OFF-DIAGONAL ENERGY SENSITIVITY FOR MIXED-PRECISION POST-TRAINING QUANTIZATION OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed-precision post-training quantization (MP-PTQ) is a promising approach for improving resource efficiency under constrained computational and memory budgets through effective bit allocation. MP-PTQ methods based on independently optimized layer-wise or module-wise surrogate objectives are computationally efficient, but primarily rely on local quantization errors and therefore do not explicitly account for cross-layer interactions under the task loss. In contrast, search-based methods treat mixed-precision allocation as a black-box optimization problem, leading to costly exploration of a large search space. To address these limitations, we propose Off-Diagonal Energy Sensitivity (ODES), which captures interactions within and across projections through the off-diagonal structure of the task-loss Hessian. Starting from a second-order expansion of the task loss, we separate its off-diagonal interaction term. Then, we derive a sufficient condition under which ODES-guided precision reallocation reduces the second moment of the off-diagonal task-loss contribution. To estimate ODES at LLM scale without materializing the Hessian, we develop a compressed probe-centered (CPC) estimator. CPC reduced estimator state memory by approximately $126\times$ while producing ODES estimates that closely matched those obtained with the direct Welford estimator. Using the estimated ODES ranking, we reduce mixed-precision bit allocation to a one-dimensional search over the number of reallocated projection pairs and evaluate these candidates using HQQ as a lightweight quantization proxy. This reduced the total allocation time by 36–50% relative to search-based methods while maintaining comparable performance in downstream accuracy and perplexity. ODES-guided allocations also improved over uniform allocation with RTN, HQQ, and GPTQ, indicating that the ODES ranking transfers across different quantizers.

1 INTRODUCTION

With the rapid advancement of large language models (LLMs), there is growing demand for efficient model deployment on devices with limited memory and computational resources (Dettmers et al., 2022; Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023; Yao et al., 2022; Zhu et al., 2024). Mixed-precision post-training quantization (MP-PTQ) enables more efficient quantization by assigning bit-widths adaptively to projections with different sensitivities (Huang et al., 2025; Lee et al., 2024; Shen et al., 2020). Existing fine-grained allocation methods typically rely on sensitivity proxies such as activation statistics or Hessian-based measures derived from local reconstruction objectives (Chen et al., 2024; Huang et al., 2025; Zhang et al., 2025). These sensitivities do not explicitly capture interactions across layers under the task loss. In contrast, search-based MP-PTQ methods such as AMQ and EvoPress employ evolutionary search to optimize mixed-precision bit-width configurations (Lee et al., 2025; Sieberling et al., 2025). Configuration-level methods can reflect component interactions through configuration-level candidate evaluations, but treat these evaluations as black-box feedback.

We introduce Off-Diagonal Energy Sensitivity (ODES), a projection-level sensitivity measure that summarizes interactions between quantization perturbations and provides a ranking for precision

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.

allocation. ODES aims to suppress interaction-induced loss fluctuations by guiding precision reallocation toward reducing the second moment of the off-diagonal loss contribution.

Our contributions are summarized as follows: **(1)** We propose ODES, a projection-level sensitivity measure and derive a sufficient condition to guide the extent of precision reallocation. **(2)** We develop a matrix-free compressed probe-centered (CPC) estimator that substantially reduces estimator state memory while achieving high consistency with direct Welford projection rankings at LLM scale. **(3)** We reduce ODES-guided mixed-precision allocation to a one-dimensional search over the number of reallocated projection pairs, reducing total allocation time by 36% and 50% relative to AMQ and EvoPress, respectively **(4)** We empirically observe that the theory-guided boundary K_{th} bounds the effective reallocation range, and that ODES-guided reallocation reduces window-level loss variation relative to a diagonal-curvature baseline.

2 Related Work

Bit allocation. Uniform-precision PTQ methods for LLMs, such as GPTQ, AWQ, OmniQuant, and LeanQuant (Frantar et al., 2023; Lin et al., 2024; Shao et al., 2024; Zhang & Shrivastava, 2025), and extreme low-bit methods such as QuIP, AQLM, and VPTQ (Chee et al., 2023; Egiazarian et al., 2024; Liu et al., 2024), assign the same bit-width to every projection; mixed-precision allocation is complementary to these quantizers. Mixed-precision allocation methods differ in their allocation granularity and the information used to assess sensitivity. Fine-grained methods use activation statistics or reconstruction-based sensitivity measures (Chen et al., 2024; Dettmers et al., 2024; Huang et al., 2025; Kim et al., 2024; Lee et al., 2024; Zhang et al., 2025), while configuration-level methods evaluate combinations of bit assignments (Lee et al., 2025; Sieberling et al., 2025; Wang et al., 2019; Yao et al., 2021a). AMQ performs projection-wise bit allocation using sensitivity-based search space pruning, a learned quality predictor, and iterative multi-objective search based on NSGA-II (Deb et al., 2002; Lee et al., 2025). To reduce search cost, AMQ uses HQQ (Badri & Shaji, 2023) as a proxy and transfers the resulting bit-width configurations to PTQ methods such as GPTQ and AWQ (Frantar et al., 2023; Lin et al., 2024). EvoPress formulates non-uniform compression as an evolutionary optimization problem, motivated by the observation that compression effects across projections are not necessarily independent (Sieberling et al., 2025). Recent work has also coupled mixed-precision allocation more tightly with the underlying PTQ procedure. APTQ extends GPTQ-style reconstruction to attention outputs and uses the trace of an approximate reconstruction Hessian to guide mixed-precision allocation (Guan et al., 2024).

Hessian-Aware Weight Quantization Sensitivity. Second-order sensitivity analysis of weight perturbations dates back to Optimal Brain Damage and Optimal Brain Surgeon (Hassibi & Stork, 1993; LeCun et al., 1990), and underlies modern layer-wise PTQ and pruning (Frantar & Alistarh, 2022; Li et al., 2021; Nagel et al., 2020). HAWQ and HAWQ-V2 estimate layer-wise quantization sensitivity in CNNs from the diagonal Hessian blocks, using the dominant eigenvalue and the average Hessian trace, respectively (Dong et al., 2019; 2020), and Q-BERT applies the same idea to Transformer layers (Shen et al., 2020). These layer-wise sensitivities do not explicitly model interactions represented by cross-layer Hessian blocks, although such interactions have been observed to shape the quantization loss landscape (Nahshan et al., 2021). CLADO addresses this limitation by estimating cross-layer sensitivity from forward loss differences under candidate quantization perturbations (Deng et al., 2025). For HAWQ-V2 and CLADO, however, mixed-precision bit allocation depends on candidate-specific quantization perturbations. Candidate-dependent error terms complicate sensitivity reuse when quantization outcomes also depend on the calibration context and surrounding bit assignments. ODES summarizes squared off-diagonal entries of the task-loss Hessian at the pretrained weights, including entries within and across projections, without requiring candidate-specific quantized weights for score computation.

3 Off-Diagonal Energy Sensitivity for Bit Allocation

This section derives ODES from the second moment of the off-diagonal interaction term of the task loss (Sections 3.1–3.2) and presents matrix-free estimators for it (Section 3.3).

3.1 PRELIMINARIES

Let $\mathcal{D} = \{x_d\}_{d=1}^{D}$ be a dataset of token windows for an LLM. The vector $\mathbf{w}^\star \in \mathbb{R}^P$ contains the pretrained weights of the M linear projections selected for quantization. Let $\mathbf{w} = Q(\mathbf{w}^\star)$ denote the quantized weights and $\boldsymbol{\delta} = \mathbf{w} - \mathbf{w}^\star$ the weight perturbations. For the task loss $L(\mathbf{w}; x_d)$, given by the window-level negative log-likelihood (NLL), define $\mathbf{H}_d = \nabla^2_{\mathbf{w}} L(\mathbf{w}^\star; x_d)$, with all other model parameters fixed. With the expectation taken over a uniformly sampled window from $\mathcal{D}$, the second-order approximation of the task loss gives (LeCun et al., 1990; Nagel et al., 2020)

$$\mathbb{E}_d\big[L(\mathbf{w}^\star + \boldsymbol{\delta}; x_d) - L(\mathbf{w}^\star; x_d)\big] \approx \frac{1}{D}\sum_{d=1}^{D}\Big[\underbrace{\tfrac{1}{2}\sum_{i=1}^{P} H_{d,ii}\delta_i^2}_{D_d(\boldsymbol{\delta})} + \underbrace{\sum_{1\le i<j\le P} H_{d,ij}\delta_i\delta_j}_{C_d(\boldsymbol{\delta})}\Big]. \tag{1}$$

The pretrained weights $\mathbf{w}^\star$ are assumed to locally minimize the mean task loss, so the first-order term vanishes (Dong et al., 2020). Higher-order terms are neglected for sufficiently small perturbations (Nagel et al., 2020). The terms $D_d(\boldsymbol{\delta})$ and $C_d(\boldsymbol{\delta})$ denote the diagonal and off-diagonal contributions, respectively.

3.2 OFF-DIAGONAL INTERACTION

We assume that the quantization perturbations are independent and zero-mean, with $\mathbb{E}_\delta[\delta_i] = 0$ and $\mathbb{E}_\delta[\delta_i^2] = \sigma_i^2$, following the standard additive quantization-noise model (Sripad & Snyder, 1977; Widrow & Kollár, 2008). Let $\boldsymbol{\sigma}^2 := (\sigma_1^2, \ldots, \sigma_P^2)^\top$ denote the variance vector of the quantization perturbations. Hence, for $i \neq j$, $\mathbb{E}_\delta[\delta_i\delta_j] = 0$, and therefore $\mathbb{E}_\delta[C_d(\boldsymbol{\delta})] = 0$. This cancellation holds only in expectation, while $C_d(\boldsymbol{\delta})$ can be nonzero for an individual perturbation. We therefore consider its second moment,

$$J_{\text{off}}(\boldsymbol{\sigma}^2) := \frac{1}{D}\sum_{d=1}^{D}\mathbb{E}_\delta\left[C_d(\boldsymbol{\delta})^2\right] = \frac{1}{D}\sum_{d=1}^{D}\sum_{i<j} H_{d,ij}^2\sigma_i^2\sigma_j^2. \tag{2}$$

