PPCC: PRICE-PRESERVING CALIBRATION CORESETS FOR POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibration coreset selection is essential for efficient post-training quantization (PTQ) of large language models, as limited calibration data must support reliable layer-wise choices of bitwidths, quantization granularity, and pre-quantization transformations. However, existing criteria based on data diversity or activation coverage do not directly characterize how calibration errors affect these decisions. To address this, we propose Price-Preserving Calibration Coresets (PPCC), a framework that reinterprets calibration selection through the preservation of pairwise quantization price gaps. This perspective decomposes price-gap distortion into distribution shift, finite-pool estimation, and coreset compression, where the compression term, which coreset selection controls, motivates a price-weighted moment-matching objective. Building on this foundation, inspired by traditional herding algorithms, we introduce an efficient greedy algorithm with a correction term derived by minimizing the expected terminal discrepancy. Theoretically, we establish finite-pool concentration bounds and prove that the resulting price-weighted squared moment discrepancy is no larger than its expectation under uniform random selection at the same budget. Empirical results show that PPCC recovers the full-pool price minimizer at all three tested budgets on Llama-3.2-1B and improves average downstream task scores in most configuration selection settings, with perplexity reductions of up to 35.2% over the best baseline on Llama-3.1-8B.

1 INTRODUCTION

Post-training quantization (PTQ) relies on a small calibration coreset to estimate the statistics used by methods such as GPTQ, AWQ, and SmoothQuant (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023). In budget-constrained quantization, these statistics also guide the selection of PTQ configurations, including bitwidths, granularities, and pre-quantization transformations (Qiu et al., 2026). Given a finite calibration pool $\mathcal{U}$ of size M, selecting a coreset $\mathcal{S}$ of size $N \ll M$ is therefore important for obtaining a high-quality quantized model with limited calibration cost (He et al., 2025).

Existing approaches characterize calibration quality through different properties of the selected data. COLA (He et al., 2025) empirically examines how the calibration distribution, sample count, and sequence length influence quantization performance. ZipCal (Monaco et al., 2026) emphasizes token-frequency statistics and lexical diversity, while COVERCAL (Shihab et al., 2026) formulates sample selection as weighted set cover over outlier channels. We provide a discussion of related work on calibration selection in Appx. A.

Despite these advances, the connection between calibration quality and quantization configuration decisions remains insufficiently understood. PricePTQ (Qiu et al., 2026) provides an efficient price-based framework for evaluating different PTQ configurations and then obtaining the optimal configurations that perform best. However, it treats the calibration set as fixed and does not analyze how calibration selection affects these comparisons. Moreover, criteria based on data diversity or coverage do not directly specify which statistics distinguish the feasible configurations. This motivates a question: *what should a calibration coreset preserve so that it supports the same optimal configuration decisions?*

Our **first** contribution is to formalize calibration coreset selection through configuration decision preservation (Sec. 2, left panel of Fig. 1). We focus on preserving the relative prices that determine

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.

configuration preference instead of estimating every configuration's absolute price equally well. Specifically, we formulate calibration selection as the Price-Preserving Calibration Coreset (PPCC) problem by minimizing the worst contraction of deployment price gaps over the calibration coreset (Sec. 2.1). To make these comparisons analytically tractable, we derive a curvature-weighted second-order price representation that separates activation statistics from candidate responses (Prop. 2.2).

Building on this foundation, our **second** contribution is an analytical error bound that decomposes price-gap distortion into distribution shift, finite-pool estimation, and coreset compression (Thm. 3.1, middle panel of Fig. 1). For fixed reference candidate responses, the distribution shift term identifies how activation covariance mismatch affects candidate comparisons (Prop. 3.3). The finite-pool term yields a high-probability concentration bound under sub-Gaussian activation models, together with a sufficient pool size depending logarithmically on the number of configuration pairs (Thm. 3.4). The coreset term is bounded by sensitivity-weighted activation moment mismatch and an explicit off-diagonal residual (Thm. 3.5). Crucially, the coreset term is the only term which is directly controlled by sample selection.

Inspired by herding algorithms (Welling, 2009), our **third** contribution is PPCC (Alg. 1, right panel of Fig. 1), an efficient algorithm derived from this moment-matching bound (Thm. 3.5 and Lem. 4.1). The correction is obtained by minimizing the conditional expected terminal J_2 under a uniform completion of the remaining budget. Theoretically, PPCC is a derandomization of uniform sampling, where its price-weighted squared moment discrepancy is provably no larger than the expected discrepancy of a uniform random subset of the same size (Thm. 4.2). Empirically, PPCC achieves the highest configuration ranking agreement and recovers the full-pool price minimizer at all three tested budgets on Llama-3.2-1B (Sec. 5.1). Besides, on pre-deployment configuration selection experiments, PPCC outperforms Random, ZipCal, and CoverCal in average downstream task score in most configuration selection settings, with perplexity reductions of up to 35.2% over the best baseline on Llama-3.1-8B (Sec. 5.2 and Appx. D).

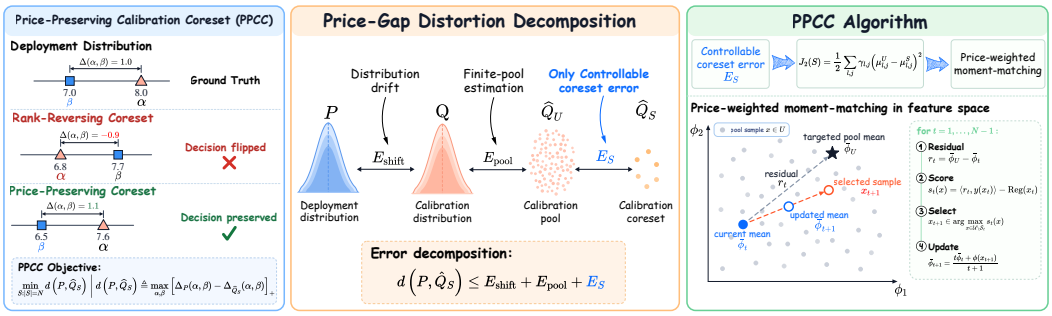

Figure 1: Overview of the Price-Preserving Calibration Coreset (PPCC) framework, including the problem setting, the price-gap distortion decomposition, and the proposed PPCC method (Alg. 1).

2 Price-Preserving Calibration Coresets

For configuration selection, a calibration coreset should preserve the configuration preferred under the deployment distribution. This decision depends on pairwise objective gaps, where calibration errors that shrink these gaps enough to reverse a comparison can change the selected configuration. We first formalize this principle as the Price-Preserving Calibration Coreset (PPCC) problem (Sec. 2.1), then derive a per-sample price approximation that links configuration comparisons to activation statistics (Sec. 2.2). The left panel of Fig. 1 gives an intuitive view of this distinction, where contracting a positive deployment price gap can reverse the preferred configuration, whereas preserving the gap preserves the decision.

2.1 Problem Formulation

Consider a full-precision reference model π_0 with L quantizable linear layers. For layer $l \in [L]$, let $\mathbf{W}_l$ denote its weight matrix, with input width n_l, and let $\mathcal{A}_l$ be a finite set of candidate PTQ configurations. A choice $\alpha_l \in \mathcal{A}_l$ specifies the bitwidth, numerical finite format, granularity, clipping

or scaling rule, and the pre-quantization transformation. The corresponding global configuration is $\alpha = (\alpha_1, \ldots, \alpha_L) \in \mathcal{A} \triangleq \prod_{l=1}^{L} \mathcal{A}_l$. Moreover, each local choice incurs a deployment cost $\kappa_l(\alpha_l)$, which may represent memory, latency or hardware compatibility.

For a sample $\mathbf{x}$ drawn from the deployment distribution P encountered by the quantized model in the real world, the full-precision model and the model quantized under configuration α induce output distributions $\pi_0(\cdot \mid \mathbf{x})$ and $\pi_\alpha(\cdot \mid \mathbf{x})$, respectively. Thus, the population objective is defined as the expected forward KL divergence between the full-precision and quantized models:

$$\mathcal{J}_P(\alpha) \triangleq \mathbb{E}_{\mathbf{x}\sim P}\left[D_{\mathrm{KL}}\left(\pi_0(\cdot \mid \mathbf{x}) \,\|\, \pi_\alpha(\cdot \mid \mathbf{x})\right)\right] , \tag{1}$$

which uses the full-precision model as reference and measures output distortion from quantization.

Given a global budget C, the deployment-optimal configuration is therefore

$$\alpha_P^* \in \arg\min_{\alpha\in\mathcal{F}} \mathcal{J}_P(\alpha) , \qquad \mathcal{F} \triangleq \left\{ \alpha \in \mathcal{A} : \sum_{l=1}^{L} \kappa_l(\alpha_l) \le C \right\} , \tag{2}$$

where $\mathcal{F}$ is the feasible configuration family under the budget constraint.

In practice, the expectation in (1) is ideal and unavailable, since the deployment distribution P is unknown. Only a noisy calibration coreset $\mathcal{S}$ is available for post-training quantization, which may lead to sub-optimal configuration selection. Therefore, a high-quality calibration coreset is crucial for preserving the optimal configuration selection and ensuring the performance of the quantized model.

To better formalize practical calibration coreset selection, let Q denote the calibration distribution, which need not coincide with P. The selector has access only to a finite calibration pool $\mathcal{U} = \{\mathbf{x}_i\}_{i=1}^{M}$, $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} Q$ and selects a calibration coreset $\mathcal{S} \subseteq \mathcal{U}$ containing at most N samples, where $M \gg N$. The pool and the selected coreset induce the uniform empirical distributions:

$$\widehat{Q}_{\mathcal{U}} \triangleq \frac{1}{M}\sum_{\mathbf{x}\in\mathcal{U}} \delta_{\mathbf{x}} , \qquad \widehat{Q}_{\mathcal{S}} \triangleq \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{S}} \delta_{\mathbf{x}} , \tag{3}$$

where $\delta_{\mathbf{x}}$ denotes the Dirac point mass at $\mathbf{x}$.

Replacing P with $\widehat{Q}_{\mathcal{S}}$ in (1) gives empirical optimal configuration selection under the objective:

$$\alpha_{\mathcal{S}}^* \in \arg\min_{\alpha\in\mathcal{F}} \mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\alpha) , \qquad \mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\alpha) \triangleq \mathbb{E}_{\mathbf{x}\sim\widehat{Q}_{\mathcal{S}}}[D_{\mathrm{KL}}(\pi_0(\cdot \mid \mathbf{x}) \,\|\, \pi_\alpha(\cdot \mid \mathbf{x}))] . \tag{4}$$

The most direct requirement for successful calibration is decision preservation, namely $\alpha_{\mathcal{S}}^* = \alpha_P^*$, which may hold even when the empirical objective substantially distorts the relative ordering among all remaining configurations. This condition captures the final decision of interest, but it constrains only the winning configuration for one particular candidate set. Besides, the population optimal configuration α_P^* is unknown, so the selector cannot directly verify whether $\alpha_{\mathcal{S}}^* = \alpha_P^*$.

We therefore study the stronger requirement that the calibration coreset preserve the complete ordering induced by the target objective. Specifically, $\mathcal{S}$ is said to be order preserving over $\mathcal{F}$ if

$$\mathcal{J}_P(\alpha) < \mathcal{J}_P(\beta) \quad \Longleftrightarrow \quad \mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\alpha) < \mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\beta), \quad \forall \alpha, \beta \in \mathcal{F} , \tag{5}$$

which is sufficient for $\alpha_{\mathcal{S}}^* = \alpha_P^*$, since taking $\alpha = \alpha_P^*$ shows that $\mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\alpha_P^*)$ is smaller than $\mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\beta)$ for every $\beta \neq \alpha_P^*$.

To quantify how well a calibration coreset preserves the configuration ordering, we consider the distortion of pairwise objective gaps. For any two distinct configurations $\alpha, \beta \in \mathcal{F}$, define the population and empirical gaps as

$$\Delta_P(\alpha, \beta) \triangleq \mathcal{J}_P(\alpha) - \mathcal{J}_P(\beta), \qquad \Delta_{\widehat{Q}_{\mathcal{S}}}(\alpha, \beta) \triangleq \mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\alpha) - \mathcal{J}_{\widehat{Q}_{\mathcal{S}}}(\beta) , \tag{6}$$

and the one-sided pairwise price-gap distortion of a calibration coreset $\mathcal{S}$ is defined as:

$$d(P, \widehat{Q}_{\mathcal{S}}) \triangleq \sup_{(\alpha,\beta)\in\mathcal{O}_P} \left[\Delta_P(\alpha, \beta) - \Delta_{\widehat{Q}_{\mathcal{S}}}(\alpha, \beta)\right]_+ , \quad \mathcal{O}_P \triangleq \left\{(\alpha, \beta) \in \mathcal{F}^2 : \Delta_P(\alpha, \beta) > 0\right\} , \tag{7}$$

where $[\cdot]_+ \triangleq \max\{0, \cdot\}$ denotes the ReLU function and $\mathcal{O}_P$ is the set of configuration pairs ordered by the deployment distribution. Thus, $(\alpha, \beta) \in \mathcal{O}_P$ means that α has a higher deployment price than β. For every pair in $\mathcal{O}_P$, the quantity inside the ReLU function is the contraction of its deployment price gap on the selected coreset. Gap expansion contributes zero because it reinforces the configuration ordering instead of threatening it. By contrast, if the coreset reverses an ordered pair, so that $\Delta_{\widehat{Q}_\mathcal{S}}(\alpha, \beta) \leq 0$, the resulting distortion is at least the original deployment gap.

