ARSO: RECONSTRUCTION-AWARE ROUNDING WITH JOINT SCALE OPTIMIZATION FOR WEIGHT-ONLY LLM QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) compresses large language models (LLMs) into low-precision representations to reduce memory and accelerate inference without retraining. In weight-only PTQ, existing methods either rely on round-to-nearest (RTN), whose local decisions ignore reconstruction error, or employ costly iterative rounding optimization. Moreover, quantization scales are often reused for dequantization, coupling integer-grid construction with weight reconstruction. We identify reconstruction freedom as an important optimization resource. Scale Optimization (SO) exploits scale freedom by deriving reconstruction-optimal dequantization scales for fixed integer codes; eliminating this scale freedom further reveals a directional criterion over integer weights, enabling Reconstruction-Aware Rounding (AR). In sequential quantization, AR exploits future compensation freedom through a conditional reconstruction metric, while J-SO exploits prefix scale freedom by jointly updating processed-group scales. This stage-wise exploitation naturally yields ARSO, a PTQ method with single-pass reconstruction-aware rounding that achieves the lowest perplexity in 32 of 36 settings and the highest average zero-shot accuracy across all six models at 2 bits among the evaluated methods.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong capabilities across a wide range of tasks, but their parameter counts and memory footprints make deployment costly. Weight-only post-training quantization (PTQ) converts model weights to low-bit representations, such as INT4, without retraining. During memory-bound autoregressive decoding, reducing weight traffic can improve serving throughput and make deployment on resource-constrained devices more feasible (Kim et al., 2024; Lin et al., 2024).

A typical weight-only PTQ pipeline first selects a scale to construct an integer grid, maps each full-precision weight to an integer code, and then dequantizes the resulting codes. Although the ultimate objective is to preserve model outputs on calibration activations, the central discrete decision—the assignment of integer codes—is often made according to much simpler local criteria. Round-to-nearest (RTN) assigns each weight to its nearest grid point based only on scalar distance. Optimization-based approaches such as AdaRound (Nagel et al., 2020) and BRECQ (Li et al., 2021) instead relax discrete rounding variables and optimize them against reconstruction objectives, producing more informed integer assignments at the cost of iterative optimization.

Recent methods improve quantization accuracy through several complementary mechanisms (Frantar et al., 2023; Lin et al., 2024; Shao et al., 2024; Ashkboos et al., 2024; Liu et al., 2025), yet standard group-wise linear quantization still leaves part of the reconstruction flexibility underused. Consider a dequantized group parameterized as

$$q = s\, w_{\mathrm{int}}, \tag{1}$$

where w_{int} denotes the integer codes and s the dequantization scale. Standard affine quantization typically reuses the grid-defining scale for dequantization, thereby coupling integer-grid selection with reconstruction magnitude. These two roles, however, serve different objectives: the former

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.

determines the integer assignments, whereas the latter determines how the fixed integer codes are reconstructed in real-valued space. Consequently, a scale that induces favorable integer assignments need not be the reconstruction-optimal scale for those assignments.

We present **ARSO**, which naturally follows from this view by exploiting the reconstruction freedoms that remain readily adjustable throughout the quantization process. We first introduce *Scale Optimization* (SO), which exploits scale freedom by deriving reconstruction-optimal dequantization scales in closed form for fixed integer codes. The quantization scale and the dequantization scale consequently assume distinct roles: the former selects an integer grid, whereas the latter analytically fits the magnitude of that fixed grid to the calibration reconstruction objective.

After substituting the optimal dequantization scale back into the reconstruction objective, the remaining discrete problem is governed by the activation-aware directional alignment between the full-precision weights and the candidate integer weights. This yields *Reconstruction-Aware Rounding* (AR), an iteration-free criterion that directly ranks integer candidates according to the reconstruction objective.

In sequential quantization, additional reconstruction freedom arises from variables that remain readily adjustable as quantization decisions are progressively fixed. Specifically, the unquantized future groups retain continuous compensation freedom and can absorb part of the current reconstruction error. Meanwhile, although the integer codes of processed groups are costly to revise, their dequantization scales remain readily adjustable, providing prefix scale freedom. AR accounts for future compensation freedom by eliminating the optimal continuous adjustment of the unquantized groups, while J-SO exploits the prefix scale freedom by jointly updating the reconstruction scales of processed groups. ARSO thus naturally exploits the reconstruction freedom available at each stage of the quantization process. Figure 1 summarizes how ARSO arises from eliminating the reconstruction freedoms available during quantization.

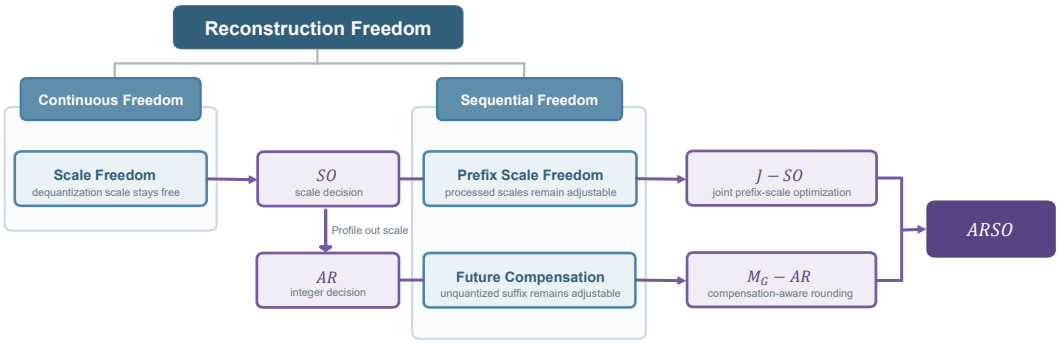


Figure 1: Overview of reconstruction freedom in ARSO.

Our contributions are threefold:

(1) We identify reconstruction freedom as an important optimization resource in weight-only quantization. In particular, dequantization scales remain adjustable after integer codes are fixed, motivating Scale Optimization (SO), which derives their reconstruction-optimal values in closed form.

(2) By analytically eliminating scale freedom, we reveal that quantization naturally decomposes into two functionally distinct decisions: integer weights determine the reconstruction direction, while dequantization scales determine the reconstruction magnitude. This decoupling reduces reconstruction-aware integer selection to a directional problem, yielding AR, a rounding procedure that is simultaneously reconstruction-aware and iteration-free.

(3) We extend this principle to sequential quantization by exploiting two additional freedoms introduced by progressive parameter fixing, leading to ARSO. ARSO improves quantization by optimizing scales and integer codes directly, without introducing additional codebooks, auxiliary parameters, or inference-time transformations.

2 Related Work

2.1 LLM Quantization

Quantization converts high-precision values into low-precision representations. The round-to-nearest (RTN) linear quantization with bit width b can be formulated as

$$Q(X) = s\left(\text{clamp}\left(\text{round}(X/s) + z, 0, 2^b - 1\right) - z\right), \tag{2}$$

where $s = \frac{\max(X)-\min(X)}{2^b-1}$ and $z = -\text{round}(\min(X)/s)$. We focus on weight-only PTQ with group-wise quantization, where each group of g consecutive weights along the input dimension has separate s and z.

