. Prior work has observed that activation outliers in LLMs often exhibit such a channel-wise persistent structure across tokens (Xiao et al., 2023; Raman et al., 2025). Motivated by this observation, we represent each persistent outlier channel by a token-shared level and assign the remaining token-dependent variation to a regular component. This decomposition serves as an analytical abstraction that separates a sparse persistent component, whose coordinates may coherently interfere after rotation, from the remaining activation variation, thereby enabling separate analyses of Hadamard rotation and channel-wise scaling.

To formalize this abstraction, let $\mathcal{C}_{\mathrm{co}}(\mathbf{X}) \subseteq [d]$ be the set of persistent channel-wise outlier (CO) channels detected on calibration data. For each CO channel $k \in \mathcal{C}_{\mathrm{co}}(\mathbf{X})$, let L_k denote its persistent

CO level. We decompose the activation as $\mathbf{X} = \mathbf{X}^{\text{co}} + \mathbf{X}^{\text{reg}}$, where $\mathbf{X}^{\text{co}}$ is defined by

$$(\mathbf{X}^{\text{co}})_{t,k} = \begin{cases} L_k, & k \in \mathcal{C}_{\text{co}}(\mathbf{X}), \\ 0, & k \notin \mathcal{C}_{\text{co}}(\mathbf{X}), \end{cases} \qquad t \in [T], \tag{3}$$

and $\mathbf{X}^{\text{reg}} = \mathbf{X} - \mathbf{X}^{\text{co}}$. We refer to this decomposition as the persistent sparse CO model. Under this model, $\mathbf{X}^{\text{co}}$ captures the persistent activation component shared across tokens, while $\mathbf{X}^{\text{reg}}$ contains both the token-dependent deviations of the CO channels and the activations of the remaining channels.

We consider transformations of the form $\mathbf{P} = \mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ is a positive diagonal scaling matrix, with each $\lambda_k > 0$ denoting the scaling coefficient for channel k, $\mathbf{D}$ is a diagonal sign matrix, and $\mathbf{H}$ is the normalized Hadamard matrix. The normalized Hadamard matrix is orthogonal, has entries of equal magnitude $1/\sqrt{d}$, and admits an $O(d \log d)$ fast transform; combining it with random diagonal signs gives the randomized Hadamard transform commonly used for coordinate flattening and incoherence processing (Ailon & Chazelle, 2006; Tseng et al., 2024). The diagonal scaling redistributes channel magnitudes between activations and weights, while the signed Hadamard transformation disperses activation coordinates across the transformed channel dimension. Since the corresponding inverse transformation is absorbed into the weights, these operations preserve the full-precision linear-layer output.

4 Error Decomposition and Residual Bounds for Quantization

We first derive an exact decomposition of the local objective in Section 3 to separate the roles of weight compensation and transformation design. We then use residual bounds to guide practical configurations: CO bounds motivate signed Hadamard rotation and sign-based probability amplification, while a second-moment surrogate of the regular bound yields an L_2 scaling rule.

4.1 Local Quantization Error Decomposition and Residual Bound

First, we exactly decompose the local quantization error and then bound the orthogonal residual to clarify the distinct roles of weight optimization, rotation, and scaling.

Theorem 1 (Local quantization error decomposition and residual bound)**.** *Define* $(\mathbf{V}^\star)^\top := \mathbf{V}^\top - \widetilde{\mathbf{Z}}^\dagger \mathbf{A}\mathbf{V}^\top$ *and* $\mathbf{\Pi}_{\widetilde{\mathbf{Z}}} := \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\dagger$*, where* $\mathbf{\Pi}_{\widetilde{\mathbf{Z}}}$ *is the orthogonal projector onto the column space of* $\widetilde{\mathbf{Z}}$*. Then, the local quantization error* $\mathcal{J}(\mathbf{P}, \widetilde{\mathbf{V}})$ *in Eq. (2) admits the following exact decomposition and upper bound:*

$$\mathcal{J}(\mathbf{P}, \widetilde{\mathbf{V}}) = \mathcal{J}_{\text{AGWC}}(\mathbf{P}, \widetilde{\mathbf{V}}) + \frac{1}{T} \left\| \left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right) \mathbf{A}\mathbf{V}^\top \right\|_F^2 \tag{4}$$

$$\leq \mathcal{J}_{\text{AGWC}}(\mathbf{P}, \widetilde{\mathbf{V}}) + \frac{d}{4q_A^2 T} \left(\sqrt{\|\mathbf{W}\mathbf{\Lambda}\|_2^2 \, \mathcal{J}_{\text{co}}(\mathbf{\Lambda}, \mathbf{D})} + \sqrt{\mathcal{J}_{\text{reg}}(\mathbf{\Lambda}, \mathbf{D})} \right)^2, \tag{5}$$

where

$$\mathcal{J}_{\text{AGWC}}(\mathbf{P}, \widetilde{\mathbf{V}}) = \frac{1}{T} \left\| \widetilde{\mathbf{Z}} \left(\widetilde{\mathbf{V}} - \mathbf{V}^\star \right)^\top \right\|_F^2, \tag{6}$$

$$\mathcal{J}_{\text{co}}(\mathbf{\Lambda}, \mathbf{D}) = \sum_{t=1}^{T} \left\| (\mathbf{X}^{\text{co}})_{t,:} \mathbf{\Lambda}^{-1} \mathbf{D}\mathbf{H}^\top \right\|_\infty^2, \tag{7}$$

$$\mathcal{J}_{\text{reg}}(\mathbf{\Lambda}, \mathbf{D}) = \|\mathbf{W}\mathbf{\Lambda}\|_2^2 \sum_{t=1}^{T} \left\| (\mathbf{X}^{\text{reg}})_{t,:} \mathbf{\Lambda}^{-1} \mathbf{D}\mathbf{H}^\top \right\|_\infty^2. \tag{8}$$

The proof is provided in Appendix B.1.

For full-column-rank $\widetilde{\mathbf{Z}}$, this projection identity matches CoreQ's mismatch decomposition (Cha et al., 2026). CoreQ uses this decomposition to limit overfitting to calibration data by adjusting how much the weights compensate for input errors. We instead bound the transformation-dependent residual through persistent CO and regular activation quantities to clarify how rotation and scaling complement weight compensation.

We call $\mathcal{J}_{\mathrm{AGWC}}$ in Eq. (6) the *activation-guided weight compensation* (AGWC) term: the activation discrepancy $\mathbf{A}$ guides the continuous target $\mathbf{V}^\star$, and $\mathcal{J}_{\mathrm{AGWC}}$ measures the reconstruction error of the quantized weights relative to this target. Geometrically, this term lies in the column space of $\widetilde{\mathbf{Z}}$ and can therefore be affected by changing the quantized weights, whereas the second term lies in the orthogonal complement of this space. Thus, for a fixed transformation, weight optimization can reduce the AGWC term but cannot change the orthogonal residual. The CO and regular quantities are not additional terms in the exact decomposition; they appear only in an upper bound on the orthogonal residual. Theorem 1 thus separates the scope of weight compensation from the residual error requiring transformation design. This distinction guides how weight optimization, rotation, and scaling can be combined.