Lemma 2.1 (Top-1 Preservation Sufficient Condition). *Suppose that α_P^* is unique and define its deployment margin by*

$$\gamma_P^* \triangleq \min_{\beta \in \mathcal{F} \setminus \{\alpha_P^*\}} \Delta_P(\beta, \alpha_P^*) > 0 \,,$$

and if $d(P, \widehat{Q}_\mathcal{S}) < \gamma_P^*$, we have $\alpha_\mathcal{S}^* = \alpha_P^*$.

We provide the proof in Appx. B.1. Lem. 2.1 shows that sufficiently small pairwise gap distortion guarantees the selection consistency. This observation motivates the Price-Preserving Calibration Coreset (PPCC) problem, which selects a calibration coreset $\mathcal{S}$ from the accessible pool $\mathcal{U}$ that minimizes the pairwise price-gap distortion:

$$\mathcal{S}^* = \arg\min_{\substack{\mathcal{S} \subseteq \mathcal{U} \\ |\mathcal{S}| = N}} d(P, \widehat{Q}_\mathcal{S}) \,. \tag{8}$$

However, the distortion $d(P, \widehat{Q}_\mathcal{S})$ is not directly computable since the KL objective in (1) involves heavy computation and the deployment distribution P and $\mathcal{O}_P$ are unknown. In Sec. 2.2, we introduce a per-sample price representation that transfers the KL objective to a curvature-weighted quantization price whose sample dependence enters only through layerwise input covariances. Besides, the distortion in (8) decomposes into three error terms that are analyzed through the finite calibration pool and the selected coreset in Sec. 3.

2.2 Per-Sample Quantization Price

For a local candidate α_l, let $\mathbf{T}_l(\alpha_l)$ be its invertible input-side transformation, with $\mathbf{T}_l(\alpha_l) = \mathbf{I}$ for untransformed candidates. The candidate-coordinate weight and activation are

$$\mathbf{W}_l(\alpha_l) = \mathbf{W}_l \mathbf{T}_l(\alpha_l)^{-1}, \qquad \mathbf{x}_l(\alpha_l) = \mathbf{T}_l(\alpha_l)\mathbf{x}_l \,. \tag{9}$$

Quantizing $\mathbf{W}_l(\alpha_l)$ in this coordinate gives $\widehat{\mathbf{W}}_l(\alpha_l)$ and the perturbation

$$\Delta\mathbf{W}_l(\alpha_l) \triangleq \widehat{\mathbf{W}}_l(\alpha_l) - \mathbf{W}_l(\alpha_l) \,. \tag{10}$$

Let $\mathbf{H}_{z,l} = \nabla^2_{z_l} \mathcal{J}_P(\alpha) \succeq 0$ denote the reference-side output curvature. For samples of a common sequence length T, the reference activation matrix $\mathbf{x}_l \in \mathbb{R}^{n_l \times T}$ defines the input covariance

$$\mathbf{C}_l(\mathbf{x}) \triangleq \frac{1}{T} \mathbf{x}_l \mathbf{x}_l^\top \,. \tag{11}$$

These quantities yield a curvature-weighted quadratic approximation to the KL objective:

Proposition 2.2 (Per-Sample Quantization Price). *Assume that the forward-KL objective is three times differentiable along the perturbation path from the full-precision model to configuration α. For $R \in \{P, Q, \widehat{Q}_\mathcal{U}, \widehat{Q}_\mathcal{S}\}$, its Taylor expansion gives*

$$\mathcal{J}_R(\alpha) = \rho_R(\alpha) + r_R(\alpha) \,, \qquad \rho_R(\alpha) \triangleq \mathbb{E}_{\mathbf{x} \sim R}\left[p(\mathbf{x}; \alpha)\right] \,, \qquad p(\mathbf{x}; \alpha) \triangleq \sum_{l=1}^{L} p_l(\mathbf{x}; \alpha_l) \,.$$

The first-order term vanishes at the full-precision reference, and the retained quadratic price is

$$p_l(\mathbf{x}; \alpha_l) \triangleq \frac{1}{2} \left\langle \mathbf{M}_l(\alpha_l), \mathbf{C}_l(\mathbf{x}) \right\rangle \,, \quad \mathbf{M}_l(\alpha_l) \triangleq \mathbf{T}_l(\alpha_l)^\top \Delta\mathbf{W}_l(\alpha_l)^\top \mathbf{H}_{z,l} \Delta\mathbf{W}_l(\alpha_l) \mathbf{T}_l(\alpha_l) \,,$$

where $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$. The factors of $\mathbf{T}_l(\alpha_l)$ express all candidate responses in the shared reference coordinate of $\mathbf{C}_l(\mathbf{x})$. The residual $r_R(\alpha)$ collects omitted cross-layer, curvature-approximation, and higher-order terms.

Remark. We provide the proof in Appx. B.2. The representation separates candidate responses $\mathbf{M}_l(\alpha_l)$ from sample statistics $\mathbf{C}_l(\mathbf{x})$. In the residual analysis of Appx. C.1, a trace proxy of the quadratic price attains Spearman correlations of 0.91 and 0.96 with the empirical forward KL on Qwen3-0.6B and Llama-3.2-1B. We thus neglect residual r_R and use the quadratic price ρ_R in place of $\mathcal{J}_R$ for the remainder of the analysis.

3 PRICE-GAP DISTORTION ANALYSIS

Directly computing the pairwise price-gap distortion $d(P, \widehat{Q}_{\mathcal{S}})$ is impossible since the deployment distribution P is unknown. Thus, we first decompose the distortion into three terms (Sec. 3.1), including the distribution shift term (Sec. 3.2), the finite-pool estimation term (Sec. 3.3), and the coreset compression term (Sec. 3.4), which the selector controls and which guides the algorithm design in Sec. 4. The middle panel of Fig. 1 visualizes these three terms.

3.1 PRICE-PRESERVING ERROR DECOMPOSITION

For any distribution R appearing below, we write $\Delta_R(\alpha, \beta)$ for the corresponding configuration price gap, following the notation in Sec. 2.1.

Theorem 3.1 (Error Decomposition). *For any distribution $R, R' \in \{P, Q, \widehat{Q}_{\mathcal{U}}, \widehat{Q}_{\mathcal{S}}\}$, analogous to (7), we define the distributional distortion:*

$$d(R, R') \triangleq \sup_{(\alpha,\beta)\in\mathcal{O}_P} [\Delta_R(\alpha, \beta) - \Delta_{R'}(\alpha, \beta)]_+ \ .$$

Thus, for any $\mathcal{S} \subseteq \mathcal{U}$ with $|\mathcal{S}| = N$, we have

$$d(P, \widehat{Q}_{\mathcal{S}}) \leq \underbrace{d(P, Q)}_{E_{\text{shift}}} + \underbrace{d(Q, \widehat{Q}_{\mathcal{U}})}_{E_{\text{pool}}} + \underbrace{d(\widehat{Q}_{\mathcal{U}}, \widehat{Q}_{\mathcal{S}})}_{E_{\mathcal{S}}} \ .$$

We provide the proof in Appx. B.3. E_{shift} reflects the mismatch between P and Q, while E_{pool} arises from observing only M samples from Q. Conditional on the realized pool, only the deterministic compression error $E_{\mathcal{S}}$ is directly controlled by selection. Besides, to isolate changes in candidate evaluation from changes caused by quantizer fitting, we use a shared reference response family:

Assumption 3.2 (Fixed Candidate Responses). For every layer l and local candidate α_l, the reference response matrix $\mathbf{M}_l(\alpha_l)$ is specified before coreset selection and is used unchanged when evaluating prices under P, Q, $\widehat{Q}_{\mathcal{U}}$, and $\widehat{Q}_{\mathcal{S}}$. In particular, it is not refitted when the selected coreset changes. Sample dependence enters the reference prices only through $\mathbf{C}_l(\mathbf{x})$.

Assump. 3.2 decouples coreset selection from quantizer fitting, so that all candidate configurations are compared under a common response family. It can be realized by constructing reference responses from fixed model weights or an auxiliary reference set and then freezing them.

3.2 DISTRIBUTION SHIFT

Proposition 3.3 (Distribution Shift Expression). *For any distribution R and a configuration pair $(\alpha, \beta) \in \mathcal{F}^2$, define the layer covariance*

$$\mathbf{C}_l^R \triangleq \mathbb{E}_{\mathbf{x}\sim R}[\mathbf{C}_l(\mathbf{x})] \ , \qquad \mathbf{D}_l^{\alpha,\beta} \triangleq \mathbf{M}_l(\alpha_l) - \mathbf{M}_l(\beta_l) \ ,$$

the distribution-shift term satisfies:

$$E_{\text{shift}} = \frac{1}{2} \sup_{(\alpha,\beta)\in\mathcal{O}_P} \left[\sum_{l=1}^{L} \left\langle \mathbf{D}_l^{\alpha,\beta}, \mathbf{C}_l^P - \mathbf{C}_l^Q \right\rangle \right]_+ .$$

Remark. We provide the proof in Appx. B.4. Prop. 3.3 shows that the distribution-shift term is governed by the price-sensitive mismatch between the layerwise input covariances induced by the deployment distribution P and the calibration distribution Q. If P were known, this characterization would provide a principled criterion for choosing Q by minimizing this mismatch over an admissible family of calibration distributions. This perspective is consistent with empirical findings that calibration data substantially affect post-training compression and that pretraining-like or self-generated data are effective calibration sources (Ji et al., 2025; Bandari et al., 2024; Williams et al., 2025; Williams & Aletras, 2024; Paglieri et al., 2024). In our experiments, we therefore use C4 as a standard pretraining-like calibration source, following its common use in PTQ and calibration studies (Frantar et al., 2023; Lin et al., 2024; Bandari et al., 2024).

3.3 FINITE-POOL ESTIMATION ERROR

The finite-pool estimation term E_{pool} in Thm. 3.1 is induced by the finite-size calibration pool $\mathcal{U}$. For distinct configurations $\alpha, \beta \in \mathcal{F}$, write $q(\mathbf{x}; \alpha, \beta) = p(\mathbf{x}; \alpha) - p(\mathbf{x}; \beta)$. The finite-pool term is the largest downward estimation error in a fixed price gap:

$$E_{\mathrm{pool}} = \sup_{(\alpha,\beta)\in\mathcal{O}_P} \left[\mathbb{E}_{\mathbf{x}\sim Q}[q(\mathbf{x};\alpha,\beta)] - \frac{1}{M}\sum_{i=1}^{M} q(\mathbf{x}_i;\alpha,\beta) \right]_+ . \tag{12}$$

Thus, a price difference is a signed quadratic form in the reference activations, whose concentration is governed by the activation tails. We model each activation coordinate as sub-Gaussian, a standard condition that covers Gaussian and bounded activations, where heavier channels such as activation outliers enter only through larger scale constants in ν and b. Thm. 3.4 below gives a uniform concentration bound on the finite-pool estimation term E_{pool}:

Theorem 3.4 (Informal, Finite-Pool Concentration). *Let $K = |\mathcal{F}| \geq 2$, let $\mathcal{P}$ denote the unordered pairs of distinct feasible configurations, and let $J = |\mathcal{P}| = K(K-1)/2$. Under Assump. 3.2, suppose that each activation coordinate is marginally sub-Gaussian under Q. Then there exist constants $\nu > 0$ and $b > 0$ such that, for every $\eta \in (0,1)$, with probability at least $1-\eta$:*

$$E_{\mathrm{pool}} \leq B_{\mathrm{pool}}(M,\eta) \triangleq \sqrt{\frac{2\nu^2}{M}\log\frac{J}{\eta}} + \frac{b}{M}\log\frac{J}{\eta} .$$

For any $\varepsilon_{\mathrm{pool}} > 0$, a sufficient pool size for $E_{\mathrm{pool}} \leq \varepsilon_{\mathrm{pool}}$ with probability at least $1-\eta$ is:

$$M \gtrsim \left(\frac{\nu^2}{\varepsilon_{\mathrm{pool}}^2} + \frac{b}{\varepsilon_{\mathrm{pool}}} \right) \log\frac{J}{\eta} .$$

Remark. The formal version is given in Thm. B.2, and the proof is provided in Appx. B.5. Thm. 3.4 shows that the finite-pool estimation term E_{pool} concentrates around its mean with high probability, and the bound depends logarithmically on the number of configuration pairs J and inversely on the pool size M. Equivalently, Thm. 3.4 gives a sufficient pool size guarantee.

