SOFTWATER: THE RATE-DISTORTION GEOMETRY OF SOFTMAX QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization methods typically optimize weighted mean-squared error (WMSE) on linear-layer outputs, yet the final softmax layer is fundamentally different: its relevant distortion is the change in the predictive distribution. We formulate softmax-layer quantization as a rate-distortion problem under KL divergence and show that, to second order, its geometry is class-aware, coupling feature covariance with class-specific softmax curvature. This distinction matters because output heads are often left in high precision even though, in modern language models, they can account for 15–30% of all parameters, and over half of the stored bytes once the body is quantized. We introduce *SoftWater*, a practical quantizer derived from this KL geometry. A separable approximation reduces the otherwise prohibitive joint class-feature weighting to a single feature-side factorization together with inexpensive class-dependent rescaling, allowing the head to be quantized using statistics collected in one calibration pass. The resulting allocation assigns precision according to the output class frequency discounted by its variance across inputs, which is especially important in highly imbalanced output spaces such as LLM vocabularies. Across five language models from 1B to 32B parameters, SoftWater outperforms the released WaterSIC quantizer, near-optimal under WMSE, at matched head rates on 59 of 60 evaluation points, reducing head-induced KL by 6.5–8.3$\times$ at 2 bits despite using none of WaterSIC's additional refinements. Because the class-side statistic comes from calibration data, matching calibration to the deployment domain gives the lowest KL on that domain throughout. On Llama-3.2-1B-Instruct with quantized bodies, where the tied head is also the embedding, a 2-bit head removes 45–60% of stored bytes for a 2.9–3.7% perplexity increase and a 4-bit head is near-lossless. These results show that treating the softmax layer according to its own distortion geometry can make aggressive end-to-end model quantization practical.

1 INTRODUCTION

Post-training quantization (PTQ) reduces the precision of trained weights while minimizing model output distortion. For linear layers the standard objective is a weighted mean-squared error (WMSE) on the layer output, $\mathbb{E}\|(W - \hat{W})X\|^2$, targeted by sequential rounding Frantar et al. (2023a;b); Zhang et al. (2026); Chen et al. (2026), lattice codebooks Chee et al. (2024); Tseng et al. (2024; 2025); Savkin et al. (2025), or learned clipping and rotations Shao et al. (2024); Liu et al. (2025b); Ashkboos et al. (2024); Shao et al. (2025b). Most recently, WaterSIC Lifar et al. (2026) brought the WMSE problem within 0.255 bits of its information-theoretic limit. Under WMSE the error weight is $I \otimes \mathbb{E}_X[XX^\top]$: rate is allocated across input features, and every output row is treated alike.

A softmax layer breaks this homogeneity. Its output is a probability distribution, so the distortion that matters is the KL divergence between the original and quantized outputs, and a second-order expansion of that divergence gives the error weight $\mathbb{E}[(\mathrm{diag}(p) - pp^\top) \otimes XX^\top]$. The class-side factor is the softmax curvature, which varies by orders of magnitude across classes whenever the output distribution is imbalanced. The correct rate allocation for a softmax layer is therefore *class-aware*: class k gets a grid scaled by $1/\bar{\lambda}_k^2$, where $\bar{\lambda}_k = \mathbb{E}[p_k(1 - p_k)]$ is the output class frequency discounted by its variance across inputs, on top of the usual feature-side scaling by $|\ell_{ii}|^{-1}$. Frequent,

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.

low-variance classes get fine grids, and rare ones get coarse grids. WMSE is the special case in which every class is assumed equally likely.

The gap is especially large in the output head of a language model, whose tokens follow Zipfian distributions Piantadosi (2014). Ignoring the gap has a high cost: in small models with modern vocabularies, the head holds 15–30% of all parameters (Table 1), yet PTQ pipelines leave it in fp16, so a "2-bit" Llama-3.2-1B stores nearly 5 bits per weight. Head inference cost is instead reduced through orthogonal approaches Grave et al. (2017); Jean et al. (2015); Liu et al. (2025a); Shao et al. (2025a); Tranheden et al. (2026).

Table 1: Share of total parameters held by the output head (vocab × hidden) in modern small LLMs.

Model	Vocab	Hid	Head	total
Gemma-3-1B	262k	1152	302M	30.2%
Qwen3-0.6B	151k	1024	156M	26.1%
Gemma-2-2B	256k	2304	590M	22.6%
Llama-3.2-1B	128k	2048	263M	21.3%
Qwen3-1.7B	151k	2048	311M	18.1%
Gemma-3-4B	262k	2560	671M	15.6%

Encoding against the joint error weight would need a $Kn \times Kn$ Cholesky, infeasible as Kn is routinely on the order of hundreds of millions. However, a dither argument shows the cross-class terms contribute nothing to the expected distortion, and a separability assumption, which we test empirically, reduces the weight to $\bar{\lambda} \otimes \mathbb{E}_X[XX^\top]$: one $n \times n$ factorization rescaled per class, encodable by successive interference cancellation from two statistics read off a single forward pass. The raw allocation is coarsest on classes with vanishing calibration probability, exactly where the expansion is least trustworthy, so we smooth the calibration distribution with a prior that caps the grid spacing on those rows. The prior mixing coefficient interpolates between two regimes: at one end the calibration data sets the allocation per class, at the other every class gets equal allocation and we recover WaterSIC. The resulting method, *SoftWater*, contains the WMSE treatment of the softmax layer as the uniform-class case.

Our contributions are:

- **Class-aware geometry of softmax quantization.** We pose softmax-layer quantization as a rate-distortion problem under output KL, derive the induced error weight, and show that it separates into a feature-side and a class-side factor, the latter rescaling the lattice by class frequency and variance (§3–4.3).
- **The SoftWater algorithm.** A class-rescaled SIC lattice with a smoothing prior that keeps the allocation inside the regime where the second-order expansion is valid, costing one extra K-vector in the same calibration pass and leaving the decode format unchanged (§4.4).
- **LLM head quantization.** Across five models from 1B to 32B parameters, SoftWater outperforms the released WaterSIC quantizer at matched head rates on 59 of 60 test points while using none of that pipeline's refinements, transfers unchanged to released GuidedQuant bodies Kim et al. (2025), and cuts whole-model size by up to 60%. The class-side statistic carries domain information not captured by activation covariance alone, so head quantization can be targeted at its deployment domain (§5).

2 Prior work

We review the WMSE formulation of linear layer quantization, the WaterSIC algorithm whose SIC encoder we reuse, and recent work beyond MSE objectives.

2.1 Linear layer quantization

A linear layer takes an activation $X \in \mathbb{R}^n$ and a weight matrix $W \in \mathbb{R}^{K \times n}$ and outputs $z = WX$. Quantization looks for $\hat{W} \in \mathbb{R}^{K \times n}$ minimizing the distortion

$$D = \frac{1}{Kn}\mathbb{E}\|(\hat{W} - W)X\|_F^2 = \frac{1}{Kn}\sum_{k=1}^{K}(\hat{W}_k - W_k)\Sigma_X(\hat{W}_k - W_k)^\top, \tag{1}$$

where W_k and $\hat{W}_k$ are rows and $\Sigma_X = \mathbb{E}_X[XX^\top]$. Most calibration-based PTQ methods target (1): GPTQ and its successors by sequential rounding (Frantar et al., 2023a;b; Zhang et al., 2026; Chen

et al., 2026), QuIP-style methods by incoherence processing and structured codebooks (Chee et al., 2024; Tseng et al., 2024; 2025; Savkin et al., 2025).

2.2 THE WATERSIC ALGORITHM

(Lifar et al., 2026) solve (1) within 0.255 bits of the information-theoretic limit (Lifar et al., 2026, Theorem 3.3) by allocating rates *per column* from Σ_X. Their algorithm, WaterSIC, restricts $\hat{W}_k$ to a scaled integer lattice $\mathbb{Z}^{1\times n}A$, where $A = \mathrm{diag}(\alpha_1, \cdots, \alpha_n)$ with $|A|^{1/n} = \bar{\alpha}$: the lattice has the same point density $\bar{\alpha}^{-n}$ as $\bar{\alpha}\mathbb{Z}^{1\times n}$, but each α_i sets the grid resolution of column i, which entropy coding turns into a per-column bit rate. The objective reads

$$D = \frac{1}{Kn}\sum_{k=1}^{K} \|(\hat{W}_k A - W_k)L\|^2, \tag{2}$$

where L is the lower triangular Cholesky factor of $\Sigma_X = LL^\top$, and $\hat{W}_k$ is restricted to $\mathbb{Z}^{1\times n}$. WaterSIC approximately solves this NP-hard problem with successive interference cancellation (SIC), exploiting the triangular structure of L: columns are encoded in sequence by scalar equations that account for previously quantized entries. All rows are processed identically, as (2) decouples them.

To set the scales A, WaterSIC assumes the rows W_k are iid $\mathcal{N}(0, \sigma_W^2 I)$ and $\sigma_W/\bar{\alpha} \to \infty$. Then the quantization error approaches a uniform distribution (Lifar et al., 2026, Lemma 3.2, Appendix B), and with per-column entropy coding the rate approaches, as K grows,

$$R_{\text{SIC}} \approx R_{\text{High-Rate}}(D, \Sigma_X) + \frac{1}{2}\log\left(\frac{2\pi e}{12}\right) + \frac{1}{2}\log\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(\alpha_i \ell_{ii})^2}{\prod_{i=1}^{n}(\alpha_i \ell_{ii})^{2/n}}\right), \tag{3}$$

where $R_{\text{High-Rate}}(D, \Sigma_X)$ is the optimal rate and ℓ_{ii} is the i-th diagonal entry of L. The choice

$$\alpha_i = c_{\text{WS}}/|\ell_{ii}|, \quad i = 1, \cdots, n, \tag{4}$$

kills the third term and attains the $(1/2)\log(2\pi e/12) \approx 0.255$ bit gap; here $c_{\text{WS}} > 0$ sets the lattice density, and $c_{\text{WS}} = \bar{\alpha}|L|^{1/n}$ recovers the density $\bar{\alpha}^{-n}$.

The released pipeline adds several modifications (Lifar et al., 2026, see §4). We keep the two that any unequal-rate scheme needs: *entropy coding*, which turns the integer codes into a bitstream and is what makes unequal per-column rates realizable, and *rate assignment*, which tunes c_{WS} by the secant method since the post-coding rate cannot be set directly. The rest are quality refinements orthogonal to rate allocation, and we do not use them, so the gains in §5 are attributable to the allocation alone and are in principle additive with them.

2.3 QUANTIZATION METRICS BEYOND MSE

WMSE is the standard metric in model quantization (e.g., Nagel et al., 2020; Frantar et al., 2023a), mainly because it is tractable: it decouples rows, admits tools like Cholesky decompositions, connects directly to rate-distortion theory, and needs forward passes only. But it is a proxy for output behavior, and recent work has looked for better metrics; (Lifar et al., 2026) list approximating a target perplexity or KL loss as their first open direction. (Lee et al., 2026) swap MSE for the cross entropy between full-precision and quantized logits in the final block *before* the head. (Kim et al., 2025) weight the layerwise objective by end-loss guidance, made tractable by assuming Fisher coefficients coupling different output channels vanish, so the objective again decouples across rows; this is the end-to-end analogue of the cross-row decorrelation we derive in §4.3, which for the softmax layer we can justify rather than assume, and their released pipeline produces the quantized bodies of §5. (Tseng et al., 2026) build YAQA on tractable Hessian sketches of end-to-end KL, also decomposing the Hessian as a Kronecker product, but target adaptive rounding for body layers, whereas we derive the layer-local KL geometry of the softmax layer and its rate allocation. A separate line modifies the softmax layer itself to cut inference cost (Grave et al., 2017; Jean et al., 2015; Liu et al., 2025a; Shao et al., 2025a; Tranheden et al., 2026); we keep the architecture and quantize the existing layer, which composes with those approaches.