4.2 Weight Compensation through the Lens of Error Decomposition

Theorem 1 provides a common perspective on existing weight-reconstruction and compensation methods. When $\mathbf{A} = \mathbf{0}$, $\mathbf{V}^\star = \mathbf{V}$ and $\mathcal{J}_{\mathrm{AGWC}}$ reduces to the standard GPTQ-style reconstruction objective (Frantar et al., 2023). Thus, shared-input reconstruction can be viewed as the special case in which no activation discrepancy needs to be compensated. When the input discrepancy is nonzero, however, shared-input reconstruction, even with quantized activations, does not explicitly compensate for this discrepancy.

QEP (Arai & Ichikawa, 2025) corrects weights before applying a quantizer such as GPTQ; with full compensation, matching inputs and reference outputs, and an undamped invertible Hessian, its target equals $\mathbf{V}^\star$. With GPTQ, the corrected weights are quantized sequentially, updating the remaining weights. GPTAQ (Li et al., 2025) derives sequential weight updates from a reconstruction objective of the form in Eq. (2), incorporating input-error compensation through channel-wise residual components for efficiency. For fixed inputs and $\mathbf{P}$, Eq. (4) makes $\mathcal{J}$ and $\mathcal{J}_{\mathrm{AGWC}}$ equivalent up to a weight-independent constant over the same feasible set, without implying identical algorithmic outputs. The orthogonal residual remains unchanged by weight compensation, so the CO and regular quantities in its bound guide complementary rotation and scaling.

In practice, we adopt GPTAQ with activation quantization enabled before weight calibration to account for local and propagated input errors against full-precision reference outputs (Li et al., 2025).

4.3 Randomized Hadamard Rotation for $\mathcal{J}_{\mathrm{co}}$

We now analyze the rotation-dependent CO quantity $\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda}, \mathbf{D})$ in the residual upper bound of Theorem 1 and clarify how randomized Hadamard rotation controls this quantity.

Theorem 2 (Hadamard rotation bounds)**.** *Under the persistent sparse CO model defined in Eq. (3), let*

$$N_{\mathrm{co}}(\mathbf{X}) = \left|\mathcal{C}_{\mathrm{co}}(\mathbf{X})\right|, \qquad L_{\Lambda,\max} = \max_{k \in \mathcal{C}_{\mathrm{co}}(\mathbf{X})} \frac{|L_k|}{\lambda_k}. \tag{9}$$

Then, the following three bounds hold:

Fixed rotation. *If $\mathbf{D} = \mathbf{I}$, then*

$$\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda}, \mathbf{I}) \le \frac{T N_{\mathrm{co}}(\mathbf{X})^2}{d} L_{\Lambda,\max}^2. \tag{10}$$

Randomized rotation. *If $\mathbf{D} = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_d)$ with independent Rademacher signs, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda}, \mathbf{D}) \le \frac{2T N_{\mathrm{co}}(\mathbf{X})}{d} L_{\Lambda,\max}^2 \left(\log(2d) + \log\frac{1}{\delta}\right). \tag{11}$$

Sign-based probability amplification. *Let $\mathcal{D}_{N_s} = \{\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(N_s)}\}$ be a set of N_s independently sampled random sign matrices. Then, with probability at least $1-\delta$,*

$$\min_{\mathbf{D} \in \mathcal{D}_{N_s}} \mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda}, \mathbf{D}) \le \frac{2T N_{\mathrm{co}}(\mathbf{X})}{d} L_{\Lambda,\max}^2 \left(\log(2d) + \frac{1}{N_s}\log\frac{1}{\delta}\right). \tag{12}$$

The proof is provided in Appendix B.2.

Theorem 2 identifies interference among persistent outlier channels as a mechanism governing the CO quantity in the residual bound. Its fixed and randomized cases quantify this mechanism for existing Hadamard rotation schemes (Tseng et al., 2024; Ashkboos et al., 2024); our proposed third case provides a tighter high-probability bound on the minimum CO quantity over independent sign configurations. Because each entry of a normalized Hadamard matrix has magnitude $1/\sqrt{d}$, each CO channel contributes to every rotated coordinate at this scale; however, under a fixed rotation, contributions from multiple CO channels can constructively align, yielding the worst-case dependence $N_{\mathrm{co}}(\mathbf{X})^2$. Random signs convert these contributions into Rademacher sums and reduce this dependence to $N_{\mathrm{co}}(\mathbf{X})$ up to logarithmic factors in d, while independent repetitions further tighten the high-probability bound on the minimum CO quantity through probability amplification. Appendix A provides an interference-based interpretation of the $\mathcal{O}(N_{\mathrm{co}}(\mathbf{X})^2)$ dependence under fixed rotation and its reduction to $\mathcal{O}(N_{\mathrm{co}}(\mathbf{X}))$ under random signs. Since both the number of persistent CO channels $N_{\mathrm{co}}(\mathbf{X})$ and their scaled maximum magnitude $L_{\Lambda,\max}$ depend on the model, layer, and activation distribution, these quantities provide one possible explanation for why different models exhibit different levels of accuracy degradation under low-bit quantization.

4.4 Deriving L_2 Scaling from a Bound on $\mathcal{J}_{\mathrm{reg}}$

We next consider the regular quantity $\mathcal{J}_{\mathrm{reg}}$ in the residual bound. Unlike persistent COs, the regular component is not dominated by a small number of shared large coordinates. After random-sign Hadamard rotation, its maximum coordinate can be controlled by its second-moment energy.

Proposition 1 (L_2 scaling from second-moment balancing)**.** *Let* $\mathbf{D}$ *be a random sign matrix with independent Rademacher signs. For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\mathcal{J}_{\mathrm{reg}}(\mathbf{\Lambda}, \mathbf{D}) \leq \frac{2\log(2dT/\delta)}{d} \|\mathbf{W}\mathbf{\Lambda}\|_2^2 \left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2. \tag{13}$$

Consequently, the regular quantity $\mathcal{J}_{\mathrm{reg}}$ *is controlled by* $\|\mathbf{W}\mathbf{\Lambda}\|_2^2 \left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2$*. Using the Frobenius surrogate* $\mathcal{R}_2(\mathbf{\Lambda}) = \|\mathbf{W}\mathbf{\Lambda}\|_F^2 \left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2$*, we denote by* $\lambda_k^{(2)}$ *the channel-wise scaling coefficient obtained by minimizing* $\mathcal{R}_2(\mathbf{\Lambda})$*. This gives*

$$\lambda_k^{(2)} = \sqrt{\frac{\|(\mathbf{X}^{\mathrm{reg}})_{:,k}\|_2}{\|(\mathbf{W})_{:,k}\|_2}}. \tag{14}$$

The proof is provided in Appendix B.3.

The resulting rule has the same L_2 balancing form as QuIP's rescaling (Chee et al., 2023). Note that QuIP motivates this scaling through a heuristic surrogate for output error in weight-only quantization, whereas we derive it from a bound on the activation-quantization residual. This derivation also enables a common interpretation of L_2 and SmoothQuant-style L_∞ scaling, as discussed in Section 4.5.

Proposition 1 explains why L_2 statistics are natural for the regular component. The second-moment bound on $\mathcal{J}_{\mathrm{reg}}$ combines the activation-side energy $\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\|_F^2$ with the weight-side amplification factor $\|\mathbf{W}\mathbf{\Lambda}\|_2^2$. In particular, the weight-side spectral norm $\|\mathbf{W}\mathbf{\Lambda}\|_2$ measures the worst-case amplification induced by the scaled weight matrix and therefore reflects not only individual channel magnitudes but also cross-channel interactions through the geometry of the weight columns. Directly optimizing the resulting objective can therefore account for such interactions in principle. However, the objective is not separable across channels, making it difficult to obtain a simple closed-form channel-wise scaling rule.