Existing affine quantization methods typically use one scale both to construct an integer grid and to dequantize the resulting codes. This parameterization couples two roles with different objectives: grid construction determines the discrete integer representation, whereas dequantization determines the continuous reconstruction magnitude. Recent work has also revisited the optimization of quantization scales and grids (Amboage et al., 2026; Kim et al., 2026). In contrast, SO explicitly separates the grid-construction scale from the dequantization scales and derives the reconstruction-optimal dequantization scales in closed form.

Scaling- and transformation-based methods reshape weights or activations into more quantization-friendly representations, including SmoothQuant (Xiao et al., 2023), AWQ (Lin et al., 2024), QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2025), and FlatQuant (Sun et al., 2025). A complementary direction improves low-bit quantization by increasing the expressiveness of the representation through vector, additive, or structured codebooks (Chee et al., 2023; Tseng et al., 2024a; Egiazarian et al., 2024; Van Baalen et al., 2024; Liu et al., 2024; Tseng et al., 2024b). ARSO instead retains standard scalar group-wise affine quantization and improves its reconstruction solely by selecting better dequantization scales and integer codes.

2.2 Optimization-Based Quantization

Another line of PTQ research improves quantization by iteratively optimizing quantization variables against reconstruction objectives. AdaRound (Nagel et al., 2020) replaces hard rounding decisions with continuous relaxation variables and optimizes them under a layer-wise reconstruction objective, while BRECQ (Li et al., 2021) extends reconstruction-based optimization to blocks of layers. More recently, OmniQuant (Shao et al., 2024) adapts this paradigm to LLMs by jointly optimizing learnable weight clipping and equivalent transformations through block-wise calibration. Recent reconstruction-aware PTQ methods further optimize rounding or quantization variables through iterative procedures (Dadgarnia et al., 2026; Zhang et al., 2026). These methods improve integer assignment but require iterative optimization, whereas AR derives reconstruction-aware rounding analytically without learned rounding variables.

2.3 Compensation-Based Quantization

Second-order PTQ methods exploit curvature information to reduce the effect of discrete weight perturbations. Optimal Brain Compression (OBC) (Frantar et al., 2022) extends OBS-style second-order correction to post-training compression, while GPTQ (Frantar et al., 2023) scales sequential second-order weight quantization to large language models. Let w_p and q_p denote the current full-precision and quantized weight columns, respectively, and let R be an upper-triangular factor satisfying

$$H^{-1} = R^\top R. \tag{3}$$

After fixing q_p, GPTQ propagates its quantization residual to the remaining unquantized columns according to

$$\Delta W_{:,p:} = -\frac{w_p - q_p}{R_{p,p}} R_{p,p:}, \tag{4}$$

which provides an efficient second-order correction under the local quadratic reconstruction objective. Subsequent methods enrich the reconstruction objective or compensation procedure; TurboBoA (Kim et al., 2026), for example, incorporates attention-aware information while retaining an efficient sequential compensation framework.

3 METHOD

Problem formulation. We consider weight-only post-training quantization of a linear layer with weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Given calibration activations $X \in \mathbb{R}^{N \times d_{\text{in}}}$, we minimize the layer-wise reconstruction error

$$\mathcal{L}(W, Q) = \left\| XW^\top - XQ^\top \right\|_F^2 . \tag{5}$$

Let $H = X^\top X$ denote the input-feature Hessian. For one output channel, the reconstruction objective can be written as

$$\mathcal{L}(w, q) = (w - q)H(w - q)^\top. \tag{6}$$

We partition the input dimension into n_g consecutive groups of size g, and write $w = [w_0, \ldots, w_{n_g - 1}]$, where $w_i \in \mathbb{R}^{1 \times g}$; q and w_{int} are partitioned analogously. The corresponding Hessian block between groups i and j is denoted by $H_{i,j} \in \mathbb{R}^{g \times g}$. For asymmetric quantization, we use centered integer codes $w_{\text{int},i} = \bar{w}_{\text{int},i} - z_i$, such that a dequantized group can be written as a scale multiplied by $w_{\text{int},i}$. At step k, let G_k denote the current group, $P_{k-1} = \{G_0, \ldots, G_{k-1}\}$ the processed prefix, and $U_k = \{G_{k+1}, \ldots, G_{n_g - 1}\}$ the unprocessed suffix. After G_k is fixed, the processed prefix becomes $P_k = P_{k-1} \cup G_k$. When the step index is clear, we simply write G, P, and U.

3.1 SCALE OPTIMIZATION

Standard group-wise affine quantization commonly uses the same scale α_i to construct an integer grid and to dequantize the resulting integer codes:

$$\begin{aligned} w_{\text{int},i} &= \text{clamp}\left(\text{round}\left(\frac{w_i}{\alpha_i} \right) + z_i, 0, 2^b - 1 \right) - z_i, \\ q_i &= \alpha_i w_{\text{int},i}. \end{aligned} \tag{7}$$

However, these two uses have different objectives. We therefore retain the integer codes produced by the original quantizer but introduce an independent dequantization scale s_i:

$$q_i = s_i w_{\text{int},i}. \tag{8}$$

The dequantization scales therefore constitute continuous scale freedom conditioned on the chosen integer codes. For one output channel, let $\boldsymbol{s} = [s_0, \ldots, s_{n_g - 1}]^\top$. Thus we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{s}) &= \sum_{i,j} \left(s_i w_{\text{int},i} - w_i \right) H_{i,j} \left(s_j w_{\text{int},j} - w_j \right)^\top \\ &= \boldsymbol{s}^\top \mathbf{H}_s \boldsymbol{s} - 2\boldsymbol{b}^\top \boldsymbol{s} + C, \end{aligned} \tag{9}$$

where $C = wHw^\top$ is independent of $\boldsymbol{s}$, and

$$\begin{aligned} (\mathbf{H}_s)_{i,j} &= w_{\text{int},i} H_{i,j} w_{\text{int},j}^\top, \\ b_i &= \sum_j w_{\text{int},i} H_{i,j} w_j^\top. \end{aligned} \tag{10}$$

The matrix $\mathbf{H}_s$ is positive semidefinite, so Eq. (9) defines a convex quadratic in $\boldsymbol{s}$. Its unconstrained minimizer satisfies the normal equation

$$\mathbf{H}_s \boldsymbol{s}^* = \boldsymbol{b}. \tag{11}$$

In implementation, we add a small diagonal regularizer ϵI to $\mathbf{H}_s$ solely for numerical stability.

We refer to this closed-form procedure as *Scale Optimization* (SO), which gives the globally optimal use of the available scale freedom conditioned on the fixed integer codes. This decouples dequantization from grid construction. Although s_i is unrestricted during optimization, the resulting representation can always be converted into the standard affine representation with positive scales by reversing the integer grid. Specifically, any $s_i < 0$ can be converted exactly by replacing $(s_i, \bar{w}_{\text{int},i}, z_i)$ with $(-s_i,\ 2^b - 1 - \bar{w}_{\text{int},i},\ 2^b - 1 - z_i)$, which leaves the dequantized weights unchanged.

3.2 Reconstruction-Aware Rounding

3.2.1 Integer Direction Is All You Need

Existing rounding methods face a trade-off between efficiency and reconstruction quality: RTN makes independent local decisions without considering reconstruction error, whereas optimization-based methods typically introduce costly iterative optimization.