3 PROBLEM SETUP

We consider a softmax layer that takes an input $X \in \mathbb{R}^n$ with n features and produces a distribution $p = \mathrm{softmax}(z) \in \mathbb{R}^K$ over K classes, where $z = WX$ are the logits, $W \in \mathbb{R}^{K \times n}$ are the weights with rows $W_k \in \mathbb{R}^{1 \times n}$, and

$$p_k = \mathrm{softmax}(z)_k = \frac{\exp(z_k)}{\sum_{l=1}^{K} \exp(z_l)}.$$

We treat X as a random vector drawn from the distribution of layer inputs induced by the data the model is deployed on, and estimate expectations with respect to it from a calibration sample. Both z and p are therefore functions of X; we suppress the dependence except where it matters.

We replace W by a matrix $\hat{W}$ with entries represented with a budget of R bits per weight. For LLM heads, K is the vocabulary size and W can hold a large fraction of all parameters (Table 1), often tied to the input embedding. Let $\Delta = \hat{W} - W$ and $\delta = \Delta X$ denote the weight and logit perturbations, with rows $\Delta_k \in \mathbb{R}^{1 \times n}$ and entries $\delta_k = \Delta_k X$. The class distribution of the quantized layer is then $q \in \mathbb{R}^K$, defined by $q_k = \mathrm{softmax}(z + \delta)_k$. Our goal is to find, for a given rate R, the $\hat{W}$ that minimizes the distortion between p and q, which §4.2 formalizes as an expected KL divergence.

4 SOFTWATER: PER-COLUMN AND PER-CLASS RATE ALLOCATION

The KL geometry of a softmax layer induces rate allocation along two axes: input features and output classes. SoftWater allocates along both. A separability approximation makes the resulting lattice encodable by SIC at $O(n^3)$ cost. We derive its objective, reduce it to a tractable problem, add a smoothing step that keeps the proxy accurate, and contrast linear layer and softmax quantization.

4.1 LINEAR LAYER VS SOFTMAX QUANTIZATION

In §2.2 we saw that SIC's error geometry over the Cholesky factor L, combined with entropy coding, points to grid spacing proportional to $|\ell_{ii}|^{-1}$. Our goal in this section is to show that, for softmax quantization, the grid spacing should instead be proportional to $(\bar{\lambda}_k^{1/2}|\ell_{ii}|)^{-1}$, where

$$\bar{\lambda}_k = \mathbb{E}_X[p_k(1 - p_k)] = \mathbb{E}_X[p_k](1 - \mathbb{E}_X[p_k]) - \mathrm{Var}[p_k].$$

That is, the grid spacing is inversely proportional to the square root of class frequency discounted by cross-context variance, the spread of p_k across contexts. Therefore, through $(\bar{\lambda}_k^{1/2}|\ell_{ii}|)^{-1}$,

> *SoftWater corrects for feature geometry, class frequency and class variance: frequent, low-variance classes get more bits than rare, high-variance ones.*

How far SoftWater's spacings depart from WaterSIC's is determined by the class distribution, not by the algorithm. For a near-uniform distribution the $\bar{\lambda}_k$ are nearly equal and SoftWater recovers WaterSIC. For the heavy tails of natural-language vocabularies, $\mathbb{E}_X[p_k]$ spans many orders of magnitude and the spread in $\bar{\lambda}_k^{-1/2}$ is large, bounded only by the smoothing floor of §4.4. This is why the class side, not the feature side, is where a softmax layer's rate budget is won or lost. Solving the softmax problem through WMSE therefore amounts to assuming identically distributed classes.

4.2 OBJECTIVE DERIVATION

Linear layer quantization measures distortion by the output WMSE. A softmax output has more structure: it lies on the probability simplex and is a distribution. We therefore use the KL divergence, and look for $\hat{W}$ minimizing $D_{\mathrm{KL}}(p \,\|\, q) = \sum_k p_k(\log p_k - \log q_k)$. Writing

$$\log p_k = z_k - \log \sum_{l=1}^{K} \exp(z_l), \quad \log q_k = z_k + \delta_k - \log \sum_{l=1}^{K} \exp(z_l + \delta_l),$$

and substituting gives $D_{\mathrm{KL}}(p \,\|\, q) = \log \sum_{l=1}^{K} \exp(z_l + \delta_l) - \log \sum_{l=1}^{K} \exp(z_l) - p^\top \delta$. Let $f(\delta) = D_{\mathrm{KL}}(p \,\|\, q)$, with $\nabla f(\delta) = \mathrm{softmax}(z+\delta) - p$ and $\nabla^2 f(\delta) = \mathrm{diag}(q) - qq^\top$. Since $f(0) = 0$ and $\nabla f(0) = 0$, a second-order Taylor expansion at $\delta = 0$ gives

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) &= f(0) + \nabla f(0)^\top \delta + \tfrac{1}{2}\delta^\top \nabla^2 f(0)\delta + O(\|\delta\|^3) \\ &\approx \tfrac{1}{2}\mathrm{vec}(\Delta)^\top (\nabla^2 f(0) \otimes XX^\top)\mathrm{vec}(\Delta), \end{aligned} \tag{5}$$

where $\mathrm{vec}(\Delta) \in \mathbb{R}^{Kn}$ stacks the rows of Δ and $\nabla^2 f(0) = \mathrm{diag}(p) - pp^\top$. The expansion (5) holds as long as the logit perturbation stays small; §4.4 enforces this. The objective is then

$$\mathbb{E}_X\left[\tfrac{1}{2}\mathrm{vec}(\Delta)^\top (\nabla^2 f(0) \otimes XX^\top)\mathrm{vec}(\Delta)\right] = \tfrac{1}{2}\mathrm{vec}(\Delta)^\top \mathbb{E}_X[\nabla^2 f(0) \otimes XX^\top]\mathrm{vec}(\Delta). \tag{6}$$

The error weight is thus $\mathbb{E}_X[\nabla^2 f(0) \otimes XX^\top]$, a $Kn \times Kn$ matrix coupling classes and features. Encoding against it would need a Cholesky factorization of that $Kn \times Kn$ matrix, costing $O(K^3 n^3)$ time and $O(K^2 n^2)$ memory. Instead, we obtain a tractable approximation exploiting its structure.

4.3 A TRACTABLE PROXY

The obstacle is that the expectation couples $\nabla^2 f(0)$ and $XX^\top$, giving sample-dependent off-diagonal terms and a full matrix. If we can decouple them, then $\mathrm{chol}(A \otimes B) = \mathrm{chol}(A) \otimes \mathrm{chol}(B)$ splits the factorization into two smaller ones.

Under SIC, rows are quantized independently against the same L, with sequential correction along coordinates within each row: errors are coupled *within* a row, not *across* rows. Per-entry subtractive dither makes this precise. Under the dither ensemble the errors are zero-mean and uncorrelated across rows, and the dither is drawn independently of the data, so the expectations over X and Δ factor. The cross-terms $-p_k p_l$, $k \neq l$, then contribute nothing, and the class-side diagonal is exact:

$$\mathbb{E}_{X,\Delta}[\mathrm{vec}(\Delta)^\top (\nabla^2 f(0) \otimes XX^\top)\mathrm{vec}(\Delta)] = \mathbb{E}_{X,\Delta}[\mathrm{vec}(\Delta)^\top (\lambda \otimes XX^\top)\mathrm{vec}(\Delta)], \tag{7}$$

where λ is the softmax curvature

$$\lambda = \mathrm{diag}(p \odot (1 - p)). \tag{8}$$

Replacing the error weight in (6) by its class-side diagonal gives

$$\tfrac{1}{2}\mathrm{vec}(\Delta)^\top \mathbb{E}_X[\lambda \otimes XX^\top]\mathrm{vec}(\Delta), \tag{9}$$

again a deterministic function of the perturbation. Dithering is thus the device that identifies (9) as the right target; empirically, undithered errors already behave as if uncorrelated across rows, so we drop it from the final scheme and leave it as an optional tweak.

The matrix $\mathbb{E}_X[\lambda \otimes XX^\top]$ is block diagonal, so K Cholesky decompositions of size n replace one of size Kn. Except for small K that is still too expensive, so we assume λ and $XX^\top$ are *separable*.

Assumption 1 (Separability). *The softmax curvature λ and $XX^\top$ are uncorrelated:*

$$\mathbb{E}_X[\lambda \otimes XX^\top] = \mathbb{E}_X[\lambda] \otimes \mathbb{E}_X[XX^\top]. \tag{10}$$

We confirm the validity of separability in §5.6. It lets the two Kronecker factors be estimated independently: the covariance Σ_X separately from the expected curvature

$$\bar{\lambda} = \mathbb{E}_X[\lambda] = \mathbb{E}_X[\mathrm{diag}(p \odot (1 - p))]. \tag{11}$$

Moreover, $\bar{\lambda}$ is just K scalars, accumulated from the output distributions the calibration forward pass already produces: SoftWater needs no backward pass and no second pass over the data, and costs one K-vector alongside the $n \times n$ covariance, both from the same forward pass. The Cholesky factor L of Σ_X is computed once and rescaled K times to give the factor $\mathcal{L} = \bar{\lambda}^{1/2} \otimes L$ of $\bar{\lambda} \otimes \Sigma_X$. The distortion metric is then

$$D = \tfrac{1}{2}\mathrm{vec}(\Delta)^\top \mathcal{L}\mathcal{L}^\top \mathrm{vec}(\Delta), \tag{12}$$

a WMSE over the joint index (k, i), which SIC minimizes directly on $\mathcal{L}$.

The induced grid spacing for class k and column i is

$$\alpha_{(k-1)n+i} = c_{\mathrm{SW}}/(\bar{\lambda}_k^{1/2}|\ell_{ii}|) \quad k = 1, \cdots, K, \quad i = 1, \cdots, n, \tag{13}$$

where $c_{\mathrm{SW}} > 0$ again sets the point density. Letting

$$\mathcal{B} = \mathrm{diag}(\bar{\lambda}_1^{-1/2}, \cdots, \bar{\lambda}_K^{-1/2}), \qquad \text{and} \qquad \mathcal{A} = \mathrm{diag}(|\ell_{11}|^{-1}, \cdots, |\ell_{nn}|^{-1}),$$

and choosing

$$c_{\mathrm{SW}} = \bar{\alpha}|\mathcal{B}|^{-1/K}|\mathcal{A}|^{-1/n}, \tag{14}$$

gives point density $\bar{\alpha}^{-Kn}$. SoftWater therefore uses $c_{\mathrm{SW}}\mathcal{B}\mathbb{Z}^{K\times n}\mathcal{A}$, rescaled on the class side as well as the feature side. The hyperparameter c_{SW} is the sole rate knob: it scales the lattice uniformly, raises the post-coding rate monotonically, and is set by secant search (§2.2).

Remark. We notice that $\mathcal{A}$ and $\mathcal{B}$ are not just column and row rescalers; they change the grid spacing itself, redistributing rate across classes through the coupling with c_{SW}. In contrast, plain column and row rescalers applied after coding change reconstruction but not rate.