3.4 CORESET COMPRESSION AND MOMENT MATCHING

$E_{\mathcal{S}}$ in Thm. 3.1 is the compression error induced by coreset selection. Define the pool and coreset covariances and their residual:

$$\mathbf{C}_l^U \triangleq \frac{1}{M}\sum_{\mathbf{x}\in\mathcal{U}} \mathbf{C}_l(\mathbf{x}) , \qquad \mathbf{C}_l^S \triangleq \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{S}} \mathbf{C}_l(\mathbf{x}) , \qquad \Delta\mathbf{C}_l(\mathcal{S}) \triangleq \mathbf{C}_l^U - \mathbf{C}_l^S , \tag{13}$$

which yields a tractable moment bound:

Theorem 3.5 (Price-Weighted Moment Upper Bound). *For layer l and input channel j, define the sample moment $c_{l,j}(\mathbf{x}) \triangleq [\mathbf{C}_l(\mathbf{x})]_{jj}$, its full-pool target and coreset estimate are:*

$$\mu_{l,j}^U \triangleq \frac{1}{M}\sum_{\mathbf{x}\in\mathcal{U}} c_{l,j}(\mathbf{x}) , \qquad \mu_{l,j}^S \triangleq \frac{1}{N}\sum_{\mathbf{x}\in\mathcal{S}} c_{l,j}(\mathbf{x}) . \tag{14}$$

For every $\mathcal{S} \subseteq \mathcal{U}$, we have:

$$E_{\mathcal{S}} \leq R_{\mathrm{off}}(\mathcal{S}) + J_1(\mathcal{S}) , \qquad J_1(\mathcal{S}) \triangleq \sum_{l,j} \gamma_{l,j} \left| \mu_{l,j}^U - \mu_{l,j}^S \right| ,$$

where the off-diagonal residual $R_{\mathrm{off}}(\mathcal{S})$ and the worst-case sensitivity $\gamma_{l,j}$ are defined as:

$$R_{\mathrm{off}}(\mathcal{S}) \triangleq \frac{1}{2}\sup_{\alpha,\beta\in\mathcal{F}} \left| \sum_{l=1}^{L} \left\langle \mathrm{off}(\mathbf{D}_l^{\alpha,\beta}), \mathrm{off}(\Delta\mathbf{C}_l(\mathcal{S})) \right\rangle \right| , \quad \gamma_{l,j} \triangleq \frac{1}{2}\sup_{\alpha,\beta\in\mathcal{F}} \left| [\mathbf{D}_l^{\alpha,\beta}]_{jj} \right| ,$$

where $\mathrm{off}(\mathbf{A})$ sets the diagonal of $\mathbf{A}$ to zero.

Remark. We provide the proof in Appx. B.6. Thm. 3.5 shows that the coreset compression term $E_{\mathcal{S}}$ can be upper bounded by the off-diagonal residual $R_{\text{off}}(\mathcal{S})$ and the weighted absolute moment discrepancy $J_1(\mathcal{S})$. On Llama-3.2-1B, $R_{\text{off}}(\mathcal{S})$ is about $12\times$ smaller than $J_1(\mathcal{S})$, with median ratios between 0.085 and 0.088 across all evaluated budgets (Appx. C.3), supporting $J_1(\mathcal{S})$ as a tractable selector surrogate. Besides, the moment discrepancy $\mu^U_{l,j} - \mu^S_{l,j}$ indicates the coreset selection problem is actually a moment matching problem, which requires the coreset to match the moments of the full calibration pool, further discussed in Sec. 4. Specifically, the sensitivity $\gamma_{l,j}$ can be computed directly without enumerating all configuration pairs:

$$\gamma_{l,j} = \frac{1}{2}\left(\max_{a\in\mathcal{F}}[\mathbf{M}_l(a)]_{jj} - \min_{a\in\mathcal{F}}[\mathbf{M}_l(a)]_{jj}\right) . \tag{15}$$

4 Price-Weighted Moment Matching

Thm. 3.5 shows that a calibration coreset should match its moments $\mu^S_{l,j}$ to the full calibration pool moments $\mu^U_{l,j}$, weighted by the sensitivities $\gamma_{l,j}$. We first use this bound to construct a coreset selection algorithm inspired by the herding algorithm.

Connection to Herding (Welling, 2009). Herding is a greedy algorithm for constructing a discrete approximation to a target distribution by sequentially selecting points that minimize the discrepancy between the empirical and target means in a reproducing kernel Hilbert space (RKHS). Since Thm. 3.5 turns the PPCC problem (8) into a moment-matching problem, we can apply herding to select a coreset that approximates the full calibration pool moments. The absolute moment discrepancy J_1 provides the direct compression bound, but its coordinatewise form does not yield a single Euclidean terminal potential for a fixed-size selection problem. We therefore use the quadratic moment objective:

$$J_2(\mathcal{S}) \triangleq \frac{1}{2}\sum_{l,j}\gamma_{l,j}\left(\mu^U_{l,j} - \mu^S_{l,j}\right)^2 , \qquad \Gamma_{\text{tot}} \triangleq \sum_{l,j}\gamma_{l,j} , \tag{16}$$

that still upper bounds the coreset compression error $E_{\mathcal{S}}$ in Thm. 3.5:

Lemma 4.1 (Price-Weighted Moment Upper Bound by J_2)**.** *For every nonempty $\mathcal{S} \subseteq \mathcal{U}$:*

$$J_1(\mathcal{S}) \le \sqrt{2\Gamma_{\text{tot}} J_2(\mathcal{S})}, \qquad E_{\mathcal{S}} \le R_{\text{off}}(\mathcal{S}) + \sqrt{2\Gamma_{\text{tot}} J_2(\mathcal{S})} .$$

The proof is in Appx. B.7. Furthermore, flatten the layer-channel coordinates into the feature vector:

$$\boldsymbol{\phi}(\mathbf{x}) \triangleq \left[\sqrt{\gamma_{l,j}}c_{l,j}(\mathbf{x})\right]_{l,j} \in \mathbb{R}^D, \qquad D = \sum_l n_l , \tag{17}$$

the feature map induces a price-aware linear kernel:

$$k_{\text{PPCC}}(\mathbf{x},\mathbf{x}') = \langle\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\mathbf{x}')\rangle = \sum_{l,j}\gamma_{l,j}c_{l,j}(\mathbf{x})c_{l,j}(\mathbf{x}') , \tag{18}$$

which forms the RKHS in which herding is performed.

Algorithm Design. We select a coreset $\mathcal{S}$ of size N from the calibration pool $\mathcal{U}$ by adding one sample at a time. At iteration t, let $\mathcal{S}_t$ contain t distinct pool elements, including $\mathcal{S}_0 = \varnothing$, and define

$$\overline{\boldsymbol{\phi}}_U \triangleq \frac{1}{M}\sum_{\mathbf{x}\in\mathcal{U}}\boldsymbol{\phi}(\mathbf{x}) , \qquad \overline{\boldsymbol{\phi}}_t \triangleq \frac{1}{t}\sum_{\mathbf{x}\in\mathcal{S}_t}\boldsymbol{\phi}(\mathbf{x}) , \qquad \mathbf{r}_t \triangleq \overline{\boldsymbol{\phi}}_U - \overline{\boldsymbol{\phi}}_t . \tag{19}$$

Since the coreset has a fixed size and is constructed without replacement, the terminal discrepancy depends on the composition of the remaining pool. To account for this finite-population effect, let

$$\mathbf{y}(\mathbf{x}_i) \triangleq \boldsymbol{\phi}(\mathbf{x}_i) - \overline{\boldsymbol{\phi}}_U, \qquad \mathbf{h}_t \triangleq \sum_{\mathbf{x}\in\mathcal{S}_t}\mathbf{y}(\mathbf{x}), \qquad Q_t \triangleq \sum_{\mathbf{x}\in\mathcal{U}\setminus\mathcal{S}_t}\|\mathbf{y}(\mathbf{x})\|_2^2 . \tag{20}$$

For a partial coreset $\mathcal{S}_t$, define $\Phi_t(\mathcal{S}_t)$ as the conditional expected terminal J_2 after uniformly completing the remaining $N-t$ positions from $\mathcal{U}\setminus\mathcal{S}_t$. A finite-population calculation gives

$$\Phi_t(\mathcal{S}_t) = \frac{M-N}{2N^2(M-t)(M-t-1)}\left[(M-N-1)\|\mathbf{h}_t\|_2^2 + (N-t)Q_t\right] . \tag{21}$$

Minimizing the one-step look-ahead objective $\Phi_{t+1}(\mathcal{S}_t \cup \{\mathbf{x}\})$ is therefore equivalent to

$$\mathbf{x}_{t+1} \in \arg\max_{\mathbf{x}\in\mathcal{U}\setminus\mathcal{S}_t} \left\{ \langle \mathbf{r}_t, \mathbf{y}(\mathbf{x})\rangle - \frac{\lambda_t}{2t}\|\mathbf{y}(\mathbf{x})\|_2^2 \right\}, \qquad \lambda_t \triangleq \frac{M-2N+t}{M-N-1}, \tag{22}$$

where the first sample simply minimizes $\|\mathbf{y}(\mathbf{x})\|_2^2$.

Overall, we summarize the complete PPCC algorithm in Alg. 1 and the right panel of Fig. 1 illustrates the complete algorithm. Given the moment features $\{c_{l,j}(\mathbf{x})\}$ and the sensitivities $\{\gamma_{l,j}\}$, Alg. 1 runs in $O(NMD)$ time with $O(MD)$ memory complexity, and is not required to form an $M \times M$ kernel matrix.

Algorithm 1 Price-Preserving Calibration Coreset (PPCC) Algorithm

Require: Pool $\mathcal{U}$ of size $M \geq 4$, budget $1 \leq N \leq \lfloor M/2 \rfloor$, moments $\{c_{l,j}(\mathbf{x})\}$, sensitivities $\{\gamma_{l,j}\}$
Ensure: A size-N subset $\mathcal{S}_N$ with equal weights and no repeated pool elements

Construct features by (17); $\overline{\boldsymbol{\phi}}_U \leftarrow M^{-1}\sum_{\mathbf{x}\in\mathcal{U}} \boldsymbol{\phi}(\mathbf{x})$
$\mathbf{y}(\mathbf{x}) \leftarrow \boldsymbol{\phi}(\mathbf{x}) - \overline{\boldsymbol{\phi}}_U$; $e(\mathbf{x}) \leftarrow \|\mathbf{y}(\mathbf{x})\|_2^2$
Choose $\mathbf{x}_1 \in \arg\min_{\mathbf{x}\in\mathcal{U}} e(\mathbf{x})$
$\mathcal{S}_1 \leftarrow \{\mathbf{x}_1\}$; $\overline{\boldsymbol{\phi}}_1 \leftarrow \boldsymbol{\phi}(\mathbf{x}_1)$
for $t = 1, \ldots, N-1$ **do**
  $\mathbf{r}_t \leftarrow \overline{\boldsymbol{\phi}}_U - \overline{\boldsymbol{\phi}}_t$
  Choose $\mathbf{x}_{t+1}$ by (22), using cached $e(\mathbf{x})$ for $\|\mathbf{y}(\mathbf{x})\|_2^2$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{\mathbf{x}_{t+1}\}$; $\overline{\boldsymbol{\phi}}_{t+1} \leftarrow \big(t\overline{\boldsymbol{\phi}}_t + \boldsymbol{\phi}(\mathbf{x}_{t+1})\big)/(t+1)$
end for
return $\mathcal{S}_N$

Theoretical Analysis The following theorem guarantees that PPCC attains a price-weighted moment discrepancy no larger than the expected discrepancy of uniform random selection:

Theorem 4.2 (Moment-Matching Guarantee). *Suppose $M \geq 4$ and $1 \leq N \leq \lfloor M/2 \rfloor$. Let $\mathcal{S}_N^{\mathrm{PPCC}}$ be the coreset returned by Alg. 1, and let $e_{\min} \triangleq \min_{\mathbf{x}\in\mathcal{U}} \|\mathbf{y}(\mathbf{x})\|_2^2$ and $V_\phi \triangleq M^{-1}\sum_{\mathbf{x}\in\mathcal{U}} \|\mathbf{y}(\mathbf{x})\|_2^2$. Then the conditional completion potential is nonincreasing and*

$$J_2(\mathcal{S}_N^{\mathrm{PPCC}}) \leq \Phi_1 \leq \mathbb{E}[J_2(\mathcal{S}_N^{\mathrm{rand}}) \mid \mathcal{U}] .$$

Moreover,

$$\Phi_1 = \mathbb{E}[J_2(\mathcal{S}_N^{\mathrm{rand}}) \mid \mathcal{U}] - \frac{(M-N)(M-2N)}{2N^2(M-1)(M-2)} (V_\phi - e_{\min}) .$$

Thus the inequality is strict whenever $M > 2N$ and $e_{\min} < V_\phi$.

Remark. The proof is given in Appx. B.8. The theorem interprets PPCC as a derandomization of uniform sampling where each step selects the candidate that minimizes the expected terminal discrepancy under random completion. This yields a coreset whose J_2 is no larger than the expected J_2 of uniform random sampling at the same budget. Moreover, $V_\phi - e_{\min}$ identifies a source of strict improvement. When $M > 2N$, initializing with a sample closer to the pool mean than average reduces the completion potential, and subsequent selections preserve this gain.