SO allows the scale freedom of each integer candidate to be analytically eliminated before comparison, leaving only the quality of its integer direction. We first consider a single group, with full-precision weight $w \in \mathbb{R}^{1\times g}$, corresponding calibration activations X_G, and a candidate centered integer-code vector $u \in \mathbb{R}^{1\times g}$. For fixed u, the optimal unrestricted dequantization scale is given by the one-dimensional SO problem

$$c^*(u) = \arg\min_{c\in\mathbb{R}} \left\| X_G w^\top - c X_G u^\top \right\|_2^2 = \frac{\left\langle X_G w^\top, X_G u^\top \right\rangle}{\|X_G u^\top\|_2^2}. \tag{12}$$

Substituting $c^*(u)$ back into the reconstruction objective gives

$$\begin{aligned} \mathcal{L}^*(u) &= \left\| X_G w^\top \right\|_2^2 - \frac{\left[\left\langle X_G w^\top, X_G u^\top \right\rangle\right]^2}{\|X_G u^\top\|_2^2} \\ &= \left\| X_G w^\top \right\|_2^2 \left[1 - \cos^2\left(X_G w^\top, X_G u^\top\right)\right]. \end{aligned} \tag{13}$$

Since $\left\| X_G w^\top \right\|_2^2$ is constant across integer candidates, we obtain

$$\arg\min_u \mathcal{L}^*(u) = \arg\max_u \cos^2\left(X_G w^\top, X_G u^\top\right) \tag{14}$$

$$= \arg\max_u \frac{\left(w H_{GG} u^\top\right)^2}{u H_{GG} u^\top}. \tag{15}$$

Thus, once the scale freedom is eliminated, the remaining discrete choice is purely an integer-direction problem: scale optimization absorbs the candidate magnitude, while reconstruction quality is determined by its activation-aware direction.

3.2.2 Group-wise Rounding

For multiple groups, the entire scale vector can still be eliminated in closed form. However, doing so leaves a globally coupled discrete optimization over the integer directions of all groups. Since solving this combinatorial problem is intractable at LLM scale, we determine the integer codes sequentially.

When quantizing the current group G, the remaining suffix U has not yet been discretized and therefore retains continuous compensation freedom to absorb part of the error introduced by G. Eliminating this freedom induces the conditional metric

$$M_G = H_{GG} - H_{GU} H_{UU}^{-1} H_{UG}. \tag{16}$$

Hence, M_G measures precisely the component of the current-group error that cannot be absorbed by the future compensation freedom. Applying the same scale-elimination argument to the compensated target w_G under the conditional metric M_G yields the compensation-aware integer objective

$$\boxed{\eta(u) = \frac{\left(w_G M_G u^\top\right)^2}{u M_G u^\top}} \tag{17}$$

The derivation of Eq. (16) and its equivalent inverse-Cholesky form are provided in Appendix A.

Based on Eq. (17), AR sequentially determines the rounding direction of each weight. At coordinate t, all other coordinates are held at their current integer values, and the two adjacent integer candidates (round-down and round-up) are evaluated by the corresponding whole-group integer vector directions. The candidate achieving a larger $\eta(\cdot)$ score is selected as the rounding decision. In implementation, a small threshold δ is introduced to avoid numerically insignificant flips.

3.3 JOINTLY UPDATED SEQUENTIAL COMPENSATION

The preceding derivations directly determine the two quantization variables—the integer codes and scales—from the reconstruction objective. When these decisions are made sequentially, the progressive fixing process leaves additional reconstruction freedom: future groups remain available for compensation, while the dequantization scales of processed groups remain adjustable. This naturally extends AR and SO to the sequential setting, yielding the compensation-based method ARSO: AR accounts for future compensation freedom through the conditional metric introduced in Sec. 3.2.2, while joint SO exploits the remaining prefix scale freedom.

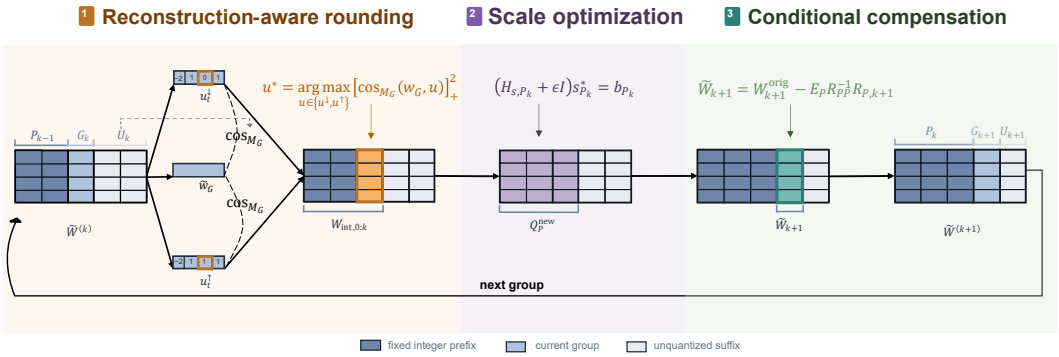


Figure 2: One sequential step of ARSO. M_G-aware AR selects the current integer codes, J-SO jointly re-optimizes the processed-prefix scales, and the updated prefix residual compensates the next group.

Algorithm 1 ARSO

Require: Full-precision weights W^{orig}, Hessian H, groups $\{G_k\}_{k=0}^{n_g-1}$

Initialize the current weights $\widetilde{W} \leftarrow W^{\text{orig}}$
for $k = 0, \ldots, n_g - 1$ **do**
  Set $W_{\text{int},k} \leftarrow \text{RTN}(\widetilde{W}_{G_k})$
  for $t = 1, \ldots, g$ **do**
    Use AR to determine $(W_{\text{int},k})_t$ from the two adjacent candidates at t
  end for
  Freeze the integer codes $W_{\text{int},k}$
  Set $P \leftarrow G_{0:k}$ and solve $\boldsymbol{s}_P^* \leftarrow (\mathbf{H}_{s,P} + \epsilon I)^{-1} \boldsymbol{b}_P$
  Update $Q_P^{\text{new}} \leftarrow \boldsymbol{s}_P^* \odot W_{\text{int},P}$ and $E_P \leftarrow W_P^{\text{orig}} - Q_P^{\text{new}}$
  if $k < n_g - 1$ **then**
    $\widetilde{W}_{k+1} \leftarrow W_{k+1}^{\text{orig}} - E_P R_{PP}^{-1} R_{P,k+1}$
  end if
end for
return W_{int} and the final optimized scales

At step k, AR first determines the integer codes of G_k under the conditional metric derived in Sec. 3.2.2. After these codes are fixed, J-SO jointly optimizes the dequantization scales of the processed prefix $P_k = G_{0:k}$. Here, $\mathbf{H}_{s,P}$ and $\boldsymbol{b}_P$ follow Eq. (10) with all group indices restricted to P. The optimized scales yield the updated prefix reconstruction Q_P^{new}, whose residual is $E_P = W_P^{\text{orig}} - Q_P^{\text{new}}$. The refreshed residual is then used to compensate the next group:

$$W_{k+1}^{\text{opt}} = W_{k+1}^{\text{orig}} - E_P R_{PP}^{-1} R_{P,k+1}. \tag{18}$$

We next justify the compensation target in Eq. (18). Under the global quadratic reconstruction objective, the optimal suffix state conditioned on the fixed prefix residual is

$$W_U^{\text{opt}} = W_U^{\text{orig}} - E_P R_{PP}^{-1} R_{PU}, \tag{19}$$

where R is an upper-triangular factor satisfying $H^{-1} = R^\top R$.