To obtain a tractable channel-wise rule, we use the Frobenius surrogate $\mathcal{R}_2(\mathbf{\Lambda})$. This relaxation replaces the worst-case coupled amplification by the total second-moment energy of the scaled weight columns. The surrogate is a product of sums of squared channel-wise quantities: $\lambda_k\|(\mathbf{W})_{:,k}\|_2$ on the weight side and $\|(\mathbf{X}^{\mathrm{reg}})_{:,k}\|_2/\lambda_k$ on the activation side. The equality condition in the Cauchy–Schwarz inequality yields the L_2 scaling rule in Proposition 1. The proposition is stated in terms of the regular component $\mathbf{X}^{\mathrm{reg}}$ to characterize the error component targeted by the scaling rule. In practice, since $\mathbf{X}^{\mathrm{reg}}$ is not explicitly identified, we approximate its channel-wise second-moment statistics using the empirical full activations, as described in Section 4.6.

4.5 Comparing L_2 Scaling with SmoothQuant-Style L_∞

Our residual analysis connects L_2 and SmoothQuant-style L_∞ scaling through successive relaxations of the same regular-term bound, providing a common interpretation not established by QuIP's heuristic derivation.

From Proposition 1, after random-sign Hadamard rotation, the regular quantity satisfies $\mathcal{J}_{\mathrm{reg}}(\mathbf{\Lambda}, \mathbf{D}) \lesssim \|\mathbf{W}\mathbf{\Lambda}\|_2^2 \left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2$. Our L_2 scaling is obtained from a Frobenius surrogate of this second-moment bound. In contrast, a SmoothQuant-style interpretation can be obtained by further upper bounding the same quantity using L_∞ channel statistics:

$$\begin{aligned}\|\mathbf{W}\mathbf{\Lambda}\|_2^2 \left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2 &\le \|\mathbf{W}\mathbf{\Lambda}\|_F^2 \left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2 \\ &\le d'T \left(\sum_{k=1}^{d} \lambda_k^2 \left\|(\mathbf{W})_{:,k}\right\|_\infty^2\right)\left(\sum_{k=1}^{d} \frac{\|(\mathbf{X}^{\mathrm{reg}})_{:,k}\|_\infty^2}{\lambda_k^2}\right). \end{aligned} \tag{15}$$

This relaxation is looser because it first replaces the spectral norm of the scaled weight matrix by its Frobenius norm and then replaces channel-wise second-moment quantities by their maximum magnitudes. We denote by $\lambda_k^{(\infty)}$ the channel-wise scaling coefficient obtained by minimizing the resulting L_∞ surrogate. Then, up to a common positive scalar,

$$\lambda_k^{(\infty)} = \sqrt{\frac{\|(\mathbf{X}^{\mathrm{reg}})_{:,k}\|_\infty}{\|(\mathbf{W})_{:,k}\|_\infty}}. \tag{16}$$

This rule is stated in terms of the regular component $\mathbf{X}^{\mathrm{reg}}$ to characterize the error component targeted by the scaling. In practice, since $\mathbf{X}^{\mathrm{reg}}$ is not explicitly identified, we approximate its channel-wise maximum statistics using the empirical full activations, recovering the familiar SmoothQuant-style scaling rule.

Our framework thus explains both scaling rules through their retained statistics: the L_2 surrogate preserves channel-wise second moments, whereas the SmoothQuant-style L_∞ surrogate further relaxes them to channel maxima. This provides a bound-based rationale for using L_2 statistics after randomized Hadamard rotation.

4.6 From Theory to Practice: L2-SmoothRot

The preceding analysis leads to L2-SmoothRot, a simple configuration with three components: signed online rotation, sign-based probability amplification, and L_2 channel scaling.[1] Theorem 2 motivates the first two, while Proposition 1 derives the scaling rule from a second-moment surrogate. We implement these choices without backpropagation within the QuaRot framework, following SmoothRot for scaling placement and fusion.

Signed online rotation (SOR). Theorem 2 shows that the sign structure of a Hadamard rotation is itself important: a fixed Hadamard transform can retain constructive interference among persistent CO channels, whereas random signs suppress this interference through Rademacher cancellation. This suggests that the random-sign principle should be applied not only to offline rotations, but also to online Hadamard rotations that directly act on activations before quantization. In the original QuaRot, random signs are used only for offline rotation, whereas the online Hadamard rotations are fixed (Ashkboos et al., 2024). Motivated by this distinction, we insert sign flips before the online Hadamard rotations at the attention output and FFN down projections. We also apply the same orthogonal signed rotation $\mathbf{R} = \mathbf{D}\mathbf{H}^\top$ to queries and keys within each head after RoPE, before key quantization, preserving unquantized attention scores since $(\mathbf{Q}\mathbf{R})(\mathbf{K}\mathbf{R})^\top = \mathbf{Q}\mathbf{K}^\top$.

Sign-based probability amplification (SPA). Independent sign randomization provides probability amplification, tightening the high-probability bound on the minimum CO quantity over sign configurations (Theorem 2). Motivated by this result and the observed rotation variability (Liu et al., 2025), we generate independent sign configurations and evaluate them using validation perplexity

[1] Code will be publicly released upon acceptance. The repository URL is withheld to preserve anonymity during double-blind review.

Table 1: W4A4 results with GPTAQ weight quantization and KV4 across model families and scales. We report WikiText-2 perplexity (PPL) and average zero-shot accuracy over six tasks. Full results are in Table 3.

Model	FP reference		W4A4							
	W16A16		QuaRot		SmoothRot		SpinQuant		L2-SmoothRot	
	PPL ↓	Avg. ↑	PPL ↓	Avg. ↑	PPL ↓	Avg. ↑	PPL ↓	Avg. ↑	PPL ↓	Avg. ↑
Llama-7B	5.67	69.39	6.12	66.17	6.13	66.16	6.19	66.18	**6.09**	**66.67**
Llama-13B	5.05	71.78	5.42	69.49	5.42	69.09	5.44	69.31	**5.37**	**69.91**
Llama2-7B	5.47	69.82	5.96	66.32	5.98	66.45	6.06	65.98	**5.94**	**66.59**
Llama2-13B	4.86	72.57	5.24	69.87	5.26	70.15	5.26	70.23	**5.23**	**70.50**
Llama3-8B	5.94	73.30	7.45	65.71	7.57	65.85	7.33	**68.36**	**7.20**	68.10
Llama3.2-1B	9.41	59.80	13.81	49.58	14.13	48.63	13.33	50.37	**12.92**	**50.71**
Llama3.2-3B	7.55	68.20	9.27	61.12	9.35	59.91	9.19	**62.03**	**9.08**	60.94
Mistral-7B-v0.3	5.35	73.73	5.73	70.55	5.77	69.80	5.74	**71.17**	**5.71**	71.00

computed solely through forward passes. Each configuration jointly specifies sign patterns for offline, attention-output, FFN-down, and QK rotations, generated from a single random seed. Each category shares its sign vector across layers, and QK additionally shares its vector across all query and key heads. We screen the resulting rotation candidates with round-to-nearest (RTN) quantization and choose among the GPTAQ-reevaluated finalists using validation perplexity (Section 5).