Because a deployed quantized model corresponds to a single realization of $\boldsymbol{\delta}$, the relevant quantity is how far C_d can deviate from its zero mean: for a uniformly sampled window, Chebyshev's inequality gives $\Pr(|C_d(\boldsymbol{\delta})| \ge t) \le J_{\text{off}}(\boldsymbol{\sigma}^2)/t^2$, so J_{off} controls the interaction-induced loss shift that a particular quantized model can incur. Appendix A.2 separates the diagonal and off-diagonal quadratic fluctuations and examines the diagonal contribution to the reduction bound. We next consider how $J_{\text{off}}(\boldsymbol{\sigma}^2)$ changes under precision reallocation. Let $\mathcal{I}_+$ and $\mathcal{I}_-$ denote the parameter sets whose precision is increased and decreased by one bit, respectively, with all remaining parameters unchanged. We assume a common baseline variance $\sigma_i^2 = \sigma^2$ and the idealized one-bit response $\delta_i' = \delta_i/2$ for $i \in \mathcal{I}_+$ and $\delta_i' = 2\delta_i$ for $i \in \mathcal{I}_-$. Let $\boldsymbol{\sigma}'^2$ denote the variance vector corresponding to $\boldsymbol{\delta}'$. For $s \in \{+, -\}$, define $R_s := \frac{1}{D}\sum_{d=1}^{D}\sum_{i\in\mathcal{I}_s}\sum_{j\neq i} H_{d,ij}^2$. Under these assumptions (Appendix A.1),

$$J_{\text{off}}(\boldsymbol{\sigma}^2) - J_{\text{off}}(\boldsymbol{\sigma}'^2) \ge \frac{15\sigma^4}{32}\left(R_+ - 16R_-\right). \tag{3}$$

Therefore, $R_+ > 16R_-$ is sufficient for $J_{\text{off}}(\boldsymbol{\sigma}'^2) < J_{\text{off}}(\boldsymbol{\sigma}^2)$: transferring one bit of precision from $\mathcal{I}_-$ to $\mathcal{I}_+$ reduces the interaction second moment. The factor 16 reflects an asymmetry: a one-bit demotion multiplies the perturbation variance of each affected coordinate by 4, so an interaction between two demoted coordinates grows sixteenfold, whereas a promotion can remove at most the original contribution. Precision should therefore move toward projections that are strongly coupled to the rest of the network and away from those that are weakly coupled. This condition motivates a projection-level sensitivity based on the task-loss Hessian. For projection m with parameter set $\mathcal{I}_m$ and $p_m = |\mathcal{I}_m|$, we define its off-diagonal energy at window d and its per-parameter sensitivity as

$$E_{m,d}^{\text{off}} := \sum_{i\in\mathcal{I}_m}\sum_{j\neq i} H_{d,ij}^2, \qquad S_m := \frac{1}{p_m}\mathbb{E}_d\left[E_{m,d}^{\text{off}}\right]. \tag{4}$$

Although the sufficient condition depends on the total incident energy $p_m S_m$, we rank projections by the normalized quantity S_m to compare them on a per-parameter basis and avoid favoring larger

projections solely due to parameter count (Appendix A.3). Let m_k^+ and m_k^- denote the k-th highest- and lowest-ranked projections, respectively, and consider each pair separately, with $\mathcal{I}_+ = \mathcal{I}_{m_k^+}$ and $\mathcal{I}_- = \mathcal{I}_{m_k^-}$. The sufficient condition for a one-bit precision exchange between this pair is

$$\frac{1}{D}\sum_{d=1}^{D} E^{\text{off}}_{m_k^+,d} > \frac{16}{D}\sum_{d=1}^{D} E^{\text{off}}_{m_k^-,d}, \tag{5}$$

which Section 4.2.1 examines empirically.

3.3 Matrix-Free Profiling

Since materializing the full $P \times P$ Hessian is infeasible at LLM scale, we introduce three matrix-free estimators of the fixed-window sensitivity $s_m := E^{\text{off}}_{m,d}/p_m$ using Hessian–vector products (HVPs) (Pearlmutter, 1994; Yao et al., 2020). We write $H := \mathbf{H}_d$ and let $N \geq 2$ denote the number of probes per window.

Hutchinson–Welford estimation. For independent Rademacher probes $z_r \in \{-1, +1\}^P$, we use the stochastic probing construction $y_r = z_r \odot Hz_r$ (Avron & Toledo, 2011; Bekas et al., 2007; Hutchinson, 1990; Yao et al., 2021b). Since $\mathrm{Var}_z(y_{r,i}) = \sum_{j \neq i} H_{ij}^2$, the coordinate-wise sample variance gives the unbiased estimator

$$\widehat{s}_m^{\text{W}} = \frac{1}{p_m(N-1)} \sum_{i \in \mathcal{I}_m} \sum_{r=1}^{N} (y_{r,i} - \bar{y}_i)^2, \qquad \bar{y}_i = \frac{1}{N}\sum_{r=1}^{N} y_{r,i}. \tag{6}$$

We compute this statistic online using Welford updates (Chan et al., 1983; Welford, 1962).

Scalar covariance estimation. To avoid parameter-wise running statistics, define the projection average $a_{r,m} = p_m^{-1} \sum_{i \in \mathcal{I}_m} y_{r,i}$ and the global quadratic form $q_r = z_r^\top H z_r$. Then

$$s_m = \frac{1}{2}\,\mathrm{Cov}_z(a_{r,m}, q_r), \qquad \widehat{s}_m^{\text{C}} = \frac{1}{2}\widehat{\mathrm{Cov}}(a_{r,m}, q_r). \tag{7}$$

Here, $\widehat{\mathrm{Cov}}$ denotes the sample covariance over the N probes.

Compressed probe-centered estimation. The memory cost of direct Welford estimation arises from maintaining parameter-wise running means. CPC decomposes the Hutchinson–Welford statistic into the projection mean and centered probe responses, and replaces the squared norm of the mean centered response by a random sketch, so that no parameter-wise means are stored. Specifically,

$$\widehat{s}_m^{\text{W}} = V_{a,m} + \frac{N}{N-1}\left(\bar{B}_m - \frac{1}{p_m}\sum_{i \in \mathcal{I}_m} \bar{u}_{m,i}^2\right), \qquad \widehat{s}_m^{\text{CPC}} = V_{a,m} + \frac{N}{N-1}\left(\bar{B}_m - \frac{\|g_m\|_2^2}{p_m}\right). \tag{8}$$

Here $u_{r,m,i} = y_{r,i} - a_{r,m}$ are the probe-centered responses, $\bar{a}_m = N^{-1}\sum_r a_{r,m}$, $V_{a,m} = (N-1)^{-1}\sum_r (a_{r,m} - \bar{a}_m)^2$, $\bar{B}_m = (Np_m)^{-1}\sum_r \sum_{i \in \mathcal{I}_m} u_{r,m,i}^2$, and $\bar{u}_{m,i} = N^{-1}\sum_r u_{r,m,i}$. The sketch vector is $g_m = R_m \bar{u}_m$, where $\bar{u}_m = (\bar{u}_{m,i})_{i \in \mathcal{I}_m}$ and $R_m \in \mathbb{R}^{k_m \times p_m}$, $k_m \ll p_m$, is a fixed randomized sketch such that $\|R_m \bar{u}_m\|_2^2 \approx \|\bar{u}_m\|_2^2$ (Achlioptas, 2003; Johnson & Lindenstrauss, 1984; Woodruff, 2014).

Restoring the window index, let $\widehat{s}^X_{m,d}$ estimate $s_{m,d} = E^{\text{off}}_{m,d}/p_m$ for $X \in \{\text{W}, \text{C}, \text{CPC}\}$. Averaging over the D windows as in Eq. 4 gives $\widehat{S}^X_m = D^{-1}\sum_{d=1}^{D} \widehat{s}^X_{m,d}$, which we use to rank projections. Estimator derivations and implementation details are provided in Appendix B.

4 Experimental Results

Experimental setup. We evaluated models from the Llama-2 (Touvron et al., 2023), Llama-3.1/3.2 (Grattafiori et al., 2024), and Qwen3 (Yang et al., 2025) families. We used C4 (Raffel et al., 2020) for ODES profiling. The C4 samples used for perplexity evaluation were disjoint from those used for ODES profiling, quantizer calibration, and configuration search. Profiling used $D = 64$

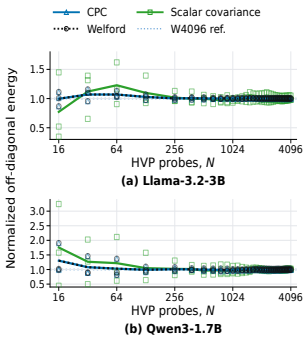


Figure 1: Convergence of the off-diagonal energy estimates on (a) Llama-3.2-3B and (b) Qwen3-1.7B, averaged over all projections.

Table 1: Fidelity was evaluated against independent Welford reference runs, each computed with $N = 4{,}096$ probes. Metrics were averaged over two reference comparisons; CPC metrics were additionally averaged over three sketch seeds. ρ denotes Spearman rank correlation, and Top/Bottom 10% denote percentage overlaps with the corresponding sets. State denotes estimator state memory, and Time denotes wall-clock time per HVP probe.

Method	Rel. L_1 (%) ↓	Rel. L_2 (%) ↓	ρ ↑	Top 10% ↑	Bot. 10% ↑	State (MiB) ↓	Time ↓ (s/probe)
Llama-3.2-1B							
Welford	2.75	3.75	0.999	100.00	100.00	1,856.081	0.816
Scalar cov.	74.02	66.63	0.534	77.78	5.56	0.050	0.822
CPC	2.75	3.76	0.999	100.00	100.00	14.818	0.832
Llama-3.2-3B							
Welford	6.99	8.35	0.999	98.33	100.00	5,376.122	1.978
Scalar cov.	116.57	76.98	0.398	63.33	1.67	0.076	2.014
CPC	6.99	8.35	0.999	98.33	100.00	42.478	2.043
Qwen3-1.7B							
Welford	21.50	23.01	0.999	96.67	98.33	2,688.063	1.179
Scalar cov.	150.54	70.57	0.284	65.00	0.00	0.041	1.171
CPC	21.49	23.01	0.999	96.67	98.33	21.243	1.182

1,024-token windows. We evaluated perplexity (PPL) on C4, WikiText-2 (Merity et al., 2017), and Penn Treebank (PTB) (Marcus et al., 1993). Downstream performance is reported as the average accuracy over BoolQ (Clark et al., 2019), LAMBADA (Paperno et al., 2016), ARC-Easy and ARC-Challenge (Clark et al., 2018), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), and MMLU (Hendrycks et al., 2021), evaluated with the Language Model Evaluation Harness (Gao et al., 2024).

4.1 Matrix-Free Profiling

We evaluated CPC against direct Welford and scalar covariance on C4, with profiling performed on a single NVIDIA H100 for Llama-3.2-1B, Llama-3.2-3B, and Qwen3-1.7B (settings in Appendix B.4).

Probe-budget evaluation. Figure 1 compares the three estimators over HVP probe budgets from $N = 16$ to $4{,}096$ on Llama-3.2-3B and Qwen3-1.7B. The values were averaged over projections and normalized with respect to the high-budget Welford reference at $N = 4{,}096$. CPC and direct Welford used the same Rademacher probe stream at each budget. Across both models, CPC closely tracked direct Welford over the entire range. In contrast, scalar covariance showed substantially larger deviations at small N, particularly on Qwen3-1.7B, although the discrepancy decreased as N increased. **Fidelity and cost at $N = 64$.** At $N = 64$ (Table 1), the relative L_1 and L_2 errors of CPC and direct Welford differed by at most 0.01 percentage points, and CPC achieved a Spearman correlation of 0.999, with top- and bottom-10% overlaps of 96.67–100% and 98.33–100%, respectively. In contrast, scalar covariance exhibited substantially larger relative errors and lower rank agreement at the same probe budget. Direct Welford required $1{,}856.1$–$5{,}376.1$ MiB of estimator state, whereas CPC required 14.8–42.5 MiB, corresponding to approximately a 126-fold reduction, while maintaining similar measured time per probe. The direct Welford state is essentially an additional FP16 copy of all quantized projection weights, which CPC replaces with block sketches. Since CPC at $N = 64$ already preserved the high-budget Welford ranking, we fixed $N = 64$ for all subsequent profiling.