Although Eq. (19) characterizes the conditional optimum of the entire suffix, sequential quantization only requires the immediately next group to attain this optimum, because only this group is quantized before the processed-prefix reconstruction and its residual are refreshed again, which gives Eq. (18).

The derivation of Eq. (19) and its connection to GPTQ compensation are provided in Appendix B.

4 EXPERIMENTS

We evaluate ARSO on LLaMA-2-7B, LLaMA-2-13B (Touvron et al., 2023), LLaMA-3-8B (Grattafiori et al., 2024), Qwen3-4B, Qwen3-8B, and Qwen3-14B (Yang et al., 2025). Unless otherwise specified, we use group-wise asymmetric quantization with a group size of 128 while keeping activations in FP16. For calibration, we use 128 sequences of length 2048, consisting of 43 WikiText-2, 43 C4, and 42 RedPajama (Weber et al., 2024) samples, which are merged and shuffled with a fixed random seed of 0. Calibration and evaluation data are strictly disjoint: no samples from the evaluation splits are used for calibration. We report perplexity (PPL) on WikiText-2 (Merity et al., 2017) and C4 (Raffel et al., 2020), and zero-shot accuracy on BoolQ (Clark et al., 2019), ARC-Challenge (ARC-C), ARC-Easy (ARC-E) (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2020), and MMLU (Hendrycks et al., 2021). AR uses a fixed numerical threshold $\delta = 10^{-6}$ across all experiments to suppress numerically insignificant rounding changes. For all baselines, we use the official implementations and keep the quantization bit width, group size, calibration data, and evaluation pipeline consistent whenever supported by the released code. SpinQuant results on Qwen3 are unavailable because the released implementation does not support the evaluated Qwen3 architectures. All quantization experiments are conducted on a single NVIDIA RTX A6000 GPU.

4.1 MAIN RESULTS

Table 1 compares ARSO with GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), SpinQuant (Liu et al., 2025), and OmniQuant (Shao et al., 2024) across all models under 2-, 3-, and 4-bit quantization. ARSO achieves the lowest perplexity in 32 of the 36 model–dataset–bit settings, including 11 of 12 settings at both 2 and 3 bits and 10 of 12 settings at 4 bits, with particularly pronounced gains under aggressive low-bit quantization. ARSO also substantially outperforms OmniQuant while reducing offline quantization time by up to $14.2\times$.

Table 1: Perplexity of weight-only quantized models with group size 128. Lower is better. FP16 denotes the full-precision reference. The best quantized result is shown in bold. "–" denotes unavailable results.

		WikiText-2						C4					
Bits	Method	L2-7B	L2-13B	L3-8B	Q3-4B	Q3-8B	Q3-14B	L2-7B	L2-13B	L3-8B	Q3-4B	Q3-8B	Q3-14B
	FP16	5.472	4.884	6.138	13.665	9.728	8.645	6.973	6.468	8.931	16.635	13.300	12.018
2	GPTQ	33.75	17.86	415.04	136.57	56.37	24.28	38.06	23.04	415.07	152.48	61.92	31.02
	AWQ	2.30e5	1.25e5	1.70e6	3.71e7	1.62e9	4.31e7	1.73e5	9.67e4	2.11e6	3.14e7	1.83e9	4.69e7
	SpinQuant	89.49	–	220.63	–	–	–	68.03	–	235.35	–	–	–
	OmniQuant	11.55	8.52	573.32	48.62	25.50	18.33	14.68	**11.05**	588.59	53.95	32.31	23.10
	ARSO	**10.59**	**8.06**	**18.79**	**28.49**	**17.09**	**12.97**	**14.20**	11.31	**29.37**	**34.77**	**24.67**	**18.88**
3	GPTQ	6.31	5.38	33.37	15.73	11.12	9.72	7.95	7.09	20.76	18.48	14.80	13.13
	AWQ	6.14	5.29	8.07	16.85	11.23	9.94	7.75	**6.94**	11.47	20.04	14.88	13.26
	SpinQuant	6.40	–	8.05	–	–	–	8.07	–	11.77	–	–	–
	OmniQuant	6.14	5.36	8.87	17.33	11.67	10.03	7.81	7.03	12.59	20.32	15.53	13.58
	ARSO	**6.00**	**5.24**	**7.45**	**14.92**	**10.65**	**9.38**	**7.70**	6.95	**10.94**	**18.35**	**14.46**	**12.94**
4	GPTQ	5.69	4.97	11.56	14.22	10.20	8.84	7.22	6.57	12.11	**16.96**	13.74	12.26
	AWQ	5.60	**4.96**	6.53	14.68	9.99	8.86	**7.11**	**6.56**	9.45	17.55	13.65	12.31
	SpinQuant	5.61	–	6.53	–	–	–	7.15	–	9.49	–	–	–
	OmniQuant	5.61	4.97	6.62	14.72	10.15	8.96	7.12	6.57	9.58	17.63	13.81	12.40
	ARSO	**5.58**	**4.96**	**6.45**	**13.84**	**9.88**	**8.81**	**7.11**	6.57	**9.37**	16.99	**13.54**	**12.19**

Table 2: Zero-shot accuracy (%) of 2-bit weight-only quantized models with group size 128. Higher is better. Best results are in bold. "–" denotes unavailable results.

LLaMA-2-7B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg.
FP16	79.33	45.05	73.78	76.16	78.73	69.46	41.70	66.31
GPTQ	57.95	25.60	32.03	36.08	55.55	51.22	23.76	40.31
AWQ	37.83	26.62	25.76	26.09	50.76	50.12	25.51	34.67
SpinQuant	54.01	27.05	26.94	28.65	52.50	49.57	23.32	37.43
OmniQuant	62.94	24.32	41.33	48.56	63.49	53.91	23.35	45.41
ARSO	**66.18**	**29.35**	**49.66**	**51.19**	**68.12**	**60.22**	**26.61**	**50.19**

LLaMA-2-13B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg.
FP16	82.14	48.81	76.52	79.62	80.30	72.53	52.40	70.33
GPTQ	51.68	24.74	38.09	37.47	57.34	49.96	22.93	40.32
AWQ	62.17	27.82	27.02	26.03	50.16	49.01	26.89	38.44
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	65.32	29.78	46.46	56.84	65.45	53.99	23.69	48.79
ARSO	**77.22**	**35.15**	**61.78**	**60.86**	**72.03**	**65.59**	**28.71**	**57.34**

LLaMA-3-8B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg.
FP16	81.71	54.10	77.65	79.09	80.96	73.40	62.05	72.71
GPTQ	48.81	23.72	27.78	26.17	51.09	49.64	23.22	35.77
AWQ	62.17	24.83	25.00	26.46	50.54	50.43	24.65	37.73
SpinQuant	38.04	25.85	27.15	26.46	50.44	50.83	23.24	34.57
OmniQuant	37.86	22.44	27.44	27.58	54.95	50.99	22.95	34.89
ARSO	**70.95**	**27.90**	**48.32**	**50.93**	**66.81**	**62.35**	**28.47**	**50.82**