On the separability assumption. Although $\lambda_k = \lambda_k(X)$ depends on X, separability does not require independence between class probabilities and the input. It requires only that the class-dependent curvature be weakly correlated with the second-order geometry of X. The error in the k-th block is

$$\mathbb{E}_X[\lambda_k(X)XX^\top] - \bar{\lambda}_k\Sigma_X = \mathbb{E}_X\left[(\lambda_k(X) - \bar{\lambda}_k)(XX^\top - \Sigma_X)\right], \tag{15}$$

so the approximation is accurate when examples with greater curvature on class k do not have systematically different feature covariance. More directly, for the k-th row Δ_k of the perturbation,

$$\Delta_k\left(\mathbb{E}_X[\lambda_k(X)XX^\top] - \bar{\lambda}_k\Sigma_X\right)\Delta_k^\top = \mathrm{Cov}_X\left(\lambda_k(X), (\Delta_k X)^2\right). \tag{16}$$

The proxy therefore need not match each block in every direction. It is enough that class uncertainty be weakly correlated with feature energy along the error directions the algorithm produces. This is plausible when representations are normalized and their geometry is stable across contexts. If $\|X\|^2$ is constant, for instance, the exact and separable blocks have the same trace,

$$\mathrm{Tr}\left(\mathbb{E}_X[\lambda_k(X)XX^\top]\right) = \bar{\lambda}_k\,\mathrm{Tr}(\Sigma_X),$$

so separability preserves the total curvature per class and drops only class-dependent anisotropy.

4.4 A TRACTABLE AND ACCURATE PROXY: SMOOTHING

The proxy (12) is accurate only if the above assumptions hold, which we verify in §5, and if the Taylor expansion of $D_{\mathrm{KL}}(p \,\|\, q)$ stays accurate, which requires a small logit perturbation δ.

Grid spacing under our metric is proportional to $(\bar{\lambda}_k^{1/2}|\ell_{ii}|)^{-1}$, so rare classes get coarser grids. If a class is absent from the calibration data, then $p_k \ll 1$ and $\bar{\lambda}_k^{1/2} \ll 1$, the grid is far too coarse, the weights are badly perturbed, and (5) breaks down. We therefore smooth p to lower bound p_k, so that uncovered classes thus keep logit perturbations inside the regime where (5) is valid, by computing

$$\tilde{\lambda} = \mathbb{E}_X[\mathrm{diag}(\tilde{p} \odot (1 - \tilde{p}))], \tag{17}$$

where, given $\epsilon \in (0, 1)$, $\tilde{p}$ mixes p with a uniform distribution over the K classes:

$$\tilde{p} = (1 - \epsilon)p + \epsilon/K \tag{18}$$

Since $\tilde{p}_k \geq \epsilon/K$, $\tilde{\lambda}_k \geq \frac{\epsilon}{K}(1 - \frac{\epsilon}{K})$, while $\tilde{\lambda}_k \leq 1/4$ always. At $\epsilon = 0$, $\tilde{\lambda}$ comes from the calibration data alone (11), while at $\epsilon = 1$, every class get equal weight, recovering WaterSIC. We fix $\epsilon = 0.1$ as default, and $\tilde{\lambda}$ replaces $\bar{\lambda}$ in the grid spacings (13), the rescalers $\mathcal{B}$, and the constant c_{SW}.

5 EXPERIMENTS

We evaluate SoftWater on five models from 1B to 32B parameters: head-only against an fp16 body (§5.1), then with quantized bodies (§5.2) and downstream tasks (§5.4). Three further experiments target calibration at a deployment domain (§5.3), ablate the class-side statistic (§5.5), and test separability (§5.6). Our arm is the plain allocation, with no rescaler optimization, LMMSE shrinkage or finetuning; the baseline is the released WaterSIC pipeline minus finetuning. See A for full details.

Algorithm 1 SoftWater weight-only quantization of softmax

Require: $W \in \mathbb{R}^{K\times n}$, $\Sigma_X \succeq 0$, $\tilde{\lambda} = \mathbb{E}_X[\mathrm{diag}(\tilde{p} \odot (1-\tilde{p}))]$, $\bar{\alpha} > 0$.
Ensure: $\alpha \in \mathbb{R}^n_+$, $\beta \in \mathbb{R}^K_+$, $B_1, \ldots, B_n \in \{0,1\}^*$ s.t. $\hat{W} = \mathrm{diag}(\beta)\, Z_{\mathrm{SIC}}\, \mathrm{diag}(\alpha)$.

Compute lower-triangular $L \in \mathbb{R}^{n\times n}$ s.t. $\Sigma_X = LL^\top$ ▷ Cholesky
$\beta_k \leftarrow \frac{(\prod_l \tilde{\lambda}_l^{1/2})^{1/K}}{\tilde{\lambda}_k^{1/2}}, \quad \forall k \in [K]$ ▷ class-side scales
$\alpha_i \leftarrow \bar{\alpha} \cdot \frac{|L|^{1/n}}{|\ell_{ii}|}, \quad \forall i \in [n]$ ▷ feature-side water filling; ℓ_{ii} is the ith diagonal element of L
$R \leftarrow WL$
for $i = n : 1$ **do**
  $Z_{\mathrm{SIC},:,i} \leftarrow \mathrm{round}\big(R_{:,i} \oslash (\alpha_i \ell_{ii}\, \beta)\big)$ ▷ $\oslash$: elementwise division by the K-vector $\alpha_i \ell_{ii}\, \beta$
  $R \leftarrow R - \alpha_i \big(\beta \odot Z_{\mathrm{SIC},:,i}\big) \cdot L_{i,:}$ ▷ $\odot$: elementwise product; $L_{i,:}$ is the ith row of L
end for
for $i = 1 : n$ **do**
  $B_i \leftarrow \mathrm{EC}\big(Z_{\mathrm{SIC},:,i}\big)$ ▷ entropy coding of the ith column
end for

5.1 Head-only isolation

We quantize the heads of five models to 2, 3 and 4 bits and keep their bodies in fp16; Llama-3.2-1B and Qwen3-1.7B tie head and embedding, so a separate fp16 embedding is kept. The baseline is the official WaterSIC release without finetuning, so both arms are purely layerwise, with nominal damping $\delta = 10^{-6}\,\mathrm{mean}(\mathrm{diag}\Sigma_X)$. Both are calibrated on 128 sequences of 1024 tokens from WikiText2 (WT2), and we measure perplexity (PPL) and $D_{\mathrm{KL}}\left(p_{\text{base}} \,\|\, p_{\text{quant}}\right)$ on WT2 and C4.

Tables 2 and 5, in A.1, show the results. Every arm lands within 0.005 bits of its target rate, and SoftWater wins 59 of 60 cells; in all models a 2-bit SoftWater head outperforms a 3-bit WaterSIC head on WT2 KL. The KL advantage is $6.5\times$–$8.3\times$ on WT2 at 2 bits and nearly flat in head rate, but smaller out of domain ($3.5\times$–$5.1\times$ on C4); perplexity behaves differently, with the gap to fp16 nearly closing at 4 bits.

Table 2: Head-only quantization, fp16 body. KL is measured against the all-fp16 model on identical token streams. WS: WaterSIC w/o FT; SW: SoftWater (ours). Vocabulary size in parentheses. Best per head rate in bold; full results in Table 5.

Method	Head	WT2 PPL	WT2 KL	C4 PPL	C4 KL
Llama-3.2-1B (128,256)					
fp16	16	11.313	0	15.614	0
WS	2	12.049	0.0619	17.283	0.0996
SW	2	**11.419**	**0.0096**	**16.035**	**0.0283**
WS	3	11.469	0.0136	15.965	0.0218
SW	3	**11.342**	**0.0021**	**15.704**	**0.0057**
WS	4	11.356	0.0032	15.713	0.0053
SW	4	**11.320**	**0.0005**	**15.640**	**0.0014**
Qwen3-32B (151,936)					
fp16	16	8.452	0	12.269	0
WS	2	8.756	0.0584	13.029	0.0811
SW	2	**8.498**	**0.0070**	**12.479**	**0.0160**
WS	3	8.500	0.0125	12.439	0.0168
SW	3	**8.464**	**0.0015**	**12.315**	**0.0034**
WS	4	8.477	0.0030	12.302	0.0040
SW	4	**8.457**	**0.0004**	**12.275**	**0.0008**

5.2 Quantized body

We next evaluate two models with quantized bodies: Llama-3.2-1B-Instruct, keeping the tie so the quantized head also serves as the input embedding, and Qwen2.5-7B-Instruct, whose untied head is quantized while the embedding stays in fp16. Since the WaterSIC repository has no quantized bodies, we use released GuidedQuant Kim et al. (2025) and VPTQ Liu et al. (2024) checkpoints (A), and both quantizers calibrate on hidden states of the frozen quantized-body model, so Σ_X and $\tilde{\lambda}$ reflect what the head sees at deployment.

Tables 3 and 7 show the results; SoftWater is better in 71 of 72 cells. On Llama-3.2-1B-Instruct the head is 52–68% of the stored bytes once the body is quantized, so quantizing it is the only remaining way to shrink the model: a 2-bit SoftWater head removes 45–60% of the bytes for a 2.9–3.7% WT2 perplexity increase, against 9.7–17.6% for WaterSIC, and a 4-bit head costs 0.2–0.3% while still removing at least 38%. On Qwen2.5-7B-Instruct the head is 19–27% of the bytes, so savings are 14–24%, but the geometry is the same: a 2-bit SoftWater head costs under 1% perplexity against 3.4–4.5% for WaterSIC. In both models the head dominates as the body improves: on the 4-bit bodies a 2-bit head raises WT2 KL by 102% (Llama) and 24% (Qwen) under SoftWater, against 396% and 186% under the released quantizer.

Table 3: Perplexity (PPL) on WikiText2 (WT2) and C4, and KL divergence from baseline: fp16-head, body quantized at 2 and 4 bits. Body and head columns give respective rates. WS: WaterSIC w/o FT. Model size and reduction (Δsize) after quantizing the tied head. Best value in bold.

Llama-3.2-1B-Instruct

Body	Method	Head	Size (MiB)	Δsize	WT2 PPL (↓)	WT2 KL (↓)	C4 PPL (↓)	C4 KL (↓)
2	fp16	16	737.3	0.0%	26.81	0.497	38.54	0.511
	WS	2	297.0	−59.7%	31.53 (+17.6%)	0.671 (+34.9%)	47.14 (+22.3%)	0.732 (+43.3%)
	SW (ours)	2	297.0	−59.7%	**27.80 (+3.7%)**	**0.536 (+7.9%)**	**42.39 (+10.0%)**	**0.607 (+18.8%)**
	WS	3	327.7	−55.6%	27.60 (+3.0%)	0.533 (+7.2%)	40.75 (+5.7%)	0.570 (+11.7%)
	SW (ours)	3	327.7	−55.6%	**27.03 (+0.8%)**	**0.506 (+1.7%)**	**39.49 (+2.5%)**	**0.536 (+4.9%)**
	WS	4	358.4	−51.4%	27.16 (+1.3%)	0.510 (+2.6%)	39.03 (+1.3%)	0.524 (+2.6%)
	SW (ours)	4	358.4	−51.4%	**26.89 (+0.3%)**	**0.501 (+0.7%)**	**38.99 (+1.2%)**	**0.521 (+2.1%)**
4	fp16	16	962.6	0.0%	16.53	0.028	24.54	0.028
	WS	2	532.5	−44.7%	18.14 (+9.7%)	0.138 (+396.2%)	28.11 (+14.5%)	0.178 (+531.5%)
	SW (ours)	2	532.5	−44.7%	**17.01 (+2.9%)**	**0.056 (+101.9%)**	**26.60 (+8.4%)**	**0.106 (+276.9%)**
	WS	3	563.2	−41.5%	16.83 (+1.8%)	0.052 (+88.9%)	25.26 (+2.9%)	0.063 (+122.3%)
	SW (ours)	3	563.2	−41.5%	**16.64 (+0.7%)**	**0.034 (+24.1%)**	**25.12 (+2.3%)**	**0.052 (+84.6%)**
	WS	4	594.0	−38.3%	16.65 (+0.7%)	0.036 (+28.0%)	**24.70 (+0.6%)**	0.037 (+32.4%)
	SW (ours)	4	594.0	−38.3%	**16.58 (+0.3%)**	**0.030 (+9.3%)**	24.77 (+0.9%)	**0.035 (+25.3%)**

5.3 Domain-targeted calibration

Rate allocation varies across domains, so we target calibration to five themes and evaluate in and out of domain; Table 8 in the appendix gives the corpora, splits and fields. We calibrate on 1.05M tokens rather than 131k, since at 131k most of the vocabulary is never observed and the domains would not be distinguishable through $\tilde{\lambda}$.