5 Experiments

In this section, we evaluate the proposed PPCC algorithm by examining its ability to preserve configuration rankings and selection decisions (Sec. 5.1), and its impact on downstream performance in pre-deployment configuration selection for post-training quantization (Sec. 5.2).

5.1 Configuration Ranking and Decision Preservation

To evaluate whether PPCC preserves configuration rankings and selection decisions, we compare its performance against random selector, ZipCal (Monaco et al., 2026), and CoverCal (Shihab et al., 2026) baseline methods (see Appx. E for details). We perform the evaluation on Llama-3.2-1B

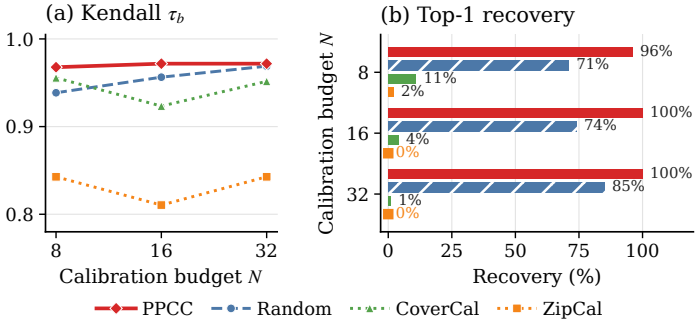

Figure 2: Configuration ranking and decision preservation on Llama-3.2-1B: (a) Kendall τ_b; (b) Top-1 recovery rate of the optimal configuration.

using a C4 pool of $M = 9646$ sequences at calibration budgets $N \in \{8, 16, 32\}$. The candidate configuration set consists of $K = 32$ distinct bitwidth allocations, each assigning INT3 to eight of the model's sixteen transformer blocks and INT4 to the remaining eight, with a fixed group size of 128.

As shown in Fig. 2, PPCC achieves the highest ranking agreement across the three calibration sizes, with Kendall's τ_b ranging from 0.9677 to 0.9718. Besides, PPCC recovers the optimal configuration in 96%, 100%, and 100% of cases at $N = 8, 16, 32$, respectively, compared with 11%, 4%, and 1% for CoverCal, almost 0% throughout for ZipCal, and Random succeeds in 71-85$\%$ of its draws. Consistent with Lem. 2.1, rank swaps among near-tied configurations do not change the selected optimum as long as the distortion stays below the deployment margin γ_P^*.

5.2 Pre-Deployment Configuration Selection

We then evaluate PPCC in the pre-deployment configuration selection for post-training quantization (PTQ), which aims to select the best layer-wise quantization configuration within average bitwidth constraints, as formulated in (2). We perform the evaluation on Llama-3.1-8B Base, comparing against Random, ZipCal (Monaco et al., 2026), and CoverCal (Shihab et al., 2026). Each calibration selector method chooses a coreset of $N = 32$ sequences from a C4 pool of $M = 9646$ sequences of length $T = 2048$. We consider bitwidth, transformation, granularity, and mixed (selecting from all three types of configuration changes) selection. The detailed experimental setup is in Appx. E.

Table 1: Pre-deployment configuration selection on Llama-3.1-8B Base. Avg. averages PIQA, WiC, ARC-C, and HellaSwag downstream tasks. Bold/underlined scores are best/second-best. Time is the calibration selection overhead in minutes on a single H100 GPU.

	Bitwidth			Transformation			Granularity			Mixed		
Method	PPL↓	Avg.↑	Time	PPL↓	Avg.↑	Time	PPL↓	Avg.↑	Time	PPL↓	Avg.↑	Time
Random	14.34	<u>60.47</u>	—	8.73	61.54	—	16.64	62.10	—	8.63	60.85	—
CoverCal	<u>10.96</u>	59.65	16.05	8.66	<u>61.97</u>	16.05	<u>13.13</u>	61.88	16.05	**8.38**	<u>61.17</u>	16.05
ZipCal	32.26	59.90	0.15	<u>8.62</u>	61.73	0.15	19.57	**63.97**	0.15	8.60	59.47	0.15
PPCC *(ours)*	**8.71**	**61.88**	11.80	**8.60**	**62.16**	11.93	**8.51**	<u>62.47</u>	11.82	<u>8.47</u>	**63.01**	12.40

As shown in Tab. 1, PPCC achieves the highest average task score and the lowest PPL in most of the selection settings. Specifically, in bitwidth and granularity selection, it reduces PPL by 20.5% and 35.2%, respectively, relative to the best baseline in each setting. In mixed selection, PPCC ranks first on all four downstream tasks, exceeding the strongest baseline by 1.840 percentage points in average score. In these settings, the configurations selected with PPCC coresets yield better downstream performance of the quantized model. Besides, although PPCC costs more selection time than model-free ZipCal since it does not require forward passes through the model, PPCC is more efficient than CoverCal, and achieves better performance than both baselines. Per-task results for Llama-3.1-8B Base and additional results on Qwen3-8B Base and Instruct are provided in Appx. D.

6 Conclusion

PPCC connects calibration selection with PTQ decisions through a three-term decomposition of price-gap distortion. The resulting moment bound motivates a herding-inspired selector whose quadratic moment discrepancy is provably no larger than that of uniform random sampling in expectation, and strictly smaller under a mild non-degeneracy condition. Experiments show accurate configuration rankings, recovery of the optimal configuration at all tested budgets, and downstream gains.

AI USE STATEMENT

We used AI tools to polish the writing and formatting of the manuscript. We did not use AI tools to formulate the research idea, develop or prove the theoretical arguments, design or implement the experiments, or draft the initial manuscript. The authors take full responsibility for the final content of this work, including all text, claims and code.

ETHICS STATEMENT

This work involves no human-subject data collection, releases no new human-subject dataset, and introduces no user-facing system. The experiments use existing public calibration and evaluation datasets and pretrained models identified in the paper.

REPRODUCIBILITY STATEMENT

The main paper specifies the PPCC formulation, algorithm, theoretical assumptions, and price-gap analysis, while Appx. B provides the proofs. Appx. E documents calibration pool construction, baselines, candidate configuration spaces, and quantization and evaluation settings. The datasets and pretrained models used in the experiments are publicly available and identified in the paper.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Proc. NeurIPS*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proc. ICML*, pp. 1355–1362, 2012. URL https://arxiv.org/abs/1203.4523.

Abhinav Bandari, Lu Yin, Cheng-Yu Hsieh, Ajay Kumar Jaiswal, Tianlong Chen, Li Shen, Ranjay Krishna, and Shiwei Liu. Is C4 dataset optimal for pruning? an investigation of calibration data for LLM pruning. In *Proc. EMNLP*, pp. 18089–18099. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1004. URL https://aclanthology.org/2024.emnlp-main.1004/.

Yutian Chen, Max Welling, and Alexander J. Smola. Super-samples from kernel herding. In *Proc. UAI*, pp. 109–116, 2010.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *Proc. ICLR*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Bowei He, Lihao Yin, Hui-Ling Zhen, Shuqi Liu, Han Wu, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. Preserving LLM capabilities through calibration data curation: From analysis to optimization. In *Proc. NeurIPS*, volume 38, 2025. doi: 10.52202/085713-1961. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Sifan Zhou, Jiangyong Yu, Chen Xu, Zhixuan Chen, Zhe Jiang, and Zhihang Yuan. OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. In *Proc. ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5cebc89b113920dbff7c79854ba765a3-Paper-Conference.pdf.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In *Proc. ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4466–4475. PMLR, 2021. URL https://proceedings.mlr.press/v139/hubara21a.html.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proc. ICML*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Beware of calibration data for pruning large language models. In *Proc. ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ede933e10afa991a10b6f36b6522129-Abstract-Conference.html.

Janghwan Lee, Minsoo Kim, Seungcheol Baek, Seok Joong Hwang, Wonyong Sung, and Jungwook Choi. Enhancing computation efficiency in large language models through weight and activation quantization. In *Proc. EMNLP*, pp. 14726–14739. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.910. URL https://aclanthology.org/2023.emnlp-main.910/.

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 *Proc. MLSys*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Francesco Pio Monaco, Elia Cunegatti, Flavio Vella, and Giovanni Iacca. Frequency matters: Fast model-agnostic data curation for pruning and quantization. In *Proc. EMNLP*, 2026. URL https://arxiv.org/abs/2603.16105.

Jaehoon Oh and Dokwan Oh. Beyond fixed-length calibration for post-training compression of LLMs. In *Proc. EMNLP Findings*, pp. 19355–19366. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1054. URL https://aclanthology.org/2025.findings-emnlp.1054/.

Davide Paglieri, Saurabh Dash, Tim Rocktäschel, and Jack Parker-Holder. Outliers and calibration sets have diminishing effect on quantization of modern LLMs, 2024. URL https://arxiv.org/abs/2405.20835.

Nilesh Prasad Pandey, Markus Nagel, Mart van Baalen, Yin Huang, Chirag Patel, and Tijmen Blankevoort. A practical mixed precision algorithm for post-training quantization, 2023. URL https://arxiv.org/abs/2302.05397.

Junbin Qiu, Jian Mu, Weitong Zhang, and Yao Shu. Predicting quantization price for selecting ptq configurations before deployment, 2026. URL https://arxiv.org/abs/2609.28270.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Coverage-based calibration for post-training quantization via weighted set cover over outlier channels, 2026. URL https://arxiv.org/abs/2604.24008.

Max Welling. Herding dynamical weights to learn. In *Proc. ICML*, pp. 1121–1128, Montreal Quebec Canada, June 2009. ACM. ISBN 978-1-60558-516-1.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proc. ACL*, pp. 10100–10118. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.544. URL https://aclanthology.org/2024.acl-long.544/.

Miles Williams, George Chrysostomou, and Nikolaos Aletras. Self-calibration for language model quantization and pruning. In *Proc. NAACL*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.509. URL https://aclanthology.org/2025.naacl-long.509/.

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 *Proc. ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

A Related Work

Post-Training Quantization (PTQ). PTQ methods improve low-precision operators through reconstruction and equivalent transformations. GPTQ (Frantar et al., 2023) compensates weight errors using approximate second-order information, while AWQ (Lin et al., 2024) protects salient weight channels through scaling informed by activation statistics. SmoothQuant (Xiao et al., 2023) redistributes quantization difficulty between weights and activations, QuaRot (Ashkboos et al., 2024) suppresses outliers through rotations that preserve full-precision computation, and OSTQuant (Hu et al., 2025) learns orthogonal and scaling transformations to improve distribution fitting. These methods improve individual quantization schemes, while choosing among their deployment configurations requires a separate selection criterion.

PTQ Configuration Selection. Configuration selection combines layerwise choices under deployment constraints. Integer-programming methods (Hubara et al., 2021) optimize bitwidth assignments under accuracy or compression budgets, while Pandey et al. (2023) use output signal-to-quantization-noise ratios to guide a search over hardware-compatible mixed-precision configurations. These approaches allocate precision within specified quantization schemes. PricePTQ (Qiu et al., 2026) broadens the comparison to heterogeneous bitwidths, granularities, and transformations by pricing candidate-induced layer-output error with reference-model curvature. However, selecting configurations from estimated scores leaves open whether the same choice is preserved after compressing the calibration pool. Because small price-gap errors can reverse closely competing choices, PPCC studies the calibration statistics needed to preserve the preferred configuration.

Calibration Coreset Selection. Calibration studies document substantial variation in downstream performance across data choices (Williams & Aletras, 2024). Self-calibration (Williams et al., 2025) uses model-generated text to approximate the pre-training distribution, while sequence-length-aware and state-aware length calibration (Lee et al., 2023; Oh & Oh, 2025) adjust context lengths and token-state composition. These approaches improve data construction, leaving subset selection for configuration comparisons as a separate problem.

Within a pool, COLA (He et al., 2025) selects representative activation patterns through clustering, while ZipCal (Monaco et al., 2026) greedily increases lexical coverage without target-model forward passes. COVERCAL (Shihab et al., 2026) optimizes weighted outlier-channel coverage, with a greedy $(1 - 1/e)$ approximation guarantee and a loss bound under a stylized clipping model. These criteria capture useful calibration properties but do not directly bound pairwise configuration price-gap distortion.

Herding and kernel herding (Welling, 2009; Chen et al., 2010) match feature expectations through a process connected to conditional-gradient optimization (Bach et al., 2012; Jaggi, 2013). PPCC uses this foundation to select equally weighted samples without replacement in sensitivity-weighted activation-moment space. Its finite-population correction makes it a derandomization of uniform sampling, with a price-weighted moment discrepancy provably no larger than that of random selection in expectation. PPCC further links coreset moment discrepancy to a margin-based condition for preserving the price minimizer, accounting explicitly for distribution shift, finite-pool error, and an off-diagonal residual.

