KronQ: LLM Quantization via Kronecker-Factored Hessian

Donghyun Lee[1], **Yuhang Li**[2], **Ruokai Yin**[2], **Priyadarshini Panda**[1]
[1]University of Southern California, [2]Yale University
donghyun.lee.1@usc.edu

https://github.com/Intelligent-Computing-Lab-Panda/KronQ

Abstract

Post-training quantization (PTQ) is a widely adopted technique for compressing large language models (LLMs) without retraining. Existing second-order PTQ methods, including GPTQ, construct quantization objectives exclusively from input activation statistics, effectively assuming that all output channels contribute equally to the layer-wise reconstruction objective. We propose KronQ, a PTQ framework that challenges this assumption by introducing the gradient covariance into the quantization pipeline. Under the Kronecker-factored Hessian approximation, the quantization loss depends jointly on both the activation and gradient covariances, and KronQ exploits this at two complementary levels. (1) KronQ introduces bidirectional incoherence processing, extending the existing input-side random rotation to the output dimension using the gradient covariance, reducing weight magnitude variance across both input and output dimensions. (2) KronQ derives a new sensitivity metric for inter-layer mixed-precision allocation, driven by the gradient and activation Hessian traces. Notably, in the case of 2-bit weight-only quantization on LLaMA-3-70B, while GPTQ and GPTAQ diverge or produce degenerate quantizations (>2000 perplexity on WikiText-2), KronQ achieves 7.93 perplexity.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of natural language understanding and generation tasks (Brown et al., 2020; Touvron et al., 2023; Grattafiori et al., 2024). However, their growing parameter counts, reaching hundreds of billions in recent models, pose significant deployment challenges, as the memory and computational requirements far exceed what is available on commodity and edge hardware. Quantization (Gholami et al., 2022; Liang et al., 2021) provides an effective solution by compressing model weights and activations to low-bit representations, reducing both memory footprint and inference latency. Among quantization techniques, post-training quantization (PTQ) has attracted particular interest due to its practicality. It requires no retraining and operates on a pretrained model using only a small calibration dataset.

PTQ methods for LLMs span several paradigms (Zhao et al., 2025): (1) Compensation-based methods apply the Optimal Brain Surgeon (OBS) principle (Hassibi & Stork, 1992; LeCun et al., 1989) layer-by-layer, using the input activation covariance as a proxy Hessian to compensate for rounding error in weight quantization (Frantar & Alistarh, 2022; Frantar et al., 2022; Kim et al., 2024; Li et al., 2025; Tseng et al., 2025; Kim et al., 2025). (2) Rotation-based methods apply orthogonal transforms to weight matrices and activations to suppress outliers before quantization (Chee et al., 2023; Tseng et al., 2024; Ashkboos et al., 2024; Liu et al., 2024). (3) Salience-based methods identify sensitive weight channels via activation magnitudes and protect them through per-channel scaling (Lin et al., 2024; Xiao et al., 2023). These paradigms are complementary and are often combined in practice. In this work, we focus on the compensation-based family, in particular GPTQ (Frantar et al., 2022) and its successors, combined with orthogonal transforms from the rotation-based paradigm. These methods have become the dominant framework for LLM weight quantization and are widely integrated into production serving stacks, including HuggingFace Transformers (Wolf et al.,

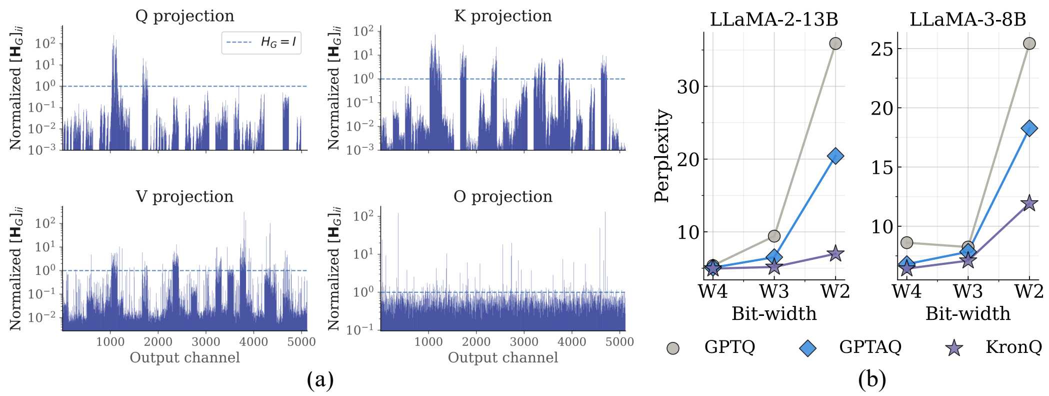


Figure 1: (a) Normalized diagonal entries of the gradient covariance H_G for Q, K, V, and O projections in LLaMA-2-13B. Diagonal entries vary by orders of magnitude, revealing heterogeneous output-side sensitivity. (b) WikiText-2 perplexity of GPTQ, GPTAQ, and KronQ on LLaMA-2-13B and LLaMA-3-8B across W4/W3/W2 weight-only quantization.

2020), GPTQModel (ModelCloud.ai & qubitium@modelcloud.ai, 2024), and vLLM (Kwon et al., 2023).

However, compensation-based methods share a fundamental limitation: the quantization objective is characterized solely through the input activation covariance $\mathbf{H}_X$, which captures only the input-side second-order information of the weight space. Under the Kronecker-factored approximation (K-FAC) (Martens & Grosse, 2015), the full weight Hessian factorizes as $\mathbf{H} \approx \mathbf{H}_X \otimes \mathbf{H}_G$, where $\mathbf{H}_G = \mathbb{E}[\mathbf{g}\mathbf{g}^\top]$ is the gradient covariance with g as the output gradient. Yet compensation-based methods built on the GPTQ-style (Frantar et al., 2022) column-wise OBS solver leave $\mathbf{H}_G$ unused, implicitly treating all output directions as equally important by assuming $\mathbf{H}_G = \mathbf{I}$. As shown in Figure 1(a), output channels vary substantially in gradient magnitude, making this a suboptimal approximation.

We propose KronQ, a post-training quantization method built on the Kronecker-factored Hessian approximation $\mathbf{H} \approx \mathbf{H}_X \otimes \mathbf{H}_G$. By incorporating the gradient covariance $\mathbf{H}_G$ into the quantization objective, KronQ accounts for output-side Hessian information, a dimension entirely overlooked by existing compensation-based methods. Our contributions are as follows.

1. **Kronecker-Factored Quantization Error.** We extend the layer-wise quantization objective to incorporate $\mathbf{H}_G$ under the K-FAC approximation and introduce bidirectional incoherence processing. Notably, $\mathbf{H}_G$ cancels in the quantization update, preserving the efficiency of the base quantizer.
2. **Inter-Layer Mixed Precision via Joint Hessian Traces.** We derive a sublayer sensitivity metric from $\mathrm{tr}(\mathbf{H}_G) \cdot \mathrm{tr}(\mathbf{H}_X)$ that differentiates sublayers sharing identical input statistics, enabling optimal bit-width allocation across sublayers.
3. **Empirical Validation.** We evaluate KronQ on LLaMA-2 and LLaMA-3 from 7B to 70B across weight-only and weight-and-activation settings at W2/W3/W4, achieving consistent state-of-the-art results with the largest gains at 2-bit, as shown in Figure 1(b).

2 Related Work

Compensation-Based Post-Training Quantization. The Optimal Brain Surgeon (OBS) framework (Hassibi & Stork, 1992; LeCun et al., 1989) provides the theoretical basis for Hessian-guided weight perturbation. GPTQ (Frantar et al., 2022) scales this to LLMs by performing column-wise OBS updates using the input activation covariance $\mathbf{H}_X$ as a proxy Hessian, establishing the de facto standard for one-shot LLM quantization. BoA (Kim et al., 2024) extends GPTQ with attention-aware Hessians that capture inter-layer interactions within the attention module. GPTAQ (Li et al., 2025) improves upon GPTQ by correcting for input drift caused by prior-layer quantization errors that accumulate during sequential layer-wise quantization, substantially reducing perplexity at ultra-low bit-widths. Two recent methods go beyond input-only statistics. GuidedQuant (Kim et al., 2025) reweights $\mathbf{H}_X$ with end-loss gradient saliency but ignores cross-output-channel structure. YAQA (Tseng et al., 2025), the most closely related to our work, adopts the same factorization $\mathbf{H} \approx \mathbf{H}_X \otimes \mathbf{H}_G$ but folds $\mathbf{H}_G$ into a modified LDLQ solver via costly power-iterated Hessian sketches. In contrast, KronQ obtains $\mathbf{H}_G$ from a single backward pass and keeps the solver exactly GPTAQ, using $\mathbf{H}_G$ only for incoherence processing and mixed-precision allocation.

Rotation-Based Post-Training Quantization. QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2024) apply orthogonal transforms to weight matrices and activations to suppress outliers, enabling low-bit weight-and-activation quantization. QuIP (Chee et al., 2023) establishes the theoretical foundation, showing that incoherent weight and Hessian matrices yield lower quantization error, and QuIP# (Tseng et al., 2024) extends this with the randomized Hadamard transform and lattice-based vector quantization for practical 2-bit quantization. More recent methods explore learnable transformations, including Kronecker-structured affine maps (Sun et al., 2024) and jointly optimized orthogonal and scaling transforms (Hu et al., 2025).

Mixed-Precision Quantization. Mixed-precision methods differ in granularity. Intra-layer methods allocate bit-widths within a layer: SliM-LLM (Huang et al., 2024) uses weight saliency and CMPQ (Chen et al., 2024) uses activation norms. Inter-layer methods assign one bit-width per sublayer: HAWQ-V2 (Dong et al., 2020) ranks layers by Hessian traces, while AMQ (Lee et al., 2025), Q-Palette (Lee & Song, 2026), and HIGGS (Malinovskii et al., 2025) rely on evolutionary search, dynamic programming, and a linearity theorem, respectively. All use only $\mathbf{H}_X$, which is shared across Q, K, V and cannot differentiate them, whereas the KronQ score $\mathrm{tr}(\mathbf{H}_G) \cdot \mathrm{tr}(\mathbf{H}_X)$ breaks this degeneracy via the output-side factor at the cost of a single backward pass.

3 Preliminary

3.1 Compensation-based Quantization

Let $\mathbf{X} \in \mathbb{R}^{d_{\text{in}} \times n}$ be a matrix of n calibration inputs stacked as columns, $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ the weight matrix of a linear layer with input dimension d_{in} and output dimension d_{out}, and $\widehat{\mathbf{W}}$ its quantized counterpart. The standard layer-wise PTQ objective minimizes the squared output reconstruction error:

$$\min_{\widehat{\mathbf{W}}} \|\mathbf{W}\mathbf{X} - \widehat{\mathbf{W}}\mathbf{X}\|_F^2 \;=\; \mathrm{tr}\Big[(\mathbf{W} - \widehat{\mathbf{W}})\,\mathbf{H}_X\,(\mathbf{W} - \widehat{\mathbf{W}})^\top\Big], \tag{1}$$

where $\mathbf{H}_X = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ is the input activation covariance, which serves as a proxy for the layer-wise Hessian. GPTQ (Frantar et al., 2022) solves this objective greedily column by column using the OBS (Hassibi & Stork, 1992) update. The compensation applied to the remaining unquantized columns after quantizing column p is given by

$$\Delta\mathbf{W}_{:,p+1:} = -\frac{\mathbf{W}_{:,p} - \widehat{\mathbf{W}}_{:,p}}{[\mathbf{H}_X^{-1}]_{pp}} \cdot [\mathbf{H}_X^{-1}]_{p+1:,p}, \tag{2}$$

where $[\mathbf{H}_X^{-1}]_{pp}$ is the p-th diagonal entry of the inverse Hessian and $[\mathbf{H}_X^{-1}]_{p+1:,p}$ is its p-th column restricted to rows $p+1$ onward. This distributes the quantization error of $\mathbf{W}_{:,p}$ optimally to the remaining weights. However, the layer-wise framework causes quantization errors to accumulate across layers, as the objective in Equation (1) is evaluated on activations already degraded by prior-layer quantization. GPTAQ (Li et al., 2025) addresses this by correcting for input drift via asymmetric calibration:

$$\min_{\widehat{\mathbf{W}}} \|\mathbf{W}\widetilde{\mathbf{X}} - \widehat{\mathbf{W}}\mathbf{X}\|_F^2 \;=\; \min_{\widehat{\mathbf{W}}} \mathrm{tr}\Big[\Delta\mathbf{W}\,\mathbf{H}_X\,\Delta\mathbf{W}^\top - \mathbf{W}\,\Delta\mathbf{X}\mathbf{X}^\top\,\Delta\mathbf{W}^\top\Big], \tag{3}$$

where $\Delta\mathbf{W} = \mathbf{W} - \widehat{\mathbf{W}}$, $\mathbf{X}$ is the input from the running quantized model, $\widetilde{\mathbf{X}}$ is the corresponding activation of the full-precision model, and $\Delta\mathbf{X} = \widetilde{\mathbf{X}} - \mathbf{X}$ is the input drift. Despite this correction, GPTAQ inherits the same proxy Hessian $\mathbf{H}_X$, leaving the output-side gradient statistics unaccounted for.