Qwen3-4B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg.
FP16	85.02	53.84	78.54	68.45	74.92	66.14	68.41	70.76
GPTQ	44.22	23.89	27.27	26.95	50.65	49.01	23.52	35.07
AWQ	37.83	26.54	24.92	26.52	50.65	48.93	26.89	34.61
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	63.49	**29.27**	**48.61**	41.85	**63.71**	55.80	30.88	47.66
ARSO	**68.84**	28.92	44.57	**42.95**	62.95	**57.93**	**31.58**	**48.25**

Qwen3-8B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg.
FP16	86.73	56.23	80.93	74.94	77.53	68.11	73.04	73.93
GPTQ	45.20	26.28	28.11	30.94	50.98	50.51	23.66	36.53
AWQ	62.17	27.39	26.01	25.94	51.52	50.43	22.95	38.06
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	73.61	**33.96**	**59.43**	48.64	68.61	56.75	26.57	52.51
ARSO	**74.46**	33.28	57.45	**53.98**	**70.78**	**60.30**	**47.57**	**56.83**

Qwen3-14B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg.
FP16	89.33	60.41	82.74	78.87	79.92	73.09	77.13	77.36
GPTQ	55.90	25.85	33.29	43.26	57.40	49.72	26.16	41.66
AWQ	62.17	26.37	25.34	25.77	50.87	49.64	24.65	37.83
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	76.97	**44.03**	**71.42**	58.88	74.37	63.22	49.47	62.62
ARSO	**84.43**	42.24	66.29	**62.37**	**74.43**	**68.67**	**58.58**	**65.29**

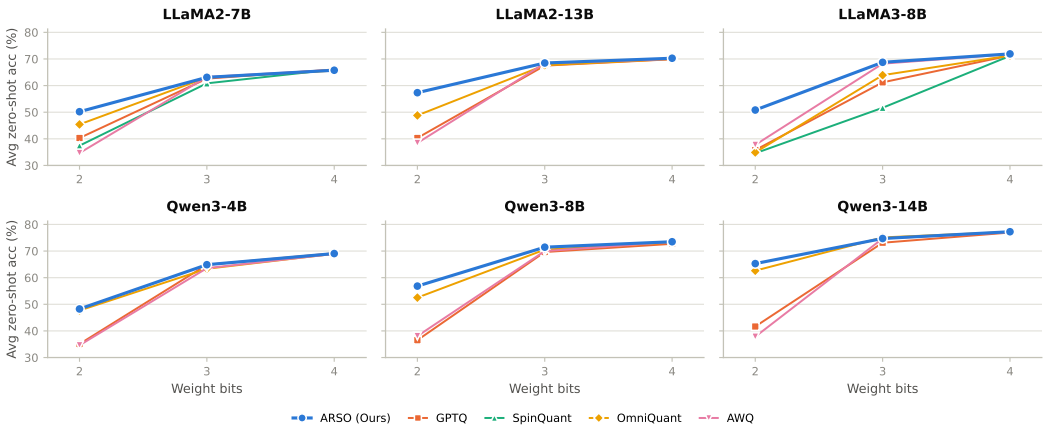


Figure 3: Average zero-shot accuracy across seven downstream tasks under 2-, 3-, and 4-bit weight-only quantization.

Table 2 and Fig. 3 show that these reconstruction gains also translate to downstream performance. ARSO achieves the highest average zero-shot accuracy across the six models at all three bit widths, outperforming the strongest competing method with complete model coverage by 6.14, 0.55, and 0.12 percentage points at 2, 3, and 4 bits, respectively. The substantially larger gain at 2 bits indicates that exploiting reconstruction freedom becomes particularly valuable under aggressive quantization. Complete 3- and 4-bit task-wise zero-shot results are provided in Appendix C.

4.2 Ablation Study

We conduct ablation experiments to examine the contributions of the reconstruction freedoms exploited by ARSO. Starting from a controlled sequential RTN baseline with the same compensation schedule as ARSO, we progressively introduce SO, J-SO, and compensation-aware AR.

Baseline $\rightarrow$ SO evaluates the benefit of exploiting scale freedom after integer codes are fixed, while SO $\rightarrow$ J-SO examines whether this freedom should be preserved and repeatedly re-optimized during sequential quantization. Adding compensation-aware AR yields the complete ARSO, which fur-

Table 3: Ablation study on Qwen3-8B with group size 128. Lower perplexity is better.

Method	2-bit		3-bit		4-bit	
	WT2	C4	WT2	C4	WT2	C4
Baseline	64.920	67.214	11.152	14.886	10.169	13.730
+ SO	18.736	27.673	10.761	14.610	9.898	13.587
+ J-SO	17.978	25.891	10.756	14.584	10.034	13.592
+ Comp.-aware AR (ARSO)	**17.092**	**24.668**	**10.651**	**14.460**	**9.884**	**13.543**

ther exploits future reconstruction freedom when making integer decisions. These three transitions correspond directly to the reconstruction freedoms identified in our formulation.

The gains are most pronounced under aggressive 2-bit quantization. At 3–4 bits, where the baseline quantization error is substantially smaller, the differences between variants narrow, while the complete ARSO remains consistently competitive.

4.3 Efficiency

We finally evaluate the offline quantization cost of ARSO on LLaMA-3-8B. Notably, AR turns reconstruction-aware rounding from an iterative optimization problem into a single-pass discrete decision process. For each group, AR makes a single pass over the coordinates, evaluating two adjacent integer candidates with the profiled score in Eq. (17). Its numerator and denominator are updated incrementally rather than recomputed after each rounding decision.

Table 4: End-to-end quantization time (seconds) on LLaMA-3-8B. Lower is better.

Method	2-bit	4-bit
GPTQ	733	724
AWQ	912	1224
SpinQuant	1596	1604
OmniQuant	20678	10268
ARSO (Ours)	1460	1378

ARSO requires additional preprocessing time compared with GPTQ because it performs reconstruction-aware integer selection and repeatedly refreshes the prefix scales. Nevertheless, at 2 and 4 bits, ARSO requires only about $2.0\times$ and $1.9\times$ the quantization time of GPTQ, respectively, while remaining faster than SpinQuant and $14.2\times$ and $7.5\times$ faster than OmniQuant. Importantly, this additional cost is incurred only once during offline quantization. ARSO preserves the standard group-wise affine representation after quantization and introduces neither additional inference-time parameters nor additional computation.

5 Conclusion

We introduced ARSO, a weight-only PTQ method built upon the principle of reconstruction freedom. Across six LLaMA and Qwen models under 2–4-bit quantization, ARSO achieves the lowest perplexity in 32 of 36 model–dataset–bit settings among the evaluated methods and obtains the highest average zero-shot accuracy across the six models at every evaluated bit width. Notably, ARSO extracts higher accuracy from the same affine quantization representation by improving only the integer-code and dequantization-scale decisions. The gains are most pronounced under aggressive 2-bit quantization, suggesting that reconstruction freedom provides an important optimization resource when the discrete representation is most constrained.