B Proofs

B.1 Proof of Lem. 2.1

Proof. For any competitor $\beta \neq \alpha_P^*$, uniqueness of α_P^* gives $\Delta_P(\beta, \alpha_P^*) \geq \gamma_P^* > 0$, so $(\beta, \alpha_P^*) \in \mathcal{O}_P$. By the definition of $d(P, \widehat{Q}_\mathcal{S})$ and $u \leq [u]_+$,

$$\Delta_{\widehat{Q}_\mathcal{S}}(\beta, \alpha_P^*) \geq \Delta_P(\beta, \alpha_P^*) - d(P, \widehat{Q}_\mathcal{S}) \geq \gamma_P^* - d(P, \widehat{Q}_\mathcal{S}) > 0 \,. \tag{23}$$

Hence $\mathcal{J}_{\widehat{Q}_\mathcal{S}}(\beta) > \mathcal{J}_{\widehat{Q}_\mathcal{S}}(\alpha_P^*)$ for every $\beta \neq \alpha_P^*$, and therefore $\alpha_\mathcal{S}^* = \alpha_P^*$. □

B.2 Proof of Prop. 2.2

Proof. Fix a distribution $R \in \{P, Q, \widehat{Q}_{\mathcal{U}}, \widehat{Q}_{\mathcal{S}}\}$ and a configuration α. Express both models in the candidate coordinates associated with α, and let θ and $\widehat{\theta}_\alpha$ denote the full-precision and quantized parameter vectors, respectively, with all parameters outside the L quantized layers held fixed. Because the transformations in (9) are function preserving in full precision, $\pi_\theta = \pi_0$ and $\pi_{\widehat{\theta}_\alpha} = \pi_\alpha$.

For a parameter vector θ', define the reference-supervised cross-entropy

$$\mathcal{L}_R(\theta') \triangleq \mathbb{E}_{\mathbf{x}\sim R}\left[-\sum_y \pi_0(y \mid \mathbf{x}) \log \pi_{\theta'}(y \mid \mathbf{x})\right] . \tag{24}$$

The entropy of $\pi_0(\cdot \mid \mathbf{x})$ is independent of θ', so the forward KL is exactly

$$\mathcal{J}_R(\alpha) = \mathcal{L}_R(\widehat{\theta}_\alpha) - \mathcal{L}_R(\theta) . \tag{25}$$

At the full-precision reference, the score identity gives

$$\begin{aligned} \nabla_{\theta'} \mathcal{L}_R(\theta')|_{\theta'=\theta} &= -\mathbb{E}_{\mathbf{x}\sim R}\left[\sum_y \pi_0(y \mid \mathbf{x}) \nabla_\theta \log \pi_0(y \mid \mathbf{x})\right] \\ &= -\mathbb{E}_{\mathbf{x}\sim R}\left[\sum_y \nabla_\theta \pi_0(y \mid \mathbf{x})\right] = -\mathbb{E}_{\mathbf{x}\sim R}\left[\nabla_\theta \sum_y \pi_0(y \mid \mathbf{x})\right] = 0 . \end{aligned} \tag{26}$$

Consequently, the Taylor expansion around the full-precision reference has no first-order term. We identify its retained layerwise block below and defer the exact residual decomposition to Appx. C.1.

Fix layer l and token t. Let $\mathbf{x}_{l,t}$ denote the activation in the shared reference input coordinate and

$$\mathbf{x}_{l,t}(\alpha_l) = \mathbf{T}_l(\alpha_l)\mathbf{x}_{l,t} \tag{27}$$

its representation in the candidate coordinate. Write

$$\mathbf{w}_l = \mathrm{vec}(\mathbf{W}_l(\alpha_l)), \qquad \boldsymbol{\delta}_l = \mathrm{vec}(\Delta\mathbf{W}_l(\alpha_l)), \qquad \mathbf{z}_{l,t} = \mathbf{W}_l(\alpha_l)\mathbf{x}_{l,t}(\alpha_l) . \tag{28}$$

Let $\mathcal{L}_{\mathbf{x},l,t}(\mathbf{z}_{l,t})$ denote the reference-supervised loss viewed as a function of this layer output with the remaining reference computation fixed. With $\mathbf{J}_{l,t} \triangleq \nabla_{\mathbf{w}_l}\mathbf{z}_{l,t}$, $\mathbf{g}_{l,t} \triangleq \nabla_{\mathbf{z}_{l,t}}\mathcal{L}_{\mathbf{x},l,t}$, and $\mathbf{H}_{z,l}(\mathbf{x}, t) \triangleq \nabla^2_{\mathbf{z}_{l,t}}\mathcal{L}_{\mathbf{x},l,t}$, the second-order chain rule gives

$$\nabla^2_{\mathbf{w}_l}\mathcal{L}_{\mathbf{x},l,t} = \mathbf{J}_{l,t}^\top \mathbf{H}_{z,l}(\mathbf{x}, t)\mathbf{J}_{l,t} + \sum_i [\mathbf{g}_{l,t}]_i \nabla^2_{\mathbf{w}_l}[\mathbf{z}_{l,t}]_i = \mathbf{J}_{l,t}^\top \mathbf{H}_{z,l}(\mathbf{x}, t)\mathbf{J}_{l,t} , \tag{29}$$

because every coordinate of $\mathbf{z}_{l,t} = \mathbf{W}_l(\alpha_l)\mathbf{x}_{l,t}(\alpha_l)$ is linear in $\mathbf{w}_l$. Moreover,

$$\mathbf{J}_{l,t}\boldsymbol{\delta}_l = \Delta\mathbf{W}_l(\alpha_l)\mathbf{x}_{l,t}(\alpha_l) = \Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{x}_{l,t} . \tag{30}$$

Hence the corresponding quadratic action is:

$$\frac{1}{2}\boldsymbol{\delta}_l^\top \mathbf{J}_{l,t}^\top \mathbf{H}_{z,l}(\mathbf{x}, t)\mathbf{J}_{l,t}\boldsymbol{\delta}_l = \frac{1}{2}\big(\Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{x}_{l,t}\big)^\top \mathbf{H}_{z,l}(\mathbf{x}, t)\big(\Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{x}_{l,t}\big) . \tag{31}$$

The retained PricePTQ approximation keeps the layer- and token-diagonal quadratic actions and replaces the sample-dependent output Hessian by the fixed PSD curvature $\mathbf{H}_{z,l}$ from the proposition. Let $\mathbf{X}_l(\mathbf{x}) = [\mathbf{x}_{l,1}, \ldots, \mathbf{x}_{l,T}]$, so that $\mathbf{C}_l(\mathbf{x}) = T^{-1}\mathbf{X}_l(\mathbf{x})\mathbf{X}_l(\mathbf{x})^\top$. The retained contribution of layer l is:

$$\begin{aligned} p_l(\mathbf{x}; \alpha_l) &= \frac{1}{2T}\sum_{t=1}^{T}\big(\Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{x}_{l,t}\big)^\top \mathbf{H}_{z,l}\big(\Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{x}_{l,t}\big) \\ &= \frac{1}{2T}\,\mathrm{Tr}\left(\mathbf{H}_{z,l}\Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{X}_l(\mathbf{x})\mathbf{X}_l(\mathbf{x})^\top\mathbf{T}_l(\alpha_l)^\top\Delta\mathbf{W}_l(\alpha_l)^\top\right) \\ &\overset{(a)}{=} \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{T}_l(\alpha_l)^\top\Delta\mathbf{W}_l(\alpha_l)^\top\mathbf{H}_{z,l}\Delta\mathbf{W}_l(\alpha_l)\mathbf{T}_l(\alpha_l)\mathbf{C}_l(\mathbf{x})\right) \\ &\overset{(b)}{=} \frac{1}{2}\left\langle \mathbf{M}_l(\alpha_l), \mathbf{C}_l(\mathbf{x})\right\rangle , \end{aligned} \tag{32}$$

where (a) follows by cyclic invariance of the trace, and (b) uses the definition of $\mathbf{M}_l(\alpha_l)$ in Prop. 2.2. Thus the two factors of $\mathbf{T}_l(\alpha_l)$ pull the candidate-coordinate quadratic response back to the shared reference input coordinate. When a candidate has no explicit transformation, $\mathbf{T}_l(\alpha_l) = \mathbf{I}$ and the expression reduces to the usual untransformed form.

Summing over layers and taking expectation over R gives:

$$\rho_R(\alpha) = \mathbb{E}_{\mathbf{x}\sim R}\left[p(\mathbf{x};\alpha)\right] = \frac{1}{2}\sum_{l=1}^{L}\left\langle \mathbf{M}_l(\alpha_l), \mathbf{C}_l^R \right\rangle, \qquad \mathbf{C}_l^R \triangleq \mathbb{E}_{\mathbf{x}\sim R}\left[\mathbf{C}_l(\mathbf{x})\right] . \tag{33}$$

This proves the per-sample representation and shows explicitly that, once $\mathbf{M}_l(\alpha_l)$ is fixed, all sample dependence is linear in $\mathbf{C}_l(\mathbf{x})$. Taylor's theorem therefore gives $\mathcal{J}_R(\alpha) = \rho_R(\alpha) + r_R(\alpha)$, where $r_R(\alpha)$ collects the exact-versus-retained Hessian mismatch, discarded cross-layer terms, and higher-order terms. Their explicit decomposition and relative bound are given in Appx. C.1. □

B.3 Proof of Thm. 3.1

Proof. Fix $(\alpha,\beta) \in \mathcal{O}_P$ and insert the two intermediate gaps:

$$\begin{aligned}\Delta_P(\alpha,\beta) - \Delta_{\widehat{Q}_\mathcal{S}}(\alpha,\beta) =& \left[\Delta_P(\alpha,\beta) - \Delta_Q(\alpha,\beta)\right] + \left[\Delta_Q(\alpha,\beta) - \Delta_{\widehat{Q}_\mathcal{U}}(\alpha,\beta)\right] \\ &+ \left[\Delta_{\widehat{Q}_\mathcal{U}}(\alpha,\beta) - \Delta_{\widehat{Q}_\mathcal{S}}(\alpha,\beta)\right] .\end{aligned} \tag{34}$$

Applying $[a+b+c]_+ \le [a]_+ + [b]_+ + [c]_+$ and the definition of $d(R,R')$ in Thm. 3.1 gives

$$\begin{aligned}\left[\Delta_P(\alpha,\beta) - \Delta_{\widehat{Q}_\mathcal{S}}(\alpha,\beta)\right]_+ &\le \left[\Delta_P(\alpha,\beta) - \Delta_Q(\alpha,\beta)\right]_+ + \left[\Delta_Q(\alpha,\beta) - \Delta_{\widehat{Q}_\mathcal{U}}(\alpha,\beta)\right]_+ \\ &\quad + \left[\Delta_{\widehat{Q}_\mathcal{U}}(\alpha,\beta) - \Delta_{\widehat{Q}_\mathcal{S}}(\alpha,\beta)\right]_+ \\ &\le d(P,Q) + d(Q,\widehat{Q}_\mathcal{U}) + d(\widehat{Q}_\mathcal{U},\widehat{Q}_\mathcal{S}) .\end{aligned} \tag{35}$$

All three distortions use the same index set $\mathcal{O}_P$, so the last bound holds for every $(\alpha,\beta) \in \mathcal{O}_P$. Taking the supremum over this set on the left-hand side proves Thm. 3.1. □

B.4 Proof of Prop. 3.3

Proof. By (33), the population price gap is

$$\Delta_P(\alpha,\beta) - \Delta_Q(\alpha,\beta) = \frac{1}{2}\sum_{l=1}^{L}\left\langle \mathbf{D}_l^{\alpha,\beta}, \mathbf{C}_l^P - \mathbf{C}_l^Q \right\rangle . \tag{36}$$

Substitution into $d(P,Q)$ gives the distribution-shift expression in Prop. 3.3.

Furthermore,

$$\begin{aligned}E_{\text{shift}} &\overset{(a)}{\le} \frac{1}{2}\sup_{(\alpha,\beta)\in\mathcal{O}_P}\sum_{l=1}^{L}\left|\left\langle \mathbf{D}_l^{\alpha,\beta}, \mathbf{C}_l^P - \mathbf{C}_l^Q \right\rangle\right| \\ &\overset{(b)}{\le} \frac{1}{2}\sum_{l=1}^{L}\left(\sup_{\alpha,\beta\in\mathcal{F}}\|\mathbf{D}_l^{\alpha,\beta}\|_F\right)\|\mathbf{C}_l^P - \mathbf{C}_l^Q\|_F = \frac{1}{2}\sum_{l=1}^{L} G_l\|\mathbf{C}_l^P - \mathbf{C}_l^Q\|_F ,\end{aligned} \tag{37}$$

where (a) uses $[\cdot]_+ \le |\cdot|$ and (b) uses Cauchy-Schwarz inequality. Here $G_l \triangleq \sup_{\alpha,\beta\in\mathcal{F}}\|\mathbf{D}_l^{\alpha,\beta}\|_F$. □

B.5 Proof of Thm. 3.4

For distinct configurations, let $q(\mathbf{x};\alpha,\beta) = p(\mathbf{x};\alpha) - p(\mathbf{x};\beta)$. The finite-pool error and its quadratic representation are

$$E_{\text{pool}} = \sup_{(\alpha,\beta)\in\mathcal{O}_P}\left[\mathbb{E}_{\mathbf{x}\sim Q}[q(\mathbf{x};\alpha,\beta)] - \frac{1}{M}\sum_{i=1}^{M} q(\mathbf{x}_i;\alpha,\beta)\right]_+ , \tag{38}$$

By Prop. 2.2,

$$q(\mathbf{x};\alpha,\beta)=\frac{1}{2}\sum_{l=1}^{L}\left\langle \mathbf{D}_l^{\alpha,\beta},\mathbf{C}_l(\mathbf{x})\right\rangle=\frac{1}{2T}\sum_{l=1}^{L}\sum_{t=1}^{T}\mathbf{x}_{l,t}^{\top}\mathbf{D}_l^{\alpha,\beta}\mathbf{x}_{l,t}\,, \tag{39}$$

where $\mathbf{x}_{l,t}$ is the t-th column of the reference activation matrix $\mathbf{x}_l$, and $\mathbf{D}_l^{\alpha,\beta}$ is defined in Prop. 3.3.