3.2 Incoherence Processing

Incoherence processing (Chee et al., 2023) is a preprocessing technique that reduces quantization error by spreading weight magnitudes uniformly across all dimensions. Formally, the incoherence of $\mathbf{W}$ and its proxy Hessian $\mathbf{H}_X$ are measured by

$$\mu(\mathbf{W}) = \frac{\sqrt{d_{\text{out}} d_{\text{in}}}}{\|\mathbf{W}\|_F} \max_{i,j} |\mathbf{W}_{ij}|, \qquad \mu(\mathbf{H}_X) = \sqrt{d_{\text{in}}} \cdot \max_{i,j} |\mathbf{Q}_{ij}|, \tag{4}$$

where $\mathbf{Q}$ denotes the eigenvector matrix of $\mathbf{H}_X$, and lower μ leads to tighter quantization error bounds. Intuitively, this is achieved by randomizing the spectral directions of $\mathbf{W}$ and $\mathbf{H}_X$, so that rounding sensitivity is no longer concentrated along specific coordinate axes. To this end, random orthogonal matrices $\mathbf{U} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ and $\mathbf{V} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ are applied:

$$\mathbf{W} \leftarrow \mathbf{U}\mathbf{W}\mathbf{V}^\top, \quad \mathbf{H}_X \leftarrow \mathbf{V}\mathbf{H}_X\mathbf{V}^\top. \tag{5}$$

After quantization, the rotations must be applied online during inference. QuIP (Chee et al., 2023) instantiates $\mathbf{U}$ and $\mathbf{V}$ as Kronecker-structured random orthogonal matrices, introducing $\Theta(d_{\text{in}}^{3/2} + d_{\text{out}}^{3/2})$ inference overhead per layer. QuIP# (Tseng et al., 2024) replaces these with randomized Hadamard transforms, reducing this to $\Theta(d_{\text{in}} \log d_{\text{in}} + d_{\text{out}} \log d_{\text{out}})$. However, this incoherence process accounts only for input-side Hessian information through $\mathbf{H}_X$, leaving the output-side entirely unaddressed.

4 Method

4.1 Kronecker-Factored Quantization Error

The standard layer-wise PTQ objective in Equation (1) uses $\mathbf{H}_X$ as a proxy for the full Hessian $\mathbf{H}$, discarding the output-side statistics entirely. We recover this missing factor via the Kronecker-factored approximation (Martens & Grosse, 2015; van der Ouderaa et al., 2023). For a linear layer $\mathbf{y} = \mathbf{W}\mathbf{x}$, the per-sample gradient factorizes as $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{g}\mathbf{x}^\top$, where $\mathbf{g} = \partial \mathcal{L}/\partial \mathbf{y}$. The empirical Fisher approximation of $\mathbf{H}$ (Kunstner et al., 2019) gives:

$$\mathbf{H} \;=\; \mathbb{E}\left[\text{vec}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right) \text{vec}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right)^\top\right] \;=\; \mathbb{E}\left[(\mathbf{x} \otimes \mathbf{g})(\mathbf{x} \otimes \mathbf{g})^\top\right] \;=\; \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \otimes \mathbf{g}\mathbf{g}^\top\right]. \tag{6}$$

Since the joint expectation $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \otimes \mathbf{g}\mathbf{g}^\top] \in \mathbb{R}^{d_{\text{out}} d_{\text{in}} \times d_{\text{out}} d_{\text{in}}}$ is intractable, we apply the K-FAC independence assumption $\mathbf{x} \perp\!\!\!\perp \mathbf{g}$ (Martens & Grosse, 2015; Botev et al., 2017), under which the expectation factorizes:

$$\mathbf{H} \;\approx\; \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \otimes \mathbb{E}[\mathbf{g}\mathbf{g}^\top] \;=:\; \mathbf{H}_X \otimes \mathbf{H}_G, \tag{7}$$

reducing complexity from $\mathcal{O}(d_{\text{in}}^2 d_{\text{out}}^2)$ to $\mathcal{O}(d_{\text{in}}^2 + d_{\text{out}}^2)$. Substituting (7) into (1) yields the Kronecker-factored quantization objective:

$$\min_{\widehat{\mathbf{W}}} \; \text{tr}\left[\mathbf{H}_G \, \Delta\mathbf{W} \, \mathbf{H}_X \, \Delta\mathbf{W}^\top\right]. \tag{8}$$

Most existing PTQ methods implicitly set $\mathbf{H}_G = \mathbf{I}$. However, as shown in Figure 1(a), output channels vary substantially in gradient magnitude, making this approximation suboptimal. KRONQ retains $\mathbf{H}_G$, estimated via a single backward pass over the calibration set and stored as a $d_{\text{out}} \times d_{\text{out}}$ matrix. Building on GPTAQ (Li et al., 2025), we further correct for input drift by incorporating the asymmetric correction from Equation (3):

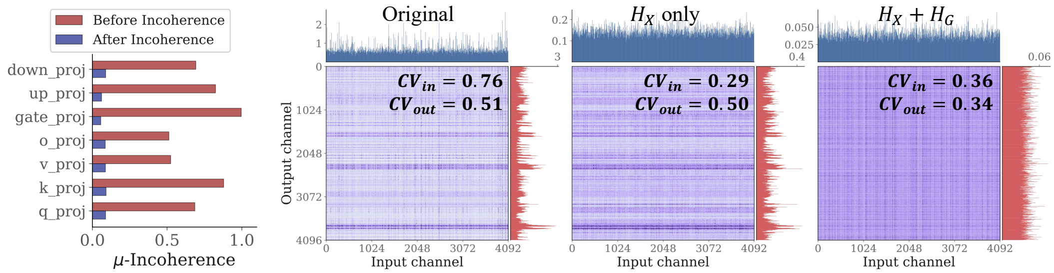
$$\min_{\widehat{\mathbf{W}}} \text{tr}\left[\mathbf{H}_G \left(\Delta\mathbf{W} \, \mathbf{H}_X \, \Delta\mathbf{W}^\top - \mathbf{W} \, \Delta\mathbf{X}\mathbf{X}^\top \, \Delta\mathbf{W}^\top\right)\right]. \tag{9}$$

The optimal weight update under (9) is characterized in the following proposition.

Proposition 1 (KronQ Weight Compensation) *Under the column-wise OBS update applied to (9), $\mathbf{H}_G$ cancels algebraically, and the weight compensation after quantizing column p reduces to:*

$$\Delta\mathbf{W}_{:,p+1:} = -\delta_p \, [\mathbf{H}_X^{-1}]_{p,p+1:} + \mathbf{W}_{:,p} \cdot [\mathbf{P}]_{p,p+1:}, \tag{10}$$

where $\delta_p = (\mathbf{W}_{:,p} - \widehat{\mathbf{W}}_{:,p}) / [\mathbf{H}_X^{-1}]_{pp}$ is the scaled quantization error and $\mathbf{P} = \alpha \left((\Delta\mathbf{X}\mathbf{X}^\top \mathbf{H}_X^{-\top}) \cdot \text{triu}\right) \mathbf{H}_X^{-1}$ is the GPTAQ asymmetric correction matrix.



(a) (b)

Figure 2: (a) μ-incoherence of $\mathbf{H}_G$ before and after incoherence preprocessing across sublayers of LLaMA-2-7B. (b) Weight magnitude distribution of `Q_proj` under three configurations: original weights, after input-side incoherence ($\mathbf{H}_X$ only), and after bidirectional incoherence ($\mathbf{H}_X + \mathbf{H}_G$), where CV_{in} and CV_{out} denote the coefficient of variation of column and row norms, respectively.

The derivation of Proposition 1 is provided in Appendix A.1. KronQ inherits the computational efficiency of GPTAQ while exploiting $\mathbf{H}_G$ for bidirectional incoherence processing and mixed-precision allocation.

4.2 Bidirectional Incoherence Processing

Chee et al. (2023) and Tseng et al. (2024) show that quantization error can be reduced when the proxy Hessian is incoherent—its eigenvectors are not aligned with the coordinate axes—and that $\mathbf{H}_X$ is highly coherent in practice, motivating column-side incoherence processing. Under the KronQ objective (9), $\mathbf{H}_G$ appears as the complementary output-side Hessian factor. This raises a natural question: ***is $\mathbf{H}_G$ also coherent?***

We answer this via the incoherence measure $\mu(\mathbf{H}_G) = \sqrt{d_{\text{out}}} \cdot \max_{i,j} |\mathbf{Q}_{ij}|$ from Equation (4). Here, $\mu/\sqrt{d_{\text{out}}} \to 1$ indicates maximal coherence and $\mu/\sqrt{d_{\text{out}}} \to 0$ indicates incoherence. Figure 2(a) presents $\mu(\mathbf{H}_G)/\sqrt{d_{\text{out}}}$ before and after incoherence processing across sublayers of LLaMA-2-7B. $\mu(\mathbf{H}_G)/\sqrt{d_{\text{out}}}$ reaches up to 0.99, but it drops below 0.10 for all sublayers after incoherence processing, confirming that output-side rotation effectively incoherences $\mathbf{H}_G$. This motivates bidirectional incoherence processing (BiIP). Extending the input-side diagonal rescaling of Chee et al. (2023) to both column and row directions, we apply:

$$\mathbf{W} \leftarrow \mathbf{S}_G \mathbf{W} \mathbf{S}_X, \qquad \mathbf{S}_X = \mathrm{diag}\left(\frac{[\mathbf{H}_X]_{jj}}{\|\mathbf{W}_{:,j}\|^2}\right)^{1/4}, \quad \mathbf{S}_G = \mathrm{diag}\left(\frac{[\mathbf{H}_G]_{ii}}{\|\mathbf{W}_{i,:}\|^2}\right)^{1/4}, \tag{11}$$

where the output-side $\mathbf{S}_G$ term, derived from $\mathbf{H}_G$, is novel. Since the subsequent quantization operates in the rescaled weight space, the Hessians are transformed accordingly as $\mathbf{H}_X \leftarrow \mathbf{S}_X^{-1}\mathbf{H}_X\mathbf{S}_X^{-1}$ and $\mathbf{H}_G \leftarrow \mathbf{S}_G^{-1}\mathbf{H}_G\mathbf{S}_G^{-1}$, ensuring the KronQ objective remains consistent. We then apply orthogonal transforms:

$$\mathbf{W} \leftarrow \mathbf{U}\mathbf{W}\mathbf{V}^\top, \quad \mathbf{H}_X \leftarrow \mathbf{V}\mathbf{H}_X\mathbf{V}^\top, \quad \mathbf{H}_G \leftarrow \mathbf{U}\mathbf{H}_G\mathbf{U}^\top, \quad \Delta\mathbf{X}\mathbf{X}^\top \leftarrow \mathbf{V}\Delta\mathbf{X}\mathbf{X}^\top\mathbf{V}^\top, \tag{12}$$

instantiating $\mathbf{U}$ and $\mathbf{V}$ as randomized Hadamard transforms (Tseng et al., 2024) to make both $\mathbf{H}_G$ and $\mathbf{H}_X$ incoherent with high probability. Figure 2(b) shows that incoherencing $\mathbf{H}_X$ alone leaves the output-channel coefficient of variation (CV_{out}) nearly unchanged, whereas bidirectional incoherence processing reduces both CV_{in} to 0.36 and CV_{out} to 0.34. At inference, BiIP introduces no additional overhead beyond QuIP# (Tseng et al., 2024). The diagonal rescaling $\mathbf{S}_X$ and $\mathbf{S}_G$ are reverted as elementwise operations with negligible cost, while $\mathbf{U}$ and $\mathbf{V}$ introduce $\Theta(d_{\text{in}} \log d_{\text{in}} + d_{\text{out}} \log d_{\text{out}})$ per layer. The KronQ objective (Equation (9)) stays invariant under these transformations, as we formalize in Theorem 1 with proof in Appendix A.2.

Theorem 1 *For any orthogonal* $\mathbf{U} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ *and* $\mathbf{V} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$, KronQ *objective* $\mathrm{tr}\left[\mathbf{H}_G\left(\Delta\mathbf{W}\,\mathbf{H}_X\,\Delta\mathbf{W}^\top - \mathbf{W}\,\Delta\mathbf{X}\mathbf{X}^\top\,\Delta\mathbf{W}^\top\right)\right]$ *is invariant under the transformation* (12).

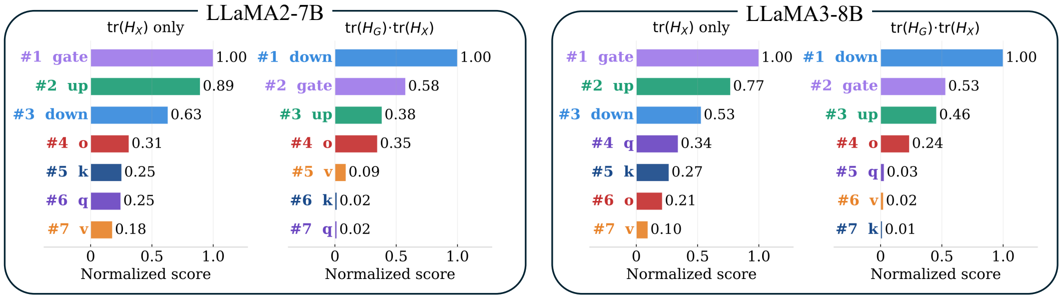


Figure 3: Sublayer sensitivity rankings under the KronQ score $\mathrm{tr}(\mathbf{H}_G) \cdot \mathrm{tr}(\mathbf{H}_X)$ and the activation-only score $\mathrm{tr}(\mathbf{H}_X)$ on LLaMA-2-7B and LLaMA-3-8B.