In A.3, Tables 9 and 10 show the in-domain and full results. The targeting claim holds exactly on KL: for every evaluation domain and head rate, the lowest KL among the ten arms is the SoftWater head calibrated on that domain, and at 2 bits matching calibration lowers KL by $2.0\times$ to $5.0\times$ over the best mismatched arm. On the diagonal SoftWater also outperforms the released quantizer everywhere at 2 bits, by $1.9\times$ to $6.5\times$ in KL, and wins 41 of 50 stream cells overall; seven of the nine losses are the code column under non-code calibration, the expected cost of sharpening allocation toward calibration-frequent tokens.

5.4 Zero-shot tasks

Perplexity and KL measure exactly what SoftWater controls, so we now check downstream accuracy. We reuse the §5.2 setup with identical head arms, and evaluate zero-shot with lm-evaluation-harness 0.4.12 on ARC-Easy, ARC-Challenge, HellaSwag, LAMBADA, WinoGrande, PIQA and OpenBookQA, reporting accuracy and normalized accuracy where defined.

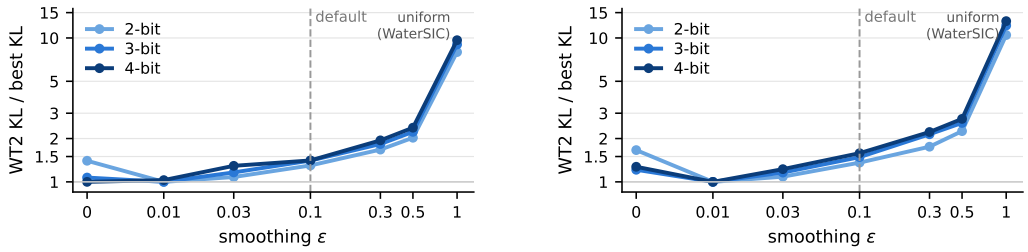
Table 4 shows 2-bit heads on all bodies. Tables 11–13 give every head rate. At a 2-bit head SoftWater roughly halves the mean accuracy lost to head quantization on every body. The gap narrows with head rate and closes at 4 bits. The difference is concentrated in LAMBADA, the only task scored by next-token prediction rather than by ranking continuations, and so the one KL controls directly: at 2 bits SoftWater gains 5.2 to 7.1 points. Sharpening the grids of frequent, low-variance classes protects exactly the token-level distribution KL measures.

Table 4: Zero-shot accuracy of Llama-3.2-1B-Instruct, 2-bit tied head. Full grid in Tables 11–13.

Body	Method	LAMB	Mean
16	fp16 head	60.1	56.72
2	fp16 head	46.3	48.86
	WS	36.8	45.89
	SW (ours)	43.9	**47.63**
3	fp16 head	59.5	55.35
	WS	53.1	52.68
	SW (ours)	58.4	**54.43**
4	fp16 head	60.1	56.30
	WS	54.0	54.39
	SW (ours)	59.2	**55.23**

5.5 Class-side statistic ablation

For most tokens $\tilde{p}_k(1-\tilde{p}_k) \approx \tilde{p}_k$ since $\tilde{p}_k$ is small, so we test if second-moment correction matters: we replace $\mathbb{E}_X[\tilde{p}_k(1-\tilde{p}_k)]$ with $\bar{p}_k = \mathbb{E}_X[\tilde{p}_k]$, and also $\tilde{\pi}_k$, the *unigram frequency* of token k smoothed the same way, which requires no forward pass and asks whether the model carries information the corpus does not.



(a) Llama-3.2-1B.

(b) Qwen3-32B.

Figure 1: WT2 KL/best KL at that head rate over several ϵ values in (18).

Table 14 shows the results. SoftWater and the $\bar{p}_k$-variant are close in domain, but SoftWater wins every out-of-domain KL cell and 12 of 15 in-domain ones, so the second-moment correction is small but systematic and is what makes the allocation transfer off-domain. Both variants beat $\tilde{\pi}_k$ at every rate and model, by roughly $2\times$ in KL, so the model's own marginal carries information the corpus count does not. Since $\tilde{\lambda}_k$ and $\bar{p}_k$ cost the same, SoftWater is the best of the three.

5.6 Separability

To test Assumption 1, we compare the true distortion $\frac{1}{2}\sum_k \Delta_k \mathbb{E}_X[\lambda_k(X)XX^\top]\Delta_k^\top$ with the proxy $\frac{1}{2}\sum_k \tilde{\lambda}_k \Delta_k \Sigma_X \Delta_k^\top$ that SoftWater optimizes; their ratio minus one, ρ, carries a rate penalty of $\frac{1}{2}\log_2(1+\rho)$ bits and includes the deliberate bias from smoothing. With heads calibrated on WT2 and both sums computed on WT2 and C4, Table 15a shows the proxy underestimates the true distortion by at most 10%, a penalty of at most 0.07 bits. To ask whether a better Kronecker factorization exists, we run the power iteration of (Tseng et al., 2026) (Appendix B) on $\mathbb{E}_X[\tilde{\lambda}(X)\otimes XX^\top]$, whose fixed point is the Frobenius-optimal factorization; from three initializations, $(\tilde{\lambda}, \Sigma_X)$ (SoftWater), (I, Σ_X) (WaterSIC) and (I, I), Table 15b shows the SoftWater factors are already close to optimal.

5.7 Smoothing ϵ ablation

We conclude the experiments with an ablation of the smoothing coefficient ϵ in (18). Figure 1 shows the KL ratio for each ϵ over the best KL at that head rate on WT2 for two models. In the range $[0, 0.3]$, the ratio is mostly flat, quickly rising as ϵ approaches 1, corresponding to the uniform class prior that applying WaterSIC to softmax quantization would implicitly assume.

6 Conclusion

A softmax layer is not a linear layer with a nonlinearity attached: its output is a distribution, and the distortion that matters is the change in that distribution. Taking KL divergence as the metric, we showed that the rate-distortion geometry of softmax quantization is class-aware, weighting error by class-specific curvature as well as feature covariance, and that the WMSE treatment is the special case in which every class is equally likely. A separability approximation makes this geometry tractable, reducing it to one feature-side factorization with class-dependent rescaling from statistics gathered in a single calibration pass, and a smoothing prior keeps the allocation inside the regime where the second-order analysis holds. The resulting method, SoftWater, is the practical consequence of this analysis. On LLM heads, where highly-imbalanced vocabularies make the class-side gap largest, it outperforms the near-optimal WMSE quantizer WaterSIC at every head rate across five models from 1B to 32B parameters, halves the zero-shot accuracy lost to a 2-bit head, and targets a deployment domain through its calibration statistic, so that the head, which can hold over half the stored bytes if left in high precision, can be quantized aggressively with little loss. A limitation bound these results: the separable surrogate is accurate to within 10% on the models we measure but not guaranteed in general. The analysis itself is not specific to language models: it applies to any linear-softmax layer with a fixed class dimension, including vision and speech classifiers and MoE routers, whose skewed expert utilization mirrors the class imbalance we exploit.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks. Additionally, we used generative AI tools to suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyze existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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, October 2024. URL http://arxiv.org/abs/2404.00456. arXiv:2404.00456 [cs].

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-Bit Quantization of Large Language Models With Guarantees, January 2024. URL http://arxiv.org/abs/2307.13304. arXiv:2307.13304 [cs].

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The Geometry of LLM Quantization: GPTQ as Babai's Nearest Plane Algorithm, May 2026. URL http://arxiv.org/abs/2507.18553. arXiv:2507.18553 [cs.LG].

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, March 2023a. URL http://arxiv.org/abs/2210.17323. arXiv:2210.17323 [cs].

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning, January 2023b. URL http://arxiv.org/abs/2208.11580. arXiv:2208.11580 [cs].

Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for GPUs, June 2017. URL http://arxiv.org/abs/1609.04309. arXiv:1609.04309 [cs.CL].

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1–10, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1001. URL http://aclweb.org/anthology/P15-1001.

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens JS Schaefer, Deokjae Lee, Yeonhong Park, Jae W. Lee, and Hyun Oh Song. GuidedQuant: Large Language Model Quantization via Exploiting End Loss Guidance, September 2025. URL http://arxiv.org/abs/2505.07004. arXiv:2505.07004 [cs.LG].

Jung Hyun Lee, June Yong Yang, Jungwook Choi, and Eunho Yang. LFQ: Logit-aware Final-block Quantization for Boosting the Generation Quality of Low-Bit Quantized LLMs, May 2026. URL http://arxiv.org/abs/2605.29756. arXiv:2605.29756 [cs.AI].

Egor Lifar, Semyon Savkin, Or Ordentlich, and Yury Polyanskiy. WaterSIC: Information-Theoretically (Near) Optimal Linear Layer Quantization, June 2026. URL http://arxiv.org/abs/2603.04956. arXiv:2603.04956 [cs.LG].

Dong Liu, Yanxuan Yu, and Ben Lengerich. CSV-Decode: Certifiable Sub-Vocabulary Decoding for Efficient Large Language Model Inference, November 2025a. URL http://arxiv.org/abs/2511.21702. arXiv:2511.21702 [cs.CL].

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, October 2024. URL http://arxiv.org/abs/2409.17066. arXiv:2409.17066 [cs.AI].

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations, February 2025b. URL http://arxiv.org/abs/2405.16406. arXiv:2405.16406 [cs].

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or Down? Adaptive Rounding for Post-Training Quantization, June 2020. URL http://arxiv.org/abs/2004.10568. arXiv:2004.10568 [cs].

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, October 2014. ISSN 1531-5320. doi: 10.3758/s13423-014-0585-6. URL https://doi.org/10.3758/s13423-014-0585-6.

Semyon Savkin, Eitan Porat, Or Ordentlich, and Yury Polyanskiy. NestQuant: Nested Lattice Quantization for Matrix Products and LLMs, July 2025. URL http://arxiv.org/abs/2502.09720. arXiv:2502.09720 [cs.LG].

Jintian Shao, Hongyi Huang, Jiayi Wu, YiMing Cheng, ZhiYu Wu, You Shan, and MingKai Zheng. VQ-Logits: Compressing the Output Bottleneck of Large Language Models via Vector Quantized Logits, May 2025a. URL http://arxiv.org/abs/2505.10202. arXiv:2505.10202 [cs.CL].