AI USE STATEMENT

In this work, we used generative AI tools to assist with discussions and verification of methodological details, as well as method implementation and debugging. We did not use generative AI tools to derive mathematical results, autonomously conduct experiments, or generate experimental data or results. Additionally, we used generative AI tools for language editing, literature search assistance, and figure design. We reviewed all AI-assisted work. All methodological details were independently reviewed by the authors, AI-assisted code was reviewed and tested through the reported experiments, mathematical arguments were verified by the authors, and references and related-work claims were checked against their original sources. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce ARSO throughout the main paper and appendix. The complete method, including the closed-form scale optimization, reconstruction-aware rounding criterion, sequential compensation, and overall algorithm, is described in Sec. 3. Experimental configurations, including the evaluated models, quantization settings, calibration data, random seed, evaluation benchmarks, and hardware, are specified in Sec. 4. Detailed derivations of the compensation-aware rounding metric and sequential compensation are provided in the appendix. The source code used to reproduce the reported experiments is included in the anonymous supplementary material.

REFERENCES

Juan Amboage, Pablo Monteagudo-Lago, Ian Colbert, Giuseppe Franco, and Nicholas Fraser. Optimal post-training quantization scales and where to find them. *arXiv*, 2026. doi: 10.48550/arXiv.2606.10890. URL https://arxiv.org/abs/2606.10890.

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 *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-3180. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

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

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

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv*, 2018. doi: 10.48550/arXiv.1803.05457. URL https://arxiv.org/abs/1803.05457.

Alireza Dadgarnia, Soroush Tabesh, Mahdi Nikdan, Michael Helcig, Eldar Kurtic, Maximilian Kleinegger, and Dan Alistarh. GSQ: Highly-accurate low-precision scalar quantization for

LLMs via gumbel-softmax sampling. *arXiv*, 2026. doi: 10.48550/arXiv.2604.18556. URL https://arxiv.org/abs/2604.18556.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12284–12303. PMLR, 2024. URL https://proceedings.mlr.press/v235/egiazarian24a.html.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1caf09c9f4e6b0150b06a07e77f2710c-Abstract-Conference.html.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2210.17323. URL https://arxiv.org/abs/2210.17323.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Junhan Kim, Yeo Jeong Park, Seungwoo Son, Chungman Lee, Ho-young Kim, Joonyoung Kim, and Yongkweon Jeon. TurboBoA: Faster and exact attention-aware quantization without backpropagation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/dd3bd4e35cdd224ed2153a996f41077f-Abstract-Conference.html.

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

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

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

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

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e5b1c0d4866f72393c522c8a00eed4eb-Abstract-Conference.html.

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

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

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

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

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c6483c8a68083af3383f91ee0dc6db95-Abstract-Conference.html.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57587–57613. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25l.html.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023. doi: 10.48550/arXiv.2307.09288. URL https://arxiv.org/abs/2307.09288.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656. PMLR, 2024a. URL https://proceedings.mlr.press/v235/tseng24a.html.

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with trellises and incoherence processing. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-1904. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6de2e84b8da47bb2eb5e2ac96c63d2b0-Abstract-Conference.html.

Mart Van Baalen, Andrey Kuzmin, Ivan Koryakovskiy, Markus Nagel, Peter Couperus, Cedric Bastoul, Eric Mahurin, Tijmen Blankevoort, and Paul Whatmough. GPTVQ: The blessing of dimensionality for LLM quantization. *arXiv*, 2024. doi: 10.48550/arXiv.2402.15319. URL https://arxiv.org/abs/2402.15319.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

An Yang et al. Qwen3 technical report. *arXiv*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

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

Shihao Zhang, Haoyu Zhang, Ian Colbert, and Rayan Saab. Qronos: Correcting the past by shaping the future... in post-training quantization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8d22a09f18d9fc85dcbe6f482e13e1d7-Abstract-Conference.html.

A Derivation of the Compensation-Aware Rounding Metric

We derive the conditional metric used by AR when the reconstruction error of the current group can be partially absorbed by the remaining unquantized groups. We use the notation defined in Sec. 3. At the time G is quantized, the preceding groups have already been fixed and their reconstruction error has been propagated forward. We therefore use w_G to denote the current compensated real-valued target of group G.

For a centered integer-code candidate u and its dequantization scale c, the reconstruction error introduced in the current group is

$$e_G = w_G - cu. \tag{20}$$

The future groups remain unquantized and can still be continuously adjusted. Let e_U denote the corresponding change in their reconstruction residual. Partitioning the Hessian over the current and future groups gives

$$H_{G\cup U, G\cup U} = \begin{bmatrix} H_{GG} & H_{GU} \\ H_{UG} & H_{UU} \end{bmatrix}. \tag{21}$$

For fixed c and u, the smallest reconstruction loss attainable after optimally adjusting the future groups is

$$\mathcal{L}_{\mathrm{cond}}(c,u) = \min_{e_U} \begin{bmatrix} e_G & e_U \end{bmatrix} \begin{bmatrix} H_{GG} & H_{GU} \\ H_{UG} & H_{UU} \end{bmatrix} \begin{bmatrix} e_G^\top \\ e_U^\top \end{bmatrix}. \tag{22}$$

Expanding the quadratic form yields

$$\mathcal{L}(c,u,e_U) = e_G H_{GG} e_G^\top + 2 e_G H_{GU} e_U^\top + e_U H_{UU} e_U^\top. \tag{23}$$

Since H_{UU} is positive definite after damping, minimizing Eq. (23) with respect to e_U gives

$$e_U^* = -e_G H_{GU} H_{UU}^{-1}. \tag{24}$$

This is the optimal continuous adjustment of the future groups for absorbing the error produced in the current group. Substituting Eq. (24) back into the quadratic objective gives

$$\begin{aligned} \mathcal{L}_{\mathrm{cond}}(c,u) &= e_G \left(H_{GG} - H_{GU} H_{UU}^{-1} H_{UG} \right) e_G^\top \\ &= (w_G - cu) M_G (w_G - cu)^\top, \end{aligned} \tag{25}$$

where

$$\boxed{M_G = H_{GG} - H_{GU} H_{UU}^{-1} H_{UG}.} \tag{26}$$

The difference between H_{GG} and M_G is important. The local metric H_{GG} measures the reconstruction loss caused by the current-group error before any future adjustment. In contrast, M_G measures only the loss that remains after the future groups have been optimally adjusted. It is therefore the effective reconstruction metric seen by the current group under sequential linear compensation.

We next eliminate the dequantization scale. For fixed integer codes u, Eq. (25) can be written as

$$\mathcal{L}_{\mathrm{cond}}(c,u) = w_G M_G w_G^\top - 2c\, w_G M_G u^\top + c^2 u M_G u^\top. \tag{27}$$

Minimizing over the unrestricted reconstruction scale gives

$$c^*(u) = \frac{w_G M_G u^\top}{u M_G u^\top}. \tag{28}$$

Substituting Eq. (28) into Eq. (27) yields

$$\mathcal{L}^*_{\mathrm{cond}}(u) = w_G M_G w_G^\top - \frac{\left(w_G M_G u^\top \right)^2}{u M_G u^\top}. \tag{29}$$

The first term is independent of the integer candidate. Consequently, minimizing the reconstruction loss after jointly eliminating the future compensation and the dequantization scale is equivalent to

$$\boxed{\begin{aligned} \arg\min_u \mathcal{L}^*_{\mathrm{cond}}(u) &= \arg\max_u \eta(u), \\ \eta(u) &= \frac{\left(w_G M_G u^\top \right)^2}{u M_G u^\top}. \end{aligned}} \tag{30}$$

Thus, the proposed score is not an additional heuristic objective. It is the profiled reconstruction objective obtained after accounting for both optimal scale selection and optimal continuous compensation by the future groups.