L_2 scaling (L2S). We use the scaling rule derived by minimizing the second-moment surrogate in Proposition 1, with empirical activation statistics as a proxy for the regular component. Following SmoothRot, we apply scaling only to the FFN down-projection input. For the down-projection input $\mathbf{X}_{\text{down}}$ and weight $\mathbf{W}_{\text{down}}$, we use the transformation $\mathbf{X}_{\text{down}}\mathbf{W}_{\text{down}}^{\top} = (\mathbf{X}_{\text{down}}\mathbf{\Lambda}^{-1})(\mathbf{W}_{\text{down}}\mathbf{\Lambda})^{\top}$. Since the regular component is not explicitly separated in practice, we use the empirical full-activation norm as a readily computable proxy:

$$\lambda_k^{(2)} = \sqrt{\frac{\|(\mathbf{X}_{\text{down}})_{:,k}\|_2}{\|(\mathbf{W}_{\text{down}})_{:,k}\|_2}}. \tag{17}$$

As in SmoothRot, the scaling factors can be merged into the corresponding up- and down-projection weights and therefore introduce no additional inference-time overhead.

Overall, the decomposition in Theorem 1 and the bounds in Theorem 2 and Proposition 1 provide a unified rationale for combining weight compensation with rotation and scaling. Their effects are coupled, since changing the transformation can alter both terms of the exact decomposition.

5 Experiments

Tasks and baselines. We evaluate Llama-7B/13B (Touvron et al., 2023a), Llama2-7B/13B (Touvron et al., 2023b), Llama3-8B (Grattafiori et al., 2024), Llama3.2-1B/3B (Meta AI, 2024), and Mistral-7B-v0.3 (Jiang et al., 2023; Mistral AI, 2024) under W4A4 quantization. We evaluate perplexity (PPL) on 64 samples each from WikiText-2 (WT2) test and a C4 validation subset. For zero-shot evaluation, we report accuracy on PIQA (Bisk et al., 2020), ARC-Easy (ARC-E) and ARC-Challenge (ARC-C) (Clark et al., 2018), HellaSwag (HS) (Zellers et al., 2019), WinoGrande (WG) (Sakaguchi et al., 2021), and LAMBADA (LB) (Paperno et al., 2016). We compare our method, L2-SmoothRot, with QuaRot (Ashkboos et al., 2024), SmoothRot (Czakó et al., 2025), and SpinQuant (Liu et al., 2025), applying the same GPTAQ weight-quantization settings to the transformed models produced by all four methods.

Quantization and calibration. Following the experimental settings of SmoothRot (Czakó et al., 2025), we simulate W4A4 decoder linear layers with KV4: weights use per-output-channel symmetric quantization, activations per-token symmetric quantization, and K/V per-token, per-head asymmetric quantization. Activation and K/V clipping ratios are 0.9 and 0.95. GPTAQ uses 128 WikiText-2

Table 2: Component ablation on Llama3-8B with GPTAQ (W4A4KV4). Parentheses indicate the residual-bound term motivating each component.

Method				Llama3-8B
QuaRot	SOR ($\mathcal{J}_{\text{co}}$)	SPA ($\mathcal{J}_{\text{co}}$)	L2S ($\mathcal{J}_{\text{reg}}$)	PPL ↓
✓				7.45
✓	✓			7.30
✓		✓		7.40
✓			✓	7.32
✓	✓	✓	✓	**7.20**

train samples. Calibration includes local activation and propagated weight, activation, and K/V errors against corresponding unquantized-path outputs. Scaling statistics use the untransformed, unquantized model with 512 WikiText-2 train samples. Following the rotation-training protocol of SpinQuant (Liu et al., 2025), we train rotations on WikiText-2 train for 100 steps with frozen 16-bit weights and 4-bit activations and K/V, then quantize the rotated weights with GPTAQ.

SPA protocol. For L2-SmoothRot, we evaluate 10 rotation candidates using RTN perplexity on WikiText-2 validation, reevaluate the top 3 with GPTAQ on the same validation split, and use the candidate with the lowest GPTAQ validation perplexity. Appendix C further evaluates SPA by varying the candidate budget and comparing screening strategies.

Main results. Table 1 summarizes the main W4A4 results across eight models from two families and from 1B to 13B parameters; Table 3 in Appendix D reports C4 perplexity and per-task accuracies. L2-SmoothRot achieves the lowest WikiText-2 perplexity on all eight models and the highest average zero-shot accuracy on five. As shown in Table 3, it also achieves the lowest C4 perplexity on every model, indicating that the perplexity improvements are not limited to WikiText-2. On Llama3-8B and Mistral-7B-v0.3, its average accuracy is within 0.26 and 0.17 percentage points of SpinQuant, respectively; on Llama3.2-3B, it improves perplexity over SpinQuant while trailing its average accuracy by 1.09 points. Overall, these results demonstrate the practical value of the transformation design guided by our analysis, with consistent perplexity gains and competitive downstream accuracy across model families and scales without backpropagation.

Component ablation. Table 2 ablates the proposed components on Llama3-8B under W4A4 quantization with GPTAQ. Starting from QuaRot (7.45), adding SOR, SPA, or L2S individually reduces WikiText-2 perplexity to 7.30, 7.40, and 7.32, respectively, with SOR providing the largest individual gain. Combining all components further reduces perplexity to 7.20. These gains are consistent with their complementary motivations: SOR and SPA primarily target $\mathcal{J}_{\text{co}}$, whereas L2S targets $\mathcal{J}_{\text{reg}}$, although their effects on quantization error need not be independent.

6 Conclusion

In this paper, we analyzed local weight-activation quantization error through an exact decomposition into an activation-guided weight compensation term and an orthogonal residual. Bounding the residual through persistent channel-wise outlier and regular activation quantities clarifies how rotation and scaling complement weight optimization. Within this framework, our bounds explain how random signs suppress constructive interference among persistent outlier channels and how probability amplification through independent sign randomization tightens the high-probability bound on the minimum channel-wise outlier quantity. For scaling, minimizing a second-moment surrogate of the regular-term bound yields an L_2 rule, while a further relaxation recovers a SmoothQuant-style L_∞ rule. Combining signed online rotation, sign-based probability amplification, and L_2 scaling achieved performance competitive with gradient-trained SpinQuant across eight Llama and Mistral models without backpropagation. Our analysis provides a principled basis for designing output-preserving transformations for low-bit weight-activation quantization.

AI USE STATEMENT

Generative AI tools were used to assist with writing and polishing the manuscript. We reviewed and revised all AI-assisted text.

REFERENCES

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pp. 557–563, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595931341. doi: 10.1145/1132516.1132597. URL https://doi.org/10.1145/1132516.1132597.

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/df2034a516cbd617a96492cc476276c9-Abstract-Conference.html.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 100213–100240. Curran Associates, Inc., 2024. doi: 10.52202/079017-3180. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b5b939436789f76f08b9d0da5e81af7c-Paper-Conference.pdf.

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.

Seohyeon Cha, Huancheng Chen, Dongjun Kim, Haoran Zhang, Kevin Chan, Gustavo de Veciana, and Haris Vikalo. CoreQ: Learning-free mismatch correction and successive rounding for quantization, 2026. URL https://arxiv.org/abs/2602.05902v2. Version 2.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429. Curran Associates, Inc., 2023. doi: 10.52202/075280-0196. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0df38cd13520747e1e64e5b123a78ef8-Paper-Conference.pdf.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Patrik Czakó, Gábor Kertész, and Sándor Szénási. Smoothrot: Combining channel-wise scaling and rotation for quantization-friendly llms. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 6461–6466, 2025. doi: 10.1109/SMC58881.2025.11342731.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. URL https://arxiv.org/abs/2210.17323.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi (Jack) Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,

Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu (Sid) Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2282952.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. GPTAQ: Efficient finetuning-free quantization for asymmetric calibration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36690–36706. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25dn.html.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e5b1c0d4866f72393c522c8a00eed4eb-Paper-Conference.pdf.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024.