Chee et al. (2023) shows that incoherence processing reduces the worst-case rounding loss under LDLQ by tightening the bound on $\mathrm{tr}(D)$, the diagonal factor of the LDL decomposition of the proxy Hessian. Under the Kronecker approximation, this analysis extends to both factors simultaneously.

Proposition 2 (KronQ-LDLQ Optimality) *Let $\mu_X = \mu(\mathbf{H}_X)$ and $\mu_G = \mu(\mathbf{H}_G)$ denote the incoherence of $\mathbf{H}_X$ and $\mathbf{H}_G$ after BiIP, and let $k_X = \mathrm{rank}(\mathbf{H}_X)$ and $k_G = \mathrm{rank}(\mathbf{H}_G)$. Then:*

$$L_{\mathrm{worst}}(\textit{KronQ-LDLQ}) \leq \frac{\mu_X^2\, \mu_G^2\, k_X\, k_G}{4\, d_{\mathrm{in}}\, d_{\mathrm{out}}} \cdot \mathrm{tr}(\mathbf{H}_X)\, \mathrm{tr}(\mathbf{H}_G). \tag{13}$$

This improves over GPTQ-LDLQ by a factor of $\mathrm{tr}(D_G)/d_{\mathrm{out}} < 1$, where D_G is the diagonal factor of the LDL decomposition of $\mathbf{H}_G$, which is small whenever $\mathbf{H}_G$ is approximately low-rank. The full derivation is given in Appendix A.3. The overall algorithms are summarized in Algorithm 1 and 2.

4.3 Mixed-Precision Allocation via Joint Hessian Traces

Mixed-precision quantization allocates a bit budget across sublayers by ranking them according to a sensitivity metric. Under the second-order approximation, the expected quantization loss for layer ℓ is proportional to $\epsilon_\ell^2 \cdot \mathrm{tr}(\mathbf{H}^{(\ell)})$ (Dong et al., 2020), where ϵ_ℓ^2 denotes the per-element rounding error variance at bit-width b_ℓ. Since the goal is to rank sublayers rather than compute absolute loss values, ϵ_ℓ^2 is treated as a constant within each layer (Dong et al., 2020), and sensitivity reduces to $\mathrm{tr}(\mathbf{H}^{(\ell)})$. Existing methods approximate $\mathbf{H}^{(\ell)} \approx \mathbf{H}_X^{(\ell)}$, yielding the activation-only score $\mathrm{tr}(\mathbf{H}_X^{(\ell)})$. Under the Kronecker factorization $\mathbf{H} \approx \mathbf{H}_X \otimes \mathbf{H}_G$, this extends naturally to both factors:

$$\mathbb{E}[\mathcal{L}_\ell] \;\propto\; \mathrm{tr}(\mathbf{H}_G^{(\ell)}) \cdot \mathrm{tr}(\mathbf{H}_X^{(\ell)}). \tag{14}$$

We define the KronQ sensitivity score as $s_\ell = \mathrm{tr}(\mathbf{H}_G^{(\ell)}) \cdot \mathrm{tr}(\mathbf{H}_X^{(\ell)})$, computed post-BiIP, and allocate higher bit-widths to sublayers with larger s_ℓ. Within a transformer attention block, the query (Q), key (K), and value (V) projections share the same input, so $\mathbf{H}_X^{(Q)} = \mathbf{H}_X^{(K)} = \mathbf{H}_X^{(V)}$: any activation-only metric assigns identical sensitivity to all three. The KronQ score breaks this degeneracy via $\mathbf{H}_G$, which differs across Q, K, and V as they receive different downstream gradients. As shown in Figure 3, the sublayer rankings under $\mathrm{tr}(\mathbf{H}_G) \cdot \mathrm{tr}(\mathbf{H}_X)$ differ substantially from those under $\mathrm{tr}(\mathbf{H}_X)$ alone. The KronQ score yields strictly better perplexity–bit-width tradeoffs, as we quantify in Section 5.3.

5 Experiments

5.1 Experimental Setup

We evaluate KronQ on the LLaMA family, from LLaMA-2-7B/13B/70B (Touvron et al., 2023) to LLaMA-3-8B/70B (Grattafiori et al., 2024). We compare against GPTQ (Frantar et al., 2022) and GPTAQ (Li et al., 2025) as primary baselines. We consider three regimes: (i) weight-only quantization (WxA16), (ii) group quantization (WxA16, g=128), and (iii)

Table 1: Weight-only quantization (WxA16) WikiText-2 perplexity ↓ and zero-shot accuracy ↑ on three reasoning benchmarks (PiQA, ArcC, WG), evaluated on LLaMA-2-7B. NaN indicates diverged quantization, and – indicates results not available. Avg is reported only when all three tasks are available.

Bits	Method	L2-7B	L2-13B	L2-70B	L3-8B	L3-70B	PiQA	ArcC	WG	Avg
FP16	–	5.47	4.88	3.32	6.14	2.85	78.5	40.6	67.3	62.1
4	OmniQuant	5.74	5.02	3.47	–	–	77.1	37.9	67.0	60.7
	AWQ	5.83	5.06	3.48	7.11	–	76.0	–	–	–
	BoA	5.62	5.10	–	6.56	–	78.4	43.9	69.1	63.8
	SpinQuant	5.58	5.00	3.43	6.49	3.49	77.9	43.3	67.5	62.9
	OSTQuant	5.64	4.94	3.41	6.53	3.19	78.9	44.5	68.6	64.0
	QEP	5.75	5.03	3.49	6.65	–	76.3	–	–	–
	QuIP	8.43	5.14	3.83	7.00	–	–	–	–	–
	QuIP#	5.66	5.00	3.42	–	–	77.2	40.4	67.5	61.7
	GPTQ	5.82	5.37	3.50	8.61	27.49	77.2	41.2	68.7	62.4
	GPTAQ	5.69	5.08	3.46	6.80	399.46	76.8	41.6	68.4	62.3
	KronQ	**5.56**	**4.95**	**3.40**	**6.42**	**3.25**	**78.1**	**45.7**	**68.9**	**64.2**
3	OmniQuant	6.58	5.58	3.92	–	–	73.6	35.3	63.6	57.5
	AWQ	15.30	6.45	4.36	11.80	–	64.7	–	–	–
	BoA	6.01	5.83	–	7.78	–	78.4	40.3	67.6	62.1
	QEP	6.15	5.35	3.81	7.70	–	72.5	–	–	–
	QuIP	12.05	5.50	4.14	8.29	–	–	–	–	–
	QuIP#	6.19	5.34	3.71	–	–	76.1	38.1	65.1	59.8
	GPTQ	6.74	9.41	5.47	8.24	2.6e3	76.0	36.5	65.7	59.4
	GPTAQ	6.21	6.52	3.93	7.84	1.6e4	75.4	37.2	65.3	59.3
	KronQ	**5.84**	**5.18**	**3.66**	**7.09**	**4.41**	**77.1**	**42.6**	**67.6**	**62.4**
2	OmniQuant	37.73	17.22	7.81	–	–	57.5	21.6	51.5	43.5
	AWQ	NaN	NaN	NaN	NaN	–	50.7	–	–	–
	BoA	12.76	18.33	–	21.70	–	64.9	28.3	54.5	49.2
	QEP	11.97	8.42	5.87	27.33	–	50.5	–	–	–
	QuIP	65.59	11.23	6.54	70.52	–	54.6	19.4	51.8	41.9
	QuIP#	12.30	7.60	4.87	–	–	68.0	29.2	59.0	52.1
	GPTQ	31.11	35.89	9.54	25.43	2.6e3	63.0	28.4	53.9	48.4
	GPTAQ	NaN	20.43	6.41	18.27	NaN	NaN	NaN	NaN	NaN
	KronQ	**8.15**	**6.99**	**5.14**	**11.92**	**7.93**	**68.8**	**29.1**	**61.0**	**52.9**

weight-and-activation quantization (WxA4). For activation quantization, we utilize the QuaRot (Ashkboos et al., 2024) framework. All other quantization configurations follow GPTAQ (Li et al., 2025). Perplexity is measured on WikiText-2 (Merity et al., 2016) and zero-shot accuracy is reported on seven commonsense reasoning benchmarks: PiQA (Bisk et al., 2020), Arc-Easy (ArcE), Arc-Challenge (ArcC) (Clark et al., 2018), HellaSwag (HS) (Zellers et al., 2019), WinoGrande (WG) (Sakaguchi et al., 2021), BoolQ (Clark et al., 2019), and OpenBookQA (OBQA) (Mihaylov et al., 2018). All experiments use 128 calibration samples from WikiText-2 with 2048 context lengths and are run on A100 GPUs. $\mathbf{H}_G$ is precomputed via a single backward pass prior to calibration and loaded layer-by-layer on demand, incurring no additional overhead during quantization. Results for prior works are taken from their respective papers, except for OmniQuant (Shao et al., 2023) zero-shot accuracy, which we reproduce using the official implementation.

5.2 Results on Uniform Precision

Weight-only Quantization. Table 1 shows WikiText-2 perplexity and zero-shot accuracy on three reasoning benchmarks (PiQA, ArcC, WG) for per-channel weight-only quantization at W4/W3/W2. The comparison baselines include OmniQuant (Shao et al., 2023), AWQ (Lin et al., 2024), BoA (Kim et al., 2024), SpinQuant (Liu et al., 2024), OSTQuant (Hu et al., 2025), QEP (Arai & Ichikawa, 2025), QuIP (Chee et al., 2023), QuIP# (Tseng et al., 2024), GPTQ (Frantar et al., 2022), and GPTAQ (Li et al., 2025). For QuIP#, we report the no-E8P, no-finetuning variant to match our scalar quantization setting. KronQ achieves the lowest perplexity across nearly all settings, with the most significant gains at W2 and W3 where activation-covariance-only methods degrade severely. The zero-shot accuracy improvements mirror the perplexity gains, with KronQ achieving consistently higher accuracy across all benchmarks and bit-widths. On LLaMA-3-70B, GPTQ and GPTAQ

Table 2: Wikitext-2 perplexity and zero-shot accuracy for weight-only group quantization (g=128) at W2.

Model	Method	Bits	Wiki2↓	PiQA	ArcE	ArcC	HS	WG	BoolQ	OBQA	Avg↑
LLaMA-2-7B	OmniQuant	2	11.06	65.8	45.0	28.0	49.3	54.6	61.2	34.2	48.3
	GPTQ	2	274.00	57.8	38.5	26.3	29.3	53.1	42.1	28.2	39.3
	GPTAQ	2	23.19	59.3	39.7	26.3	31.0	55.7	41.0	29.0	40.3
	KronQ	**2**	**7.61**	**69.9**	**52.8**	**32.1**	**56.8**	**63.9**	**65.4**	**35.2**	**53.7**
LLaMA-2-13B	OmniQuant	2	8.26	69.0	57.1	32.7	56.5	53.1	63.7	35.6	52.5
	GPTQ	2	8.40	67.6	43.6	29.4	49.3	61.3	42.1	35.0	46.9
	GPTAQ	2	7.17	65.1	46.4	29.5	47.6	60.2	57.0	32.8	48.4
	KronQ	**2**	**6.51**	**72.6**	**68.9**	**38.1**	**63.9**	**65.9**	**66.0**	**36.8**	**58.9**
LLaMA-3-8B	GPTQ	2	16.41	55.2	34.2	20.9	45.1	58.7	59.3	29.4	43.3
	GPTAQ	2	13.08	59.5	37.9	23.0	50.2	60.1	60.7	31.6	46.1
	KronQ	**2**	**10.76**	**69.0**	**56.3**	**34.4**	**53.3**	**63.3**	**65.9**	**34.2**	**53.8**

Table 3: Wikitext-2 perplexity and zero-shot accuracy for weight & activation quantization.

Model	Method	Bits	Wiki2↓	PiQA	ArcE	ArcC	HS	WG	BoolQ	OBQA	Avg↑
LLaMA-2-7B	GPTQ	W2A4	36.74	54.0	31.7	22.7	31.1	51.1	57.3	26.4	39.2
	GPTAQ	W2A4	10.91	61.6	44.7	25.1	43.3	55.5	62.1	30.2	46.1
	KronQ	**W2A4**	**9.38**	**64.7**	**46.0**	**26.3**	**46.2**	**58.3**	**64.3**	**34.6**	**48.6**
LLaMA-2-13B	GPTQ	W2A4	12.55	60.9	42.7	25.1	40.6	55.6	62.1	28.8	45.1
	GPTAQ	W2A4	8.41	66.5	52.0	29.4	49.4	58.6	62.5	35.2	50.5
	KronQ	**W2A4**	**7.77**	**67.2**	**55.4**	**29.0**	**51.4**	**59.4**	**62.5**	**33.8**	**51.2**
LLaMA-3-8B	GPTQ	W2A4	32.79	56.5	36.7	21.8	34.6	52.8	60.0	27.2	41.4
	GPTAQ	W2A4	19.14	59.9	40.4	24.3	41.0	53.8	62.6	27.8	44.3
	KronQ	**W2A4**	**16.47**	**60.7**	**39.7**	**25.3**	**41.4**	**53.4**	**62.6**	**30.0**	**44.7**

diverge or fail to produce valid quantizations, while KronQ achieves 4.41 and 7.93 at W3 and W2, respectively. We attribute this to the anomalously large weight outliers unique to LLaMA-3-70B, as further analyzed in Appendix D. The remaining zero-shot results, comparison with gradient-based methods, and generalization to Mistral-7B are presented in Appendix C.1, C.2, and C.3, respectively.