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, March 2024. URL http://arxiv.org/abs/2308.13137. arXiv:2308.13137 [cs.LG].

Yuantian Shao, Yuanteng Chen, Peisong Wang, Jianlin Yu, Jing Lin, Yiwu Yao, Zhihui Wei, and Jian Cheng. DartQuant: Efficient Rotational Distribution Calibration for LLM Quantization, November 2025b. URL http://arxiv.org/abs/2511.04063. arXiv:2511.04063 [cs].

Wilhelm Tranheden, Shahnawaz Ahmed, Devdatt Dubhashi, Jonna Matthiesen, and Hannes von Essen. FlashHead: Efficient Drop-In Replacement for the Classification Head in Language Model Inference, March 2026. URL http://arxiv.org/abs/2603.14591. arXiv:2603.14591 [cs.LG].

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even Better LLM Quantization with Hadamard Incoherence and Lattice Codebooks, June 2024. URL http://arxiv.org/abs/2402.04396. arXiv:2402.04396 [cs].

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with Trellises and Incoherence Processing, June 2025. URL http://arxiv.org/abs/2406.11235. arXiv:2406.11235 [cs].

Albert Tseng, Zhaofeng Sun, and Christopher De Sa. Model-Preserving Adaptive Rounding, June 2026. URL http://arxiv.org/abs/2505.22988. arXiv:2505.22988 [cs.LG].

Charles F. Van Loan. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, 123(1-2):85–100, November 2000. ISSN 03770427. doi: 10.1016/S0377-0427(00)00393-9. URL https://linkinghub.elsevier.com/retrieve/pii/S0377042700003939.

Shihao Zhang, Haoyu Zhang, Ian Colbert, and Rayan Saab. Qronos: Correcting the Past by Shaping the Future... in Post-Training Quantization, February 2026. URL http://arxiv.org/abs/2505.11695. arXiv:2505.11695 [cs.LG].

A EXPERIMENT DETAILS

The official WaterSIC released implementation was downloaded from `https://github.com/egorlifar/watersic`. Quantized bodies quantized under GuidedQuant Kim et al. (2025) in the QTIP trellis format Tseng et al. (2025), downloaded from `https://huggingface.co/jusjinuk/Llama-3.2-1B-Instruct-2bit-GuidedQuant-QTIP`, and decoded to dense fp16 with a standalone reimplementation verified against the reference forward. The quantized Qwen2.5-7B-Instruct bodies are the VPTQ-community releases of VPTQ (Liu et al., 2024), downloaded from Hugging Face: `https://huggingface.co/VPTQ-community`.

A.1 HEAD-ONLY ISOLATION

Table 5: Head-only isolation: the body is kept in fp16 and only the head matrix is quantized; tied models (Llama-3.2-1B, Qwen3-1.7B) retain an fp16 copy of the embedding. KL is measured against the all-fp16 model on identical token streams. WS: WaterSIC w/o FT. Parentheses after each model name give its vocabulary size V; on PPL cells, the increase over the fp16 row of the same model; on SW KL cells, the WS/SW KL ratio at that rate. Best value per head group in bold.

Method	Head	WT2 PPL (↓)	WT2 KL (↓)	C4 PPL (↓)	C4 KL (↓)
Llama-3.2-1B (128,256)					
fp16	16	11.313	0	15.614	0
WS	2	12.049 (+6.5%)	0.06192	17.283 (+10.7%)	0.09957
SW (ours)	2	**11.419 (+0.9%)**	**0.00957 (6.5×)**	**16.035 (+2.7%)**	**0.02834 (3.5×)**
WS	3	11.469 (+1.4%)	0.01364	15.965 (+2.2%)	0.02176
SW (ours)	3	**11.342 (+0.3%)**	**0.00206 (6.6×)**	**15.704 (+0.6%)**	**0.00572 (3.8×)**
WS	4	11.356 (+0.4%)	0.00319	15.713 (+0.6%)	0.00526
SW (ours)	4	**11.320 (+0.1%)**	**0.00047 (6.8×)**	**15.640 (+0.2%)**	**0.00138 (3.8×)**
Qwen3-1.7B (151,936)					
fp16	16	18.285	0	22.621	0
WS	2	18.876 (+3.2%)	0.06788	24.129 (+6.7%)	0.09054
SW (ours)	2	**18.460 (+1.0%)**	**0.00900 (7.5×)**	**23.075 (+2.0%)**	**0.01998 (4.5×)**
WS	3	18.374 (+0.5%)	0.01448	22.916 (+1.3%)	0.02017
SW (ours)	3	**18.317 (+0.2%)**	**0.00199 (7.3×)**	**22.724 (+0.5%)**	**0.00443 (4.6×)**
WS	4	**18.277 (−0.0%)**	0.00347	22.650 (+0.1%)	0.00487
SW (ours)	4	18.298 (+0.1%)	**0.00045 (7.7×)**	**22.633 (+0.1%)**	**0.00105 (4.6×)**
Llama-3.1-8B-Instruct (128,256)					
fp16	16	8.388	0	14.189	0
WS	2	8.758 (+4.4%)	0.04583	15.462 (+9.0%)	0.07487
SW (ours)	2	**8.437 (+0.6%)**	**0.00636 (7.2×)**	**14.534 (+2.4%)**	**0.02140 (3.5×)**
WS	3	8.458 (+0.8%)	0.00940	14.439 (+1.8%)	0.01622
SW (ours)	3	**8.398 (+0.1%)**	**0.00135 (7.0×)**	**14.258 (+0.5%)**	**0.00437 (3.7×)**
WS	4	8.402 (+0.2%)	0.00221	14.255 (+0.5%)	0.00383
SW (ours)	4	**8.391 (+0.0%)**	**0.00032 (6.8×)**	**14.205 (+0.1%)**	**0.00103 (3.7×)**
Qwen3-8B (151,936)					
fp16	16	10.832	0	15.486	0
WS	2	11.081 (+2.3%)	0.04592	16.229 (+4.8%)	0.06861
SW (ours)	2	**10.898 (+0.6%)**	**0.00654 (7.0×)**	**15.689 (+1.3%)**	**0.01385 (5.0×)**
WS	3	10.877 (+0.4%)	0.00964	15.572 (+0.6%)	0.01319
SW (ours)	3	**10.848 (+0.1%)**	**0.00132 (7.3×)**	**15.523 (+0.2%)**	**0.00295 (4.5×)**
WS	4	10.854 (+0.2%)	0.00230	15.511 (+0.2%)	0.00320
SW (ours)	4	**10.836 (+0.0%)**	**0.00031 (7.3×)**	**15.501 (+0.1%)**	**0.00073 (4.4×)**
Qwen3-32B (151,936)					
fp16	16	8.452	0	12.269	0
WS	2	8.756 (+3.6%)	0.05835	13.029 (+6.2%)	0.08106
SW (ours)	2	**8.498 (+0.5%)**	**0.00703 (8.3×)**	**12.479 (+1.7%)**	**0.01596 (5.1×)**
WS	3	8.500 (+0.6%)	0.01246	12.439 (+1.4%)	0.01683
SW (ours)	3	**8.464 (+0.1%)**	**0.00147 (8.5×)**	**12.315 (+0.4%)**	**0.00341 (4.9×)**
WS	4	8.477 (+0.3%)	0.00295	12.302 (+0.3%)	0.00397
SW (ours)	4	**8.457 (+0.1%)**	**0.00035 (8.4×)**	**12.275 (+0.1%)**	**0.00081 (4.9×)**

A.2 QUANTIZED BODY

Table 6: Perplexity (PPL) on WikiText2 (WT2) and C4, and KL divergence from the baseline model, consisting of the body quantized at 2, 3 and 4 bits with GuidedQuant (BlockLDLQ, QTIP format, released checkpoints) and the head kept in fp16. Body and head columns give the respective rates. The quantized head matrix is tied (also the input embedding). WS: WaterSIC w/o FT. Parentheses give the increase over the fp16-head row of the same body. Whole-model size in MiB and reduction (Δsize) after quantizing the head. Best value per head group in bold.

Llama-3.2-1B-Instruct

Body	Method	Head	Size (MiB)	Δsize	WT2 PPL (↓)	WT2 KL (↓)	C4 PPL (↓)	C4 KL (↓)
2	fp16	16.00	737.3	0.0%	26.81	0.497	38.54	0.511
	WS	2.00	297.0	−59.7%	31.53 (+17.6%)	0.671 (+34.9%)	47.14 (+22.3%)	0.732 (+43.3%)
	SW (ours)	2.00	297.0	−59.7%	**27.80 (+3.7%)**	**0.536 (+7.9%)**	**42.39 (+10.0%)**	**0.607 (+18.8%)**
	WS	3.00	327.7	−55.6%	27.60 (+3.0%)	0.533 (+7.2%)	40.75 (+5.7%)	0.570 (+11.7%)
	SW (ours)	3.00	327.7	−55.6%	**27.03 (+0.8%)**	**0.506 (+1.7%)**	**39.49 (+2.5%)**	**0.536 (+4.9%)**
	WS	4.00	358.4	−51.4%	27.16 (+1.3%)	0.510 (+2.6%)	39.03 (+1.3%)	0.524 (+2.6%)
	SW (ours)	4.00	358.4	−51.4%	**26.89 (+0.3%)**	**0.501 (+0.7%)**	**38.99 (+1.2%)**	**0.521 (+2.1%)**
3	fp16	16.00	850.0	0.0%	17.85	0.101	26.15	0.096
	WS	2.00	409.6	−51.8%	19.99 (+12.0%)	0.230 (+126.8%)	30.84 (+17.9%)	0.269 (+180.2%)
	SW (ours)	2.00	409.6	−51.8%	**18.44 (+3.3%)**	**0.136 (+34.5%)**	**28.52 (+9.0%)**	**0.180 (+87.5%)**
	WS	3.00	440.3	−48.2%	18.23 (+2.1%)	0.128 (+26.0%)	26.96 (+3.1%)	0.135 (+40.4%)
	SW (ours)	3.00	440.3	−48.2%	**17.98 (+0.7%)**	**0.109 (+7.1%)**	**26.77 (+2.4%)**	**0.119 (+23.8%)**
	WS	4.00	471.0	−44.6%	18.08 (+1.3%)	0.113 (+11.5%)	26.45 (+1.1%)	0.107 (+11.3%)
	SW (ours)	4.00	471.0	−44.6%	**17.88 (+0.2%)**	**0.104 (+2.3%)**	**26.32 (+0.6%)**	**0.102 (+6.7%)**
4	fp16	16.00	962.6	0.0%	16.53	0.028	24.54	0.028
	WS	2.00	532.5	−44.7%	18.14 (+9.7%)	0.138 (+396.2%)	28.11 (+14.5%)	0.178 (+531.5%)
	SW (ours)	2.00	532.5	−44.7%	**17.01 (+2.9%)**	**0.056 (+101.9%)**	**26.60 (+8.4%)**	**0.106 (+276.9%)**
	WS	3.00	563.2	−41.5%	16.83 (+1.8%)	0.052 (+88.9%)	25.26 (+2.9%)	0.063 (+122.3%)
	SW (ours)	3.00	563.2	−41.5%	**16.64 (+0.7%)**	**0.034 (+24.1%)**	**25.12 (+2.3%)**	**0.052 (+84.6%)**
	WS	4.00	594.0	−38.3%	16.65 (+0.7%)	0.036 (+28.0%)	**24.70 (+0.6%)**	0.037 (+32.4%)
	SW (ours)	4.00	594.0	−38.3%	**16.58 (+0.3%)**	**0.030 (+9.3%)**	24.77 (+0.9%)	**0.035 (+25.3%)**

Table 7: Perplexity (PPL) on WikiText2 (WT2) and C4, and KL divergence from baseline: fp16-head, body quantized at 2 and 3 and 4 bits with VPTQ. Body and head columns give respective rates. WS: WaterSIC w/o FT. Parentheses give increase over the fp16-head row of the same body. Model size and reduction (Δsize) after quantizing the untied head (input embedding stays fp16). Best value per head group in bold.