4.2 ODES-Guided Mixed-Precision Allocation

We next evaluated ODES as a criterion for projection-wise mixed-precision allocation, with scores estimated by CPC at $N = 64$. For each profiling window d, CPC produces a projection-wise score vector $\widehat{\mathbf{s}}_d = (\widehat{s}_{1,d}, \ldots, \widehat{s}_{M,d})$. We ranked the 64 windows by $\|\widehat{\mathbf{s}}_d\|_2$, excluded the eight largest-norm windows (highlighted in orange in Figure 2), and averaged the remaining 56 window-wise scores to obtain the final ODES score for each projection. Let K denote the number of ranked pairs

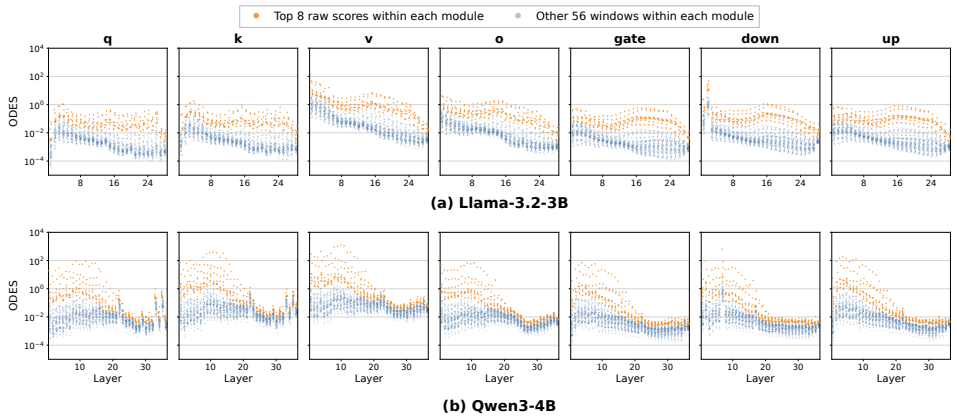


Figure 2: ODES scores are shown for all linear projections in (a) Llama-3.2-3B and (b) Qwen3-4B. The eight windows highlighted in orange have the largest score vector ℓ_2 norms among the 64 profiling windows.

included in the reallocation. Starting from a uniform $b_0 = 3$-bit allocation, we assigned $b_0 + 1$ bits to the K highest-ranked projections $m_1^+, \ldots, m_K^+$ and $b_0 - 1$ bits to the K lowest-ranked projections $m_1^-, \ldots, m_K^-$, with all remaining projections at b_0 bits.

Once the ranking is fixed, the allocation is determined solely by K, so we searched along the one-dimensional sequence of configurations induced by the ODES ranking with $K \in \{1, \ldots, 20\}$. Each candidate was quantized with HQQ (Badri & Shaji, 2023) as a lightweight proxy and scored by its perplexity on C4 windows sampled with a different random seed from the evaluation set, averaged over three seeds; we selected the candidate with the lowest proxy perplexity and denote it by K_{HQQ}. This configuration was then applied to the target PTQ method without further search. Because the ranking depends only on the pretrained weights, the CPC profile is computed once per model and reused for every quantizer and every candidate K; the remaining cost is the evaluation of mutually independent candidates. For reference, we also report the configuration at the theory-guided boundary K_{th} (Section 4.2.1). All configurations were compared under approximately matched parameter-weighted bit budgets (Appendix C.1).

4.2.1 Theory-Guided Allocation Boundary

We next examined whether the ODES-guided allocation exhibited the sensitivity behavior predicted by our analysis. We swept the reallocation extent K and compared ODES with MeanDiag, a diagonal-curvature baseline motivated by HAWQ-V2 (Dong et al., 2020; Yao et al., 2020) that ranks projections by the mean diagonal Hessian entry and applies the same reallocation procedure. Figure 3 summarizes the results across the quantizers considered in the subsequent experiments, with each quantizer evaluated separately and aggregated only at the final stage.

The first row compares $E^{\mathrm{off}}_{m_k^+}$ with $16E^{\mathrm{off}}_{m_k^-}$. We define K_{th} as the first value of K for which $16E^{\mathrm{off}}_{m_k^-} > E^{\mathrm{off}}_{m_k^+}$. The remaining rows show how downstream accuracy, window-NLL variation, and perplexity change around this boundary, where the normalized window-NLL variance is

$$\bar{V}_q(K) = \frac{1}{|\mathcal{S}_q(K)|} \sum_{s \in \mathcal{S}_q(K)} \mathrm{Var}_d\left[\ell_{q,s,d}(K)\right], \qquad \widetilde{V}(K) = \frac{1}{|\mathcal{Q}_K|} \sum_{q \in \mathcal{Q}_K} \frac{\bar{V}_q(K)}{\bar{V}_q(0)}, \tag{9}$$

where $\ell_{q,s,d}(K)$ is the mean target-token NLL of evaluation window d under quantizer q and seed s, $\mathcal{S}_q(K)$ denotes the set of seeds available for quantizer q at K, and $\mathcal{Q}_K$ denotes the set of quantizers included at K. In all four models, and in the four additional models of Figure C.1, the downstream-accuracy optimum did not exceed K_{th}. Moreover, across three models except Qwen3-8B, ODES showed a decreasing trend in normalized window-NLL variance as K increased, and it remained generally lower than MeanDiag. For the Llama models with larger K_{th}, the reduction in window-NLL variance persisted over a broader range of K. In contrast, for the Qwen models with smaller

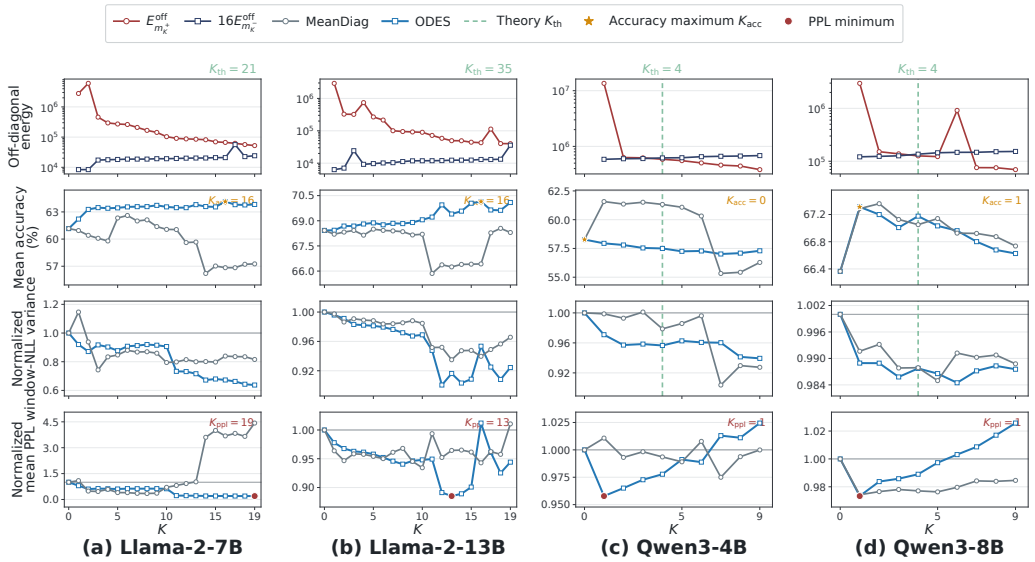
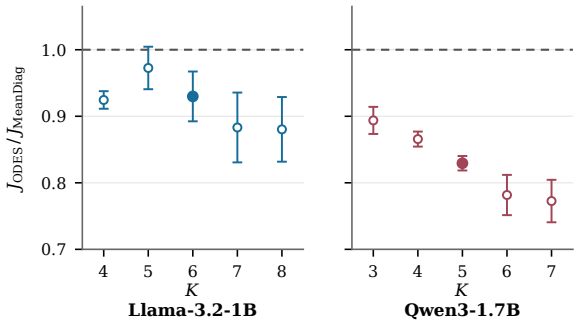


Figure 3: The first row compares the off-diagonal energies of the paired high- and low-ODES projections, with the green dashed line marking the boundary of the sufficient condition. The remaining rows show downstream accuracy, normalized window-NLL variance, and normalized mean PPL for ODES and MeanDiag. The normalized metrics were computed relative to the corresponding uniform allocation at $K = 0$. The star and circle mark the accuracy maximum (K_{acc}) and the PPL minimum (K_{ppl}), respectively. Results are shown for (a) Llama-2-7B, (b) Llama-2-13B, (c) Qwen3-4B, and (d) Qwen3-8B.

K_{th}, the effect was confined to a much narrower range and could diminish beyond the boundary. Taken together, these trends suggest that K_{th} can serve as a useful signal for identifying the effective range of precision reallocation before its benefit begins to diminish.



(a) Controlled fluctuation across neighboring K.

(b) Theory-motivated anchor points

Model	K	$\frac{J_{\mathrm{ODES}}}{J_{\mathrm{MeanDiag}}}$	Reduction
Llama-3.2-1B	6	0.9298 ± 0.0374	7.02%
Qwen3-1.7B	5	0.8294 ± 0.0109	17.06%

Mean $\pm$ SD across 3 probe seeds.

Figure 4: **Controlled assessment of off-diagonal fluctuation reduction.** **(a)** Ratio $J_{\mathrm{ODES}}/J_{\mathrm{MeanDiag}}$ across two theory-motivated local ranges of K. For Llama-3.2-1B, $K = 4$–8 spans the theory-guided boundary $K_{\mathrm{th}} = 5$; for Qwen3-1.7B, $K = 3$–7 lies within its boundary $K_{\mathrm{th}} = 12$. Error bars denote SD across three independent probe seeds, and the dashed horizontal line indicates equal fluctuation. **(b)** Representative anchor points from the two regimes.

Controlled assessment of off-diagonal fluctuation reduction. Because practical PTQ perturbations affect both the diagonal and off-diagonal quadratic terms, window-NLL variability alone cannot isolate the off-diagonal fluctuation targeted by ODES. We therefore considered a controlled perturbation model in which the perturbation magnitudes were held fixed and only their signs were resampled independently: $\delta_i = a_i s_i$, where $s_i \in \{-1, +1\}$. For a fixed window d, the diagonal

quadratic term is invariant to the signs, yielding

$$J_d = \mathrm{Var}_s\left[\frac{1}{2}\boldsymbol{\delta}^\top H_d \boldsymbol{\delta}\right] = \mathrm{Var}_s\left[\sum_{i<j} H_{d,ij} a_i a_j s_i s_j\right],$$

which isolates the sign-induced off-diagonal quadratic fluctuation. Using eight fixed C4 windows, 64 sign probes per window, and three independent probe-seed replicas with paired inputs and signs across methods, ODES yielded lower mean J than MeanDiag throughout both tested local ranges, with reductions of 7.02% (Llama-3.2-1B) and 17.06% (Qwen3-1.7B) at the anchor points in Fig. 4(b).