Group Quantization. Group quantization generally achieves better performance than per-channel quantization due to its more optimized scaling factors and zero points. Table 2 reports results under weight-only 2-bit group quantization with 128 column blocks. KronQ maintains its lead over GPTQ and GPTAQ. For example, on LLaMA-2-7B, GPTQ degrades to 274.0 PPL under group quantization while KronQ achieves 7.61. The perplexity comparison with OmniQuant (Shao et al., 2023) and AWQ (Lin et al., 2024) and the results of 3 and 4-bit are shown in Appendix C.4 and C.5, respectively.

Weight-and-Activation Quantization. For activation quantization, we apply QuaRot (Ashkboos et al., 2024) framework to suppress outliers. Table 3 presents W2A4 results on LLaMA-2-7B/13B and LLaMA-3-8B. KronQ consistently outperforms both baselines across all settings. The gains are again largest at W2A4, where KronQ reduces PPL from 36.74 to 9.38 on LLaMA-2-7B. The comparison with QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2024) and other bit-width results are presented in Appendix C.6 and C.7, respectively.

Generalization to Newer Models and Harder Benchmarks. To verify that KronQ's gains are not specific to the LLaMA family or to perplexity, we evaluate recent model families: Gemma-3-12B (Gemma Team, 2025), DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025), and Phi-4-mini-instruct (Abouelenin et al., 2025). Table 4 reports weight-only WikiText-2 perplexity at W4 and W2 against GPTQ (Frantar et al., 2022) and GPTAQ (Li et al., 2025). KronQ wins all eight settings, and the margin widens sharply at W2, confirming that output-side curvature is particularly valuable in the ultra-low-bit regime. Beyond perplexity, Table 5 evaluates four harder benchmarks, GPQA-Diamond (Rein et al., 2023), MMLU (Hendrycks et al., 2020), AIME-2024 (Maxwell-Jia, 2025), and LiveCodeBench (Jain et al., 2025) on reasoning models, DeepSeek-R1-Distill-Llama-8B and Gemma-3-12B-IT at W4. KronQ outperforms both baselines on all four benchmarks. The gains are largest on generative reasoning tasks such as LiveCodeBench, where KronQ nearly doubles the GPTAQ scores on both models. Full detailed evaluation settings are provided in Appendix C.8.

Table 4: WikiText-2 perplexity (↓) on newer 2025 model families (WxA16, W4/W2).

Model	Bits	GPTQ	GPTAQ	KronQ
Gemma-3-12B-PT	W4	39.77	34.84	**27.97**
	W2	1.3e4	269.0	**60.10**
Gemma-3-12B-IT	W4	60.37	28.82	**25.60**
	W2	7.1e3	94.08	**54.39**
DeepSeek-R1-Distill-Llama-8B	W4	16.56	15.32	**13.58**
	W2	73.53	110.34	**27.72**
Phi-4-mini-instruct	W4	13.81	11.80	**10.17**
	W2	111.42	31.07	**19.26**

Table 5: Accuracy (↑, %) on harder benchmarks (W4A16).

Model	Benchmark	GPTQ	GPTAQ	KronQ
DeepSeek-R1-Distill-Llama-8B	LiveCodeBench	19.9	16.3	**37.3**
	AIME-2024	10.0	13.3	**16.7**
	MMLU	44.15	40.51	**53.17**
	GPQA-Diamond	28.80	27.27	**29.29**
Gemma-3-12B-IT	LiveCodeBench	7.8	9.6	**22.3**
	AIME-2024	10.0	10.0	**23.3**
	MMLU	64.82	65.21	**70.02**
	GPQA-Diamond	32.83	34.34	**38.38**

Table 6: Sublayer sensitivity rankings and WikiText-2 perplexity on LLaMA-2-7B. Each row cumulatively upgrades one additional sublayer to W3 across all layers.

Score	Ranking	Avg bits	Wiki2↓
baseline W2	-	2.00	8.15
$\mathrm{tr}(H_G)\cdot\mathrm{tr}(H_X)$	1: `down_proj`	2.17	7.22
	2: `gate_proj`	2.29	6.74
	3: `up_proj`	2.43	6.38
$\mathrm{tr}(H_X)$	1: `gate_proj`	2.17	7.45
	2: `up_proj`	2.29	6.92
	3: `down_proj`	2.43	6.38

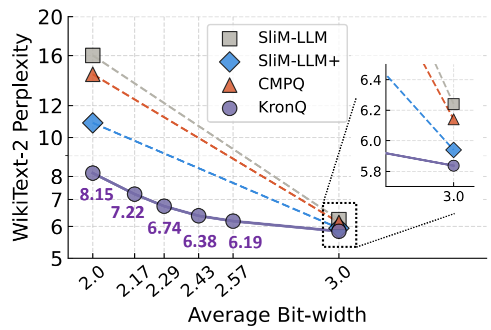


Figure 4: WikiText-2 perplexity vs. average bit-width on LLaMA-2-7B, comparing KronQ with prior mixed precision works.

5.3 Results on Mixed-precision

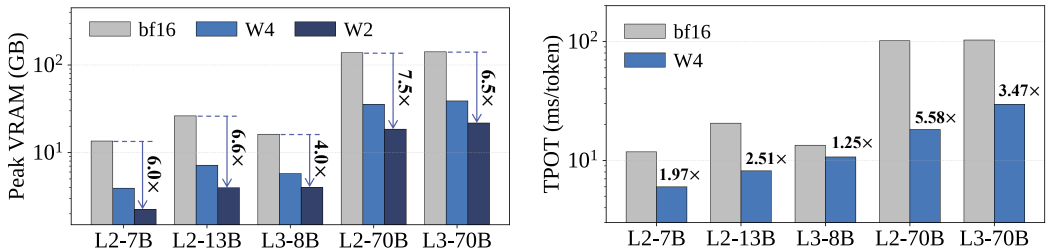
We evaluate the mixed-precision allocation strategy by incrementally upgrading the most sensitive sublayers from W2 to W3 according to the sensitivity score $s = \mathrm{tr}(\mathbf{H}_G) \cdot \mathrm{tr}(\mathbf{H}_X)$ in Equation (14), applying each upgrade across all transformer layers. Table 6 compares the sublayer rankings and WikiText-2 perplexity under the KronQ joint score versus the activation-only score $\mathrm{tr}(\mathbf{H}_X)$ on LLaMA-2-7B. The activation-only score produces suboptimal allocations: its top-1 sublayer (`gate_proj`) yields higher perplexity than the top-1 of the KronQ joint score (`down_proj`) at the same average bit-width. The KronQ joint score resolves this degeneracy via $\mathbf{H}_G$, yielding strictly better PPL–bits tradeoff. Figure 4 further compares KronQ mixed-precision against SliM-LLM (Huang et al., 2024), SliM-LLM+, and CMPQ (Chen et al., 2024) on LLaMA-2-7B, demonstrating that KronQ achieves lower perplexity than W3 baselines at only $\sim$2.6 average bits. Other architectures show the same trend in Appendix C.9, and a comparison with inter-layer mixed-precision methods is provided in Appendix C.10.

5.4 Memory and Latency

Beyond calibration cost, we measure the end-to-end inference benefit of quantization in Figure 5. On the memory side, KronQ reduces peak inference VRAM by 3.5–3.9× at W4 and 4.0–7.5× at W2 over the bf16 baseline, consistently across model scales. This compression directly translates into deployment enablement: a 70B model that requires two 80 GB A100s in bf16 (138–141 GB) fits comfortably on a single A100 at W4 (35–39 GB). On the latency side, the reduced memory traffic yields 1.25–2.51× faster decoding, measured as time per output token (TPOT), on the 7B–13B models under a matched single-GPU configuration. We report the 70B latency as well, though its bf16 baseline spans two GPUs and is included only for context.

5.5 Analysis

Ablation Study. We conduct three ablation studies to isolate the contribution of each component in KronQ, with W2 WikiText-2 perplexity shown in Table 7. First, to isolate the contribution of the base quantizer, we replace GPTAQ with a plain GPTQ base. This degrades perplexity from 8.15/6.99/11.92 to 9.81/7.95/14.52 on LLaMA-2-7B/13B and LLaMA-3-8B, confirming that drift correction and BiIP are complementary. Second, we examine the effect of diagonal rescaling ($\mathbf{S}_X$, $\mathbf{S}_G$) in BiIP. Removing the scaling while retaining



(a) Peak inference VRAM (GB) across precisions

(b) Decoding latency (TPOT, ms/token)

Figure 5: Inference efficiency of KronQ quantized models. (a) Peak VRAM and (b) decoding latency (TPOT) at batch size 1, relative to the bf16 baseline. For 70B, the bf16 baseline runs on two GPUs due to OOM on single GPU, while W4 runs on a single GPU.

Table 7: Ablation study on WikiText-2 perplexity at W2.

Model	Base		Scaling ($\mathbf{S}_X$, $\mathbf{S}_G$)		Incoherence		
	GPTQ	GPTAQ	✗	✓	$\mathbf{H}_X$ only	$\mathbf{H}_G$ only	BiIP
LLaMA-2-7B	9.81	**8.15**	8.44	**8.15**	8.47	180.43	**8.15**
LLaMA-2-13B	7.95	**6.99**	7.10	**6.99**	7.13	81.18	**6.99**
LLaMA-3-8B	14.52	**11.92**	12.94	**11.92**	12.56	14.55	**11.92**

the Hadamard rotations consistently degrades perplexity across all models, demonstrating that diagonal rescaling is a necessary complement to the orthogonal transforms. Third, we ablate the rotation directions in BiIP. Applying input-side incoherence ($\mathbf{H}_X$ only) already reduces perplexity, confirming the importance of suppressing input-side outliers (Chee et al., 2023). Output-side incoherence alone ($\mathbf{H}_G$ only) degrades severely on LLaMA-2-7B/13B as input-side outliers remain unaddressed. Combining both directions consistently achieves the lowest perplexity, demonstrating that $\mathbf{H}_G$ provides complementary output-side correction that $\mathbf{H}_X$ alone cannot capture.

Calibration Efficiency.

KronQ introduces a moderate latency overhead over GPTAQ (+8–11s/layer across models), attributable to the bidirectional Hadamard rotations in BiIP, which require dense matrix multiplications for both $\mathbf{H}_X$ and $\mathbf{H}_G$. This overhead is consistent across model sizes, reflecting that the dominant cost scales with the hidden dimension rather than the total parameter count. For memory, KronQ additionally stores $\mathbf{H}_G$ as a full $d_{\text{out}} \times d_{\text{out}}$ matrix during BiIP preprocessing, causing a temporary memory overhead above GPTAQ. However, since $\mathbf{H}_G$ cancels algebraically in the column-wise quantization updates in Proposition 1, it is released upon completion of BiIP, and the memory required for the subsequent quantization loop is identical to GPTAQ. Per-sublayer memory breakdown is provided in Appendix E.

Table 8: Calibration latency (s/layer) and memory needed to perform calibration (GiB) per layer. † represents $\mathbf{H}_G$ offload after BiIP.

Model	Method	Latency	Memory
LLaMA-2-7B	GPTQ	24.95	1.57
	GPTAQ	30.11	1.99
	KronQ	38.25	3.20 → 1.99†
LLaMA-3-8B	GPTQ	26.13	1.97
	GPTAQ	35.21	2.54
	KronQ	45.59	4.38 → 2.54†
LLaMA-2-13B	GPTQ	32.00	2.46
	GPTAQ	39.88	3.11
	KronQ	50.74	5.02 → 3.11†

6 Conclusion

We presented KronQ, a PTQ framework that incorporates the gradient covariance $\mathbf{H}_G$ into the quantization pipeline. KronQ exploits $\mathbf{H}_G$ for bidirectional incoherence processing and inter-layer mixed-precision allocation, while inheriting the computational efficiency of GPTAQ as $\mathbf{H}_G$ cancels algebraically in the quantization updates. Experiments on LLaMA-2 and LLaMA-3 from 7B to 70B show consistent gains across W2/W3/W4, with the largest improvements at 2-bit, where activation-covariance-only methods degrade severely. The limitations of this work are presented in Appendix F.

Acknowledgments

This work was supported in part by CoCoSys, a JUMP2.0 center sponsored by DARPA and SRC, the National Science Foundation (CAREER Award, Grant #2312366, Grant #2318152), the DARPA Young Faculty Award, the DoE MMICC center SEA-CROGS (Award #DE-SC0023198), and the Global Industrial Technology Cooperation Center (GITCC) program.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. *arXiv preprint arXiv:2504.09629*, 2025.

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.

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.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pp. 557–565. PMLR, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36:4396–4429, 2023.

Zihan Chen, Bike Xie, Jundong Li, and Cong Shen. Channel-wise mixed-precision quantization for large language models. *arXiv preprint arXiv:2410.13056*, 2024.

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, 2019.

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.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pp. 291–326. Chapman and Hall/CRC, 2022.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Zhihang Yuan, Jiangyong Yu, Chen Xu, Zhe Jiang, and Sifan Zhou. Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. *arXiv preprint arXiv:2501.13987*, 2025.

Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. Slim-llm: Salience-driven mixed-precision quantization for large language models. *arXiv preprint arXiv:2405.14917*, 2024.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de Las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv 2023. *arXiv preprint arXiv:2310.06825*, 2024.

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens JS Schaefer, Deokjae Lee, Yeonhong Park, Jae W Lee, and Hyun Oh Song. Guidedquant: Large language model quantization via exploiting end loss guidance. *arXiv preprint arXiv:2505.07004*, 2025.

Junhan Kim, Ho-young Kim, Eulrang Cho, Chungman Lee, Joonyoung Kim, and Yongkweon Jeon. Boa: Attention-aware post-training quantization without backpropagation. *arXiv preprint arXiv:2406.13474*, 2024.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Deokjae Lee and Hyun Oh Song. Q-palette: Fractional-bit quantizers toward optimal bit allocation for efficient llm deployment. *Advances in Neural Information Processing Systems*, 38:11525–11558, 2026.

Sangjun Lee, Seung-taek Woo, Jun-gyu Jin, Changhun Lee, and Eunhyeok Park. Amq: Enabling automl for mixed-precision weight-only quantization of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35520–35538, 2025.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. Gptaq: Efficient finetuning-free quantization for asymmetric calibration. *arXiv preprint arXiv:2504.02692*, 2025.

Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403, 2021.

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.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. Quantization hurts reasoning? an empirical study on quantized reasoning models. *arXiv preprint arXiv:2504.04823*, 2025.

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. *arXiv preprint arXiv:2405.16406*, 2024.

Vladimir Malinovskii, Andrei Panferov, Ivan Ilin, Han Guo, Peter Richtárik, and Dan Alistarh. Higgs: Pushing the limits of large language model quantization via the linearity theorem. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10857–10886, 2025.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Maxwell-Jia. AIME 2024 dataset. https://huggingface.co/datasets/Maxwell-Jia/AIME_2024, 2025. Hugging Face dataset.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

ModelCloud.ai and qubitium@modelcloud.ai. Gpt-qmodel. https://github.com/modelcloud/gptqmodel, 2024. Contact: qubitium@modelcloud.ai.

Minghai Qin. The uniqueness of llama3-70b series with per-channel quantization. *arXiv preprint arXiv:2408.15301*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

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.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, et al. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *Proceedings of machine learning research*, 235:48630, 2024.

Albert Tseng, Zhaofeng Sun, and Christopher De Sa. Model-preserving adaptive rounding. *arXiv preprint arXiv:2505.22988*, 2025.

Tycho FA van der Ouderaa, Markus Nagel, Mart Van Baalen, Yuki M Asano, and Tijmen Blankevoort. The llm surgeon. *arXiv preprint arXiv:2312.17244*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Jiaqi Zhao, Ming Wang, Miao Zhang, Yuzhang Shang, Xuebo Liu, Yaowei Wang, Min Zhang, and Liqiang Nie. Benchmarking post-training quantization in llms: Comprehensive taxonomy, unified evaluation, and comparative analysis. *arXiv preprint arXiv:2502.13178*, 2025.

A Proofs

A.1 Proof of Proposition 1 (KronQ Weight Compensation)

The general OBS optimal weight update (Hassibi & Stork, 1992) is:

$$\Delta\boldsymbol{\theta} = -\mathbf{H}^{-1}\mathbf{E}_p(\mathbf{E}_p^\top\mathbf{H}^{-1}\mathbf{E}_p)^{-1}\delta_p, \tag{15}$$

where $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W})$, $\mathbf{E}_p = \mathbf{I}\otimes e_p$ is the selection matrix for column p, and δ_p is the quantization error. Under the K-FAC approximation $\mathbf{H} = \mathbf{H}_X \otimes \mathbf{H}_G$, we have $\mathbf{H}^{-1} = \mathbf{H}_X^{-1}\otimes\mathbf{H}_G^{-1}$.

We compute the two terms in the OBS formula. First:

$$\begin{aligned}\mathbf{E}_p^\top\mathbf{H}^{-1}\mathbf{E}_p &= (\mathbf{I}\otimes e_p^\top)(\mathbf{H}_X^{-1}\otimes\mathbf{H}_G^{-1})(\mathbf{I}\otimes e_p)\\ &= [\mathbf{H}_X^{-1}]_{pp}\otimes\mathbf{H}_G^{-1}\\ &= [\mathbf{H}_X^{-1}]_{pp}\cdot\mathbf{H}_G^{-1}.\end{aligned} \tag{16}$$

Second:

$$\mathbf{H}^{-1}\mathbf{E}_p = (\mathbf{H}_X^{-1}\otimes\mathbf{H}_G^{-1})(\mathbf{I}\otimes e_p) = [\mathbf{H}_X^{-1}]_{:,p}\otimes\mathbf{H}_G^{-1}. \tag{17}$$

Substituting into the OBS formula, $\mathbf{H}_G^{-1}$ appears in both numerator and denominator and cancels algebraically:

$$\Delta\mathbf{W}_{:,p+1:} = -\frac{[\mathbf{H}_X^{-1}]_{:,p}\otimes\mathbf{H}_G^{-1}}{[\mathbf{H}_X^{-1}]_{pp}\cdot\mathbf{H}_G^{-1}}\cdot\delta_p = -\delta_p\cdot[\mathbf{H}_X^{-1}]_{p,p+1:}, \tag{18}$$

where $\delta_p = (\mathbf{W}_{:,p} - \widehat{\mathbf{W}}_{:,p})/[\mathbf{H}_X^{-1}]_{pp}$. Adding the GPTAQ asymmetric correction term (Li et al., 2025) yields (10). □

A.2 Proof of Theorem 1 (Rotation Invariance)

Denote the transformed quantities after (12) as $\mathbf{W}' = \mathbf{U}\mathbf{W}\mathbf{V}^\top$, $\widetilde{\mathbf{H}}_G = \mathbf{U}\mathbf{H}_G\mathbf{U}^\top$, $\widetilde{\mathbf{H}}_X = \mathbf{V}\mathbf{H}_X\mathbf{V}^\top$, and $(\Delta\mathbf{X}\mathbf{X}^\top)' = \mathbf{V}\Delta\mathbf{X}\mathbf{X}^\top\mathbf{V}^\top$. Note that $\Delta\mathbf{W}' = \mathbf{W}' - \widehat{\mathbf{W}}' = \mathbf{U}\Delta\mathbf{W}\mathbf{V}^\top$. We verify invariance of each term separately.

For the first term:

$$\begin{aligned}\mathrm{tr}\left[\widetilde{\mathbf{H}}_G\,\Delta\mathbf{W}'\,\widetilde{\mathbf{H}}_X\,(\Delta\mathbf{W}')^\top\right] &= \mathrm{tr}\left[\mathbf{U}\mathbf{H}_G\mathbf{U}^\top\cdot\mathbf{U}\Delta\mathbf{W}\mathbf{V}^\top\cdot\mathbf{V}\mathbf{H}_X\mathbf{V}^\top\cdot\mathbf{V}\Delta\mathbf{W}^\top\mathbf{U}^\top\right]\\ &= \mathrm{tr}\left[\mathbf{U}\mathbf{H}_G\Delta\mathbf{W}\mathbf{H}_X\Delta\mathbf{W}^\top\mathbf{U}^\top\right]\\ &= \mathrm{tr}\left[\mathbf{H}_G\,\Delta\mathbf{W}\,\mathbf{H}_X\,\Delta\mathbf{W}^\top\right],\end{aligned} \tag{19}$$

where the second step uses $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, and the third step uses the cyclic property of the trace.

For the second term:

$$\begin{aligned}\mathrm{tr}\left[\widetilde{\mathbf{H}}_G\,\mathbf{W}'\,(\Delta\mathbf{X}\mathbf{X}^\top)'\,(\Delta\mathbf{W}')^\top\right] &= \mathrm{tr}\left[\mathbf{U}\mathbf{H}_G\mathbf{U}^\top\cdot\mathbf{U}\mathbf{W}\mathbf{V}^\top\cdot\mathbf{V}\Delta\mathbf{X}\mathbf{X}^\top\mathbf{V}^\top\cdot\mathbf{V}\Delta\mathbf{W}^\top\mathbf{U}^\top\right]\\ &= \mathrm{tr}\left[\mathbf{U}\mathbf{H}_G\,\mathbf{W}\,\Delta\mathbf{X}\mathbf{X}^\top\,\Delta\mathbf{W}^\top\mathbf{U}^\top\right]\\ &= \mathrm{tr}\left[\mathbf{H}_G\,\mathbf{W}\,\Delta\mathbf{X}\mathbf{X}^\top\,\Delta\mathbf{W}^\top\right],\end{aligned} \tag{20}$$

where the second step uses $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, and the third step uses the cyclic property of the trace.

Since both terms are individually invariant, the full KronQ objective (9) is invariant under the transformation (12). □

A.3 Proof of Proposition 2 (KronQ-LDLQ Optimality)

Kronecker LDL factorisation. Let $\mathbf{H}_X = L_X D_X L_X^\top$ and $\mathbf{H}_G = L_G D_G L_G^\top$ be LDL decompositions, where $L_X = U_X + I$ and $L_G = U_G + I$ are unit lower-triangular. Using the mixed-product property of the Kronecker product:

$$\mathbf{H}_X \otimes \mathbf{H}_G = (L_X D_X L_X^\top) \otimes (L_G D_G L_G^\top) = (L_X \otimes L_G)(D_X \otimes D_G)(L_X \otimes L_G)^\top. \tag{21}$$

Since $L_X \otimes L_G$ is unit lower-triangular (the Kronecker product of two unit lower-triangular matrices is unit lower-triangular under the standard lexicographic ordering of index pairs), this is a valid LDL decomposition of $\mathbf{H}_X \otimes \mathbf{H}_G$. The diagonal factor is $D_X \otimes D_G$, and:

$$\operatorname{tr}(D_X \otimes D_G) = \sum_{i,j} [D_X]_{ii} [D_G]_{jj} = \operatorname{tr}(D_X) \cdot \operatorname{tr}(D_G). \tag{22}$$

Since the worst-case LDLQ proxy loss equals $\frac{1}{4}\operatorname{tr}(D)$ (Chee et al., 2023), it follows that:

$$L_{\text{worst}}(\text{KronQ-LDLQ}) = \frac{1}{4} \operatorname{tr}(D_X) \cdot \operatorname{tr}(D_G). \tag{23}$$

Per-factor incoherence bound. We apply Lemma 2 of Chee et al. (2023) to each Kronecker factor separately. A symmetric positive semidefinite matrix $M \in \mathbb{R}^{n \times n}$ is μ-incoherent if all eigenvectors v_i satisfy $\|v_i\|_\infty \leq \mu/\sqrt{n}$. For μ-incoherent M, QuIP Lemma 2 gives:

$$\operatorname{tr}(D_M) \;\leq\; \frac{\mu^2}{n} \operatorname{tr}(M^{1/2})^2, \tag{24}$$

where D_M is the LDL diagonal of M. The eigenvectors of $\widetilde{\mathbf{H}}_X \otimes \widetilde{\mathbf{H}}_G$ are $v_i \otimes u_j$ where v_i, u_j are eigenvectors of $\widetilde{\mathbf{H}}_X$ and $\widetilde{\mathbf{H}}_G$ respectively. Each entry satisfies $|[v_i \otimes u_j]_k| = |[v_i]_{k_1}| \cdot |[u_j]_{k_2}| \leq \frac{\mu_X}{\sqrt{d_{\text{in}}}} \cdot \frac{\mu_G}{\sqrt{d_{\text{out}}}}$, so $\widetilde{\mathbf{H}}_X \otimes \widetilde{\mathbf{H}}_G$ is $(\mu_X \mu_G)$-incoherent with ambient dimension $d_{\text{in}} d_{\text{out}}$. Applying the per-factor bound and trace invariance under orthogonal conjugation ($\operatorname{tr}(\widetilde{\mathbf{H}}_X^{1/2}) = \operatorname{tr}(\mathbf{H}_X^{1/2})$, $\operatorname{tr}(\widetilde{\mathbf{H}}_G^{1/2}) = \operatorname{tr}(\mathbf{H}_G^{1/2})$):

$$\frac{1}{4} \operatorname{tr}(D_X) \cdot \operatorname{tr}(D_G) \;\leq\; \frac{\mu_X^2\, \mu_G^2}{4\, d_{\text{in}}\, d_{\text{out}}} \cdot \operatorname{tr}(\mathbf{H}_X^{1/2})^2 \operatorname{tr}(\mathbf{H}_G^{1/2})^2. \tag{25}$$

Finally, applying $\operatorname{tr}(H^{1/2})^2 \leq \operatorname{rank}(H) \cdot \operatorname{tr}(H)$ to each factor (by Cauchy–Schwarz on the eigenvalues) gives:

$$\frac{\mu_X^2\, \mu_G^2}{4\, d_{\text{in}}\, d_{\text{out}}} \cdot \operatorname{tr}(\mathbf{H}_X^{1/2})^2 \operatorname{tr}(\mathbf{H}_G^{1/2})^2 \;\leq\; \frac{\mu_X^2\, \mu_G^2\, k_X\, k_G}{4\, d_{\text{in}}\, d_{\text{out}}} \cdot \operatorname{tr}(\mathbf{H}_X) \operatorname{tr}(\mathbf{H}_G), \tag{26}$$

where $k_X = \operatorname{rank}(\mathbf{H}_X)$ and $k_G = \operatorname{rank}(\mathbf{H}_G)$. □

Comparison with QuIP. Setting $\mathbf{H}_G = \mathbf{I}$ (rank-d_{out}, $\mu_G = 1$, $k_G = d_{\text{out}}$) recovers the QuIP (Chee et al., 2023) bound $\mu_X^2 k_X \operatorname{tr}(\mathbf{H}_X)/(4 d_{\text{in}})$. KronQ+BiIP improves this bound by the factor $\mu_G^2 k_G / d_{\text{out}}$, which can be substantially less than one when $\mathbf{H}_G$ is approximately low-rank.