Qwen2.5-7B-Instruct

Body	Method	Head	Size (MiB)	Δsize	WT2 PPL (↓)	WT2 KL (↓)	C4 PPL (↓)	C4 KL (↓)
2	fp16	16	3835.7	0.0%	10.67	0.397	15.94	0.316
	WS	2	2926.1	−23.7%	11.03 (+3.4%)	0.442 (+11.3%)	16.84 (+5.6%)	0.381 (+20.5%)
	SW (ours)	2	2926.1	−23.7%	**10.75 (+0.8%)**	**0.404 (+1.7%)**	**16.18 (+1.5%)**	**0.331 (+4.7%)**
	WS	3	2991.1	−22.0%	10.73 (+0.6%)	0.407 (+2.3%)	16.15 (+1.3%)	0.331 (+4.6%)
	SW (ours)	3	2991.1	−22.0%	**10.68 (+0.1%)**	**0.399 (+0.3%)**	**16.00 (+0.4%)**	**0.320 (+1.1%)**
	WS	4	3056.1	−20.3%	10.68 (+0.2%)	0.399 (+0.5%)	16.00 (+0.4%)	0.319 (+1.1%)
	SW (ours)	4	3056.1	−20.3%	**10.67 (+0.0%)**	**0.398 (+0.1%)**	**15.95 (+0.1%)**	**0.317 (+0.3%)**
3	fp16	16	4614.3	0.0%	8.61	0.133	13.17	0.084
	WS	2	3704.8	−19.7%	8.94 (+3.8%)	0.191 (+43.5%)	14.02 (+6.4%)	0.161 (+92.2%)
	SW (ours)	2	3704.8	−19.7%	**8.68 (+0.9%)**	**0.140 (+5.4%)**	**13.44 (+2.0%)**	**0.103 (+22.7%)**
	WS	3	3769.8	−18.3%	8.69 (+0.9%)	0.145 (+9.2%)	13.37 (+1.5%)	0.102 (+21.3%)
	SW (ours)	3	3769.8	−18.3%	**8.62 (+0.1%)**	**0.135 (+1.3%)**	**13.23 (+0.4%)**	**0.088 (+5.2%)**
	WS	4	3834.7	−16.9%	8.63 (+0.2%)	0.136 (+2.1%)	13.23 (+0.4%)	0.088 (+5.0%)
	SW (ours)	4	3834.7	−16.9%	**8.61 (+0.0%)**	**0.133 (+0.3%)**	**13.18 (+0.1%)**	**0.085 (+1.1%)**
4	fp16	16	5587.5	0.0%	8.09	0.035	12.64	0.020
	WS	2	4677.9	−16.3%	8.46 (+4.5%)	0.099 (+185.8%)	13.54 (+7.1%)	0.104 (+428.0%)
	SW (ours)	2	4677.9	−16.3%	**8.17 (+0.9%)**	**0.043 (+23.7%)**	**12.89 (+2.0%)**	**0.039 (+98.6%)**
	WS	3	4742.9	−15.1%	8.17 (+1.0%)	0.048 (+38.8%)	12.84 (+1.6%)	0.039 (+95.9%)
	SW (ours)	3	4742.9	−15.1%	**8.11 (+0.1%)**	**0.037 (+5.3%)**	**12.70 (+0.5%)**	**0.024 (+20.0%)**
	WS	4	4807.8	−14.0%	8.11 (+0.2%)	0.038 (+9.8%)	12.69 (+0.4%)	0.024 (+22.9%)
	SW (ours)	4	4807.8	−14.0%	**8.10 (+0.1%)**	**0.035 (+1.0%)**	**12.66 (+0.1%)**	**0.021 (+4.8%)**

A.3 DOMAIN-TARGETED CALIBRATION

Table 8: Domains used for calibration and evaluation. Calibration and evaluation slices are disjoint: WT2 by split, the remaining domains by offsetting the calibration stream past the evaluation slice (skip 32). Every calibration slice is 1024×1024 tokens and every evaluation slice 32×1024 tokens.

Domain	Corpus	Config	Split	Field
WT2	`wikitext`	`wikitext-2-raw-v1`	`train / test`	`text`
DeWiki	`wikimedia/wikipedia`	`20231101.de`	`train`	`text`
Code	`codeparrot/codeparrot-clean-valid`	—	`train`	`content`
Math	`open-web-math/open-web-math`	—	`train`	`text`
Law	`HFforLegal/case-law`	—	`us`	`document`

Table 9: In-domain next-token top-1 accuracy of Llama-3.2-1B-Instruct, fp16 body with a tied head quantized at 2, 3 and 4 bits. Each arm is calibrated on the domain it is evaluated on. KL: divergence from the all-fp16 model on the matching stream. WS: WaterSIC w/o FT. Best value per rate in bold.

		Top-1 accuracy (%, ↑)					KL (↓)				
Head	**Method**	**WT2**	**DeWiki**	**Code**	**Math**	**Law**	**WT2**	**DeWiki**	**Code**	**Math**	**Law**
16	fp16	45.6	52.9	66.9	54.6	44.5	0.000	0.000	0.000	0.000	0.000
2	WS	44.7	50.0	65.6	53.5	42.7	0.096	0.183	0.099	0.090	0.163
	SW (ours)	**45.4**	**52.5**	**66.3**	**54.3**	**44.1**	**0.021**	**0.028**	**0.052**	**0.028**	**0.025**
3	WS	45.4	52.3	66.7	54.3	44.1	0.023	0.040	0.024	0.023	0.034
	SW (ours)	**45.5**	**52.6**	**66.8**	**54.5**	**44.4**	**0.005**	**0.011**	**0.014**	**0.009**	**0.007**
4	WS	**45.6**	52.6	66.8	**54.5**	44.3	0.006	0.012	0.007	0.007	0.009
	SW (ours)	**45.6**	**52.9**	**66.9**	54.4	**44.5**	**0.002**	**0.003**	**0.004**	**0.002**	**0.002**

Table 10: Calibration $\times$ evaluation domain crossover on Llama-3.2-1B-Instruct, fp16 body with a tied head quantized at 2-4 bits. Each block gives the calibration domain used for both $\tilde{\lambda}$ and Σ_X; diagonal cells (calibration domain matches evaluation domain) are the targeting claim. Accuracy columns: next-token top-1 accuracy in percent, higher is better. KL columns: divergence from the all-fp16 model on the matching stream, lower is better. WS: WaterSIC w/o FT. The fp16-head row is repeated per rate as the reference. Best value per calibration block in bold; ties bold both.

		Accuracy (%, ↑)					**KL (↓)**				
Calib	**Method**	**WT2**	**DeWiki**	**Code**	**Math**	**Law**	**WT2**	**DeWiki**	**Code**	**Math**	**Law**
Head @ 2 bits											
—	fp16	45.6	52.9	66.9	54.6	44.5	0.000	0.000	0.000	0.000	0.000
WT2	WS	44.7	50.0	**64.8**	52.8	42.9	0.096	0.212	**0.151**	0.123	0.157
	SW (ours)	**45.4**	**50.5**	64.2	**53.8**	**43.6**	**0.021**	**0.200**	0.218	**0.080**	**0.086**
DeWiki	WS	43.3	50.0	**64.5**	52.1	41.6	0.172	0.183	**0.174**	0.163	0.252
	SW (ours)	**44.2**	**52.5**	64.1	**53.0**	**43.0**	**0.117**	**0.028**	0.219	**0.110**	**0.141**
Code	WS	43.4	**49.5**	65.6	53.1	42.0	0.173	**0.245**	0.099	0.121	0.210
	SW (ours)	**43.9**	**49.5**	**66.3**	**53.9**	**43.1**	**0.135**	0.272	**0.052**	**0.056**	**0.138**
Math	WS	44.1	50.2	65.4	53.5	42.7	0.135	0.194	**0.118**	0.090	0.165
	SW (ours)	**44.8**	**51.4**	**65.6**	**54.3**	**43.7**	**0.062**	**0.139**	0.119	**0.028**	**0.083**
Law	WS	43.5	49.4	**64.8**	52.7	42.7	0.144	0.262	**0.157**	0.135	0.163
	SW (ours)	**44.9**	**50.2**	64.5	**54.0**	**44.1**	**0.060**	**0.215**	0.182	**0.056**	**0.025**
Head @ 3 bits											
—	fp16	45.6	52.9	66.9	54.6	44.5	0.000	0.000	0.000	0.000	0.000
WT2	WS	45.4	**52.2**	**66.5**	54.1	43.9	0.023	0.053	**0.034**	0.028	0.040
	SW (ours)	**45.5**	**52.2**	66.4	**54.4**	**44.2**	**0.005**	**0.052**	0.049	**0.015**	**0.022**
DeWiki	WS	45.1	52.3	**66.4**	54.1	**44.0**	0.037	0.040	**0.037**	0.034	0.055
	SW (ours)	**45.2**	**52.6**	66.3	**54.3**	43.9	**0.026**	**0.011**	0.045	**0.024**	**0.033**
Code	WS	45.0	**52.2**	66.7	54.2	43.9	0.039	**0.059**	0.024	0.030	0.050
	SW (ours)	**45.2**	52.1	**66.8**	**54.4**	**44.2**	**0.027**	0.062	**0.014**	**0.013**	**0.030**
Math	WS	**45.4**	**52.2**	66.5	54.3	44.0	0.030	0.047	**0.027**	0.023	0.043
	SW (ours)	**45.4**	**52.2**	**66.6**	**54.5**	**44.2**	**0.016**	**0.041**	0.030	**0.009**	**0.021**
Law	WS	45.1	52.0	**66.4**	54.0	44.1	0.034	0.057	**0.035**	0.031	0.034
	SW (ours)	**45.5**	**52.2**	**66.4**	**54.5**	**44.4**	**0.016**	**0.056**	0.042	**0.013**	**0.007**
Head @ 4 bits											
—	fp16	45.6	52.9	66.9	54.6	44.5	0.000	0.000	0.000	0.000	0.000
WT2	WS	**45.6**	52.6	**66.8**	**54.5**	44.1	0.006	**0.015**	**0.011**	0.008	0.011
	SW (ours)	**45.6**	**52.7**	**66.8**	**54.5**	**44.4**	**0.002**	0.016	0.014	**0.005**	**0.007**
DeWiki	WS	**45.5**	52.6	**66.9**	**54.5**	44.2	0.010	0.012	**0.011**	0.010	0.014
	SW (ours)	**45.5**	**52.9**	66.7	**54.5**	**44.3**	**0.009**	**0.003**	0.015	**0.008**	**0.011**
Code	WS	45.3	52.5	66.8	**54.5**	44.2	0.011	**0.015**	0.007	0.008	0.014
	SW (ours)	**45.5**	**52.6**	**66.9**	**54.5**	**44.4**	**0.008**	0.021	**0.004**	**0.004**	**0.009**
Math	WS	45.4	**52.7**	**66.9**	**54.5**	44.3	0.008	**0.013**	**0.008**	0.007	0.013
	SW (ours)	**45.6**	**52.7**	66.8	54.4	**44.4**	**0.006**	**0.013**	0.009	**0.002**	**0.007**
Law	WS	**45.5**	52.5	**66.8**	54.4	44.3	0.009	**0.015**	**0.009**	0.008	0.009
	SW (ours)	**45.5**	**52.7**	**66.8**	**54.5**	**44.5**	**0.005**	**0.015**	0.011	**0.004**	**0.002**

A.4 ZERO-SHOT TASKS

Calibration diversity. Unlike Σ_X, the extra per-class statistic is a token frequency, so it benefits from many distinct contexts rather than many tokens. Holding the budget at 131k tokens and re-chunking into 128-token contexts instead of 1024 lifts the 2-bit-head mean by 0.35 to 0.86 points, while moving WaterSIC by at most 0.23; quadrupling the budget at context length 1024 helps by a similar amount and buys nothing that re-chunking does not. Context count, not token count, is the binding resource for $\tilde{\lambda}$.