Mistral AI. Mistral 7b v0.3. https://docs.mistral.ai/models/mistral-7b-0-3, May 2024. Model documentation.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/.

Rahul Raman, Khushi Sharma, and Sai Qian Zhang. Rethinking the outlier distribution in large language models: An in-depth study, 2025. URL https://arxiv.org/abs/2505.21670.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c6483c8a68083af3383f91ee0dc6db95-Paper-Conference.pdf.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57587–57613. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/sun25l.html.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In Ruslan

Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

Jingyang Xiang and Sai Qian Zhang. DFRot: Achieving outlier-free and massive activation-free for rotated LLMs with refined rotation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=WzGypILLDb.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

A An Interference View of Hadamard Rotation Bounds

In this section, we provide additional analysis supporting the randomized Hadamard rotation arguments in Section 4.3. We make explicit how contributions from multiple persistent CO channels can constructively interfere under a fixed Hadamard rotation and explain how random signs remove this deterministic alignment in expectation. This interpretation complements the high-probability bounds in Theorem 2.

Let $\mathbf{H} \in \mathbb{R}^{d\times d}$ be the normalized Hadamard matrix. For $d = 2^n$, identifying the indices j, k with binary vectors in $\{0,1\}^n$, we write

$$(\mathbf{H})_{j,k} = \frac{1}{\sqrt{d}}(-1)^{\langle j,k\rangle}, \tag{18}$$

where $\langle j,k\rangle$ denotes the bit-wise inner product modulo 2.

To simplify the notation, we omit channel scaling; the same argument applies by replacing L_k with L_k/λ_k. For the fixed Hadamard rotation $\mathbf{X}^{\text{co}}_{\text{had}} = \mathbf{X}^{\text{co}}\mathbf{H}^\top$, each output coordinate is

$$(\mathbf{X}^{\text{co}}_{\text{had}})_{t,j} = \frac{1}{\sqrt{d}} \sum_{k\in\mathcal{C}_{\text{co}}(\mathbf{X})} L_k(-1)^{\langle j,k\rangle}. \tag{19}$$

Thus, although each CO channel is spread across all output coordinates, the rotated coordinate is a deterministic signed sum of the CO magnitudes.

Its squared magnitude is

$$E_j := |(\mathbf{X}^{\text{co}}_{\text{had}})_{t,j}|^2 \tag{20}$$

$$= \underbrace{\frac{1}{d}\sum_{k\in\mathcal{C}_{\text{co}}(\mathbf{X})} L_k^2}_{\text{individual-channel energy}} + \underbrace{\frac{1}{d}\sum_{\substack{k,m\in\mathcal{C}_{\text{co}}(\mathbf{X})\\ k\neq m}} L_kL_m(-1)^{\langle j,k\oplus m\rangle}}_{\text{cross-channel interference}}, \tag{21}$$

where $\oplus$ denotes bit-wise XOR. The first term is independent of j, whereas the second term depends on the fixed Hadamard sign pattern. When several CO contributions have the same sign at an output coordinate, the cross terms are positive and the contributions add constructively.

The individual-channel term contains $N_{\text{co}}(\mathbf{X})$ contributions and therefore scales linearly with $N_{\text{co}}(\mathbf{X})$. In contrast, the cross-channel term contains $N_{\text{co}}(\mathbf{X})(N_{\text{co}}(\mathbf{X}) - 1)$ ordered pairs. Thus, when many

of these pairwise terms align positively, the cross-channel interference can grow on the order of $N_{\mathrm{co}}(\mathbf{X})^2$. This provides an interference-based interpretation of the quadratic dependence on the number of CO channels in the fixed-rotation bound of Theorem 2.

Now consider a random sign matrix $\mathbf{D} = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_d)$, where the ε_k are independent Rademacher variables. The corresponding coordinate energy is

$$E_j(\mathbf{D}) = \frac{1}{d}\left(\sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} \varepsilon_k L_k (-1)^{\langle j,k\rangle}\right)^2. \tag{22}$$

Expanding the square gives

$$E_j(\mathbf{D}) = \frac{1}{d}\sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} L_k^2 + \frac{1}{d}\sum_{\substack{k,m\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})\\ k\neq m}} \varepsilon_k\varepsilon_m L_k L_m (-1)^{\langle j,k\oplus m\rangle}. \tag{23}$$

Since $\mathbb{E}_{\mathbf{D}}[\varepsilon_k\varepsilon_m] = 0\,(k \neq m)$, all cross-channel interference terms vanish in expectation:

$$\mathbb{E}_{\mathbf{D}}[E_j(\mathbf{D})] = \frac{1}{d}\sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} L_k^2. \tag{24}$$

Hence, after randomization, only the individual-channel energy remains in expectation, reducing the dependence on the number of CO channels from $\mathcal{O}(N_{\mathrm{co}}(\mathbf{X})^2)$ to $\mathcal{O}(N_{\mathrm{co}}(\mathbf{X}))$. This provides an interference-based interpretation of the improvement from the fixed to the randomized bound in Theorem 2.

Sign-based probability amplification in Theorem 2 can further be interpreted as increasing the chance of obtaining a favorable sign assignment in which the CO contributions cancel more effectively, thereby reducing the largest rotated coordinate.

B Proofs for Section 4

B.1 Proof of Theorem 1

Proof. We first prove the projection-based identity. By definition, $\widetilde{\mathbf{Z}} = \mathbf{Z} + \mathbf{A}$. Therefore,

$$\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}^\top - \mathbf{Z}\mathbf{V}^\top = \widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}^\top - \widetilde{\mathbf{Z}}\mathbf{V}^\top + \widetilde{\mathbf{Z}}\mathbf{V}^\top - \mathbf{Z}\mathbf{V}^\top \tag{25}$$

$$= \widetilde{\mathbf{Z}}\left(\widetilde{\mathbf{V}} - \mathbf{V}\right)^\top + \left(\widetilde{\mathbf{Z}} - \mathbf{Z}\right)\mathbf{V}^\top \tag{26}$$

$$= \widetilde{\mathbf{Z}}\left(\widetilde{\mathbf{V}} - \mathbf{V}\right)^\top + \mathbf{A}\mathbf{V}^\top. \tag{27}$$

Let $\mathbf{\Pi}_{\widetilde{\mathbf{Z}}} := \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\dagger$ denote the orthogonal projector onto the column space of $\widetilde{\mathbf{Z}}$, where $\widetilde{\mathbf{Z}}^\dagger$ is the Moore–Penrose pseudoinverse. We decompose $\mathbf{A}\mathbf{V}^\top$ into its projection onto this column space and its orthogonal complement:

$$\mathbf{A}\mathbf{V}^\top = \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\dagger\mathbf{A}\mathbf{V}^\top + \left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right)\mathbf{A}\mathbf{V}^\top. \tag{28}$$

Thus,

$$\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}^\top - \mathbf{Z}\mathbf{V}^\top = \widetilde{\mathbf{Z}}\left(\left(\widetilde{\mathbf{V}} - \mathbf{V}\right)^\top + \widetilde{\mathbf{Z}}^\dagger\mathbf{A}\mathbf{V}^\top\right) + \left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right)\mathbf{A}\mathbf{V}^\top. \tag{29}$$