Before proving Thm. 3.4, we first establish a useful tail bound for price differences when the activations are sub-Gaussian:

Lemma B.1 (Sub-Gaussian Activation Tail Bound). *Assume the activation $x_{l,t,j}$ is sub-Gaussian under Q for every layer l, token t, and coordinate j, i.e. there are finite constants $\kappa_{l,t,j}>0$ such that $\mathbb{E}_{\mathbf{x}\sim Q}\left[e^{x_{l,t,j}^2/\kappa_{l,t,j}^2}\right]\le 2$. Define the centered difference $Z_{\alpha,\beta}\triangleq q(\mathbf{x};\alpha,\beta)-\mathbb{E}_Q\left[q(\mathbf{x};\alpha,\beta)\right]$, we have*

$$\log\mathbb{E}_Q\left[e^{\lambda Z_{\alpha,\beta}}\right]\le\frac{\lambda^2\nu^2}{2(1-b|\lambda|)},\qquad |\lambda|<1/b\,, \tag{40}$$

where

$$R\triangleq\max_{\alpha\neq\beta}\frac{1}{2T}\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{j,k=1}^{n_l}|(\mathbf{D}_l^{\alpha,\beta})_{jk}|\kappa_{l,t,j}\kappa_{l,t,k},\quad \nu\triangleq 4R\,,\quad b\triangleq 2R\,, \tag{41}$$

If $R=0$, all price differences vanish almost surely.

Proof. For any two coordinates X,Y with scales κ_X,κ_Y, we have:

$$\mathbb{E}_Q\exp\left(\frac{|XY|}{\kappa_X\kappa_Y}\right)\overset{(a)}{\le}\left(\mathbb{E}_Q e^{X^2/\kappa_X^2}\right)^{1/2}\left(\mathbb{E}_Q e^{Y^2/\kappa_Y^2}\right)^{1/2}\overset{(b)}{\le}2\,, \tag{42}$$

where (a) comes from Hölder's inequality and (b) comes from the sub-Gaussian assumption.

Write $q=q(\mathbf{x};\alpha,\beta)$ and $Z=q-\mathbb{E}_Q[q]$, for every integer $m\ge 2$, we have:

$$\mathbb{E}_Q\left[|Z|^m\right]\overset{(a)}{\le}\mathbb{E}_Q\left[e^{|Z|/(2R)}\right]\cdot(2R)^m m!\overset{(b)}{\le}2(2R)^m m!\,, \tag{43}$$

where (a) comes from the fact that $u^m\le m!e^u$ for every $u\ge 0$, and (b) comes from:

$$\mathbb{E}_Q\left[e^{|q-\mathbb{E}_Q q|/(2R)}\right]\overset{(a)}{\le}e^{\mathbb{E}_Q[|q-\mathbb{E}_Q q|]/(2R)}\left(\mathbb{E}_Q\left[e^{|q|/R}\right]\right)^{1/2}\overset{(b)}{\le}2\,, \tag{44}$$

where (a) comes from Jensen's inequality, and (b) comes from the sub-Gaussian assumption.

Since $\mathbb{E}_Q[Z]=0$, absolute convergence of the exponential series for $2R|\lambda|<1$ and $\log u\le u-1$ yield:

$$\log\mathbb{E}_Q e^{\lambda Z}\le\sum_{m=2}^{\infty}\frac{|\lambda|^m\mathbb{E}_Q|Z|^m}{m!}\le\frac{8R^2\lambda^2}{1-2R|\lambda|}\,, \tag{45}$$

which completes the proof. If $R=0$, positivity of the scales forces every difference matrix to be zero, so every q vanishes. □

We then establish the formal finite-pool concentration theorem in Thm. B.2:

Theorem B.2 (Formal, Finite-Pool Concentration). *Let $K=|\mathcal{F}|\ge 2$ and $J=K(K-1)/2$. Under Assump. 3.2 and Lem. B.1, for every $\eta\in(0,1)$, with probability at least $1-\eta$, we have:*

$$E_{\text{pool}}\le\sqrt{\frac{2\nu^2}{M}\log\frac{J}{\eta}}+\frac{b}{M}\log\frac{J}{\eta}\,, \tag{46}$$

where

$$R\triangleq\max_{\alpha\neq\beta}\frac{1}{2T}\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{j,k=1}^{n_l}|(\mathbf{D}_l^{\alpha,\beta})_{jk}|\kappa_{l,t,j}\kappa_{l,t,k},\quad \nu\triangleq 4R\,,\quad b\triangleq 2R\,. \tag{47}$$

If $R=0$, the bound holds with $E_{\text{pool}}=0$.

Proof. Fix $(\alpha, \beta) \in \mathcal{O}_P$ and let

$$Y_i = \mathbb{E}_Q[q(\mathbf{x}_i; \alpha, \beta)] - q(\mathbf{x}_i; \alpha, \beta), \qquad \overline{Y} = \frac{1}{M} \sum_{i=1}^{M} Y_i \,. \tag{48}$$

These variables are independent and centered. For $0 < \lambda < M/b$, Lem. B.1 gives

$$\log \mathbb{E} e^{\lambda \overline{Y}} = \sum_{i=1}^{M} \log \mathbb{E} e^{(\lambda/M) Y_i} \leq \frac{\lambda^2 (\nu^2/M)}{2(1 - b\lambda/M)} \,. \tag{49}$$

Thus $\overline{Y}$ has a right-tail sub-gamma bound with variance factor $v = \nu^2/M$ and scale $c = b/M$, and consequently satisfies the Chernoff bound:

$$\Pr\left(E_{\text{pool}} > \sqrt{\frac{2\nu^2}{M} \log \frac{J}{\eta}} + \frac{b}{M} \log \frac{J}{\eta} \right) \leq \eta \,, \tag{50}$$

where for any $s > 0$ and the Chernoff's inequality

$$\Pr\left(\overline{Y} > \sqrt{2vs} + cs \right) \leq e^{-s} \,, \tag{51}$$

we choose $a = \sqrt{2s/v}$ and $\lambda = a/(1 + ca)$ and $s = \log(J/\eta)$. □

B.6 Proof of Thm. 3.5

Before the proof, we first establish an exact expression for the coreset compression term:

Lemma B.3 (Exact Coreset Compression Expression)**.** *The compression term equals:*

$$E_{\mathcal{S}} = \frac{1}{2} \sup_{(\alpha,\beta) \in \mathcal{O}_P} \left[\sum_{l=1}^{L} \left\langle \mathbf{D}_l^{\alpha,\beta}, \Delta\mathbf{C}_l(\mathcal{S}) \right\rangle \right]_+ .$$

Proof. For every configuration pair,

$$\Delta_{\widehat{Q}_\mathcal{U}}(\alpha, \beta) - \Delta_{\widehat{Q}_\mathcal{S}}(\alpha, \beta) = \frac{1}{2} \sum_{l=1}^{L} \left\langle \mathbf{D}_l^{\alpha,\beta}, \mathbf{C}_l^U - \mathbf{C}_l^S \right\rangle = \frac{1}{2} \sum_{l=1}^{L} \left\langle \mathbf{D}_l^{\alpha,\beta}, \Delta\mathbf{C}_l(\mathcal{S}) \right\rangle \,. \tag{52}$$

Taking the positive part and the supremum over $(\alpha, \beta) \in \mathcal{O}_P$ in (52) completes the proof of Lem. B.3. □

Then, we prove Thm. 3.5:

Proof. By Lem. B.3, $\mathcal{O}_P \subseteq \mathcal{F}^2$, and $[u]_+ \leq |u|$, we have

$$E_{\mathcal{S}} \leq \frac{1}{2} \sup_{\alpha,\beta \in \mathcal{F}} \left| \sum_{l=1}^{L} \left\langle \mathbf{D}_l^{\alpha,\beta}, \Delta\mathbf{C}_l(\mathcal{S}) \right\rangle \right| \,. \tag{53}$$

Decomposing each matrix into diagonal and off-diagonal parts and applying the triangle inequality gives

$$\begin{aligned} E_{\mathcal{S}} &\leq R_{\text{off}}(\mathcal{S}) + \frac{1}{2} \sup_{\alpha,\beta \in \mathcal{F}} \left| \sum_{l,j} [\mathbf{D}_l^{\alpha,\beta}]_{jj} [\Delta\mathbf{C}_l(\mathcal{S})]_{jj} \right| \\ &\leq R_{\text{off}}(\mathcal{S}) + \frac{1}{2} \sum_{l,j} \sup_{\alpha,\beta \in \mathcal{F}} |[\mathbf{D}_l^{\alpha,\beta}]_{jj}| \cdot |[\Delta\mathbf{C}_l(\mathcal{S})]_{jj}| \,. \end{aligned} \tag{54}$$

Substituting $[\Delta\mathbf{C}_l(\mathcal{S})]_{jj} = \mu_{l,j}^U - \mu_{l,j}^S$ and the definition of $\gamma_{l,j}$ in Thm. 3.5 yields the claimed upper bound. □

B.7 Proof of Lem. 4.1

Proof. For a nonempty subset $\mathcal{S} \subseteq \mathcal{U}$, let $r_{l,j} = \mu^U_{l,j} - \mu^S_{l,j}$. Then

$$J_1(\mathcal{S}) = \sum_{l,j} \sqrt{\gamma_{l,j}} \left(\sqrt{\gamma_{l,j}}|r_{l,j}|\right) \leq \sqrt{\sum_{l,j} \gamma_{l,j}} \sqrt{\sum_{l,j} \gamma_{l,j} r^2_{l,j}} = \sqrt{2\Gamma_{\mathrm{tot}} J_2(\mathcal{S})}\,, \tag{55}$$

where the inequality results from the Cauchy-Schwarz inequality.

Substitution into the upper bound in Thm. 3.5 completes the proof of Lem. 4.1. □

B.8 Proof of Thm. 4.2

Proof. Fix a partial support $\mathcal{S}_t$ and write $\mathcal{R}_t = \mathcal{U} \setminus \mathcal{S}_t$, $m_t = M - t$, and $k_t = N - t$. Let $\mathbf{h}_t = \sum_{\mathbf{x}\in\mathcal{S}_t} \mathbf{y}(\mathbf{x})$ and $Q_t = \sum_{\mathbf{x}\in\mathcal{R}_t} \|\mathbf{y}(\mathbf{x})\|_2^2$. For a uniformly sampled completion $\mathcal{T}_t \subseteq \mathcal{R}_t$ of size k_t, centeredness gives

$$\mathbb{E}\left[\sum_{\mathbf{x}\in\mathcal{T}_t} \mathbf{y}(\mathbf{x}) \mid \mathcal{S}_t\right] = -\frac{k_t}{m_t}\mathbf{h}_t\,. \tag{56}$$

The inclusion probabilities for sampling without replacement also give

$$\mathbb{E}\left[\left\|\sum_{\mathbf{x}\in\mathcal{T}_t} \mathbf{y}(\mathbf{x})\right\|_2^2 \middle| \mathcal{S}_t\right] = \frac{k_t(m_t - k_t)}{m_t(m_t-1)} Q_t + \frac{k_t(k_t-1)}{m_t(m_t-1)}\|\mathbf{h}_t\|_2^2\,. \tag{57}$$

Substituting these identities into the squared norm of the completed mean yields

$$\Phi_t(\mathcal{S}_t) = \frac{M-N}{2N^2(M-t)(M-t-1)}\left[(M-N-1)\|\mathbf{h}_t\|_2^2 + (N-t)Q_t\right]\,. \tag{58}$$

At $t = 0$, we have $\mathbf{h}_0 = \mathbf{0}$ and $Q_0 = MV_\phi$, so (58) gives

$$\Phi_0(\varnothing) = \frac{M-N}{2N(M-1)} V_\phi = \mathbb{E}[J_2(\mathcal{S}_N^{\mathrm{rand}}) \mid \mathcal{U}]\,. \tag{59}$$

This establishes the exact random-baseline identity used in the proposition. At $t = 0$, the terms in $\Phi_1(\{\mathbf{x}\})$ that depend on the first point are proportional to $\|\mathbf{y}(\mathbf{x})\|_2^2$, so the initialization in Alg. 1 minimizes Φ_1. For $1 \leq t < N$, expanding (58) after adding $\mathbf{x}$ gives, up to terms independent of $\mathbf{x}$,

$$\Phi_{t+1}(\mathcal{S}_t \cup \{\mathbf{x}\}) \propto 2(M-N-1)\langle \mathbf{h}_t, \mathbf{y}(\mathbf{x})\rangle + (M-2N+t)\|\mathbf{y}(\mathbf{x})\|_2^2\,. \tag{60}$$

Since $\mathbf{h}_t = -t\mathbf{r}_t$, minimizing this expression is exactly the score in (22).