4.3 Comparison with Search-Based Allocation

As shown in Table 2, under the same 3-bit GPTQ (Frantar et al., 2023) setting, the HQQ-selected ODES configurations improved downstream accuracy over uniform allocation on all eight models and reduced average perplexity on five of them, with no case where both metrics degraded simultaneously. Compared with AMQ and EvoPress, ODES (K_{HQQ}) achieved higher accuracy in 10 of 16 pairwise comparisons, while its average perplexity lay between those of the two search-based methods on four models and was better than both on the two Llama-2 models. This performance was achieved at substantially lower allocation cost. As shown in Table 3, the total runtime summed over the seven evaluated models was 1,323.23 minutes for ODES, compared with 2,054.24 minutes for AMQ and 2,647.00 minutes for EvoPress, corresponding to reductions of approximately 36% and 50%, respectively. The per-model reduction ranged from 25% to 56% relative to AMQ and from 26% to 78% relative to EvoPress, with the largest gains on the smaller models.

Table 2: Comparison of mixed-precision allocation methods with GPTQ under matched weighted-average bit budgets (within ± 0.01 bits).

	Allocation	Wiki2 ↓	C4 ↓	PTB ↓	Avg. PPL ↓	Avg. Acc. ↑
Llama 2 7B	Uniform	7.00	8.59	208.36	74.65	60.07
	AMQ	6.62	8.04	270.78	95.15	62.22
	EvoPress	6.54	8.65	472.90	162.70	61.10
	ODES (k_th=21)	9.31	10.40	127.34	49.02	60.86
	ODES (k_hqq=2)	7.11	8.72	146.65	54.16	61.89
Llama 3.2 1B	Uniform	27.38	27.39	64.10	39.62	46.97
	AMQ	17.60	20.47	28.17	22.08	47.68
	EvoPress	24.57	30.91	38.96	31.48	43.42
	ODES (k_th=5)	29.63	29.75	60.64	40.01	46.50
	ODES (k_hqq=2)	25.07	26.10	49.98	33.71	48.17
Llama 3.1 8B	Uniform	10.67	13.35	16.32	13.44	57.77
	AMQ	8.76	11.70	13.81	11.42	66.54
	EvoPress	10.60	14.75	16.67	14.01	61.87
	ODES (k_th=9)	11.10	13.62	16.09	13.61	64.84
	ODES (k_hqq=5)	11.04	13.58	16.25	13.62	65.99
Qwen3 4B	Uniform	19.05	20.45	34.00	24.50	58.36
	AMQ	16.51	18.92	29.63	21.69	59.67
	EvoPress	16.21	22.68	29.81	22.90	60.51
	ODES (k_th=4)	19.65	21.29	34.89	25.28	58.03
	ODES (k_hqq=1)	19.43	20.78	34.00	24.74	58.73

	Allocation	Wiki2 ↓	C4 ↓	PTB ↓	Avg. PPL ↓	Avg. Acc. ↑
Llama 2 13B	Uniform	5.83	7.42	40.05	17.77	67.78
	AMQ	5.61	7.14	62.90	25.22	68.21
	EvoPress	5.63	7.66	62.55	25.28	65.11
	ODES (k_th=35)	5.92	7.49	40.13	17.85	69.46
	ODES (k_hqq=8)	5.79	7.41	37.65	16.95	69.41
Llama 3.2 3B	Uniform	25.12	18.30	32.52	25.31	52.66
	AMQ	11.33	14.06	18.74	14.71	58.77
	EvoPress	19.70	21.98	51.24	30.97	44.59
	ODES (k_th=6)	15.87	17.69	24.71	19.42	51.79
	ODES (k_hqq=3)	15.57	17.49	24.69	19.25	53.61
Qwen3 1.7B	Uniform	30.63	30.03	62.52	41.06	44.59
	AMQ	28.28	26.28	51.63	35.40	48.87
	EvoPress	31.09	38.78	59.33	43.07	43.22
	ODES (k_th=12)	37.89	35.25	65.09	46.08	44.68
	ODES (k_hqq=4)	31.78	30.55	56.98	39.77	45.23
Qwen3 8B	Uniform	12.38	15.32	20.71	16.14	69.08
	AMQ	11.92	14.88	20.02	15.61	66.35
	EvoPress	11.85	17.80	20.18	16.61	65.80
	ODES (k_th=4)	12.75	15.57	21.04	16.46	68.91
	ODES (k_hqq=1)	12.55	15.37	20.56	16.16	69.39

Perplexity showed a different pattern. AMQ consistently improved C4 perplexity, which is expected because C4 was directly used during its configuration search, but these gains did not always transfer to other evaluation corpora. In contrast, ODES exhibited more consistent behavior across C4 and PTB on the Llama-2 models. This observation is consistent with prior findings that low-bit quantization can be sensitive to the calibration data distribution, and that matching calibration and evaluation distributions can improve perplexity on the corresponding corpus (Chen et al., 2025; Gong et al., 2024; Williams & Aletras, 2024). These results support evaluating projection sensitivity using a criterion that is less directly tied to a finite calibration set or a particular search objective.

Table 3: Total runtime comparison across models. Runtime is reported in minutes. For ODES, runtime includes CPC profiling over 64 C4 windows with $N = 64$ probes and HQQ evaluation of the 20 candidate values $K \in \{1, \ldots, 20\}$; for AMQ and EvoPress, it is the full configuration search. Llama-2-13B is omitted because its CPC profiling was run on two H100 GPUs, so its runtime is not directly comparable with the single-GPU measurements of the other models.

Method	Llama-3.2-1B	Qwen3-1.7B	Llama-3.2-3B	Qwen3-4B	Llama-2-7B	Llama-3.1-8B	Qwen3-8B
EvoPress	278.00	326.00	314.00	333.00	343.00	504.00	549.00
AMQ (NSGA-II)	141.136	174.224	228.096	309.464	396.336	389.184	415.800
ODES (CPC+20 HQQ)	**62.28**	**90.32**	**171.94**	**193.46**	**255.09**	**263.68**	**286.46**

Table 4: Model-averaged comparison of Uniform and ODES (K_{HQQ}) across eight models.

	Wiki2 ↓		C4 ↓		PTB ↓		Avg. PPL ↓		Avg. Acc. ↑	
Quantizer	Uniform	K_{HQQ}	Uniform	K_{HQQ}	Uniform	K_{HQQ}	Uniform	K_{HQQ}	Uniform	K_{HQQ}
RTN	56.82	47.21	53.00	44.62	89.57	72.31	66.46	54.71	53.55	53.73
HQQ	44.93	15.98	48.85	19.48	116.91	35.57	70.23	23.67	52.74	57.95
GPTQ	17.26	16.04	17.61	17.50	59.82	48.34	31.56	27.30	57.16	59.05
LeanQuant	11.21	11.18	13.65	13.64	24.31	24.06	16.39	16.29	63.57	63.53

4.4 Comparison across Quantizers

Table 4 shows that ODES improved average accuracy over the uniform baseline for RTN (Nagel et al., 2021), HQQ (Badri & Shaji, 2023), and GPTQ (Frantar et al., 2023), while LeanQuant (Zhang & Shrivastava, 2025) remained nearly unchanged. Averaged across quantizers, ODES yielded a 1.81 percentage-point accuracy gain while reducing mean PPL from 46.16 to 30.49. The largest gains occurred with HQQ, where ODES improved accuracy on all eight models by 1.79–9.10 percentage points and reduced mean PPL from 70.23 to 23.67; since K_{HQQ} was selected with HQQ as the proxy, part of this gain may reflect the selection itself. GPTQ, which was not used for selection, also improved accuracy on all eight models, indicating that the ranking transfers to a reconstruction-based quantizer. For LeanQuant, whose uniform 3-bit baseline was already the strongest, reallocation left accuracy essentially unchanged. Complete model-wise results at K_{th} and K_{HQQ} for all four quantizers are reported in Table C.2, and the robustness of the ODES ranking to the choice of profiling windows is analyzed in Appendix C.3.

5 Limitations

First, K_{th} should be viewed as a boundary for precision reallocation rather than the optimal K. Its derivation relies on an idealized perturbation model that practical PTQ errors need not follow, so a gap between K_{th} and K_{HQQ} is expected. Second, the one-dimensional search over K is restricted to the path defined by the ODES ranking. It also assumes paired one-bit promotion and demotion; with more bit-width choices, a single K no longer describes the search space. Third, CPC removes the model-sized running state of direct Welford estimation, but each HVP still requires a double backward pass through the full network, so profiling needs activation and gradient memory well beyond that of inference; Llama-2-13B required two H100 GPUs for this reason and is therefore omitted from the runtime comparison in Table 3. Finally, ODES uses only the off-diagonal part of the Hessian. The diagonal contribution was small in the settings we tested (Appendix A.2), but this may not hold for all models or quantizers; since diagonal statistics are available from the same CPC probes, combining the two terms is a natural extension.

6 Conclusion

We introduced ODES, a projection-level sensitivity built from the off-diagonal energy of the task-loss Hessian, together with a sufficient condition under which precision reallocation reduces the interaction second moment and the CPC estimator that makes it practical at LLM scale. The resulting ranking reduces mixed-precision allocation to a one-dimensional search whose accuracy optimum did not exceed the theory-guided boundary, and which remained competitive with search-based methods at roughly half their cost while transferring across quantizers. These results show that interaction structure at the pretrained weights carries actionable bit-allocation information before any quantization is performed.

LLM USAGE STATEMENT

In this work, we used generative AI tools (ChatGPT and Claude) for the following purposes. (i) Writing: polishing and condensing the manuscript, removing redundancy, suggesting wording for several explanatory passages (e.g., the interpretation of the sufficient condition in Section 3.2 and the limitations), checking the internal numerical consistency between tables and text, and LaTeX formatting. (ii) Retrieval and discovery: locating related work, whose relevance and citations were checked by the authors. The method, its derivations and the main experiments were conceived and conducted by the authors who take full responsibility for this paper.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. doi: 10.1016/S0022-0000(03)00025-4. URL https://www.sciencedirect.com/science/article/pii/S0022000003000254.

Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM*, 58(2):8:1–8:34, 2011. doi: 10.1145/1944345.1944349.

Hicham Badri and Appu Shaji. Half-quadratic quantization of large machine learning models. Mobius Labs blog post and software, 2023. URL https://github.com/mobiusml/hqq. https://mobiusml.github.io/hqq_blog/.

Robert A. Baston and Yuji Nakatsukasa. Stochastic diagonal estimation: Probabilistic bounds and an improved algorithm, 2022. URL https://arxiv.org/abs/2201.10684.