By the definition $(\mathbf{V}^\star)^\top = \mathbf{V}^\top - \widetilde{\mathbf{Z}}^\dagger\mathbf{A}\mathbf{V}^\top$, this becomes

$$\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}^\top - \mathbf{Z}\mathbf{V}^\top = \widetilde{\mathbf{Z}}\left(\widetilde{\mathbf{V}} - \mathbf{V}^\star\right)^\top + \left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right)\mathbf{A}\mathbf{V}^\top. \tag{30}$$

The first term belongs to the column space of $\widetilde{\mathbf{Z}}$, while the second term belongs to its orthogonal complement. Hence, by orthogonality,

$$\mathcal{J}(\mathbf{P}, \widetilde{\mathbf{V}}) = \frac{1}{T}\left\|\widetilde{\mathbf{Z}}\left(\widetilde{\mathbf{V}} - \mathbf{V}^\star\right)^\top\right\|_F^2 + \frac{1}{T}\left\|\left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right)\mathbf{A}\mathbf{V}^\top\right\|_F^2. \tag{31}$$

The first squared norm is the activation-guided weight compensation term $\mathcal{J}_{\mathrm{AGWC}}$ defined in Eq. (6). The second is independent of $\widetilde{\mathbf{V}}$ for fixed inputs and $\mathbf{P}$, and is attained as the minimum local error when unrestricted weights are set to $\mathbf{V}^\star$.

We next upper bound this orthogonal residual using the CO and regular quantities; they are not separate terms in the exact identity. For no-clipping dynamic per-token symmetric quantization, let

$$\Delta_t = \frac{\|(\mathbf{Z})_{t,:}\|_\infty}{q_A}, \qquad q_A = 2^{b_A-1} - 1. \tag{32}$$

For each entry,

$$|(\mathbf{A})_{t,j}| = |\mathcal{Q}_A((\mathbf{Z})_{t,j}) - (\mathbf{Z})_{t,j}| \le \frac{\Delta_t}{2}. \tag{33}$$

Therefore,

$$\|\mathbf{A}\|_F^2 \le \sum_{t=1}^{T}\sum_{j=1}^{d} \frac{\Delta_t^2}{4} = \frac{d}{4q_A^2}\sum_{t=1}^{T} \|(\mathbf{Z})_{t,:}\|_\infty^2 . \tag{34}$$

Since orthogonal projection does not increase the Frobenius norm,

$$\frac{1}{T}\left\|\left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right)\mathbf{A}\mathbf{V}^\top\right\|_F^2 \le \frac{1}{T}\left\|\mathbf{A}\mathbf{V}^\top\right\|_F^2 \tag{35}$$

$$\le \frac{\|\mathbf{V}\|_2^2}{T}\|\mathbf{A}\|_F^2 . \tag{36}$$

For $\mathbf{P} = \mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top$, we have $\mathbf{V} = \mathbf{W}\mathbf{P}^{-\top} = \mathbf{W}\mathbf{\Lambda}\mathbf{D}\mathbf{H}^\top$. Because $\mathbf{D}$ and $\mathbf{H}^\top$ are orthogonal matrices, $\|\mathbf{V}\|_2 = \|\mathbf{W}\mathbf{\Lambda}\|_2$. Thus,

$$\frac{1}{T}\left\|\left(\mathbf{I} - \mathbf{\Pi}_{\widetilde{\mathbf{Z}}}\right)\mathbf{A}\mathbf{V}^\top\right\|_F^2 \le \frac{d\,\|\mathbf{W}\mathbf{\Lambda}\|_2^2}{4q_A^2 T}\sum_{t=1}^{T}\left\|(\mathbf{X})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top\right\|_\infty^2 . \tag{37}$$

Finally, decompose $(\mathbf{X})_{t,:} = (\mathbf{X}^{\mathrm{co}})_{t,:} + (\mathbf{X}^{\mathrm{reg}})_{t,:}$. By the triangle inequality in ℓ_∞ and then in ℓ_2 over tokens,

$$\left(\sum_{t=1}^{T}\left\|(\mathbf{X})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top\right\|_\infty^2\right)^{1/2} \le \left(\sum_{t=1}^{T}\left\|(\mathbf{X}^{\mathrm{co}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top\right\|_\infty^2\right)^{1/2} \tag{38}$$

$$+ \left(\sum_{t=1}^{T}\left\|(\mathbf{X}^{\mathrm{reg}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top\right\|_\infty^2\right)^{1/2} . \tag{39}$$

Therefore,

$$\sum_{t=1}^{T}\left\|(\mathbf{X})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top\right\|_\infty^2 \le \left(\sqrt{\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{D})} + \sqrt{\frac{\mathcal{J}_{\mathrm{reg}}(\mathbf{\Lambda},\mathbf{D})}{\|\mathbf{W}\mathbf{\Lambda}\|_2^2}}\right)^2 . \tag{40}$$

Combining the above inequalities gives

$$\mathcal{J}(\mathbf{P},\widetilde{\mathbf{V}}) \le \mathcal{J}_{\mathrm{AGWC}}(\mathbf{P},\widetilde{\mathbf{V}}) + \frac{d}{4q_A^2 T}\left(\sqrt{\|\mathbf{W}\mathbf{\Lambda}\|_2^2\,\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{D})} + \sqrt{\mathcal{J}_{\mathrm{reg}}(\mathbf{\Lambda},\mathbf{D})}\right)^2 . \tag{41}$$

□

B.2 Proof of Theorem 2

Proof. Let the normalized Hadamard matrix be written as

$$(\mathbf{H}^\top)_{k,j} = \frac{h_{k,j}}{\sqrt{d}}, \qquad h_{k,j} \in \{-1,+1\}. \tag{42}$$

Fixed rotation. Under the persistent sparse CO model,

$$\left((\mathbf{X}^{\mathrm{co}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{H}^{\top}\right)_j = \frac{1}{\sqrt{d}} \sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} \frac{L_k}{\lambda_k} h_{k,j}. \tag{43}$$

Therefore,

$$\left|\left((\mathbf{X}^{\mathrm{co}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{H}^{\top}\right)_j\right| \le \frac{1}{\sqrt{d}} \sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} \frac{|L_k|}{\lambda_k} \le \frac{N_{\mathrm{co}}(\mathbf{X})}{\sqrt{d}} L_{\Lambda,\max}. \tag{44}$$

Taking the maximum over j and summing over T tokens gives

$$\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{I}) \le \frac{T N_{\mathrm{co}}(\mathbf{X})^2}{d} L_{\Lambda,\max}^2. \tag{45}$$

Randomized rotation. Consider a random sign matrix $\mathbf{D} = \mathrm{diag}(\varepsilon_1,\dots,\varepsilon_d)$, where ε_k are independent Rademacher variables. For a fixed coordinate j,

$$Y_j = \left((\mathbf{X}^{\mathrm{co}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^{\top}\right)_j = \frac{1}{\sqrt{d}} \sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} \frac{L_k}{\lambda_k} \varepsilon_k h_{k,j}. \tag{46}$$