The first element of a uniform completion is uniform over $\mathcal{R}_t$, hence

$$\Phi_t(\mathcal{S}_t) = \frac{1}{M-t}\sum_{\mathbf{x}\in\mathcal{R}_t} \Phi_{t+1}(\mathcal{S}_t \cup \{\mathbf{x}\})\,. \tag{61}$$

The algorithm chooses the minimum term on the right, so $\Phi_{t+1} \leq \Phi_t$ and $\Phi_N = J_2(\mathcal{S}_N^{\mathrm{PPCC}})$. Finally, evaluating (58) at $t = 1$ and using $e_{\min} = \min_{\mathbf{x}\in\mathcal{U}} \|\mathbf{y}(\mathbf{x})\|_2^2$ gives the expression for Φ_1 in Thm. 4.2. □

C Additional Discussion

C.1 Taylor Residuals in Prop. 2.2

Configuration selection requires the quadratic PricePTQ approximation to preserve rankings, so for a distribution R it suffices that the residual $r_R(\alpha)$ be too small to reverse the retained price gaps that determine the selected configuration.

Theorem C.1 (Price Residual Bound). *Fix $R \in \{P, Q, \widehat{Q}_{\mathcal{U}}, \widehat{Q}_{\mathcal{S}}\}$ and a configuration α. Let θ stack the full-precision layer weights in the candidate coordinates associated with α, and define*

$$\boldsymbol{\delta}_l \triangleq \mathrm{vec}(\Delta \mathbf{W}_l(\alpha_l)), \qquad \boldsymbol{\delta} \triangleq (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_L^\top)^\top . \tag{62}$$

Let $\mathbf{H}_{\theta,R}$ be the exact Hessian at θ of the reference-supervised objective $\mathcal{L}_R$ defined in Appx. B.2. Under column-wise vectorization, let $\widetilde{\mathbf{H}}_{\theta,R}$ be the retained block-diagonal Hessian in the candidate coordinates associated with α, with blocks

$$\widetilde{\mathbf{H}}_{ll}^R \triangleq \left[\mathbf{T}_l(\alpha_l)\mathbf{C}_l^R\mathbf{T}_l(\alpha_l)^\top\right] \otimes \mathbf{H}_{z,l} , \qquad \rho_R(\alpha) = \frac{1}{2}\boldsymbol{\delta}^\top \widetilde{\mathbf{H}}_{\theta,R}\boldsymbol{\delta} . \tag{63}$$

Suppose $B_{3,R}(\alpha)$ bounds $\|\nabla_\theta^3 \mathcal{L}_R(\theta + s\boldsymbol{\delta})\|_{\mathrm{op}}$ for every $s \in [0,1]$. Here $\|\mathcal{T}\|_{\mathrm{op}} \triangleq \sup_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=\|\mathbf{w}\|_2=1} |\mathcal{T}[\mathbf{u},\mathbf{v},\mathbf{w}]|$ denotes the multilinear operator norm induced by Euclidean norms. If $\rho_R(\alpha) > 0$, then

$$|r_R(\alpha)| \le \eta_R(\alpha)\rho_R(\alpha) , \qquad \eta_R(\alpha) \triangleq \frac{\frac{1}{2}\|\mathbf{H}_{\theta,R} - \widetilde{\mathbf{H}}_{\theta,R}\|_2\|\boldsymbol{\delta}\|_2^2 + \frac{B_{3,R}(\alpha)}{6}\|\boldsymbol{\delta}\|_2^3}{\rho_R(\alpha)} . \tag{64}$$

Proof. Write $\mathbf{H}_{lk}^R$ for the (l,k) block of $\mathbf{H}_{\theta,R}$. The score cancellation established in the proof of Prop. 2.2 removes the linear term, so Taylor's theorem gives

$$\begin{aligned} \mathcal{J}_R(\alpha) &= \frac{1}{2}\sum_{l=1}^{L} \boldsymbol{\delta}_l^\top \mathbf{H}_{ll}^R \boldsymbol{\delta}_l + \frac{1}{2}\sum_{l\neq k} \boldsymbol{\delta}_l^\top \mathbf{H}_{lk}^R \boldsymbol{\delta}_k + R_{3,R}(\alpha), \\ |R_{3,R}(\alpha)| &\le \frac{B_{3,R}(\alpha)}{6}\|\boldsymbol{\delta}\|_2^3 . \end{aligned} \tag{65}$$

The retained layerwise block derived in (33) is

$$\rho_R(\alpha) = \frac{1}{2}\sum_{l=1}^{L} \boldsymbol{\delta}_l^\top \widetilde{\mathbf{H}}_{ll}^R \boldsymbol{\delta}_l = \frac{1}{2}\sum_{l=1}^{L} \left\langle \mathbf{M}_l(\alpha_l), \mathbf{C}_l^R \right\rangle , \tag{66}$$

where the first equality uses the candidate-coordinate covariance $\mathbf{T}_l(\alpha_l)\mathbf{C}_l^R\mathbf{T}_l(\alpha_l)^\top$, and the second is the same quadratic form pulled back to the shared reference coordinate through Prop. 2.2. Subtracting (66) from (65) yields

$$\begin{aligned} r_R(\alpha) &= \frac{1}{2}\sum_{l=1}^{L} \boldsymbol{\delta}_l^\top \left(\mathbf{H}_{ll}^R - \widetilde{\mathbf{H}}_{ll}^R\right)\boldsymbol{\delta}_l + \frac{1}{2}\sum_{l\neq k} \boldsymbol{\delta}_l^\top \mathbf{H}_{lk}^R \boldsymbol{\delta}_k + R_{3,R}(\alpha) \\ &= \frac{1}{2}\boldsymbol{\delta}^\top \left(\mathbf{H}_{\theta,R} - \widetilde{\mathbf{H}}_{\theta,R}\right)\boldsymbol{\delta} + R_{3,R}(\alpha) . \end{aligned} \tag{67}$$

Applying $|\boldsymbol{\delta}^\top \mathbf{A}\boldsymbol{\delta}| \le \|\mathbf{A}\|_2\|\boldsymbol{\delta}\|_2^2$ and the remainder bound in (65) gives the numerator of (64). Dividing by $\rho_R(\alpha) > 0$ proves the result. □

The Hessian discrepancy in (64) captures both within-layer curvature mismatch and the cross-layer or cross-token couplings removed by the block-diagonal approximation, while the cubic term measures nonlocality along the path to the completed quantized model. Thus, $\eta_R(\alpha)$ measures these omitted terms relative to the retained price.

For two configurations α and β with positive retained prices, Thm. C.1 implies the ranking condition

$$\rho_R(\beta) - \rho_R(\alpha) > \eta_R(\alpha)\rho_R(\alpha) + \eta_R(\beta)\rho_R(\beta) \quad \Longrightarrow \quad \mathcal{J}_R(\alpha) < \mathcal{J}_R(\beta) . \tag{68}$$

Consequently, large price gaps are robust to noticeable absolute approximation error, and the experiment below examines this regime empirically.

Empirical Evaluation. We evaluate ranking fidelity through the isotropic trace-price proxy of PricePTQ (Qiu et al., 2026), which replaces the layer price in Prop. 2.2 by $\frac{1}{2}\sigma_l^2\omega_l^2\,\mathrm{Tr}(\mathbf{H}_{z,l})$, where σ_l^2 and ω_l^2 are the trace-normalized scales of the input covariance and the weight perturbation. For each candidate, define:

$$\widehat{r}(\alpha) \triangleq \widehat{\mathcal{J}}(\alpha) - \widehat{\rho}(\alpha)\,, \qquad e(\alpha) \triangleq \frac{|\widehat{r}(\alpha)|}{\widehat{\rho}(\alpha)}\,, \tag{69}$$

where $\widehat{\mathcal{J}}$ is the empirical forward KL and $\widehat{\rho}$ is the trace-price proxy. Each candidate quantizes one linear layer from four evenly spaced transformer blocks, namely blocks 0, 9, 18, and 27 of Qwen3-0.6B and blocks 0, 5, 10, and 15 of Llama-3.2-1B, under different bitwidths, pre-quantization transformations, and granularities, as shown in Fig. 3.

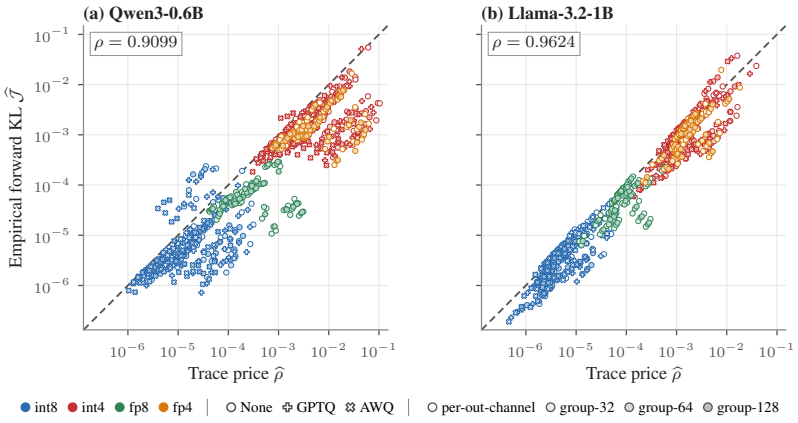

Figure 3: The empirical forward KL $\widehat{\mathcal{J}}(\alpha)$ versus the isotropic trace-price proxy $\widehat{\rho}(\alpha)$ of Qwen3-0.6B (a) and Llama-3.2-1B (b).

For both models, the trace-price proxy is a good predictor of the empirical forward KL, with most candidates lying near the $y = x$ line. Specifically, the Spearman rank correlation between the empirical forward KL and the trace-price proxy is 0.9099 for Qwen3-0.6B and 0.9624 for Llama-3.2-1B, showing that the trace-price proxy closely tracks the candidate ranking induced by the empirical forward KL.

C.2 Gaussian Activations: Explicit Tail Parameters

Jointly Gaussian activations allow more precise tail parameters than the general marginal bound in (41). The following lemma gives such parameters for (40); the concentration argument in Appx. B.5 then yields (46) with these parameters.

Stack the reference activations of one sequence as $\mathbf{a}(\mathbf{x}) = [\mathrm{vec}(\mathbf{x}_1)^\top, \ldots, \mathrm{vec}(\mathbf{x}_L)^\top]^\top$, using column-wise vectorization, and define the symmetric matrix

$$\mathbf{B}_{\alpha,\beta} = \frac{1}{2T}\,\mathrm{blockdiag}\left(\mathbf{I}_T \otimes \mathbf{D}_1^{\alpha,\beta}, \ldots, \mathbf{I}_T \otimes \mathbf{D}_L^{\alpha,\beta}\right)\,. \tag{70}$$

Then (39) becomes $q(\mathbf{x};\alpha,\beta) = \mathbf{a}(\mathbf{x})^\top \mathbf{B}_{\alpha,\beta}\mathbf{a}(\mathbf{x})$.

Lemma C.2 (Gaussian Sufficient Condition). *Suppose that, under $\mathbf{x} \sim Q$, $\mathbf{a}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \succeq 0$, and that the response matrices are fixed. For each distinct configuration pair, define*

$$\mathbf{A}_{\alpha,\beta} = \boldsymbol{\Sigma}^{1/2}\mathbf{B}_{\alpha,\beta}\boldsymbol{\Sigma}^{1/2}, \qquad \mathbf{h}_{\alpha,\beta} = \boldsymbol{\Sigma}^{1/2}\mathbf{B}_{\alpha,\beta}\boldsymbol{\mu}\,. \tag{71}$$

The centered price difference satisfies (40) with pair-specific parameters

$$\nu_{\alpha,\beta}^2 = 2\|\mathbf{A}_{\alpha,\beta}\|_F^2 + 4\|\mathbf{h}_{\alpha,\beta}\|_2^2, \qquad b_{\alpha,\beta} = 2\|\mathbf{A}_{\alpha,\beta}\|_2\,, \tag{72}$$

for all λ satisfying $2|\lambda|\|\mathbf{A}_{\alpha,\beta}\|_2 < 1$. Consequently, since $\mathcal{F}$ is finite, any positive ν, b satisfying

$$\nu^2 \geq \max_{\alpha\neq\beta} \nu_{\alpha,\beta}^2, \qquad b \geq \max_{\alpha\neq\beta} b_{\alpha,\beta} \tag{73}$$

establish (40) uniformly over $\mathcal{F}$.