Table 9: Worst-case proxy loss under LDLQ rounding.

Method	**Worst-case proxy loss**
Nearest rounding	$\frac{1}{4}\operatorname{tr}(\mathbf{H}_X) \cdot \operatorname{tr}(\mathbf{H}_G)$
GPTQ-LDLQ ($\mathbf{H}_G = \mathbf{I}$)	$\frac{1}{4}\operatorname{tr}(D_X) \cdot d_{\text{out}}$
KronQ-LDLQ (ours)	$\frac{1}{4}\operatorname{tr}(D_X) \cdot \operatorname{tr}(D_G)$

B KronQ Algorithm

Algorithm 1 Bidirectional Incoherence Pre-Processing (BiIP)

Require: FP weight $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, degraded input $\mathbf{X} \in \mathbb{R}^{d_{\text{in}} \times n}$, FP input $\tilde{\mathbf{X}} \in \mathbb{R}^{d_{\text{in}} \times n}$, $\mathbf{H}_G \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ (precomputed)

$\mathbf{H}_X \leftarrow \mathbf{X}\mathbf{X}^\top; \quad \Delta\mathbf{X}\mathbf{X}^\top \leftarrow (\tilde{\mathbf{X}} - \mathbf{X})\mathbf{X}^\top$ ▷ online statistics

sample random Hadamard $\mathbf{U} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}, \ \mathbf{V} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$

$\mathbf{S}_X \leftarrow \big(\text{diag}(\mathbf{H}_X)/\text{diag}(\mathbf{W}^\top\mathbf{W})\big)^{1/4}$ ▷ input-side rescaling

$\mathbf{S}_G \leftarrow \big(\text{diag}(\mathbf{H}_G)/\text{diag}(\mathbf{W}\mathbf{W}^\top)\big)^{1/4}$ ▷ output-side rescaling

$\mathbf{W} \leftarrow \mathbf{S}_G\, \mathbf{W}\, \mathbf{S}_X$

$\mathbf{H}_X \leftarrow \mathbf{S}_X^{-1}\mathbf{H}_X\mathbf{S}_X^{-1}; \quad \Delta\mathbf{X}\mathbf{X}^\top \leftarrow \mathbf{S}_X^{-1}\Delta\mathbf{X}\mathbf{X}^\top\mathbf{S}_X^{-1}; \quad \mathbf{H}_G \leftarrow \mathbf{S}_G^{-1}\mathbf{H}_G\mathbf{S}_G^{-1}$

$\mathbf{W} \leftarrow \mathbf{U}\mathbf{W}\mathbf{V}^\top; \quad \mathbf{H}_X \leftarrow \mathbf{V}\mathbf{H}_X\mathbf{V}^\top; \quad \Delta\mathbf{X}\mathbf{X}^\top \leftarrow \mathbf{V}\,\Delta\mathbf{X}\mathbf{X}^\top\mathbf{V}^\top$ ▷ incoherence

$\mathbf{H}_G \leftarrow \mathbf{U}\mathbf{H}_G\mathbf{U}^\top;$ **release** $\mathbf{H}_G$ ▷ $\mathbf{H}_G$ not needed after BiIP (Prop. 1)

return $\mathbf{W}, \ \mathbf{H}_X, \ \Delta\mathbf{X}\mathbf{X}^\top, \ \mathbf{S}_X, \ \mathbf{S}_G, \ \mathbf{U}, \ \mathbf{V}$

Algorithm 2 KronQ – Quantization Loop

Require: $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{H}_X \in \mathbb{R}^{n \times n}$, $\Delta\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n \times n}$, $\mathbf{S}_X$, $\mathbf{S}_G$, $\mathbf{U}$, $\mathbf{V}$ (from Alg. 1), block size B, damping λ, scaling α

$\mathbf{H}_X \leftarrow \mathbf{H}_X + \lambda \cdot \text{mean}(\text{diag}(\mathbf{H}_X))\, \mathbf{I}$ ▷ damping

$\mathbf{L} \leftarrow \textit{Inverse\_Cholesky}(\mathbf{H}_X)$

$\mathbf{P} \leftarrow \alpha\big((\Delta\mathbf{X}\mathbf{X}^\top \cdot \mathbf{L}) \odot \mathbf{M}_U\big)\mathbf{L}^\top$ ▷ GPTAQ correction; $\mathbf{M}_U$: upper-triangular mask

$\mathbf{Q} \leftarrow \mathbf{0}_{m \times n}; \quad \mathbf{E} \leftarrow \mathbf{0}_{m \times B}$

for $i = 0, B, 2B, \ldots$ **do**

  for $j = i, \ i{+}1, \ \ldots, \ i{+}B{-}1$ **do**

    $\mathbf{Q}_{:,j} \leftarrow \text{quant}(\mathbf{W}_{:,j})$

    $\mathbf{E}_{:,\,j-i} \leftarrow (\mathbf{W}_{:,j} - \mathbf{Q}_{:,j}) \ / \ \mathbf{L}_{jj}$

    $\mathbf{W}_{:,\,j:(i+B)} \leftarrow \mathbf{W}_{:,\,j:(i+B)} - \mathbf{E}_{:,\,j-i}\, \mathbf{L}^\top_{j,\,j:(i+B)} + \mathbf{W}_{:,j}\, \mathbf{P}_{j,\,j:(i+B)}$

  end for

  $\mathbf{W}_{:,\,(i+B):} \leftarrow \mathbf{W}_{:,\,(i+B):} - \mathbf{E} \cdot \mathbf{L}^\top_{i:(i+B),\,(i+B):} + \mathbf{W}_{:,\,i:(i+B)}\, \mathbf{P}_{i:(i+B),\,(i+B):}$

end for

$\mathbf{Q} \leftarrow \mathbf{U}^\top\mathbf{Q}\,\mathbf{V}; \quad \mathbf{Q} \leftarrow \mathbf{S}_G^{-1}\,\mathbf{Q}\,\mathbf{S}_X^{-1}$ ▷ revert incoherence & rescaling

return Q

C Additional Experiments

C.1 Zero-shot Accuracy on Weight-only Quantization

Table 10: Zero-shot accuracy on reasoning benchmarks (PiQA, ArcE, ArcC, HS, WG, BoolQ, OBQA) for weight-only quantization (WxA16) at W4/W3/W2. NaN indicates diverged quantization.

Model	Method	Bits	PiQA	ArcE	ArcC	HS	WG	BoolQ	OBQA	Avg↑
LLaMA-2-7B	GPTQ	4	77.2	69.2	41.2	70.5	68.7	67.3	40.8	62.1
	GPTAQ	4	76.8	70.6	41.6	70.3	68.4	68.1	42.6	62.6
	KronQ	**4**	**78.1**	**73.7**	**45.7**	**75.2**	**68.9**	**78.1**	**43.6**	**66.2**
	OmniQuant	3	74.0	62.6	36.3	67.6	63.6	65.7	39.8	58.5
	GPTQ	3	76.0	59.4	36.5	66.8	65.7	65.9	40.0	58.6
	BoA	3	78.4	69.3	40.3	72.0	67.6	74.2	41.2	63.3
	GPTAQ	3	75.4	62.0	37.2	65.9	65.3	66.1	38.4	58.6
	KronQ	**3**	**77.1**	**72.4**	**42.6**	**72.0**	**67.6**	**75.4**	**41.6**	**64.1**
	OmniQuant	2	57.2	35.3	26.1	31.7	51.5	51.1	30.6	40.5
	GPTQ	2	63.0	45.1	28.4	42.8	53.9	52.3	28.4	44.8
	BoA	2	64.9	47.7	28.3	45.5	54.5	64.7	29.6	47.9
	GPTAQ	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	KronQ	**2**	**68.8**	**55.0**	**29.1**	**52.3**	**61.0**	**62.7**	**35.0**	**52.0**

Continued on next page

Table 10 – continued

Model	Method	Bits	PiQA	ArcE	ArcC	HS	WG	BoolQ	OBQA	Avg↑
LLaMA-2-13B	GPTQ	4	77.3	73.7	45.6	67.6	71.1	71.0	43.2	64.2
	GPTAQ	4	77.9	74.4	46.5	73.1	71.3	75.5	44.8	66.2
	KronQ	**4**	**80.2**	**74.7**	**48.0**	**79.1**	**71.8**	**82.1**	**44.8**	**68.7**
	OmniQuant	3	77.7	69.0	42.7	72.8	65.9	69.0	38.8	62.3
	GPTQ	3	72.5	62.8	36.8	54.4	65.5	66.5	36.6	56.4
	BoA	3	77.4	69.3	43.5	74.7	62.5	78.9	35.2	63.1
	GPTAQ	3	75.1	65.5	39.0	62.6	64.2	67.3	38.2	58.8
	KronQ	**3**	**79.3**	**73.5**	**46.5**	**77.0**	**71.2**	**78.9**	**45.0**	**67.3**
	OmniQuant	2	62.9	44.4	28.4	49.7	52.6	62.2	33.8	47.7
	GPTQ	2	57.9	36.2	23.4	37.3	53.2	62.1	28.2	42.6
	BoA	2	63.3	46.3	29.8	49.3	52.6	62.2	27.6	47.3
	GPTAQ	2	54.8	33.1	21.4	31.7	52.0	58.0	29.4	40.1
	KronQ	**2**	**72.4**	**62.5**	**36.4**	**59.4**	**65.1**	**69.4**	**37.0**	**57.5**
LLaMA-2-70B	GPTQ	4	82.3	81.0	57.6	82.7	77.4	82.3	48.0	73.0
	GPTAQ	4	82.8	81.5	57.3	82.8	77.0	82.4	47.8	73.1
	KronQ	**4**	**82.3**	**80.3**	**57.9**	**83.8**	**79.9**	**85.4**	**49.2**	**74.1**
	OmniQuant	3	80.7	75.6	45.8	78.1	73.6	66.5	42.6	66.1
	GPTQ	3	80.4	78.9	55.8	80.2	76.2	79.1	46.0	70.9
	GPTAQ	3	81.7	77.5	54.4	79.8	75.2	81.6	45.8	70.9
	KronQ	**3**	**82.3**	**81.9**	**58.3**	**82.5**	**78.5**	**83.0**	**48.8**	**73.6**
	OmniQuant	2	69.9	55.5	31.5	55.4	53.6	64.9	33.8	52.1
	GPTQ	2	72.1	64.7	37.6	58.0	66.9	66.6	40.2	58.0
	GPTAQ	2	72.2	64.4	36.8	58.4	64.6	69.9	34.6	57.3
	KronQ	**2**	**77.4**	**75.9**	**46.9**	**72.0**	**75.0**	**79.7**	**43.2**	**67.2**
LLaMA-3-8B	GPTQ	4	79.4	75.7	49.7	75.6	73.5	77.1	44.8	68.0
	GPTAQ	4	80.0	75.7	49.7	77.2	72.5	76.5	44.4	68.0
	KronQ	**4**	**79.4**	**78.2**	**51.5**	**78.2**	**73.2**	**82.2**	**44.2**	**69.6**
	GPTQ	3	74.5	62.6	39.7	70.0	67.8	73.3	38.6	60.9
	BoA	3	77.3	72.8	45.1	72.7	71.4	78.7	42.6	65.8
	GPTAQ	3	73.5	60.1	40.4	70.8	69.9	75.7	41.0	61.6
	KronQ	**3**	**77.5**	**74.5**	**50.2**	**74.9**	**71.7**	**81.1**	**41.2**	**67.3**
	GPTQ	2	54.2	34.0	22.3	37.9	49.5	45.9	28.0	38.8
	BoA	2	59.9	44.9	26.6	43.3	55.8	60.5	29.4	45.8
	GPTAQ	2	55.0	33.1	23.6	38.1	53.4	46.8	28.0	39.7
	KronQ	**2**	**67.0**	**50.7**	**30.6**	**49.0**	**61.1**	**66.1**	**33.0**	**51.1**
LLaMA-3-70B	GPTQ	4	50.2	26.3	24.6	59.8	56.4	66.9	33.2	45.3
	GPTAQ	4	56.9	39.6	25.6	30.1	54.4	48.1	29.0	40.5
	KronQ	**4**	**84.2**	**81.8**	**61.7**	**84.8**	**79.1**	**86.8**	**48.0**	**75.2**
	GPTQ	3	52.0	25.3	26.1	26.4	49.0	41.2	28.4	35.5
	GPTAQ	3	50.0	25.9	27.5	26.2	49.7	37.9	30.0	35.3
	KronQ	**3**	**83.5**	**80.4**	**58.8**	**83.1**	**79.5**	**85.4**	**47.2**	**74.0**
	GPTQ	2	51.7	25.6	26.4	26.4	49.0	37.8	27.4	34.9
	GPTAQ	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	KronQ	**2**	**77.6**	**73.2**	**44.5**	**66.0**	**72.6**	**79.1**	**39.6**	**64.7**

C.2 Comparison with Gradient-Based Methods

YAQA (Tseng et al., 2025) shares the K-FAC form $\mathbf{H} \approx \mathbf{H}_X \otimes \mathbf{H}_G$ but folds $\mathbf{H_G}$ into a modified LDLQ solver, whereas KronQ uses $\mathbf{H_G}$ only for incoherence and bit allocation. Table 11 shows that KronQ attains lower perplexity at an order-of-magnitude lower calibration cost. GuidedQuant (Kim et al., 2025) weights the input-side Hessian by per-output-channel saliency, missing the cross-output-channel structure that $\mathbf{H_G}$ captures. As shown in Table 12, KronQ matches or exceeds its LNQ-based results with a plain uniform per-channel grid.