This is a centered Rademacher sum. By Hoeffding's inequality (Hoeffding, 1963),

$$\Pr\left(|Y_j| > u\right) \le 2\exp\left(-\frac{du^2}{2R_\Lambda^2}\right), \tag{47}$$

where $R_\Lambda^2 = \sum_{k\in\mathcal{C}_{\mathrm{co}}(\mathbf{X})} L_k^2/\lambda_k^2$. Taking a union bound over d coordinates gives

$$\Pr\left(\left\|(\mathbf{X}^{\mathrm{co}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^{\top}\right\|_\infty > u\right) \le 2d\exp\left(-\frac{du^2}{2R_\Lambda^2}\right). \tag{48}$$

Choosing

$$u^2 = \frac{2R_\Lambda^2}{d}\left(\log(2d) + \log\frac{1}{\delta}\right) \tag{49}$$

gives, with probability at least $1-\delta$,

$$\left\|(\mathbf{X}^{\mathrm{co}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^{\top}\right\|_\infty^2 \le \frac{2R_\Lambda^2}{d}\left(\log(2d) + \log\frac{1}{\delta}\right). \tag{50}$$

Because the persistent sparse CO vector is shared across tokens, summing over T tokens yields

$$\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{D}) \le \frac{2T}{d} R_\Lambda^2 \left(\log(2d) + \log\frac{1}{\delta}\right). \tag{51}$$

Since $R_\Lambda^2 \le N_{\mathrm{co}}(\mathbf{X}) L_{\Lambda,\max}^2$, the stated $N_{\mathrm{co}}(\mathbf{X})$-dependent bound follows.

Sign-based probability amplification. Consider N_s independently sampled sign matrices $\mathcal{D}_{N_s} = \{\mathbf{D}^{(1)},\dots,\mathbf{D}^{(N_s)}\}$. For a single sample,

$$\Pr\left(\mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{D}) > Tu^2\right) \le 2d\exp\left(-\frac{du^2}{2R_\Lambda^2}\right). \tag{52}$$

Hence,

$$\Pr\left(\min_{\mathbf{D}\in\mathcal{D}_{N_s}} \mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{D}) > Tu^2\right) \le \left[2d\exp\left(-\frac{du^2}{2R_\Lambda^2}\right)\right]^{N_s}. \tag{53}$$

Setting the right-hand side to δ gives

$$u^2 = \frac{2R_\Lambda^2}{d}\left(\log(2d) + \frac{1}{N_s}\log\frac{1}{\delta}\right). \tag{54}$$

Therefore, with probability at least $1-\delta$,

$$\min_{\mathbf{D}\in\mathcal{D}_{N_s}} \mathcal{J}_{\mathrm{co}}(\mathbf{\Lambda},\mathbf{D}) \le \frac{2T}{d} R_\Lambda^2 \left(\log(2d) + \frac{1}{N_s}\log\frac{1}{\delta}\right). \tag{55}$$

Using again $R_\Lambda^2 \le N_{\mathrm{co}}(\mathbf{X}) L_{\Lambda,\max}^2$ gives the stated bound. □

B.3 Proof of Proposition 1

Proof. Let $\boldsymbol{r}_t = (\mathbf{X}^{\mathrm{reg}})_{t,:}\mathbf{\Lambda}^{-1}$. For a fixed token t and coordinate j,

$$Y_{t,j} = \left(\boldsymbol{r}_t \mathbf{D}\mathbf{H}^\top\right)_j = \frac{1}{\sqrt{d}} \sum_{k=1}^{d} r_{t,k}\varepsilon_k h_{k,j}. \tag{56}$$

This is a centered Rademacher sum. By Hoeffding's inequality,

$$\Pr\left(|Y_{t,j}| > u\right) \le 2\exp\left(-\frac{du^2}{2\left\|\boldsymbol{r}_t\right\|_2^2}\right). \tag{57}$$

Taking a union bound over all T tokens and d coordinates, with probability at least $1-\delta$,

$$\left\|\boldsymbol{r}_t \mathbf{D}\mathbf{H}^\top\right\|_\infty^2 \le \frac{2\log(2dT/\delta)}{d}\left\|\boldsymbol{r}_t\right\|_2^2 \tag{58}$$

simultaneously for all t. Summing over t gives

$$\sum_{t=1}^{T}\left\|(\mathbf{X}^{\mathrm{reg}})_{t,:}\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{H}^\top\right\|_\infty^2 \le \frac{2\log(2dT/\delta)}{d}\sum_{t=1}^{T}\left\|(\mathbf{X}^{\mathrm{reg}})_{t,:}\mathbf{\Lambda}^{-1}\right\|_2^2 \tag{59}$$

$$= \frac{2\log(2dT/\delta)}{d}\left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2. \tag{60}$$

Multiplying both sides by $\|\mathbf{W}\mathbf{\Lambda}\|_2^2$ yields

$$\mathcal{J}_{\mathrm{reg}}(\mathbf{\Lambda},\mathbf{D}) \le \frac{2\log(2dT/\delta)}{d}\left\|\mathbf{W}\mathbf{\Lambda}\right\|_2^2\left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2. \tag{61}$$

Thus, the regular quantity $\mathcal{J}_{\mathrm{reg}}$ is controlled by $\|\mathbf{W}\mathbf{\Lambda}\|_2^2\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\|_F^2$ up to constants and logarithmic factors.

We derive a channel-wise scaling rule using the Frobenius surrogate

$$\mathcal{R}_2(\mathbf{\Lambda}) = \left\|\mathbf{W}\mathbf{\Lambda}\right\|_F^2\left\|\mathbf{X}^{\mathrm{reg}}\mathbf{\Lambda}^{-1}\right\|_F^2. \tag{62}$$

Let $a_k = \|(\mathbf{X}^{\mathrm{reg}})_{:,k}\|_2^2$ and $b_k = \|(\mathbf{W})_{:,k}\|_2^2$. Then,

$$\mathcal{R}_2(\mathbf{\Lambda}) = \left(\sum_{k=1}^{d}\lambda_k^2 b_k\right)\left(\sum_{k=1}^{d}\frac{a_k}{\lambda_k^2}\right). \tag{63}$$

By Cauchy–Schwarz,

$$\left(\sum_{k=1}^{d}\lambda_k^2 b_k\right)\left(\sum_{k=1}^{d}\frac{a_k}{\lambda_k^2}\right) \ge \left(\sum_{k=1}^{d}\sqrt{a_k b_k}\right)^2. \tag{64}$$

Equality holds when $\lambda_k^2 b_k \propto a_k/\lambda_k^2$, or equivalently, $\lambda_k^4 \propto a_k/b_k$. Thus, up to a common positive scalar,

$$\lambda_k = \left(\frac{a_k}{b_k}\right)^{1/4} = \sqrt{\frac{\|(\mathbf{X}^{\mathrm{reg}})_{:,k}\|_2}{\|(\mathbf{W})_{:,k}\|_2}}. \tag{65}$$

This proves the stated L_2 scaling rule. □

C Evaluation of Sign-Based Probability Amplification

Experimental settings. We examine the candidate budget and screening strategy used for SPA in L2-SmoothRot on Llama3-8B under W4A4 quantization with KV4 and GPTAQ. For a fixed pool of 100 random rotation configurations, we precompute RTN validation, GPTAQ validation, and GPTAQ test perplexities on WikiText-2. For each $N \in \{4, 10, 16, 32\}$, we sample N candidates without replacement from this pool and compare two screening strategies. Top-1 uses the candidate with the lowest RTN validation perplexity. Top-3 retains the three candidates with the lowest RTN validation perplexities and chooses among them using GPTAQ validation perplexity, as in the main experiments with $N = 10$. We record the chosen candidate's GPTAQ test perplexity, using the same candidate subset for both strategies in each trial. We repeat this procedure 1000 times for each N, reusing the precomputed candidate evaluations. These resampling trials characterize variability within the fixed pool; they do not represent 1000 independently generated candidate pools.