C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Applied Numerical Mathematics*, 57(11–12):1214–1229, 2007. doi: 10.1016/j.apnum.2007.01.003. URL https://www.sciencedirect.com/science/article/pii/S0168927407000244.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Tony F. Chan, Gene H. Golub, and Randall J. LeVeque. Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247, 1983. doi: 10.1080/00031305.1983.10483115.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0df38cd13520747e1e64e5b123a78ef8-Abstract-Conference.html.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. EfficientQAT: Efficient quantization-aware training for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10081–10100. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.498. URL https://aclanthology.org/2025.acl-long.498/.

Zihan Chen, Bike Xie, Jundong Li, and Cong Shen. Channel-wise mixed-precision quantization for large language models, 2024. URL https://arxiv.org/abs/2410.13056.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. doi: 10.1002/rsa.10073.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017.

Zihao Deng, Sayeh Sharify, Xin Wang, and Michael Orshansky. Mixed-precision quantization for deep vision models with integer quadratic programming. In *Proceedings of the 62nd ACM/IEEE Design Automation Conference*, pp. 1–7, 2025. doi: 10.1109/DAC63849.2025.11132777. URL https://arxiv.org/abs/2307.05657v2.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://openreview.net/forum?id=dXiGWqBoxaD.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Q1u25ahSuy.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian AWare quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Dong_HAWQ_Hessian_AWare_Quantization_of_Neural_Networks_With_Mixed-Precision_ICCV_2019_paper.html.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d77c703536718b95308130ff2e5cf9ee-Abstract.html.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12284–12303. PMLR, 2024.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://openreview.net/forum?id=ksVGCOlOEba.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

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. Zenodo, 2024. URL https://github.com/EleutherAI/lm-evaluation-harness. Version 0.4.3.

Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Dacheng Tao, and Xianglong Liu. LLMC: Benchmarking large language model quantization with a versatile compression toolkit. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 132–152. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.12. URL https://aclanthology.org/2024.emnlp-industry.12/.

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, 2024. URL https://arxiv.org/abs/2407.21783.

Ziyi Guan, Hantao Huang, Yupeng Su, Hong Huang, Ngai Wong, and Hao Yu. APTQ: Attention-aware post-training mixed-precision quantization for large language models. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024. doi: 10.1145/3649329.3658498. URL https://doi.org/10.1145/3649329.3658498.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25672–25692. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25aa.html.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866. URL https://doi.org/10.1080/03610919008812866.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. doi: 10.1090/conm/026/737400.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23901–23923. PMLR, 2024. URL https://proceedings.mlr.press/v235/kim24f.html.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pp. 598–605, 1990.

Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. OWQ: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13355–13364, 2024. doi: 10.1609/aaai.v38i12.29237. URL https://ojs.aaai.org/index.php/AAAI/article/view/29237.

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. 35532–35550. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1799. URL https://aclanthology.org/2025.emnlp-main.1799/.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Yifei Liu, Jicheng Wen, Yang Wang, Shengyu Ye, Li Lyna Zhang, Ting Cao, Cheng Li, and Mao Yang. VPTQ: Extreme low-bit vector post-training quantization for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8181–8196. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.467.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization, 2021. URL https://arxiv.org/abs/2106.08295.

Yury Nahshan, Brian Chmiel, Chaim Baskin, Evgenii Zheltonozhskii, Ron Banner, Alex M. Bronstein, and Avi Mendelson. Loss aware post-training quantization. *Machine Learning*, 110(11): 3245–3262, 2021. doi: 10.1007/s10994-021-06053-z.

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

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147. URL https://doi.org/10.1162/neco.1994.6.1.147.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

John K. Salmon, Mark A. Moraes, Ron O. Dror, and David E. Shaw. Parallel random numbers: As easy as 1, 2, 3. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011. URL https://www.thesalmons.org/john/random123/papers/random123sc11.pdf.

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. In *International Conference on Learning Representations*, 2024.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Q-BERT: Hessian based ultra low precision quantization of BERT. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 2020. doi: 10.1609/aaai.v34i05.6409. URL https://ojs.aaai.org/index.php/AAAI/article/view/6409.

Oliver Sieberling, Denis Kuznedelev, Eldar Kurtic, and Dan Alistarh. EvoPress: Accurate dynamic model compression via evolutionary search. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55556–55590. PMLR, 2025. URL https://proceedings.mlr.press/v267/sieberling25a.html.

A. Sripad and D. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(5):442–448, 1977. doi: 10.1109/TASSP.1977.1162977.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. HAQ: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8612–8620, 2019. doi: 10.1109/CVPR.2019.00881.

B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022. URL https://doi.org/10.1080/00401706.1962.10490022.

Bernard Widrow and István Kollár. *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511754661.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10100–10118. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.544. URL https://aclanthology.org/2024.acl-long.544/.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060. URL https://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/wNow3.pdf.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 581–590, 2020. doi: 10.1109/BigData50022.2020.9378171.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V3: Dyadic neural network quantization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11875–11886. PMLR, 2021a. URL https://proceedings.mlr.press/v139/yao21a.html.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael W. Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10665–10673, 2021b. doi: 10.1609/aaai.v35i12.17275.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Chao Zhang, Li Wang, Samson Lasaulce, and Merouane Debbah. BAQ: Efficient bit allocation quantization for large language models, 2025. URL https://arxiv.org/abs/2506.05664.

Tianyi Zhang and Anshumali Shrivastava. LeanQuant: Accurate and scalable large language model quantization with loss-error-aware grid. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ISqx8giekS.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577, 2024. doi: 10.1162/tacl_a_00704. URL https://aclanthology.org/2024.tacl-1.85/.

A Derivation of the ODES Criterion

A.1 Interaction second moment and precision reallocation

Let H_d be the symmetric task-loss Hessian on window d, and let δ_i be independent perturbations with $\mathbb{E}_\delta[\delta_i] = 0$ and $\mathbb{E}_\delta[\delta_i^2] = \sigma_i^2$. For $C_d(\boldsymbol{\delta}) = \sum_{i<j} H_{d,ij}\delta_i\delta_j$, products associated with distinct unordered pairs have zero expectation: each such product contains at least one independent, zero-mean coordinate to the first power. Consequently,

$$\mathbb{E}_\delta[C_d] = 0, \qquad \mathbb{E}_\delta[C_d^2] = \sum_{i<j} H_{d,ij}^2\sigma_i^2\sigma_j^2. \tag{A.1}$$

Writing $A_{ij} = D^{-1}\sum_d H_{d,ij}^2$ for $i \neq j$ gives

$$J_{\text{off}}(\boldsymbol{\sigma}^2) = \sum_{i<j} A_{ij}\sigma_i^2\sigma_j^2, \qquad A_{ij} = A_{ji} \geq 0. \tag{A.2}$$

Partition the coordinates into $\mathcal{I}_+$, $\mathcal{I}_-$, and $\mathcal{I}_0$, whose precision is increased, decreased, and unchanged, respectively. Assume a common baseline variance $\sigma_i^2 = \sigma^2 > 0$ and the one-bit responses $\delta_i' = \delta_i/2$ on $\mathcal{I}_+$, $\delta_i' = 2\delta_i$ on $\mathcal{I}_-$, and $\delta_i' = \delta_i$ on $\mathcal{I}_0$. Define

$$\begin{aligned}
A_{++} &= \sum_{\substack{i<j \\ i,j\in\mathcal{I}_+}} A_{ij}, \qquad A_{--} = \sum_{\substack{i<j \\ i,j\in\mathcal{I}_-}} A_{ij}, \\
A_{+-} &= \sum_{i\in\mathcal{I}_+}\sum_{j\in\mathcal{I}_-} A_{ij}, \quad A_{+0} = \sum_{i\in\mathcal{I}_+}\sum_{j\in\mathcal{I}_0} A_{ij}, \\
A_{-0} &= \sum_{i\in\mathcal{I}_-}\sum_{j\in\mathcal{I}_0} A_{ij}.
\end{aligned} \tag{A.3}$$

The variance multipliers are $1/4$, 4, and 1. The multipliers cancel on edges between $\mathcal{I}_+$ and $\mathcal{I}_-$, so the reduction is

$$\begin{aligned}
\Delta J &:= J_{\text{off}}(\boldsymbol{\sigma}^2) - J_{\text{off}}(\boldsymbol{\sigma}'^2) \\
&= \sigma^4\left(\frac{15}{16}A_{++} - 15A_{--} + \frac{3}{4}A_{+0} - 3A_{-0}\right).
\end{aligned} \tag{A.4}$$

The incident energy totals satisfy $R_+ = 2A_{++} + A_{+-} + A_{+0}$ and $R_- = 2A_{--} + A_{+-} + A_{-0}$. Substitution yields

$$\begin{aligned}
\Delta J &= \frac{15\sigma^4}{32}(R_+ - 16R_-) \\
&\quad + \sigma^4\left(\frac{225}{32}A_{+-} + \frac{9}{32}A_{+0} + \frac{9}{2}A_{-0}\right) \\
&\geq \frac{15\sigma^4}{32}(R_+ - 16R_-).
\end{aligned} \tag{A.5}$$

Thus $R_+ > 16R_-$ is sufficient for $\Delta J > 0$.

Multiple projection pairs. For K disjoint pairs changed simultaneously from the common baseline, apply the same bound to the unions of the promoted and demoted parameter sets. Since projection m has incident energy $p_m S_m$, this gives

$$J(0) - J(K) \geq \frac{15\sigma^4}{32} \sum_{k=1}^{K} \left(p_{m_k^+} S_{m_k^+} - 16 p_{m_k^-} S_{m_k^-} \right). \tag{A.6}$$

In particular, satisfying the pairwise condition for all included pairs gives a reduction relative to the common baseline.

A.2 Diagonal contribution and the scope of the interaction criterion

Separation of quadratic fluctuations. Write the second-order contribution on a fixed window as $Q_d(\boldsymbol{\delta}) = D_d(\boldsymbol{\delta}) + C_d(\boldsymbol{\delta})$, using the diagonal and off-diagonal terms in Eq. (1). In addition to independent, zero-mean perturbations, assume finite fourth moments $\mu_{4,i} = \mathbb{E}_\delta[\delta_i^4]$. The diagonal term has a mean and a variance

$$\mathbb{E}_\delta[D_d] = \frac{1}{2} \sum_i H_{d,ii} \sigma_i^2, \qquad \mathrm{Var}_\delta(D_d) = \frac{1}{4} \sum_i H_{d,ii}^2 (\mu_{4,i} - \sigma_i^4). \tag{A.7}$$

Moreover, $\mathrm{Cov}_\delta(D_d, C_d) = 0$: each product $\delta_i^2 \delta_j \delta_k$, with $j \neq k$, contains an independent, zero-mean coordinate to the first power. Consequently, defining $J_{\mathrm{diag}} = \mathbb{E}_d \mathrm{Var}_\delta(D_d)$ gives

$$J_{\mathrm{quad}} := \mathbb{E}_d \mathrm{Var}_\delta(Q_d) = J_{\mathrm{diag}} + J_{\mathrm{off}}. \tag{A.8}$$

Thus, the interaction objective depends only on off-diagonal Hessian entries, and its reduction does not require the diagonal term to vanish. Precision reallocation can nevertheless change J_{diag} through the perturbation moments. Equation (A.8) concerns centered fluctuations over perturbations at each fixed window. It is not the raw second moment of Q_d, which also includes $(\mathbb{E}_\delta D_d)^2$.