Table 11: YAQA setup (LLaMA-3.1-8B-Instruct, WikiText-2, ctx 8192, INT4 per-channel).

Method	WikiText-2 PPL
BF16	6.50
LDLQ	6.76
YAQA-A	6.71
YAQA-B	6.72
KronQ	**6.69**

Table 12: GuidedQuant setup (LLaMA-2, WikiText-2, ctx 4096).

Model	Bits	LNQ+GuidedQuant	**KronQ**
L2-7B	W3	5.57	**5.44**
L2-13B	W3	4.91	**4.84**
L2-70B	W3	3.47	**3.48**
L2-7B	W2	8.83	**7.64**
L2-13B	W2	7.26	**6.50**
L2-70B	W2	5.04	**5.07**

C.3 Comparison on Mistral-7B

Table 13 extends the weight-only quantization evaluation to Mistral-7B (Jiang et al., 2024), demonstrating that KRONQ generalizes beyond the LLaMA family.

Table 13: Weight-only quantization (WxA16) WikiText-2 perplexity ↓ on Mistral-7B.

	AWQ	QEP	QuIP	GPTQ	GPTAQ	**KRONQ**
W4	5.72	5.48	11.11	8.96	6.95	**5.32**
W3	7.90	5.84	7.11	532.46	12.89	**5.54**
W2	NaN	9.59	26.63	5900.69	27.07	**7.78**

C.4 Comparison with Previous Works on Group Quantization

Table 14 reports WikiText-2 perplexity under group quantization (g=128), where OmniQuant (Shao et al., 2023) and AWQ (Lin et al., 2024) results are taken from their papers.

Table 14: Weight-only group quantization (WxA16, g=128) WikiText-2 perplexity.

	LLaMA-2-7B			**LLaMA-2-13B**			**LLaMA-2-70B**		
Method	**W4**	**W3**	**W2**	**W4**	**W3**	**W2**	**W4**	**W3**	**W2**
RTN	5.72	6.66	4.2e3	4.98	5.51	122.08	3.46	3.97	27.27
GPTQ	5.57	6.02	274.00	4.95	5.22	8.40	3.39	3.64	5.22
AWQ	5.60	6.24	NaN	4.97	5.32	NaN	3.41	3.74	–
OmniQuant	5.58	6.03	11.06	4.95	5.28	8.26	3.40	3.78	6.55
GPTAQ	5.56	5.88	23.19	4.94	5.16	7.17	3.39	3.64	5.28
KRONQ	**5.54**	**5.77**	**7.61**	**4.93**	**5.14**	**6.51**	**3.38**	**3.60**	**4.85**

C.5 3 and 4-bit Results on Group Quantization

Table 15 provides full W2/W3/W4 results under group quantization (g=128). KRONQ maintains its lead across all bit-widths and model sizes, with particularly large margins at W2 where per-channel methods struggle.

Table 15: Group quantization (WxA16, g=128), full results (W2/W3/W4).

Model	**Method**	**Bits**	**Wiki2↓**	**PiQA**	**ArcE**	**ArcC**	**HS**	**WG**	**BoolQ**	**OBQA**	**Avg↑**
LLaMA-2-7B	GPTQ	4	5.57	78.6	72.6	44.7	75.2	69.0	77.8	44.0	66.0
	GPTAQ	4	5.56	79.1	73.8	44.5	75.2	69.0	76.4	43.6	65.9
	KRONQ	**4**	**5.54**	**78.4**	**73.1**	**44.9**	**75.7**	**69.3**	**79.4**	**43.0**	**66.2**
	GPTQ	3	6.02	76.7	68.6	41.3	71.8	67.9	72.9	41.2	62.9
	GPTAQ	3	5.88	77.3	69.7	41.2	72.4	68.7	66.9	40.6	62.4
	KRONQ	**3**	**5.77**	**77.5**	**72.0**	**43.4**	**73.3**	**67.4**	**75.1**	**40.8**	**64.2**
LLaMA-3-8B	GPTQ	4	6.41	80.3	78.2	52.0	78.3	73.6	80.9	44.4	69.7
	GPTAQ	4	6.38	80.0	76.8	51.9	78.3	73.8	78.6	44.0	69.1
	KRONQ	**4**	**6.34**	**80.4**	**78.9**	**53.2**	**78.4**	**73.3**	**80.9**	**45.8**	**70.1**
	GPTQ	3	7.42	77.9	74.2	48.8	61.2	71.5	79.2	41.6	64.9
	GPTAQ	3	7.51	77.0	73.3	46.4	72.0	72.0	76.8	42.0	65.6
	KRONQ	**3**	**6.96**	**78.7**	**74.2**	**49.1**	**74.9**	**71.7**	**81.7**	**42.4**	**67.5**
LLaMA-2-13B	GPTQ	4	4.95	80.2	76.4	48.2	78.5	72.1	79.4	45.0	68.5
	GPTAQ	4	4.94	79.7	76.5	48.0	78.7	72.7	79.0	44.0	68.4
	KRONQ	**4**	**4.93**	**80.3**	**75.7**	**48.6**	**79.0**	**72.1**	**80.8**	**45.0**	**68.8**
	GPTQ	3	5.22	79.2	74.2	46.6	76.0	70.9	78.2	43.4	66.9
	GPTAQ	3	5.16	79.1	75.3	46.4	76.1	71.1	77.7	44.4	67.2
	KRONQ	**3**	**5.13**	**79.2**	**76.3**	**48.7**	**77.5**	**72.1**	**81.6**	**44.6**	**68.6**
LLaMA-2-70B	GPTQ	4	3.39	82.8	80.9	57.2	83.5	78.1	82.7	48.0	73.3
	GPTAQ	4	3.39	82.7	80.6	57.5	83.4	77.9	83.6	48.6	73.5
	KRONQ	**4**	**3.38**	**82.7**	**79.2**	**56.2**	**84.1**	**79.8**	**85.0**	**48.4**	**73.6**
	GPTQ	3	3.64	82.3	80.0	55.6	82.5	75.5	79.1	44.6	71.4
	GPTAQ	3	3.64	82.4	80.8	55.5	82.2	76.9	83.2	47.6	72.7
	KRONQ	**3**	**3.60**	**82.2**	**80.7**	**57.1**	**83.0**	**79.0**	**83.8**	**47.8**	**73.4**
	GPTQ	2	5.22	77.7	73.3	47.1	68.4	71.0	76.0	39.0	64.6
	GPTAQ	2	5.28	78.4	75.3	46.3	69.1	73.0	75.5	41.6	65.6
	KRONQ	**2**	**4.85**	**79.2**	**75.8**	**49.8**	**74.5**	**76.6**	**78.1**	**44.2**	**68.3**
LLaMA-3-70B	GPTQ	4	3.23	83.5	84.7	62.5	84.3	79.6	85.8	47.6	75.4
	GPTAQ	4	3.17	83.2	82.9	61.3	84.3	80.7	85.4	48.4	75.2
	KRONQ	**4**	**3.15**	**83.7**	**80.7**	**60.7**	**85.0**	**80.7**	**87.0**	**47.6**	**75.1**
	GPTQ	3	4.49	82.1	80.0	58.4	82.6	78.8	85.1	44.6	73.1
	GPTAQ	3	6.56	81.7	78.8	54.6	77.5	74.4	77.0	43.4	69.6
	KRONQ	**3**	**4.20**	**82.2**	**79.5**	**59.0**	**83.1**	**78.9**	**84.9**	**47.6**	**73.6**
	GPTQ	2	16.60	74.2	68.3	42.6	40.2	69.3	74.7	38.4	58.2
	GPTAQ	2	20.62	56.4	37.9	24.7	37.3	50.7	53.6	28.2	41.3
	KRONQ	**2**	**7.65**	**79.0**	**76.4**	**49.7**	**72.2**	**73.8**	**80.7**	**43.4**	**67.9**

C.6 Comparison with Previous Works on Weight-and-activation Quantization

Table 16 reports WikiText-2 perplexity for weight-and-activation quantization. All methods apply the QuaRot (Ashkboos et al., 2024) rotation framework for activation quantization. QuaRot and SpinQuant (Liu et al., 2024) results are taken from their respective papers.

Table 16: Weight-and-activation quantization (WxA4) WikiText-2 perplexity.

Bits	Method	LLaMA-2-7B	LLaMA-2-13B	LLaMA-3-8B
W4A4	QuaRot	6.10	5.40	8.16
	SpinQuant	5.96	5.24	7.39
	GPTQ	6.04	5.29	7.78
	GPTAQ	5.87	5.17	7.39
	KronQ	**5.83**	**5.15**	**7.34**
W2A4	QuaRot	32.6	15.5	NaN
	GPTQ	36.74	12.55	32.79
	GPTAQ	10.91	8.41	19.14
	KronQ	**9.38**	**7.77**	**16.47**

C.7 3 and 4-bit Results on Weight-and-activation Quantization

Table 17 extends the weight-and-activation evaluation to W3A4/W4A4. KronQ achieves the lowest perplexity across all settings, with gains increasing at lower bit-widths.

Table 17: Weight-and-activation quantization (WxA4), W3A4/W4A4 results.

Model	Method	Bits	Wiki2↓	PiQA	ArcE	ArcC	HS	WG	BoolQ	OBQA	Avg↑
LLaMA-2-7B	GPTQ	W4A4	6.04	77.3	71.0	42.2	72.9	66.1	73.5	39.2	63.2
	GPTAQ	W4A4	5.87	77.2	71.6	43.2	73.6	67.2	74.5	41.6	64.1
	KronQ	**W4A4**	**5.83**	**77.6**	**71.1**	**42.7**	**73.6**	**67.2**	**75.8**	**42.4**	**64.3**
	GPTQ	W3A4	6.79	75.0	67.8	40.2	68.4	63.1	71.7	36.6	60.4
	GPTAQ	W3A4	6.25	75.3	68.3	40.4	70.5	65.8	71.9	40.2	61.8
	KronQ	**W3A4**	**6.19**	**75.6**	**67.9**	**41.0**	**70.5**	**65.5**	**72.5**	**40.4**	**61.9**
LLaMA-2-13B	GPTQ	W4A4	5.29	78.9	74.7	47.2	76.5	69.6	79.0	44.4	67.2
	GPTAQ	W4A4	5.17	79.3	75.2	48.6	77.0	68.8	78.7	43.4	67.3
	KronQ	**W4A4**	**5.15**	**79.2**	**75.6**	**48.4**	**77.7**	**70.5**	**80.3**	**44.2**	**68.0**
	GPTQ	W3A4	5.77	77.4	72.5	44.8	73.3	69.1	71.4	42.0	64.4
	GPTAQ	W3A4	5.48	77.7	73.1	44.6	74.0	69.9	74.7	42.8	65.3
	KronQ	**W3A4**	**5.40**	**79.1**	**75.0**	**45.8**	**75.0**	**70.2**	**77.2**	**43.8**	**66.6**
LLaMA-3-8B	GPTQ	W4A4	7.78	77.0	72.1	45.7	74.2	67.9	76.4	42.6	65.1
	GPTAQ	W4A4	7.39	77.4	72.0	47.5	75.6	69.0	77.5	42.0	65.9
	KronQ	**W4A4**	**7.34**	**77.4**	**75.3**	**48.7**	**75.1**	**70.1**	**75.8**	**42.4**	**66.4**
	GPTQ	W3A4	9.22	73.3	64.2	39.3	68.8	64.3	72.8	39.2	60.3
	GPTAQ	W3A4	8.31	74.7	68.6	42.1	70.3	65.9	73.4	40.8	62.3
	KronQ	**W3A4**	**8.25**	**74.6**	**69.5**	**42.7**	**70.0**	**67.6**	**76.0**	**38.8**	**62.7**

C.8 Harder Benchmark Evaluation Details

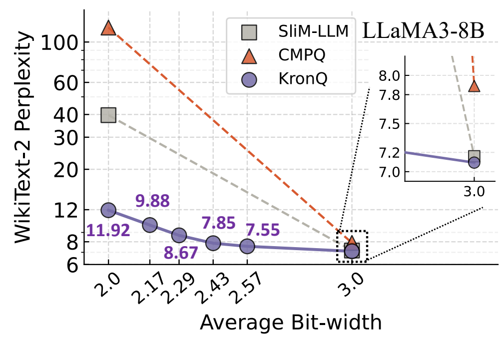
We evaluate DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) and Gemma-3-12B-IT (Gemma Team, 2025), each quantized to W4 (per-channel, asymmetric, weight-only) with KronQ and, as baselines, with GPTQ and GPTAQ under identical settings. All models are calibrated on 128 WikiText-2 sequences with no fine-tuning. MMLU, GPQA-Diamond, and AIME-2024 are evaluated through the lm-evaluation-harness (Gao et al., 2024) in the zero-shot setting, while LiveCodeBench uses its official runner with vLLM. We report log-likelihood accuracy on MMLU (57 subjects) and GPQA-Diamond (198 questions), exact-match accuracy on AIME-2024 (30 problems), and pass@1 on LiveCodeBench (`release_v5`, code-generation scenario, 32,768 maximum new tokens). For the reasoning model DeepSeek-R1-Distill-Llama-8B, we follow the decoding protocol of Liu et al. (2025), using temperature 0.6, top-p 0.95, and 32,768 maximum new tokens with stochastic sampling. We report its AIME-2024 score averaged over 8 samples (avg@8). The instruction-tuned Gemma-3-12B-IT is decoded greedily with a 2,048-token budget, and we report its AIME-2024 as pass@1.