Equivalent inverse-Cholesky form. The conditional metric can be computed from the inverse-Cholesky factor already used by sequential compensation. Let the damped Hessian of the current suffix satisfy

$$H^{-1}_{G\cup U,G\cup U} = R^{\top}_{G\cup U,G\cup U} R_{G\cup U,G\cup U}, \tag{31}$$

where $R_{G\cup U,G\cup U}$ is upper triangular. Following the current quantization order, partition it as

$$R_{G\cup U,G\cup U} = \begin{bmatrix} R_{GG} & R_{GU} \\ 0 & R_{UU} \end{bmatrix}. \tag{32}$$

By the block-inverse identity, the upper-left block of the inverse Hessian is the inverse of the Schur complement:

$$\left(H^{-1}_{G\cup U,G\cup U}\right)_{GG} = M_G^{-1}. \tag{33}$$

Meanwhile, the triangular structure in Eq. (32) gives

$$\left(H^{-1}_{G\cup U,G\cup U}\right)_{GG} = R^{\top}_{GG} R_{GG}. \tag{34}$$

Combining Eqs. (33) and (34), we obtain

$$\boxed{M_G = R^{-1}_{GG} R^{-\top}_{GG}.} \tag{35}$$

Hence, for any row vectors a and b,

$$aM_G b^{\top} = \left(aR^{-1}_{GG}\right)\left(bR^{-1}_{GG}\right)^{\top}. \tag{36}$$

The score in Eq. (30) can therefore be evaluated in the transformed space induced by R^{-1}_{GG}:

$$\eta(u) = \frac{\left(\left(w_G R^{-1}_{GG}\right)\left(uR^{-1}_{GG}\right)^{\top}\right)^2}{\left\|uR^{-1}_{GG}\right\|_2^2}. \tag{37}$$

Up to the candidate-independent factor $\left\|w_G R^{-1}_{GG}\right\|_2^2$, this is the squared cosine alignment between $w_G R^{-1}_{GG}$ and uR^{-1}_{GG}. Therefore, the inverse-Cholesky transformation maps the current group into a space where directional alignment measures precisely the reconstruction component that cannot be eliminated by optimal future compensation.

B Conditional Suffix Optimum

B.1 Derivation of the Conditional Suffix Optimum

We derive the conditional compensation rule used in Eq. (19). After the processed prefix has been jointly reconstructed by SO, its residual

$$E_P = W_P^{\text{orig}} - Q_P^{\text{new}} \tag{38}$$

provides the current reconstruction state for compensating the unprocessed weights. We partition the Hessian accordingly:

$$H = \begin{bmatrix} H_{PP} & H_{PU} \\ H_{UP} & H_{UU} \end{bmatrix}. \tag{39}$$

For a suffix state W_U, define its residual as

$$E_U = W_U^{\text{orig}} - W_U. \tag{40}$$

The global quadratic reconstruction loss can then be expanded as

$$\begin{aligned} \mathcal{L}(E_P, E_U) &= \operatorname{Tr}\left(\begin{bmatrix} E_P & E_U \end{bmatrix} H \begin{bmatrix} E_P^\top \\ E_U^\top \end{bmatrix} \right) \\ &= C_P + 2\operatorname{Tr}\left(E_P H_{PU} E_U^\top\right) + \operatorname{Tr}\left(E_U H_{UU} E_U^\top\right), \end{aligned} \tag{41}$$

where

$$C_P = \operatorname{Tr}\left(E_P H_{PP} E_P^\top\right) \tag{42}$$

is independent of E_U.

Conditioned on the fixed prefix residual E_P, minimizing Eq. (41) with respect to E_U gives

$$\begin{aligned} 0 = \frac{\partial \mathcal{L}}{\partial E_U} &= 2E_P H_{PU} + 2E_U H_{UU}, \\ E_U^* &= -E_P H_{PU} H_{UU}^{-1}. \end{aligned} \tag{43}$$

Since $E_U^* = W_U^{\text{orig}} - W_U^{\text{opt}}$, the corresponding conditional optimum of the suffix weights is

$$W_U^{\text{opt}} = W_U^{\text{orig}} + E_P H_{PU} H_{UU}^{-1}. \tag{44}$$

Therefore, the compensated suffix state is precisely the minimizer of the global quadratic reconstruction objective conditioned on the current processed-prefix reconstruction.

For an efficient implementation, let R be an upper-triangular factor satisfying

$$H^{-1} = R^\top R, \tag{45}$$

and partition it according to P and U:

$$R = \begin{bmatrix} R_{PP} & R_{PU} \\ 0 & R_{UU} \end{bmatrix}. \tag{46}$$

Since $H = R^{-1}R^{-\top}$, its off-diagonal block satisfies

$$H_{PU} = -R_{PP}^{-1} R_{PU} H_{UU}, \tag{47}$$

which gives the block identity

$$H_{PU} H_{UU}^{-1} = -R_{PP}^{-1} R_{PU}. \tag{48}$$

Substituting Eq. (48) into Eq. (44) yields

$$W_U^{\text{opt}} = W_U^{\text{orig}} - E_P R_{PP}^{-1} R_{PU}, \tag{49}$$

which is Eq. (19) in the main text. Taking the first block of U further gives

$$W_{k+1}^{\text{opt}} = W_{k+1}^{\text{orig}} - E_P R_{PP}^{-1} R_{P,k+1}, \tag{50}$$

corresponding to the next-group update in Eq. (18).

B.2 Connection to GPTQ Compensation

We first explain why sequential GPTQ compensation can be viewed as recursive conditional minimization of the original quadratic objective. Consider

$$\Delta\mathcal{L}(\delta w) = \frac{1}{2}\delta w H \delta w^\top. \tag{51}$$

After a processed prefix P is fixed, GPTQ adjusts the remaining variables U to their conditional minimizer

$$\delta w_U^* = \arg\min_{\delta w_U} \Delta\mathcal{L}(\delta w_P, \delta w_U), \tag{52}$$

so that

$$\nabla_{\delta w_U} \Delta\mathcal{L}\left(\delta w_P, \delta w_U^*\right) = 0. \tag{53}$$

Let Δu be the perturbation introduced by the next quantization decision. Since the objective is quadratic,

$$\begin{aligned}\Delta\mathcal{L}\left(\delta w_P, \delta w_U^* + \Delta u\right) = \Delta\mathcal{L}\left(\delta w_P, \delta w_U^*\right) \\ + \nabla_{\delta w_U} \Delta\mathcal{L}\left(\delta w_P, \delta w_U^*\right) \Delta u^\top + \frac{1}{2}\Delta u H_{UU} \Delta u^\top.\end{aligned} \tag{54}$$

The cross term vanishes by Eq. (53). Hence, the effect of previously quantized columns is already absorbed into the compensated state, and GPTQ can remove the fixed columns and continue on the reduced quadratic. Recursively,

$$\Delta\mathcal{L}_{\text{total}} = \sum_k \Delta\mathcal{L}_k. \tag{55}$$

Therefore,

$$\boxed{\text{sequential GPTQ compensation} \equiv \text{recursive conditional minimization.}} \tag{56}$$

Consequently, for a fixed processed prefix, the accumulated GPTQ compensation reaches the same conditional suffix state as the one-shot solution in Eq. (49).