Including the diagonal term in the reduction bound. At the common baseline, suppose $\mu_{4,i} = \kappa\sigma^4$ for all coordinates, and define

$$T := \mathbb{E}_d \sum_i H_{d,ii}^2, \qquad G := \mathbb{E}_d \sum_{i \neq j} H_{d,ij}^2, \qquad \frac{J_{\mathrm{diag}}(0)}{J_{\mathrm{off}}(0)} = \frac{\kappa - 1}{2} \frac{T}{G}, \tag{A.9}$$

where $G > 0$. For the one-bit scaling in Appendix A.1, let $T_\pm = \mathbb{E}_d \sum_{i \in \mathcal{I}_\pm} H_{d,ii}^2$. Scaling the perturbations by $1/2$ or 2 preserves κ and changes their fourth moments by factors $1/16$ or 16. The diagonal variance change is therefore

$$\Delta J_{\mathrm{diag}} := J_{\mathrm{diag}}(K) - J_{\mathrm{diag}}(0) = \frac{15(\kappa - 1)\sigma^4}{64} \left(16T_- - T_+ \right). \tag{A.10}$$

Combining this expression with Eq. (A.5), applied to the unions of the promoted and demoted coordinates, yields

$$J_{\mathrm{quad}}(0) - J_{\mathrm{quad}}(K) \geq \frac{15\sigma^4}{32}(R_+ - 16R_-) - \Delta J_{\mathrm{diag}}. \tag{A.11}$$

A positive right-hand side certifies a reduction of the centered quadratic fluctuations under this perturbation model, including the diagonal contribution. In particular, if $\Delta J_{\mathrm{diag}} \leq 0$, the original condition $R_+ > 16R_-$ also suffices for reducing J_{quad}.

Diagnostic from C4 profiles. Table A.1 evaluates Eq. (A.9) using Hessian estimates for five Llama models and independent uniform perturbations, for which $\kappa = 9/5$. This diagnostic averaged 56 of the 64 C4 profiling windows after removing the eight largest off-diagonal energy-vector norms. The diagonal statistics used $N = 64$ probes, and the off-diagonal statistics used CPC with $N = 64$ and three sketch seeds. For a projection with p_m parameters, diagonal energy was estimated as $p_m(\widehat{h}_{m,d}^2 + \max\{\widehat{v}_{m,d}, 0\})$, where $\widehat{h}_{m,d}$ is its mean diagonal estimate and $\widehat{v}_{m,d}$ is the debiased

Table A.1: Diagonal contribution on the 56-window C4 profiles. The last column uses independent, equal-variance uniform perturbations at the common baseline. Both columns report percentages.

Model	T/G (%)	$J_{\mathrm{diag}}(0)/J_{\mathrm{off}}(0)$ (%)
Llama-2-13B	0.1208	0.0483
Llama-2-7B	0.1920	0.0768
Llama-3.1-8B	0.2403	0.0961
Llama-3.2-1B	0.9528	0.3811
Llama-3.2-3B	0.2442	0.0977

estimate of variation across its diagonal entries. We summed energies across projections and then averaged across windows before forming the ratio.

The estimated diagonal fluctuation was small relative to the off-diagonal fluctuation in these profiles. For their CPC-based allocations at $K = 1, \ldots, 10$, the estimated right-hand side of Eq. (A.11) was positive in all 50 model–K settings. The estimated diagonal variance change was also negative in all 50 settings, so it reinforced the off-diagonal reduction in this calculation. This diagnostic supports the reduction after accounting for the diagonal term.

Relation to variance across windows. For a fixed quantized model, the perturbation is fixed while the evaluation window varies. Writing $g_d = \nabla L(\mathbf{w}^\star; x_d)$ and letting r_d denote the Taylor remainder gives

$$\begin{aligned} B_d &:= L(\mathbf{w}^\star; x_d) + g_d^\top \boldsymbol{\delta} + D_d(\boldsymbol{\delta}) + r_d, \\ \mathrm{Var}_d\, L(\mathbf{w}^\star + \boldsymbol{\delta}; x_d) &= \mathrm{Var}_d\, B_d + \mathrm{Var}_d\, C_d + 2\,\mathrm{Cov}_d(B_d, C_d). \end{aligned} \tag{A.12}$$

The fixed-window identity $\mathrm{Cov}_\delta(D_d, C_d) = 0$ does not imply zero covariance across windows. Likewise, stationarity of the mean loss does not make each g_d vanish. The Window-NLL variance in Appendix C.2 therefore measures the full loss across windows, whereas the interaction criterion and the diagnostic above concern perturbation-induced quadratic fluctuations.

A.3 Per-parameter normalization

Consider continuous bit offsets b_m around the common baseline, with $\sigma_i^2(\mathbf{b}) = \sigma^2 \psi(b_{m(i)})$, where $m(i)$ denotes the projection containing coordinate i. Assume ψ is differentiable near zero, with $\psi(0) = 1$ and $\psi'(0) < 0$. The interaction moment is

$$J_{\mathrm{off}}(\mathbf{b}) = \sigma^4 \sum_{i<j} A_{ij} \psi(b_{m(i)}) \psi(b_{m(j)}). \tag{A.13}$$

Because an increase of b_m costs $p_m b_m$ parameter bits, the local reduction per unit of storage is

$$-\frac{1}{p_m} \left. \frac{\partial J_{\mathrm{off}}}{\partial b_m} \right|_{\mathbf{b}=0} = -\sigma^4 \psi'(0) S_m. \tag{A.14}$$

This motivates ranking by S_m on a per-parameter basis. The finite one-bit condition still uses the total energy $p_m S_m$.

B Matrix-Free Estimation and Profiling Details

B.1 Welford and scalar covariance estimators

Fix a window and write its task-loss Hessian as H. For $N \geq 2$ independent Rademacher probes, define $y_r = z_r \odot H z_r$. Hessian–vector products are computed as the gradient of $\nabla_{\mathbf{w}} L(\mathbf{w}; x_d)^\top z_r$ at $\mathbf{w} = \mathbf{w}^\star$ (Pearlmutter, 1994). Following the stochastic probing construction (Baston & Nakatsukasa, 2022; Bekas et al., 2007; Hutchinson, 1990),

$$y_{r,i} = H_{ii} + \sum_{j \neq i} H_{ij} z_{r,i} z_{r,j}.$$

Independence of the probe coordinates gives

$$\mathbb{E}_z[y_{r,i}] = H_{ii}, \qquad \nu_i := \mathrm{Var}_z(y_{r,i}) = \sum_{j \neq i} H_{ij}^2. \tag{B.1}$$

Therefore, for $s_m = p_m^{-1} \sum_{i \in \mathcal{I}_m} \nu_i$, the unbiased sample-variance estimator is

$$\widehat{s}_m^{\mathrm{W}} = \frac{1}{p_m(N-1)} \sum_{r=1}^{N} \|y_{r,m} - \bar{y}_m\|_2^2, \qquad \mathbb{E}_z[\widehat{s}_m^{\mathrm{W}}] = s_m. \tag{B.2}$$

Bars denote averages over the N probes. Welford updates (Welford, 1962) compute this statistic online using a running coordinate mean and one summed variance accumulator per projection, requiring $O(P + M)$ running-summary storage.

Online Welford algorithm. Initialize $\mu_{0,m} = \mathbf{0}$ and $T_{0,m} = 0$ for each projection. For probe $r = 1, \ldots, N$, update

$$\begin{aligned} \Delta_{r,m} &= y_{r,m} - \mu_{r-1,m}, \\ \mu_{r,m} &= \mu_{r-1,m} + \Delta_{r,m}/r, \\ T_{r,m} &= T_{r-1,m} + \Delta_{r,m}^{\top}(y_{r,m} - \mu_{r,m}). \end{aligned} \tag{B.3}$$

The returned score is $\widehat{s}_m^{\mathrm{W}} = T_{N,m}/[p_m(N-1)]$. The last update uses the newly updated mean $\mu_{r,m}$.

For $i \neq j$, symmetry of H gives $\mathrm{Cov}_z(y_{r,i}, y_{r,j}) = H_{ij}^2$. With $a_{r,m} = p_m^{-1} \sum_{i \in \mathcal{I}_m} y_{r,i}$ and $q_r = z_r^{\top} H z_r = \sum_j y_{r,j}$,

$$\frac{1}{2}\,\mathrm{Cov}_z(a_{r,m}, q_r) = \frac{1}{2p_m} \sum_{i \in \mathcal{I}_m} \left(\nu_i + \sum_{j \neq i} H_{ij}^2 \right) = s_m. \tag{B.4}$$

Thus

$$\widehat{s}_m^{\mathrm{C}} = \frac{1}{2(N-1)} \sum_{r=1}^{N} (a_{r,m} - \bar{a}_m)(q_r - \bar{q}) \tag{B.5}$$

is also unbiased and requires $O(M)$ running-summary storage. The global q_r includes both within- and cross-projection terms.

Online scalar covariance algorithm. Initialize $\bar{q}_0 = 0$ and, for each projection, $\bar{a}_{0,m} = C_{0,m} = 0$. At probe r, update the global mean once as $\bar{q}_r = \bar{q}_{r-1} + (q_r - \bar{q}_{r-1})/r$. For each projection, update

$$\begin{aligned} \Delta_{a,r,m} &= a_{r,m} - \bar{a}_{r-1,m}, \\ \bar{a}_{r,m} &= \bar{a}_{r-1,m} + \Delta_{a,r,m}/r, \\ C_{r,m} &= C_{r-1,m} + \Delta_{a,r,m}(q_r - \bar{q}_r). \end{aligned} \tag{B.6}$$

The returned score is $\widehat{s}_m^{\mathrm{C}} = C_{N,m}/[2(N-1)]$. The last update uses the newly updated global mean $\bar{q}_r$.