Table 11: Zero-shot accuracy of Llama-3.2-1B-Instruct with a **2-bit GuidedQuant body** and a tied head quantized at 2, 3 and 4 bits. WS: WaterSIC w/o FT; SW: SoftWater. CTX: context length of 131k WT2-train tokens. lm-evaluation-harness 0.4.12, 0-shot. Best value per body block in bold.

Head	Method	CTX	Tokens	ARC-e	ARC-c	HellaS	LAMB	Wino	PIQA	OBQA	Mean
16	fp16 body	—	—	63.8	37.8	61.7	60.1	61.6	74.9	37.2	56.72
16	fp16 head	—	—	58.0	32.4	49.3	46.3	55.5	69.2	31.4	48.86
2	WS	128	131k	53.6	31.1	47.2	37.3	55.7	67.2	30.6	46.11
	WS	256	131k	52.5	29.9	47.0	36.6	56.2	67.4	30.8	45.78
	WS	1024	131k	53.2	31.5	47.4	36.8	55.3	67.7	29.4	45.89
	WS	1024	1.05M	54.0	30.7	47.2	38.4	55.6	67.1	28.8	45.97
	SW (ours)	128	131k	**56.5**	**31.7**	**48.7**	**45.4**	**56.8**	67.8	**32.6**	**48.49**
	SW (ours)	256	131k	53.4	29.7	48.1	44.9	56.4	66.9	31.6	47.30
	SW (ours)	1024	131k	54.7	31.1	47.9	43.9	55.5	**68.5**	31.8	47.63
	SW (ours)	1024	1.05M	55.7	**31.7**	48.3	44.8	56.7	68.0	30.0	47.87
3	WS	128	131k	55.1	31.3	49.1	43.7	55.8	68.8	31.0	47.83
	WS	256	131k	55.2	31.3	48.9	43.8	55.2	**69.2**	30.4	47.71
	WS	1024	131k	56.1	31.1	48.8	43.1	56.9	67.9	30.2	47.73
	WS	1024	1.05M	56.3	31.6	49.1	43.6	56.5	68.0	30.0	47.87
	SW (ours)	128	131k	57.4	32.1	**49.4**	45.9	56.7	68.7	**31.6**	48.82
	SW (ours)	256	131k	**58.0**	32.2	49.3	**46.7**	**57.5**	68.5	31.2	**49.04**
	SW (ours)	1024	131k	57.0	**32.5**	49.0	45.6	56.6	68.8	**31.6**	48.72
	SW (ours)	1024	1.05M	57.3	32.3	49.2	46.2	57.2	68.7	**31.6**	48.92
4	WS	128	131k	57.6	32.3	49.3	45.2	56.0	69.4	31.8	48.79
	WS	256	131k	57.7	32.3	49.4	45.2	55.6	**69.6**	32.0	48.82
	WS	1024	131k	57.4	31.8	**49.5**	45.5	55.7	69.2	32.0	48.74
	WS	1024	1.05M	**58.1**	32.1	49.4	45.6	55.9	68.9	32.0	48.85
	SW (ours)	128	131k	57.4	**32.6**	49.2	46.0	**56.4**	68.8	31.6	**48.86**
	SW (ours)	256	131k	56.8	32.3	49.3	46.1	55.8	69.0	31.6	48.70
	SW (ours)	1024	131k	57.3	31.7	49.2	**46.3**	56.2	69.3	31.8	48.83
	SW (ours)	1024	1.05M	57.0	31.6	49.2	46.1	56.0	69.3	**32.2**	48.77

Table 12: Zero-shot accuracy of Llama-3.2-1B-Instruct with a **3-bit GuidedQuant body** and a tied head quantized at 2, 3 and 4 bits. Details as in Table 11. Best value per head group in bold.

Head	Method	ARC-e	ARC-c	HellaS	LAMB	Wino	PIQA	OBQA	Mean
16	fp16 body	63.8	37.8	61.7	60.1	61.6	74.9	37.2	56.72
16	fp16 head	63.0	36.3	58.3	59.5	61.7	73.1	35.6	55.35
2	WS-c128	59.7	34.5	57.0	50.7	60.5	72.2	35.4	52.84
	WS-c256	59.6	34.6	56.8	51.6	59.5	71.4	**37.0**	52.94
	WS-c1024	57.9	32.8	57.2	53.1	60.5	71.8	35.6	52.68
	WS-1M	59.4	34.0	56.8	52.3	59.7	71.7	36.6	52.93
	SW-c128	**62.8**	36.6	57.2	58.5	60.2	**73.0**	35.2	54.78
	SW-c256	62.5	36.6	57.4	58.5	60.2	72.3	36.4	54.85
	SW-c1024	61.0	35.4	**57.5**	58.4	60.2	72.3	36.2	54.43
	SW-1M	61.3	**36.7**	57.3	**58.7**	**61.1**	72.6	36.4	**54.87**
3	WS-c128	63.0	36.4	57.7	58.3	**62.4**	**73.4**	36.0	**55.31**
	WS-c256	**63.1**	**37.0**	57.9	57.1	61.3	73.3	35.8	55.08
	WS-c1024	61.8	36.8	58.0	57.9	61.1	73.0	35.4	54.84
	WS-1M	62.1	36.5	58.0	58.5	61.1	72.9	35.6	54.95
	SW-c128	62.1	35.2	**58.3**	**59.3**	61.4	72.7	**37.2**	55.18
	SW-c256	62.5	35.8	**58.3**	**59.3**	60.9	72.7	36.8	55.20
	SW-c1024	62.1	36.2	58.0	59.0	61.2	73.1	35.8	55.07
	SW-1M	62.2	36.2	58.1	**59.3**	60.6	73.1	35.8	55.05
4	WS-c128	62.7	36.5	**58.3**	58.8	60.9	73.4	**37.0**	55.38
	WS-c256	62.8	36.3	58.2	59.1	**61.8**	73.4	36.4	55.43
	WS-c1024	**63.1**	36.1	58.2	59.0	61.6	73.4	35.8	55.31
	WS-1M	63.0	36.6	58.1	58.8	60.9	72.9	35.8	55.16
	SW-c128	**63.1**	36.0	58.2	59.6	61.1	72.7	35.6	55.17
	SW-c256	62.7	36.6	**58.3**	59.3	60.5	72.9	35.6	55.14
	SW-c1024	**63.1**	36.0	58.2	**59.7**	60.9	**73.5**	36.0	55.34
	SW-1M	62.8	**36.7**	58.2	59.6	61.1	73.2	36.6	**55.45**

Table 13: Zero-shot accuracy of Llama-3.2-1B-Instruct with a **4-bit GuidedQuant body** and a tied head quantized at 2, 3 and 4 bits. Details as in Table 11. Best value per head group in bold.

Head	Method	ARC-e	ARC-c	HellaS	LAMB	Wino	PIQA	OBQA	Mean
16	fp16 body	63.8	37.8	61.7	60.1	61.6	74.9	37.2	56.72
16	fp16 head	62.7	37.7	60.3	60.1	62.0	73.6	37.8	56.30
2	WS-c128	60.5	36.6	58.9	53.6	61.5	72.2	35.8	54.16
	WS-c256	61.0	37.0	59.0	53.9	60.7	72.1	36.2	54.27
	WS-c1024	60.3	37.0	59.2	54.0	60.9	72.1	**37.2**	54.39
	WS-1M	59.5	35.5	59.5	55.0	60.1	72.5	36.8	54.12
	SW-c128	62.2	37.4	59.6	59.4	61.8	73.2	36.4	55.72
	SW-c256	**63.2**	**38.7**	59.5	59.2	61.3	73.3	35.4	55.81
	SW-c1024	61.2	37.5	59.3	59.2	61.6	72.7	35.0	55.23
	SW-1M	62.5	37.8	**59.7**	**59.9**	**62.0**	**73.6**	36.2	**55.96**
3	WS-c128	62.2	38.5	**60.4**	58.7	60.9	**73.8**	37.4	55.98
	WS-c256	62.7	**38.6**	60.2	58.6	60.3	73.4	36.4	55.74
	WS-c1024	62.5	38.2	60.2	58.2	60.5	73.1	36.4	55.60
	WS-1M	62.7	37.9	60.3	58.4	60.5	**73.8**	37.0	55.79
	SW-c128	62.2	37.9	60.3	59.5	**62.0**	73.5	36.2	55.94
	SW-c256	62.5	36.9	60.2	60.0	61.8	73.4	36.6	55.93
	SW-c1024	**62.8**	37.5	60.2	**60.2**	61.1	73.3	**37.6**	56.10
	SW-1M	**62.8**	38.1	60.1	60.1	**62.0**	73.6	36.6	**56.18**
4	WS-c128	**63.2**	37.0	60.3	59.5	62.0	73.7	37.8	56.23
	WS-c256	62.6	37.5	**60.4**	59.4	62.0	73.6	38.0	56.22
	WS-c1024	63.1	37.8	60.3	59.7	62.0	**74.2**	37.8	**56.41**
	WS-1M	**63.2**	37.7	60.3	59.5	61.5	73.6	**38.4**	56.30
	SW-c128	62.5	37.8	60.2	59.8	**62.5**	73.4	37.6	56.27
	SW-c256	62.5	**38.0**	60.2	59.9	62.4	74.0	37.4	56.35
	SW-c1024	62.9	37.5	60.3	**60.3**	62.2	73.6	36.8	56.22
	SW-1M	62.4	37.8	60.3	60.2	61.9	73.7	37.8	56.29

A.5 CLASS-SIDE STATISTIC ABLATION

Table 14: Class-side statistic ablation, head-only isolation: the body is kept in fp16 and only the head matrix is quantized; tied models (Llama-3.2-1B, Qwen3-1.7B) retain an fp16 copy of the embedding. All arms differ only in the class-side statistic: $\bar{p} = \mathbb{E}_X[\tilde{p}_k]$; π is the empirical unigram frequency of token k over the calibration window; SW uses $\tilde{\lambda}_k = \mathbb{E}_X[\tilde{p}_k(1-\tilde{p}_k)]$. Every arm lands within 0.001 bits of its nominal head rate. Parentheses after each model name give its vocabulary size V; on PPL cells, the increase over the fp16 row of the same model. Best value per column per head group in bold.