When columns are organized into groups, the same GPTQ procedure can be viewed as cross-group conditional compensation followed by intra-group GPTQ compensation. ARSO retains only the cross-group part: after J-SO refreshes the processed-prefix residual, Eq. (18) places the immediately next group at the corresponding one-shot conditional optimum before it is quantized.

C FULL ZERO-SHOT RESULTS

Table 5: Zero-shot accuracy (%) of 3-bit weight-only quantized models with group size 128. Higher is better. Best results are in bold. "–" denotes unavailable results.

LLaMA-2-7B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	79.33	45.05	73.78	76.16	78.73	69.46	41.70	66.31
GPTQ	74.53	41.81	68.90	72.16	77.64	66.54	**36.72**	62.61
AWQ	75.69	42.24	**70.62**	**73.77**	77.80	**68.03**	34.44	**63.23**
SpinQuant	74.19	41.47	65.82	71.74	75.90	64.40	32.38	60.84
OmniQuant	74.01	42.15	**70.62**	72.65	**78.29**	67.09	36.50	63.05
ARSO	**75.90**	**43.43**	70.24	72.39	77.37	67.72	34.67	63.10

LLaMA-2-13B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	82.14	48.81	76.52	79.62	80.30	72.53	52.40	70.33
GPTQ	80.12	45.56	73.57	76.78	79.65	70.32	46.43	67.49
AWQ	80.21	**47.95**	74.49	**77.23**	79.65	70.72	48.28	68.36
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	79.66	46.50	73.57	76.96	**79.65**	69.22	48.01	67.65
ARSO	**80.76**	47.53	**75.17**	76.70	79.49	**71.19**	**48.55**	**68.48**

LLaMA-3-8B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	81.71	54.10	77.65	79.09	80.96	73.40	62.05	72.71
GPTQ	67.86	42.06	65.91	69.72	74.48	68.90	39.55	61.21
AWQ	80.80	47.78	**73.36**	**74.97**	**78.07**	71.27	51.38	68.23
SpinQuant	48.81	32.42	52.90	60.11	69.91	62.12	35.34	51.66
OmniQuant	74.46	41.21	63.26	73.23	76.66	68.11	50.44	63.91
ARSO	**81.28**	**47.95**	72.94	74.37	77.53	**72.45**	**54.72**	**68.75**

Qwen3-4B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	85.02	53.84	78.54	68.45	74.92	66.14	68.41	70.76
GPTQ	81.19	**45.99**	**70.83**	**63.73**	**72.74**	**62.83**	56.60	64.85
AWQ	78.62	45.31	66.08	62.53	72.42	60.22	59.81	63.57
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	81.74	42.83	60.69	62.68	72.20	62.04	61.23	63.34
ARSO	**82.26**	44.54	66.16	63.40	**72.74**	62.67	**62.36**	**64.88**

Qwen3-8B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	86.73	56.23	80.93	74.94	77.53	68.11	73.04	73.93
GPTQ	83.21	51.71	75.25	70.51	**76.93**	64.17	65.66	69.63
AWQ	84.77	50.26	75.17	70.54	75.90	66.06	67.26	69.99
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	85.38	51.71	77.19	69.39	76.06	67.01	66.97	70.53
ARSO	**85.78**	**52.56**	**77.74**	**71.01**	76.39	**68.03**	**68.66**	**71.45**

Qwen3-14B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	89.33	60.41	82.74	78.87	79.92	73.09	77.13	77.36
GPTQ	86.54	52.30	76.73	**76.05**	78.24	69.46	72.55	73.12
AWQ	87.68	56.66	79.80	75.68	78.51	71.67	72.92	74.70
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	87.83	**58.28**	**80.60**	75.45	**79.27**	72.06	72.39	**75.13**
ARSO	**88.99**	54.78	79.76	75.93	78.35	**72.06**	**73.18**	74.72

Table 6: Zero-shot accuracy (%) of 4-bit weight-only quantized models with group size 128. Higher is better. Best results are in bold. "–" denotes unavailable results.

LLaMA-2-7B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	79.33	45.05	73.78	76.16	78.73	69.46	41.70	66.31
GPTQ	78.53	**46.16**	**74.28**	75.49	**79.43**	68.27	40.25	66.06
AWQ	**80.03**	44.62	72.69	75.26	78.78	69.30	40.55	65.89
SpinQuant	77.52	45.05	74.12	**75.81**	78.02	69.77	**42.29**	**66.08**
OmniQuant	77.92	43.86	72.56	74.98	78.84	69.38	41.77	65.61
ARSO	79.14	43.52	73.78	74.78	78.13	**69.85**	41.16	65.76

LLaMA-2-13B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	82.14	48.81	76.52	79.62	80.30	72.53	52.40	70.33
GPTQ	81.25	48.12	**76.43**	**79.33**	80.41	71.98	51.74	69.90
AWQ	81.59	48.04	76.35	79.24	**80.90**	72.38	51.37	69.98
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	81.90	48.21	75.55	78.85	80.63	72.53	**52.07**	69.96
ARSO	**82.42**	**48.72**	76.22	78.96	80.47	**73.32**	51.82	**70.28**

LLaMA-3-8B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	81.71	54.10	77.65	79.09	80.96	73.40	62.05	72.71
GPTQ	**81.41**	51.37	76.39	78.13	79.33	73.72	58.71	71.29
AWQ	81.13	52.39	77.15	78.53	**79.87**	73.24	60.32	71.80
SpinQuant	80.03	52.47	**79.25**	76.83	79.16	72.53	57.76	71.15
OmniQuant	80.40	50.43	77.02	**78.65**	79.60	72.38	60.55	71.29
ARSO	78.96	**53.84**	77.36	78.45	**79.87**	**74.11**	**60.77**	**71.91**

Qwen3-4B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	85.02	53.84	78.54	68.45	74.92	66.14	68.41	70.76
GPTQ	84.50	**52.05**	74.75	**67.58**	75.24	63.46	66.71	**69.18**
AWQ	**85.29**	49.49	74.07	67.10	**75.52**	63.85	67.12	68.92
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	84.22	50.26	**75.84**	66.90	74.70	63.69	**67.14**	68.96
ARSO	85.14	50.43	**75.84**	**67.58**	74.05	**63.93**	66.45	69.06

Qwen3-8B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	86.73	56.23	80.93	74.94	77.53	68.11	73.04	73.93
GPTQ	86.30	54.44	78.24	**74.39**	76.88	67.17	71.59	72.71
AWQ	86.21	55.72	**80.18**	73.84	76.55	**69.38**	72.11	73.43
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	**86.64**	54.44	80.09	73.57	**77.42**	68.51	71.86	73.22
ARSO	**86.64**	**56.06**	79.80	73.99	76.88	68.98	**72.17**	**73.50**

Qwen3-14B

Method	BoolQ	ARC-C	ARC-E	Hella.	PIQA	Wino.	MMLU	Avg
FP16	89.33	60.41	82.74	78.87	79.92	73.09	77.13	77.36
GPTQ	**89.30**	59.90	82.41	**78.71**	**80.41**	71.82	76.16	76.96
AWQ	89.14	61.52	81.90	78.29	79.60	72.61	**76.32**	77.06
SpinQuant	–	–	–	–	–	–	–	–
OmniQuant	89.11	**61.69**	**83.38**	78.31	80.03	71.98	75.79	77.18
ARSO	89.14	61.26	83.08	78.33	79.98	**72.93**	76.09	**77.26**