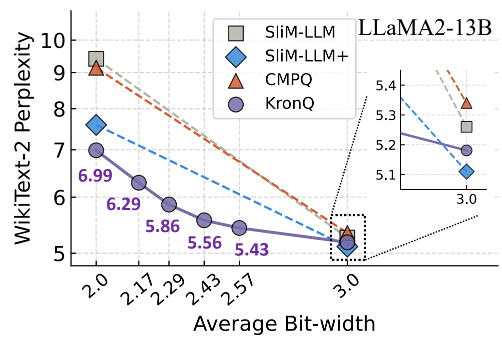
C.9 Mixed Precision Quantization on LLaMA

Table 18 and right figures extend the mixed-precision analysis of Section 5.3 to LLaMA-3-8B and LLaMA-2-13B. The KronQ joint score yields strictly better PPL–bits tradeoffs than the activation-only score on both architectures. We also compare KRONQ with SliM-LLM (Huang et al., 2024) and CMPQ (Chen et al., 2024), consistent with the LLaMA-2-7B results in the main paper.

Table 18: Sublayer sensitivity rankings and WikiText-2 perplexity on LLaMA-3-8B and LLaMA-2-13B. Each row cumulatively upgrades one additional sublayer to W3 across all layers.

Score	Ranking	Avg bits	Wiki2↓
LLaMA-3-8B			
baseline W2	-	2.00	11.92
$\text{tr}(H_G)\cdot\text{tr}(H_X)$	1: `down_proj`	2.17	9.88
	2: `gate_proj`	2.29	8.67
	3: `up_proj`	2.43	7.85
$\text{tr}(H_X)$	1: `gate_proj`	2.17	10.24
	2: `up_proj`	2.29	9.16
	3: `down_proj`	2.43	7.85
LLaMA-2-13B			
baseline W2	-	2.00	6.99
$\text{tr}(H_G)\cdot\text{tr}(H_X)$	1: `down_proj`	2.17	6.29
	2: `gate_proj`	2.29	5.86
	3: `up_proj`	2.43	5.56
$\text{tr}(H_X)$	1: `gate_proj`	2.17	6.39
	2: `up_proj`	2.29	6.00
	3: `down_proj`	2.43	5.56





C.10 Comparison with Inter-Layer Mixed-Precision Quantization

Since KronQ allocates bit-widths across sublayers, we further compare against the inter-layer methods AMQ (Lee et al., 2025), Q-Palette (Lee & Song, 2026), and HIGGS (Malinovskii et al., 2025). All comparisons are conducted under a matched data-aware setting: every method uses the same real-text calibration set (128 WikiText-2 sequences of length 2048) that KronQ uses for $\mathbf{H_G}$, so the comparison isolates the allocation methodology from calibration-data access.

For the allocation comparisons, we hold the base quantizer fixed and vary only the allocation. Table 19 holds GPTQ fixed (LLaMA-2-7B, context 2048), where KronQ's allocation beats AMQ's data-aware NSGA-II search at a matched 3.1-bit budget. Table 20 holds Q-Palette's quantizer fixed (LLaMA-3.1-8B, context 8192) and compares against Q-Palette's data-aware actual-loss allocation. KronQ's allocation achieves lower perplexity at every budget. Table 21 instead isolates the quantizer: against the data-aware GPTQ+HIGGS configuration (LLaMA-2-7B, context 4096), KronQ's scalar grid leads at every bit-width despite HIGGS using a vector quantizer.

Crucially, KronQ's allocation is a closed-form analytic score that needs only the single $\mathbf{H_G}$ backward pass already computed for BiIP, with no dedicated search or per-bit precompute. It is therefore free within the KronQ pipeline, versus roughly 5 GPU-hours for AMQ's NSGA-II search and 79 GPU-hours for Q-Palette's data-aware actual-loss term. KronQ thus matches the data-aware allocation quality at roughly one twentieth of the cost.

Table 19: KronQ vs. AMQ allocation, GPTQ fixed (L2-7B, ctx 2048).

Alloc.	Bits	PPL
AMQ	3.1	6.68
KronQ	3.1	**6.52**

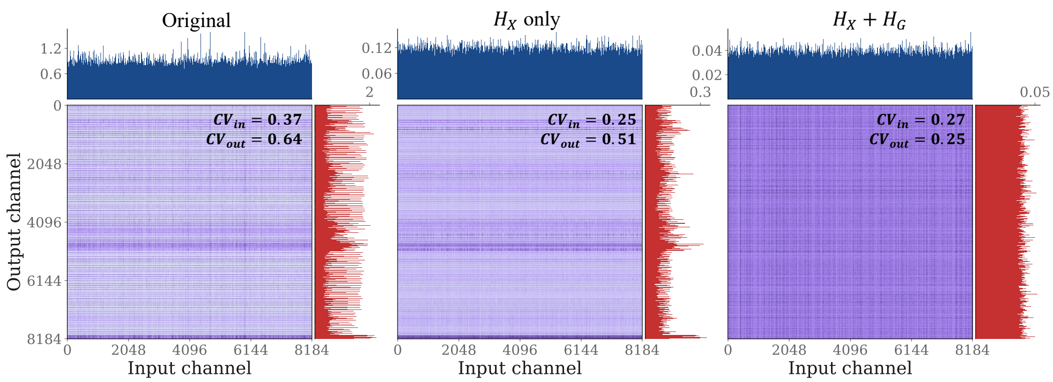
Table 20: KronQ vs. Q-Palette allocation, Q-Palette quantizer fixed (L3.1-8B, ctx 8192).

Bits	Q-Palette	KronQ
3.19	5.925	**5.896**
3.07	5.980	**5.948**
2.99	6.027	**5.989**

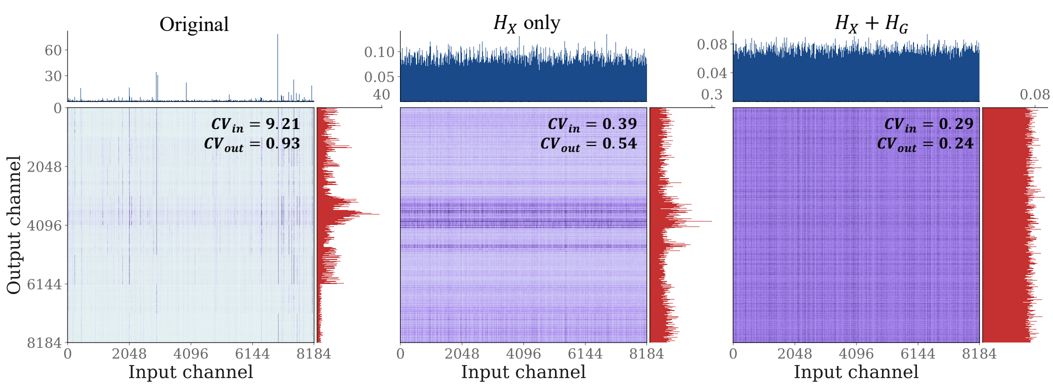
Table 21: KronQ scalar vs. HIGGS vector quantizer (L2-7B, ctx 4096).

Bits	HIGGS	KronQ
W4	5.213	**5.199**
W3	5.559	**5.444**
W2	8.637	**7.636**

D Why GPTQ and GPTAQ Fail on LLaMA-3-70B



(a) Weight distributions of LLaMA-2-70B



(b) Weight distributions of LLaMA-3-70B

Figure 6: Weight magnitude distribution of the Q projection in the first layer under three configurations: original weights, after column-side incoherence ($\mathbf{H}_X$ only), and after bidirectional incoherence ($\mathbf{H}_X + \mathbf{H}_G$). Top histograms show column norms; right histograms show row norms. $\mathrm{CV}_{\mathrm{in}}$ and $\mathrm{CV}_{\mathrm{out}}$ denote the coefficient of variation of column and row norms, respectively.

Figure 6 compares the weight magnitude distributions of LLaMA-3-70B and LLaMA-2-70B. In LLaMA-3-70B, the original weights exhibit extreme column-wise outliers ($\mathrm{CV}_{\mathrm{in}} = 9.21$) concentrated on specific input dimensions, a phenomenon previously identified by Qin (2024) as unique to the LLaMA-3/3.1-70B family. These outliers expand the quantization range by orders of magnitude, causing the column-wise OBS updates in GPTQ and GPTAQ to produce degenerate solutions. In contrast, LLaMA-2-70B shows a substantially milder distribution ($\mathrm{CV}_{\mathrm{in}} = 0.37$), consistent with its robustness to per-channel quantization. Applying $\mathbf{H}_X$-only incoherence suppresses the input-side outliers in LLaMA-3-70B ($\mathrm{CV}_{\mathrm{in}} : 9.21 \to 0.39$) but leaves the output-side heterogeneity largely unaddressed ($\mathrm{CV}_{\mathrm{out}} : 0.93 \to$

0.54). BiIP further resolves the output-side structure via $\mathbf{H}_G$, achieving $\mathrm{CV}_{\mathrm{in}} = 0.29$ and $\mathrm{CV}_{\mathrm{out}} = 0.24$, comparable to the post-BiIP values of LLaMA-2-70B (0.27 and 0.25).

E Per-Sublayer Calibration Memory

Table 22 details the memory required to perform calibration per sublayer for LLaMA-2-7B, LLaMA-3-8B, and LLaMA-2-13B, following the same convention as Li et al. (2025) (Table 9), with block slices of size $B = 128$ included in all reported values. The Cholesky factor $\mathbf{L}$ is stored as a lower-triangular matrix and the correction matrix $\mathbf{P}$ as an upper-triangular matrix. KRONQ additionally stores $\mathbf{H}_G$ as a full $d_{\mathrm{out}} \times d_{\mathrm{out}}$ matrix during BiIP; once released, the remaining memory footprint is identical to GPTAQ. The peak memory overhead is therefore transient and confined to the BiIP preprocessing step.

Table 22: Memory needed to perform calibration (GiB) per sublayer. $^{\dagger}\mathbf{H}_G$ released after BiIP, remaining memory matches GPTAQ.

Model	Method	q_proj	k_proj	v_proj	o_proj	up_proj	gate_proj	down_proj
LLaMA-2-7B	GPTQ	0.13	0.13	0.13	0.13	0.29	0.29	0.48
	GPTAQ	0.16	0.16	0.16	0.16	0.32	0.32	0.70
	KRONQ†	0.22→0.16	0.22→0.16	0.22→0.16	0.22→0.16	0.77→0.32	0.77→0.32	0.77→0.70
LLaMA-3-8B	GPTQ	0.13	0.13	0.13	0.13	0.37	0.37	0.71
	GPTAQ	0.16	0.16	0.16	0.16	0.40	0.40	1.10
	KRONQ†	0.22→0.16	0.22→0.16	0.22→0.16	0.22→0.16	1.17→0.40	1.17→0.40	1.16→1.10
LLaMA-2-13B	GPTQ	0.20	0.20	0.20	0.20	0.45	0.45	0.76
	GPTAQ	0.25	0.25	0.25	0.25	0.50	0.50	1.11
	KRONQ†	0.35→0.25	0.35→0.25	0.35→0.25	0.35→0.25	1.22→0.50	1.22→0.50	1.21→1.11

F Limitations

KRONQ requires a backward pass over the calibration set to estimate the gradient covariance $\mathbf{H}_G$ prior to quantization, introducing additional offline computation relative to activation-only methods such as GPTQ and GPTAQ. Furthermore, $\mathbf{H}_G$ is a full $d_{\mathrm{out}} \times d_{\mathrm{out}}$ matrix, incurring additional peak memory during BiIP preprocessing. Nevertheless, once $\mathbf{H}_G$ is precomputed and stored, it incurs no additional computation during quantization itself, as $\mathbf{H}_G$ cancels algebraically in the column-wise updates (Proposition 1) and is released upon completion of BiIP. During inference, KRONQ requires online reversion of the orthogonal transforms $\mathbf{U}$ and $\mathbf{V}$, introducing $\Theta(d_{\mathrm{in}} \log d_{\mathrm{in}} + d_{\mathrm{out}} \log d_{\mathrm{out}})$ overhead per layer. This is identical to QuIP# (Tseng et al., 2024) and our fused CUDA kernel keeps the per-layer cost close to that of the $\Theta(d_{\mathrm{in}} d_{\mathrm{out}})$ weight matrix-vector multiply alone.