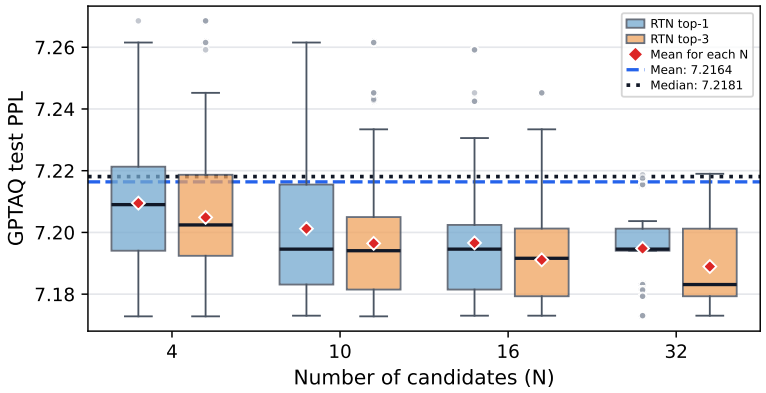

Figure 1: Evaluation of SPA on Llama3-8B under W4A4 quantization with KV4 and GPTAQ, comparing candidate budgets $N \in \{4, 10, 16, 32\}$ and top-1/top-3 screening strategies.

Results. Figure 1 shows the distributions of GPTAQ test perplexity obtained with top-1 and top-3 across candidate budgets. For both strategies, mean GPTAQ test perplexity decreases as N increases and remains below the mean over the full candidate pool. Top-3 achieves lower mean test perplexity than top-1 at every evaluated N. Top-1 also benefits from increasing the candidate budget, although its mean improvement becomes small between $N = 16$ and $N = 32$. Both strategies require N RTN evaluations, followed by one GPTAQ quantization run for top-1 or three for top-3. Thus, top-1 provides an improvement with a single GPTAQ run, while top-3 offers further gains through additional GPTAQ evaluations.

D Full Results Across Models

Table 3 reports C4 perplexity and the individual zero-shot task accuracies omitted from Table 1. All settings and baselines are identical to those in Section 5.

Table 3: Full W4A4 results with GPTAQ weight quantization and KV4 across model families and scales.

Model	Bits	Method	PPL ↓		Accuracy (%) ↑						
			WT2	C4	PIQA	ARC-E	ARC-C	HS	WG	LB	Avg.
Llama-7B	W16A16	N/A	5.67	8.03	79.16	72.85	44.62	76.19	70.01	73.51	69.39
	W4A4	QuaRot	6.12	8.84	77.53	68.01	41.38	72.82	66.77	70.54	66.17
		SmoothRot	6.13	8.87	77.58	67.72	40.27	72.66	67.88	70.83	66.16
		SpinQuant	6.19	8.95	76.44	69.07	42.32	72.75	67.01	69.51	66.18
		L2-SmoothRot	**6.09**	**8.76**	77.15	69.65	41.30	73.01	67.64	71.28	**66.67**
Llama-13B	W16A16	N/A	5.05	7.46	80.20	74.75	47.87	79.08	72.61	76.19	71.78
	W4A4	QuaRot	5.42	8.00	78.51	72.01	46.16	76.66	69.85	73.78	69.49
		SmoothRot	5.42	8.03	78.62	70.58	43.94	76.67	70.88	73.86	69.09
		SpinQuant	5.44	8.13	78.62	71.76	46.33	76.89	69.14	73.12	69.31
		L2-SmoothRot	**5.37**	**7.96**	78.29	72.18	47.01	76.68	70.72	74.60	**69.91**
Llama2-7B	W16A16	N/A	5.47	7.92	79.11	74.58	46.25	76.00	69.06	73.90	69.82
	W4A4	QuaRot	5.96	8.83	76.55	70.54	41.98	72.78	65.43	70.66	66.32
		SmoothRot	5.98	8.87	76.39	71.00	41.72	72.89	66.14	70.56	66.45
		SpinQuant	6.06	9.03	76.82	69.32	41.98	72.28	64.80	70.70	65.98
		L2-SmoothRot	**5.94**	**8.80**	77.80	71.09	41.04	72.31	66.77	70.52	**66.59**
Llama2-13B	W16A16	N/A	4.86	7.36	80.52	77.53	49.06	79.38	72.14	76.77	72.57
	W4A4	QuaRot	5.24	8.01	78.73	73.44	45.14	76.31	70.72	74.89	69.87
		SmoothRot	5.26	8.04	78.40	75.51	46.93	76.25	69.14	74.67	70.15
		SpinQuant	5.26	8.14	77.80	75.38	46.84	76.79	69.77	74.81	70.23
		L2-SmoothRot	**5.23**	**7.99**	78.51	74.54	47.70	76.93	70.64	74.69	**70.50**
Llama3-8B	W16A16	N/A	5.94	8.80	80.85	77.74	53.41	79.16	72.61	76.05	73.30
	W4A4	QuaRot	7.45	12.16	76.66	68.77	41.72	73.51	65.59	67.98	65.71
		SmoothRot	7.57	12.45	75.30	67.05	43.00	73.34	69.06	67.38	65.85
		SpinQuant	7.33	12.04	76.93	74.92	46.50	75.45	67.88	68.50	**68.36**
		L2-SmoothRot	**7.20**	**11.71**	77.09	73.32	45.56	74.85	68.35	69.45	68.10
Llama3.2-1B	W16A16	N/A	9.41	13.05	74.54	60.61	36.35	63.68	60.69	62.95	59.80
	W4A4	QuaRot	13.81	21.52	67.74	50.88	30.20	53.19	53.91	41.57	49.58
		SmoothRot	14.13	22.05	67.19	50.76	30.38	52.42	53.51	37.51	48.63
		SpinQuant	13.33	20.46	66.87	51.47	30.80	55.19	52.88	45.02	50.37
		L2-SmoothRot	**12.92**	**19.87**	68.55	51.81	31.83	54.17	53.04	44.83	**50.71**
Llama3.2-3B	W16A16	N/A	7.55	10.46	77.48	71.72	45.99	73.58	70.01	70.44	68.20
	W4A4	QuaRot	9.27	14.29	74.05	61.99	38.14	67.53	63.06	61.93	61.12
		SmoothRot	9.35	14.52	72.80	60.19	38.31	66.42	61.09	60.66	59.91
		SpinQuant	9.19	14.36	73.72	63.93	41.21	68.22	63.77	61.32	**62.03**
		L2-SmoothRot	**9.08**	**14.04**	72.47	58.54	39.33	67.94	64.40	62.97	60.94
Mistral-7B-v0.3	W16A16	N/A	5.35	8.28	82.26	78.24	52.22	80.47	73.88	75.33	73.73
	W4A4	QuaRot	5.73	8.94	80.14	75.67	47.70	77.70	69.46	72.60	70.55
		SmoothRot	5.77	8.99	80.41	73.95	45.56	77.40	68.75	72.70	69.80
		SpinQuant	5.74	8.95	80.41	75.88	49.15	77.96	70.01	73.61	**71.17**
		L2-SmoothRot	**5.71**	**8.87**	79.49	75.76	48.89	78.04	70.09	73.76	71.00