B.2 Compressed probe-centered estimation

Write $y_{r,m} = a_{r,m}\mathbf{1} + u_{r,m}$, where $\mathbf{1}^{\top} u_{r,m} = 0$, and define

$$V_{a,m} = \frac{1}{N-1} \sum_r (a_{r,m} - \bar{a}_m)^2, \quad \bar{B}_m = \frac{1}{Np_m} \sum_r \|u_{r,m}\|_2^2, \quad \bar{u}_m = \frac{1}{N} \sum_r u_{r,m}.$$

Orthogonality of the constant and centered components gives

$$\begin{aligned} \widehat{s}_m^{\mathrm{W}} &= V_{a,m} + \frac{1}{p_m(N-1)} \sum_r \|u_{r,m} - \bar{u}_m\|_2^2 \\ &= V_{a,m} + \frac{N}{N-1} \left(\bar{B}_m - \frac{\|\bar{u}_m\|_2^2}{p_m} \right). \end{aligned} \tag{B.7}$$

CPC uses a sketch $R_m \in \mathbb{R}^{k_m \times p_m}$, drawn independently of the probe stream and fixed across probes, to store $g_m = R_m \bar{u}_m$. Its estimate is

$$\widehat{s}_m^{\mathrm{CPC}} = V_{a,m} + \frac{N}{N-1}\left(\bar{B}_m - \frac{\|g_m\|_2^2}{p_m}\right). \tag{B.8}$$

The compression error is exactly

$$\widehat{s}_m^{\mathrm{CPC}} - \widehat{s}_m^{\mathrm{W}} = -\frac{N}{(N-1)p_m}\left(\|R_m \bar{u}_m\|_2^2 - \|\bar{u}_m\|_2^2\right). \tag{B.9}$$

In particular, a norm distortion of at most $\epsilon_m \|\bar{u}_m\|_2^2$ produces an additive score error of at most $N\epsilon_m \|\bar{u}_m\|_2^2/[(N-1)p_m]$. Random projections and linear sketches provide the general background for this norm approximation (Achlioptas, 2003; Dasgupta & Gupta, 2003; Johnson & Lindenstrauss, 1984; Woodruff, 2014). When $\mathbb{E}_R[R_m^\top R_m] = I$, $\mathbb{E}_R[\widehat{s}_m^{\mathrm{CPC}} \mid z_1, \ldots, z_N] = \widehat{s}_m^{\mathrm{W}}$, so CPC is unbiased in the joint expectation over probes and sketches for the fixed window.

B.3 Sketch implementation and online computation

Partition projection m into $L_m = \lceil p_m/1{,}024 \rceil$ contiguous blocks $\mathcal{B}_{m,\ell}$ of at most 1,024 coordinates. With $r_{\mathrm{sk}} = 4$ sign replicas, the sketch is

$$(R_m u)_{t,\ell} = \frac{1}{\sqrt{r_{\mathrm{sk}}}} \sum_{i \in \mathcal{B}_{m,\ell}} s_{m,t,i} u_i, \qquad k_m = r_{\mathrm{sk}} L_m. \tag{B.10}$$

Under independent Rademacher signs, $\mathbb{E}_s[s_{m,t,i} s_{m,t,j}] = \mathbf{1}_{\{i=j\}}$, which gives $\mathbb{E}_R[R_m^\top R_m] = I$. Signs are regenerated with Philox4x32-10 (Salmon et al., 2011) from fixed, projection-specific seeds, separately from the HVP probe stream; the dense matrix R_m is not stored.

The implementation can accumulate the responses before centering. Define $\widetilde{g}_m = N^{-1}\sum_r R_m y_{r,m}$, $c_m = R_m \mathbf{1}$, and $\bar{E}_{y,m} = (Np_m)^{-1}\sum_r \|y_{r,m}\|_2^2$. Then $g_m = \widetilde{g}_m - \bar{a}_m c_m$, and the equivalent estimate is

$$\widehat{s}_m^{\mathrm{CPC}} = \frac{N}{N-1}\left(\bar{E}_{y,m} - \bar{a}_m^2 - \frac{\|g_m\|_2^2}{p_m}\right). \tag{B.11}$$

This is the same estimator as Eq. B.8 in exact arithmetic.

Algorithm 1 CPC profiling for one window

Require: Fixed-window HVP operator, projections $\{\mathcal{I}_m\}$, $N \geq 2$, fixed sketches $\{R_m\}$ independent of the probes
Ensure: Projection scores $\{\widehat{s}_m^{\mathrm{CPC}}\}$

$\bar{a}_m \leftarrow 0$, $\bar{E}_{y,m} \leftarrow 0$, $\widetilde{g}_m \leftarrow \mathbf{0}$, $c_m \leftarrow R_m \mathbf{1}$ for all m
for $r = 1, \ldots, N$ **do**
  Draw a Rademacher probe z_r; $h_r \leftarrow \mathrm{HVP}(z_r)$; $y_r \leftarrow z_r \odot h_r$
  for $m = 1, \ldots, M$ **do**
    $a_{r,m} \leftarrow \mathbf{1}^\top y_{r,m}/p_m$; $E_{r,m} \leftarrow \|y_{r,m}\|_2^2/p_m$
    $\bar{a}_m \leftarrow \bar{a}_m + (a_{r,m} - \bar{a}_m)/r$
    $\bar{E}_{y,m} \leftarrow \bar{E}_{y,m} + (E_{r,m} - \bar{E}_{y,m})/r$
    $\widetilde{g}_m \leftarrow \widetilde{g}_m + (R_m y_{r,m} - \widetilde{g}_m)/r$
  end for
end for
for $m = 1, \ldots, M$ **do**
  $g_m \leftarrow \widetilde{g}_m - \bar{a}_m c_m$
  $\widehat{s}_m^{\mathrm{CPC}} \leftarrow \frac{N}{N-1}\left(\bar{E}_{y,m} - \bar{a}_m^2 - \frac{\|g_m\|_2^2}{p_m}\right)$
end for
return $\{\widehat{s}_m^{\mathrm{CPC}}\}$

Algorithm 1 summarizes the online computation. Running summaries use $O(M + \sum_m k_m)$ storage. They are reset for each window.

B.4 PROFILING EVALUATION AND RESOURCE ACCOUNTING

The estimator comparison used Llama-3.2-1B, Llama-3.2-3B, and Qwen3-1.7B on an NVIDIA H100. The projection set contained the query, key, value, attention-output, gate, up, and down projections in every Transformer block. Table B.1 summarizes the settings.

Table B.1: Profiling evaluation settings.

Item	Setting
Fidelity input	C4 window 0; 1,024 tokens
Candidate and reference	$N = 64$ candidate probes; $N = 4{,}096$ Welford reference probes
Cost-run input	128-token prefix and 1,024 target tokens
Model and HVP precision	BF16
Stored means	FP16 for all three estimators; FP32 for the Welford reference

At each probe budget, CPC and Welford used the same candidate probes. Each run was compared with two reference runs generated with the other probe seeds. Metrics were averaged over these two comparisons and, for CPC, over three sketch seeds.

Relative errors are $\mathrm{Rel}\, L_q = 100\|\widehat{\mathbf{s}} - \mathbf{s}_{\mathrm{ref}}\|_q / \|\mathbf{s}_{\mathrm{ref}}\|_q$ for $q \in \{1, 2\}$. Spearman correlation uses average ranks for ties.

Estimator memory includes its buffers and reusable temporary storage, excluding model parameters, the autograd graph, and HVP working memory. Time per probe is the elapsed time for 64 probes divided by 64, excluding model loading, warmup, and file output.

C EXPERIMENTAL DETAILS

C.1 ALLOCATION AND QUANTIZATION SETTINGS

Allocation profiling used 64 C4 windows of 1,024 tokens with $N = 64$ probes per window; the eight windows with the largest score-vector ℓ_2 norms were excluded, and each projection's score was averaged over the same remaining 56 windows. C4 (Raffel et al., 2020) was also used for calibration of activation-dependent quantizers and for configuration search where required. The GPTQ comparison used group size 128 (Frantar et al., 2023).

For bit assignments b_m, the parameter-weighted bit width is $\bar{b} = \sum_m p_m b_m / \sum_m p_m$, counting the selected projection weights. Promoting one projection and demoting another by one bit changes the budget by $p_h - p_l$; comparisons used the weighted budget of the resulting configurations. MeanDiag ranks projections by the mean diagonal entries of the task-loss Hessian and uses the same reallocation procedure.

C.2 EVALUATION AND CONFIGURATION SELECTION

Aggregation across quantizers. All quantities below were computed separately for each model and allocation rule. Let q index quantizers and s index seeds. For each configuration, window-NLL variance is

$$V_{q,s}(K) = \mathrm{Var}_d[\ell_{q,s,d}(K)]\,,$$

where $\ell_{q,s,d}(K)$ is the mean target-token NLL within evaluation window d.

Let $\mathcal{S}_q(K)$ denote the seeds included for quantizer q at K. We first averaged each metric across seeds:

$$\bar{X}_q(K) = \frac{1}{|\mathcal{S}_q(K)|} \sum_{s \in \mathcal{S}_q(K)} X_{q,s}(K), \qquad X \in \{V, P, A\},$$

where P denotes mean PPL and A denotes downstream accuracy. In particular, $\bar{V}_q(K)$ averages the window-NLL variances computed separately for individual seeds.

C.3 Robustness to the Choice of Profiling Windows

We evaluated the stability of ODES rankings across different choices of profiling windows.

For this robustness experiment, we increased the profiling pool from 64 to 128 C4 windows and excluded the 16 windows with the largest score-vector ℓ_2 norms. This preserved the one-eighth exclusion fraction used in Section 4.2. We used the ranking obtained by averaging the remaining 112 windows as the reference. To vary the choice of profiling windows, we repeatedly sampled 64 windows uniformly without replacement from this retained pool and recomputed the ODES ranking from their mean scores. We performed 10,000 repetitions per model, allowing the identities of the selected windows to vary across repetitions while keeping the subset size fixed.

We measured agreement with the reference using Spearman's rank correlation and top-10 and bottom-10 projection overlap. The latter metrics directly assess whether different window choices identify the same projections as candidates for precision promotion and demotion.

Table C.1: Agreement of ODES rankings across different choices of profiling windows. Each model was evaluated over 10,000 randomly drawn subsets of 64 windows from the 112-window pool retained after excluding the 16 largest-norm windows from 128 C4 windows. Values are means across repetitions; top-10 and bottom-10 overlap measure shared membership in sets of ten projections.

Model	Spearman ρ	Top-10 (%)	Bottom-10 (%)
Llama-3.2-1B	0.9973	98.81	92.25
Llama-3.2-3B	0.9920	91.62	85.75
Llama-2-7B	0.9947	99.99	82.24
Llama-3.1-8B	0.9918	96.92	85.23
Llama-2-13B	0.9954	100.00	86.67
Qwen3-1.7B	0.9995	94.18	91.48
Qwen3-4B	0.9977	95.44	89.10
Qwen3-8B	0.9883	89.36	79.95

Table C.1 shows that ODES rankings remained highly consistent across different choices of profiling windows. Mean Spearman correlations ranged from 0.9883 to 0.9995 across all eight models. This agreement also extended to the projections relevant to precision reallocation: mean top-10 overlap ranged from 89.36% to 100.00%, and mean bottom-10 overlap ranged from 79.95% to 92.25%.

The consistently high agreement across randomly selected window subsets suggests that ODES captures projection sensitivity patterns shared across the profiling data.

MeanDiag and ODES on additional models. Figure C.1 extends the comparison in Figure 3 to four additional models. Using the same diagnostics, it examines the relationship between the theory-guided allocation boundary and the observed accuracy, window-NLL variance, and mean PPL as the number of reallocated projection pairs K increases. Figure C.2 provides an eight-model overview of projection-wise ODES scores across profiling windows.