Proof. Fix a configuration pair and abbreviate $\mathbf{A} = \mathbf{A}_{\alpha,\beta}$ and $\mathbf{h} = \mathbf{h}_{\alpha,\beta}$. Write $\mathbf{a} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The centered difference is

$$Z = q - \mathbb{E}_Q q = \mathbf{g}^\top \mathbf{A}\mathbf{g} - \operatorname{Tr}(\mathbf{A}) + 2\mathbf{h}^\top \mathbf{g}\,. \tag{74}$$

For $2|\lambda|\|\mathbf{A}\|_2 < 1$, $\mathbf{I} - 2\lambda\mathbf{A}$ is positive definite. Completing the square in the Gaussian integral yields

$$\log \mathbb{E}e^{\lambda Z} = -\lambda \operatorname{Tr}(\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I} - 2\lambda\mathbf{A}) + 2\lambda^2 \mathbf{h}^\top (\mathbf{I} - 2\lambda\mathbf{A})^{-1}\mathbf{h}\,. \tag{75}$$

For $|u| < 1$, the logarithmic series gives

$$-\log(1-u) - u = \sum_{k=2}^{\infty} \frac{u^k}{k} \leq \frac{u^2}{2(1-|u|)}\,. \tag{76}$$

Applying this inequality to $u = 2\lambda\theta_j$, where θ_j are the eigenvalues of $\mathbf{A}$, gives

$$-\lambda \operatorname{Tr}(\mathbf{A}) - \frac{1}{2}\log\det(\mathbf{I} - 2\lambda\mathbf{A}) \leq \frac{\lambda^2\|\mathbf{A}\|_F^2}{1 - 2|\lambda|\|\mathbf{A}\|_2}\,. \tag{77}$$

Also,

$$\mathbf{h}^\top (\mathbf{I} - 2\lambda\mathbf{A})^{-1}\mathbf{h} \leq \frac{\|\mathbf{h}\|_2^2}{1 - 2|\lambda|\|\mathbf{A}\|_2}\,. \tag{78}$$

Combining the two bounds proves

$$\log \mathbb{E}e^{\lambda Z} \leq \frac{\lambda^2(\|\mathbf{A}\|_F^2 + 2\|\mathbf{h}\|_2^2)}{1 - 2|\lambda|\|\mathbf{A}\|_2}\,, \tag{79}$$

which is the asserted bound. Taking positive parameters dominating the pair-specific values gives the uniform condition. □

When $\boldsymbol{\mu} = \mathbf{0}$, the variance factor simplifies to $\nu_{\alpha,\beta}^2 = 2\|\mathbf{A}_{\alpha,\beta}\|_F^2$. The covariance $\boldsymbol{\Sigma}$ allows arbitrary correlations across channels, tokens, and layers. The parameters depend on the response differences $\mathbf{D}_l^{\alpha,\beta}$ instead of absolute candidate prices.

C.3 Empirical Evaluation of Thm. 3.5

In this section, we empirically evaluate the relative size of the off-diagonal residual in Thm. 3.5, to test whether the tractable diagonal moment term $J_1(\mathcal{S})$ is the dominant term in the empirical upper bound. We use the `meta-llama/Llama-3.2-1B` model and construct the coreset by selecting $N \in \{8, 16, 32, 64, 128\}$ samples from three independently sampled calibration pools of size $M = 2048$. For every coreset, we compute $R_{\text{off}}(\mathcal{S})/J_1(\mathcal{S})$ and summarize the ratios as shown in Fig. 4.

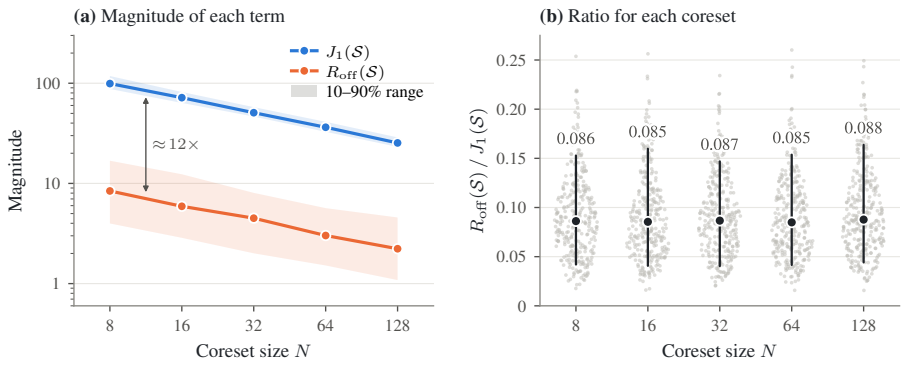

Figure 4: $J_1(\mathcal{S})$ versus $R_{\text{off}}(\mathcal{S})$. (a) Medians of both terms with shaded 10-90% ranges, and the arrow marks the gap of about $12\times$. (b) Per-support ratio $R_{\text{off}}(\mathcal{S})/J_1(\mathcal{S})$. Vertical bars show the 10-90% range, and dots mark the median, whose value is printed above each bar.

Across all evaluated budgets, the median ratio remains approximately between 0.085 and 0.088, and the upper edge of the 10-90% range remains below 0.17. Thus, $R_{\text{off}}(\mathcal{S})$ is consistently smaller than $J_1(\mathcal{S})$, and the J_1 term dominates the empirical upper bound $R_{\text{off}}(\mathcal{S}) + J_1(\mathcal{S})$, which supports using J_1 as a tractable selection surrogate.

D Additional Experiments

Tab. 2 reports the per-task results underlying the Llama-3.1-8B Base summary in Tab. 1.

Table 2: Per-task results for pre-deployment configuration selection on Llama-3.1-8B Base. Bold and underlined values mark the best and second-best scores. Time is the calibration selection overhead on a single H100 GPU.

Method	PPL↓	PIQA↑	WiC↑	ARC-C↑	HellaS.↑	Avg.↑	Time (min.)
Bitwidth selection							
Random	14.3435	75.24	**52.04**	42.58	72.01	60.467	—
CoverCal	10.9588	76.55	47.49	42.58	71.98	59.649	16.05
ZipCal	32.2627	**77.20**	49.69	41.13	71.60	59.904	0.151
PPCC *(ours)*	**8.7075**	76.93	50.94	**46.59**	**73.07**	**61.883**	11.80
Transformation selection							
Random	8.7283	77.09	50.00	45.56	73.52	61.545	—
CoverCal	8.6597	77.86	49.53	46.25	**74.23**	61.965	16.05
ZipCal	8.6245	77.20	49.69	**46.59**	73.43	61.727	0.151
PPCC *(ours)*	**8.5991**	**78.24**	**50.47**	46.33	73.62	**62.165**	11.93
Granularity selection							
Random	16.6420	76.66	50.16	47.27	74.30	62.096	—
CoverCal	13.1343	76.77	50.31	46.08	**74.35**	61.876	16.05
ZipCal	19.5696	**78.56**	**55.96**	47.10	74.26	**63.969**	0.151
PPCC *(ours)*	**8.5125**	78.02	50.47	**47.95**	73.43	62.468	11.82
Mixed selection							
Random	8.6304	76.06	51.41	42.58	73.36	60.853	—
CoverCal	**8.3830**	77.26	50.31	43.69	73.42	61.170	16.05
ZipCal	8.6038	76.71	47.96	41.30	71.91	59.470	0.151
PPCC *(ours)*	8.4725	**78.18**	**54.86**	**45.31**	**73.69**	**63.010**	12.40

Similar to Sec. 5.2, we evaluate Qwen3-8B Base using the same C4 pool and four configuration selection settings. As shown in Tab. 3, PPCC achieves the highest average score in all four settings and the lowest PPL in bitwidth selection. Its average gain over the strongest baseline reaches 1.977 percentage points in bitwidth selection. In transformation, granularity, and mixed selection, the

average gains are 0.025, 0.044, and 0.186 percentage points, while ZipCal, CoverCal, and Random, respectively, achieve lower PPL.

Table 3: Pre-deployment configuration selection on Qwen3-8B Base. Bold and underlined values mark the best and second-best scores. Time is the calibration selection overhead on a single H100 GPU.

Method	PPL↓	PIQA↑	WiC↑	ARC-C↑	HellaS.↑	Avg.↑	Time (min.)
Bitwidth selection							
Random	8.7534	77.31	53.76	46.08	**73.78**	62.732	—
CoverCal	9.0584	76.88	53.92	45.05	72.07	61.979	17.44
ZipCal	9.8294	73.50	50.63	35.92	69.09	57.286	0.129
PPCC *(ours)*	**8.6244**	**77.64**	**56.11**	**51.54**	73.55	**64.710**	12.82
Transformation selection							
Random	8.8913	77.20	50.31	49.57	**74.45**	62.884	—
CoverCal	8.8435	77.31	50.00	48.72	73.57	62.401	17.44
ZipCal	**8.7013**	77.26	53.45	**52.73**	74.18	64.404	0.129
PPCC *(ours)*	8.8984	**78.13**	**53.92**	51.28	74.39	**64.429**	12.93
Granularity selection							
Random	8.4100	78.29	56.58	49.83	75.01	64.930	—
CoverCal	**8.2953**	78.56	50.00	54.01	**75.05**	64.407	17.44
ZipCal	8.3392	77.97	**64.58**	50.85	**75.05**	67.112	0.129
PPCC *(ours)*	8.2999	**79.60**	57.68	**56.31**	75.03	**67.157**	13.00
Mixed selection							
Random	**8.4673**	77.37	50.16	47.44	**74.23**	62.298	—
CoverCal	8.4869	**77.91**	50.00	48.12	73.87	62.476	17.44
ZipCal	9.4615	73.88	**55.64**	38.31	69.43	59.317	0.129
PPCC *(ours)*	8.5112	77.42	50.78	**48.29**	74.15	**62.662**	13.88

Furthermore, we evaluate the Qwen3-8B Instruct model in non-thinking mode with $N = 64$ on the three search settings listed in Tab. 4. Different from the base model, the quantization performance of the instruct model is evaluated by reasoning tasks, including GSM8K, AIME24, AIME25, and GPQA. As shown in Tab. 4, PPCC achieves the highest average score in bitwidth and granularity selection, exceeding the strongest baseline by 0.854 and 2.163 percentage points, respectively. In transformation selection, PPCC is within 0.013 percentage points of CoverCal.

Table 4: Pre-deployment configuration selection for Qwen3-8B instruct model. Bold and underlined values mark the best and second-best scores. AIME24 and AIME25 accuracy is averaged over 16 trials. Time is the calibration selection overhead on a single H100 GPU.

Method	**GSM8K↑**	**AIME24↑**	**AIME25↑**	**GPQA↑**	**Avg.↑**	Time (min.)
Bitwidth selection						
Random	77.94	18.12	**15.62**	29.24	35.23	—
CoverCal	79.61	<u>19.17</u>	14.79	**34.38**	<u>36.98</u>	16.81
ZipCal	<u>79.68</u>	12.29	13.33	30.36	33.92	0.131
PPCC	**82.41**	**19.38**	<u>15.42</u>	<u>34.15</u>	**37.84**	13.60
Transformation selection						
Random	72.86	13.75	13.54	<u>32.81</u>	33.24	—
CoverCal	**81.73**	<u>15.42</u>	13.12	<u>32.81</u>	**35.77**	16.81
ZipCal	<u>80.14</u>	15.00	**15.21**	28.35	34.67	0.131
PPCC	79.76	**16.04**	<u>13.75</u>	**33.48**	<u>35.76</u>	13.62
Granularity selection						
Random	70.96	<u>13.54</u>	10.21	<u>33.26</u>	<u>31.99</u>	—
CoverCal	67.70	**14.58**	<u>12.71</u>	28.57	30.89	16.81
ZipCal	<u>74.30</u>	12.50	10.21	30.58	31.90	0.131
PPCC	**75.28**	12.50	**13.12**	**35.71**	**34.16**	13.26

E Experimental Setup

This section describes the calibration selectors and experimental settings used for configuration ranking in Sec. 5.1 and pre-deployment configuration selection in Sec. 5.2 and Appx. D.

E.1 Baselines

In both experiments, we compare PPCC with three calibration selection baselines:

- **Random** uniformly samples sequences from the calibration pool without replacement.
- **ZipCal** (Monaco et al., 2026) greedily selects sequences that add the most previously uncovered vocabulary, increasing lexical coverage without requiring target-model forward passes.
- **CoverCal** (Shihab et al., 2026) greedily selects sequences to maximize weighted coverage of outlier activation channels, using activation magnitudes and weight-column norms to determine channel importance.

E.2 Hyperparameters

For pre-deployment configuration selection experiment, we evaluate `Llama-3.1-8B` and `Qwen3-8B-Base` with $N = 32$, and `Qwen3-8B` in non-thinking mode with $N = 64$. Within each model and search setting, all selectors share the calibration pool, candidate space, and evaluation configuration. Tab. 5 summarizes the hyperparameters.

Table 5: Hyperparameters for pre-deployment configuration selection.

Hyperparameter	Value
Calibration source	C4 dataset (English)
Pool size M	9646 sequences
Sequence length T	2048 tokens
Calibration budget N	32 (Base), 64 (Instruct)
Bitwidth candidates	INT2, INT3, INT4
Granularity candidates	Per-output-channel, Group-wise-16 / 32 / 64 / 128
Transformation candidates	GPTQ, AWQ
Mixed candidates	All combinations of bitwidth, granularity, and transformation