Statistic	Head	WT2 PPL (↓)	WT2 KL (↓)	C4 PPL (↓)	C4 KL (↓)
Llama-3.2-1B (128,256)					
fp16	16	11.313	0	15.614	0
$\bar{p}$	2	**11.418 (+0.9%)**	**0.00952**	16.102 (+3.1%)	0.02959
π	2	11.524 (+1.9%)	0.02146	16.549 (+6.0%)	0.05830
SW (ours)	2	11.419 (+0.9%)	0.00957	**16.035 (+2.7%)**	**0.02834**
$\bar{p}$	3	11.344 (+0.3%)	0.00213	15.719 (+0.7%)	0.00613
π	3	11.356 (+0.4%)	0.00437	15.784 (+1.1%)	0.01184
SW (ours)	3	**11.342 (+0.3%)**	**0.00206**	**15.704 (+0.6%)**	**0.00572**
$\bar{p}$	4	**11.319 (+0.1%)**	0.00049	**15.624 (+0.1%)**	0.00147
π	4	11.327 (+0.1%)	0.00105	15.658 (+0.3%)	0.00280
SW (ours)	4	11.320 (+0.1%)	**0.00047**	15.640 (+0.2%)	**0.00138**
Qwen3-1.7B (151,936)					
fp16	16	18.285	0	22.621	0
$\bar{p}$	2	**18.452 (+0.9%)**	0.00951	23.138 (+2.3%)	0.02110
π	2	18.595 (+1.7%)	0.01912	23.494 (+3.9%)	0.04421
SW (ours)	2	18.460 (+1.0%)	**0.00900**	**23.075 (+2.0%)**	**0.01998**
$\bar{p}$	3	**18.310 (+0.1%)**	0.00206	22.756 (+0.6%)	0.00468
π	3	18.360 (+0.4%)	0.00389	22.823 (+0.9%)	0.00870
SW (ours)	3	18.317 (+0.2%)	**0.00199**	**22.724 (+0.5%)**	**0.00443**
$\bar{p}$	4	18.303 (+0.1%)	0.00046	22.636 (+0.1%)	0.00112
π	4	18.304 (+0.1%)	0.00091	22.669 (+0.2%)	0.00203
SW (ours)	4	**18.298 (+0.1%)**	**0.00045**	**22.633 (+0.1%)**	**0.00105**
Llama-3.1-8B-Instruct (128,256)					
fp16	16	8.388	0	14.189	0
$\bar{p}$	2	8.441 (+0.6%)	0.00658	**14.530 (+2.4%)**	0.02237
π	2	8.501 (+1.3%)	0.01356	14.822 (+4.5%)	0.04139
SW (ours)	2	**8.437 (+0.6%)**	**0.00636**	14.534 (+2.4%)	**0.02140**
$\bar{p}$	3	**8.398 (+0.1%)**	0.00140	**14.252 (+0.4%)**	0.00465
π	3	8.412 (+0.3%)	0.00274	14.293 (+0.7%)	0.00852
SW (ours)	3	**8.398 (+0.1%)**	**0.00135**	14.258 (+0.5%)	**0.00437**
$\bar{p}$	4	**8.390 (+0.0%)**	0.00033	14.207 (+0.1%)	0.00113
π	4	8.396 (+0.1%)	0.00066	14.225 (+0.3%)	0.00200
SW (ours)	4	8.391 (+0.0%)	**0.00032**	**14.205 (+0.1%)**	**0.00103**
Qwen3-8B (151,936)					
fp16	16	10.832	0	15.486	0
$\bar{p}$	2	10.915 (+0.8%)	**0.00643**	15.699 (+1.4%)	0.01484
π	2	10.965 (+1.2%)	0.01183	15.899 (+2.7%)	0.02919
SW (ours)	2	**10.898 (+0.6%)**	0.00654	**15.689 (+1.3%)**	**0.01385**
$\bar{p}$	3	**10.845 (+0.1%)**	**0.00129**	15.544 (+0.4%)	0.00328
π	3	10.853 (+0.2%)	0.00251	15.577 (+0.6%)	0.00585
SW (ours)	3	10.848 (+0.1%)	0.00132	**15.523 (+0.2%)**	**0.00295**
$\bar{p}$	4	**10.835 (+0.0%)**	0.00032	**15.495 (+0.1%)**	0.00075
π	4	10.839 (+0.1%)	0.00059	15.504 (+0.1%)	0.00132
SW (ours)	4	10.836 (+0.0%)	**0.00031**	15.501 (+0.1%)	**0.00073**
Qwen3-32B (151,936)					
fp16	16	8.452	0	12.269	0
$\bar{p}$	2	8.516 (+0.8%)	0.00709	**12.475 (+1.7%)**	0.01663
π	2	8.571 (+1.4%)	0.01461	12.703 (+3.5%)	0.03348
SW (ours)	2	**8.498 (+0.5%)**	**0.00703**	12.479 (+1.7%)	**0.01596**
$\bar{p}$	3	8.468 (+0.2%)	0.00153	**12.303 (+0.3%)**	0.00347
π	3	8.476 (+0.3%)	0.00297	12.356 (+0.7%)	0.00687
SW (ours)	3	**8.464 (+0.1%)**	**0.00147**	12.315 (+0.4%)	**0.00341**
$\bar{p}$	4	**8.457 (+0.1%)**	0.00036	12.276 (+0.1%)	0.00087
π	4	8.461 (+0.1%)	0.00072	12.289 (+0.2%)	0.00167
SW (ours)	4	**8.457 (+0.1%)**	**0.00035**	**12.275 (+0.0%)**	**0.00081**

A.6 SEPARABILITY DIAGNOSTICS

Let $Q_k = \frac{1}{2}\Delta_k \mathbb{E}_X[\lambda_k(X)XX^\top]\Delta_k^\top$ and $S_k = \frac{1}{2}\tilde{\lambda}_k \Delta_k \Sigma_X \Delta_k^\top$ be the contributions of class k to the true and proxy distortion, with Δ the perturbation SoftWater produces at each rate. The relative proxy error is

$$\rho = \frac{\sum_{k=1}^{K} Q_k}{\sum_{k=1}^{K} S_k} - 1, \tag{19}$$

and $\frac{1}{2}\log_2(1+\rho)$ is the extra rate needed to close the gap. Since S_k uses the smoothed $\tilde{\lambda}_k$ with $\epsilon = 0.1$ rather than the raw $\bar{\lambda}_k$, ρ combines the separability error of Assumption 1 with the bias smoothing introduces on purpose. Table 15a reports ρ for Llama-3.2-1B calibrated on WT2, with both sums evaluated on WT2 and C4 at 2, 3 and 4 bits; Table 15b reports the power-iteration fixed point of Appendix B from the three initializations of §5.6.

Table 15: Separability diagnostics on Llama-3.2-1B.

(a) ρ: relative proxy error (19). bits $= \frac{1}{2}\log_2(1+\rho)$. SW calibrated with WT2; Eval is the sample the ratio is measured on.

Eval	WT2			C4		
Rate	2	3	4	2	3	4
ρ	+0.10	+0.09	+0.09	+0.08	+0.08	+0.07
Bits	0.07	0.06	0.06	0.06	0.06	0.05

(b) Convergence of alternating power iteration to its fixed point $(a^\star, B^\star)$ from three initializations. SW: $a_0 = \tilde{\lambda}$, $B_0 = \Sigma_X$; WS: $a_0 = I$, $B_0 = \Sigma_X$; Id: $a_0 = I$, $B_0 = I$.

Init	**SW (ours)**			WS			Id		
Round	**0**	**1**	**2**	0	1	2	0	1	2
$\cos(a, a^\star)$	**0.99**	**1.00**	**1.00**	0.03	1.00	1.00	0.03	0.99	1.00
$\cos(B, B^\star)$	**0.98**	**1.00**	**1.00**	0.98	1.00	1.00	0.05	1.00	1.00

B HESSIAN SKETCHING

This appendix details the power iteration scheme used in §5 to compute optimized Kronecker factors of the head KL Hessian, adapted from the sketching subroutines of YAQA (Tseng et al., 2026).

Setup. We seek the best diagonal-Kronecker approximation

$$a^\star, B^\star = \underset{a \text{ diagonal},\, B}{\arg\min} \left\| \mathbb{E}_X \left[\tilde{\lambda}(X) \otimes XX^\top \right] - a \otimes B \right\|_F^2, \tag{20}$$

where $\tilde{\lambda}(X) = \mathrm{diag}(\tilde{p} \odot (1 - \tilde{p}))$ is the softmax curvature under the smoothed output distribution $\tilde{p}$ of (18) with $\epsilon = 0.1$. As noted by (Tseng et al., 2026), following (Van Loan, 2000), the best Kronecker approximation is a rank-one approximation of a rearrangement of the target matrix, so alternating least squares (ALS) on the two factors is power iteration on that rearrangement and converges to the leading singular pair, which is unique up to the scale ambiguity $(ca, B/c)$ whenever the leading singular value is simple. Since all our comparisons (cosine similarity to the fixed point, and the quantization lattice itself, through the coupling of c_{SW}) are invariant to this rescaling, we do not normalize the factors during the iteration.

Adapted updates. YAQA's Sketch A estimates the per-token Hessian with a Monte-Carlo label sampled from the model's predictive distribution and a modified backward pass. For the head layer both approximations are unnecessary: the layer's KL Hessian at a hidden state X is available in closed form as $(\mathrm{diag}(\tilde{p}) - \tilde{p}\tilde{p}^\top) \otimes XX^\top$, requiring only a forward pass, and the token-independence bias of Sketch A does not arise because no computation mixes tokens after the head. We therefore replace the sampled label with the exact per-token curvature, which preserves the expectation of their estimator while removing its sampling variance. Moreover, since a is constrained to be diagonal, the off-diagonal coupling $-\tilde{p}\tilde{p}^\top$ is Frobenius-orthogonal to the ansatz $a \otimes B$, and the diagonal of the exact class-side Hessian is precisely $\tilde{\lambda} = \tilde{p} \odot (1 - \tilde{p})$; hence each update below is the exact ALS step for (20), not an approximation of it. Writing $s(X) = \tilde{p} \odot (1 - \tilde{p}) \in \mathbb{R}^K$, one round consists of a class-side pass followed by a feature-side pass over the calibration set, mirroring the alternation order of the official YAQA release:

$$a \leftarrow \frac{\mathbb{E}_X \left[(X^\top B X)\, s(X) \right]}{\|B\|_F^2}, \qquad B \leftarrow \frac{\mathbb{E}_X \left[\langle a, s(X) \rangle\, XX^\top \right]}{\|a\|_2^2}, \tag{21}$$

where each update uses the other factor's most recent value. Each pass costs one matrix-vector product with B (or one rank-one accumulation into B) plus the head forward pass per token, and requires no backpropagation, in contrast to the general-layer setting of (Tseng et al., 2026).

Protocol. We cache the calibration hidden states once (WT2-train) and run all initializations of Table 15b on the identical cached states, so that differences reflect the initialization alone. We run 8 rounds of (21); the fixed point $(a^\star, B^\star)$ is taken as the final iterate, and convergence is monitored by the cosine similarity of consecutive iterates, which exceeds 0.999999 for every initialization from round 3 onward. Note that the SoftWater factors $(\tilde{\lambda}, \Sigma_X)$ are not merely a good initialization: round 0 of Table 15b shows they already agree with the fixed point to cosine similarity 0.976–0.992, quantifying how close the analytical separable factorization is to the Frobenius-optimal one.