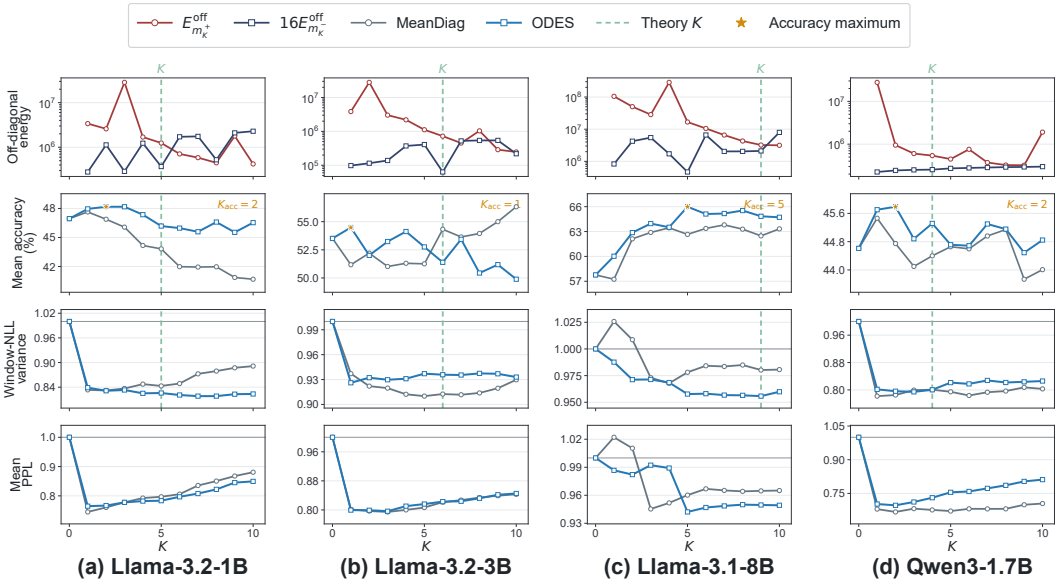


Figure C.1: Additional-model results corresponding to Figure 3 for Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, and Qwen3-1.7B. The first row compares the off-diagonal energies of the paired high- and low-ODES projections. The remaining rows show downstream accuracy, normalized window-NLL variance, and normalized mean PPL for ODES and MeanDiag as functions of K. Normalized metrics are relative to the corresponding uniform allocation at $K = 0$. Stars mark the maximum ODES accuracy, and the green vertical line marks the theory-guided boundary defined by the sufficient condition.

Table C.2: Base versus ODES at the model-level K_{th} and K_{hqq} across four quantizers.

Model	Quantizer	K_{th}	K_{hqq}	Accuracy (%) ↑					Mean PPL ↓				
				Base	$\mathrm{ODES}_{\mathrm{th}}$	Δ_{th} (pp)	$\mathrm{ODES}_{\mathrm{hqq}}$	Δ_{hqq} (pp)	Base	$\mathrm{ODES}_{\mathrm{th}}$	Δ_{th}	$\mathrm{ODES}_{\mathrm{hqq}}$	Δ_{hqq}
Llama-2-7B	RTN	21	2	62.80	63.24	+0.44	63.06	+0.26	14.73	15.74	+1.01	14.68	−0.05
	HQQ			58.11	63.44	+5.33	63.04	+4.93	26.35	15.49	−10.86	15.24	−11.11
	LeanQuant			64.86	64.81	−0.05	64.63	−0.23	13.31	13.13	−0.18	12.86	−0.45
	GPTQ			60.07	60.86	+0.80	61.89	+1.83	74.65	49.02	−25.63	54.16	−20.49
Llama-2-13B	RTN	35	8	67.99	68.44	+0.45	68.14	+0.15	16.27	15.62	−0.65	15.46	−0.81
	HQQ			66.51	68.13	+1.62	68.31	+1.79	17.34	14.90	−2.44	15.11	−2.23
	LeanQuant			69.79	69.12	−0.66	69.64	−0.15	14.74	15.71	+0.97	14.73	−0.01
	GPTQ			67.78	69.46	+1.68	69.41	+1.63	17.77	17.85	+0.08	16.95	−0.82
Llama-3.2-1B	RTN	5	2	37.02	36.53	−0.49	37.49	+0.47	323.89	481.24	+157.35	247.44	−76.45
	HQQ			41.40	46.80	+5.40	46.58	+5.18	86.22	34.70	−51.53	34.55	−51.67
	LeanQuant			52.20	52.03	−0.17	52.04	−0.16	17.43	17.87	+0.44	17.68	+0.25
	GPTQ			46.97	46.50	−0.48	48.17	+1.20	39.62	40.01	+0.38	33.71	−5.91
Llama-3.2-3B	RTN	6	3	48.01	45.14	−2.86	46.68	−1.32	37.62	41.26	+3.64	35.09	−2.53
	HQQ			48.58	58.16	+9.58	57.68	+9.10	23.86	15.49	−8.38	15.09	−8.77
	LeanQuant			62.81	63.35	+0.54	62.92	+0.10	12.39	12.70	+0.31	12.30	−0.08
	GPTQ			52.66	51.79	−0.87	53.61	+0.95	25.31	19.42	−5.89	19.25	−6.06
Llama-3.1-8B	RTN	9	5	54.14	56.56	+2.41	55.95	+1.80	28.57	20.91	−7.67	20.97	−7.60
	HQQ			60.12	65.12	+5.00	65.44	+5.32	14.72	12.09	−2.63	12.00	−2.72
	LeanQuant			70.30	70.08	−0.23	69.97	−0.33	10.17	10.33	+0.16	10.29	+0.11
	GPTQ			57.77	64.84	+7.07	65.99	+8.22	13.44	13.61	+0.16	13.62	+0.18
Qwen3-1.7B	RTN	12	4	43.10	42.12	−0.98	43.38	+0.28	58.11	60.57	+2.46	51.10	−7.00
	HQQ			34.34	42.80	+8.46	41.92	+7.59	341.19	55.75	−285.43	51.61	−289.57
	LeanQuant			53.96	54.54	+0.58	53.61	−0.35	28.51	28.60	+0.09	27.93	−0.57
	GPTQ			44.59	44.68	+0.09	45.23	+0.64	41.06	46.08	+5.02	39.77	−1.29
Qwen3-4B	RTN	4	1	52.45	51.81	−0.64	52.11	−0.34	34.21	34.60	+0.39	34.48	+0.27
	HQQ			51.81	55.29	+3.49	55.59	+3.79	31.47	28.10	−3.36	27.31	−4.16
	LeanQuant			64.19	65.05	+0.85	65.34	+1.15	19.91	19.88	−0.02	19.94	+0.03
	GPTQ			58.36	58.03	−0.33	58.73	+0.36	24.50	25.28	+0.77	24.74	+0.23
Qwen3-8B	RTN	4	1	62.91	62.50	−0.41	63.05	+0.13	18.30	18.88	+0.58	18.47	+0.18
	HQQ			61.04	64.79	+3.75	65.00	+3.96	20.69	18.71	−1.97	18.48	−2.21
	LeanQuant			70.47	70.11	−0.36	70.09	−0.38	14.66	14.68	+0.02	14.63	−0.04
	GPTQ			69.08	68.91	−0.17	69.39	+0.30	16.14	16.46	+0.32	16.16	+0.02

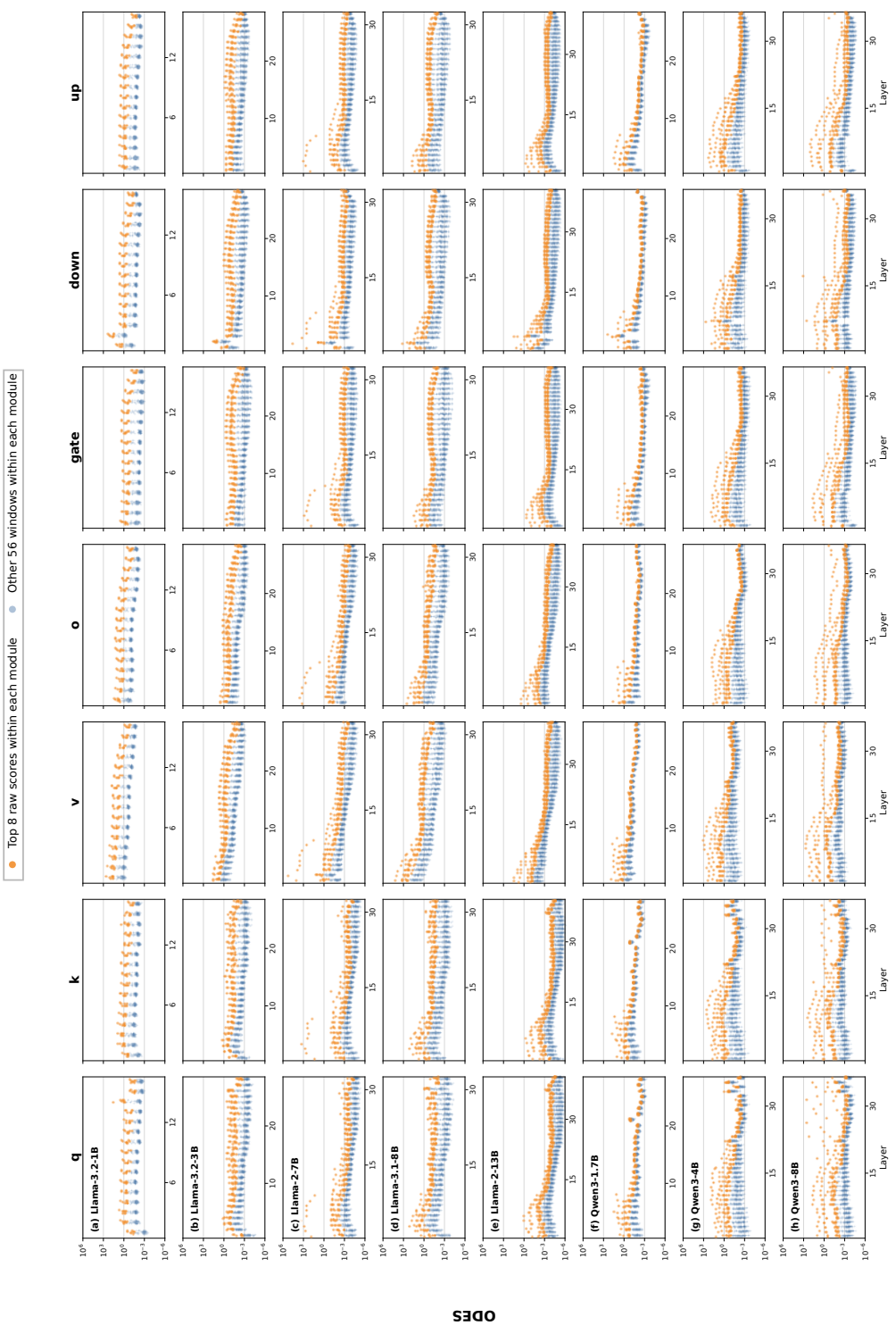


Figure C.2: Projection-wise ODES scores across profiling windows for eight models, extending Figure 2.