TASK-SENSITIVE LOW-RANK QUANTIZATION ERROR RECONSTRUCTION WITH PROPAGATION-AWARE COORDINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank quantization error reconstruction adds limited correction capacity to a fixed quantized model. Allocating that capacity requires a measure of task value, while deploying the corrections requires accounting for their responses under quantized states. We propose Propagation-Aware Task-Sensitive Low-Rank Quantization Error Reconstruction (PATR). Stage I combines input second moments with observed-token output gradients, reconstructs residuals in an active input subspace, and allocates heterogeneous ranks under a common factor-bit budget. Stage II fixes those directions and ranks, measures four single-group responses per Transformer Block under a correction-free quantized prefix, and fits their amplitudes in a suffix-sensitive metric. The Stage-I spectral gains constrain amplitude changes. Full-model Kullback–Leibler (KL) divergence selects a shared deployment step, whose amplitudes are absorbed into the existing factors. Across seven Llama and Qwen models, GPTQ and round-to-nearest (RTN) backends, and W4A8/W4A4/W3A4, complete PATR reduces WikiText2 perplexity (PPL) in all 42 matched configurations. A separate equal-budget rank-allocation study improves all 12 pairs, with a mean relative PPL reduction of 1.233%.

1 INTRODUCTION

Post-training quantization (PTQ) reduces the storage and arithmetic costs of large language models (LLMs) without full retraining. Weight reconstruction, activation-aware scaling, and equivalent transformations address errors introduced by low-precision representations (Frantar et al., 2023; Lin et al., 2024b; Xiao et al., 2023; Ashkboos et al., 2024). Under joint weight–activation (W–A) quantization, these errors also change the states passed between layers, affecting the model's final predictions.

Low-rank quantization error reconstruction (QER) adds compact corrections to quantized weights to recover part of this loss (Zhang et al., 2024; 2025; Park et al., 2026). Input-aware and saliency-guided methods use activation information to direct reconstruction toward influential residual components. Their correction capacity is still limited: increasing one projection's rank consumes resources that could be assigned elsewhere. The choice of residual directions and their distribution across projections must therefore be considered together.

Residual magnitude alone is insufficient for this choice. A direction's task importance depends on how strongly the inputs excite it and how sensitive the final loss is to its output perturbation. Task-aware quantization and two-sided low-rank modeling incorporate downstream information: YAQA uses full-model KL curvature for adaptive rounding, and OBD-LLM applies a K-FAC geometry to weight decomposition and quantization-residual compensation (Tseng et al., 2026; Li et al., 2026). For fixed-base QER, this task information can guide both the directions to reconstruct and the capacity assigned to each projection.

A task-sensitive correction constructed on a full-precision (FP) reference trajectory must also operate on states shifted by upstream quantization. Attention, MLP nonlinearities, activation quantization, and residual paths alter its finite-amplitude response. The task metric used to select a direction thus does not by itself determine its amplitude alongside other corrections. This separates the allocation of task-sensitive corrections from the coordination of their quantized-state responses.

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.

We propose PATR for a fixed quantized backbone with a constrained low-rank correction capacity. Its core component, *Task-Sensitive Correction Allocation* (Stage I), combines input second moments with observed-token negative log-likelihood (NLL) output gradients. It constructs low-rank corrections in an active input subspace and allocates ranks across projections using task-weighted spectral gains. *Quantized-State Response Coordination* (Stage II) adjusts the selected corrections using four measured single-group Block responses under correction-free quantized prefixes. The reconstruction gains constrain amplitude changes, and full-model KL selects a shared deployment step. The selected amplitudes are absorbed into the existing factors.

Our contributions are:

1. **Task-Sensitive Correction Allocation.** We construct low-rank residual corrections in a two-sided task metric and allocate capacity across projections by dynamic programming over their spectral gains under a model-wide capacity constraint.
2. **Quantized-State Response Coordination.** We fit four group amplitudes from quantized-prefix Block responses, using reconstruction gains to constrain the adjustment and full-model KL to select deployment while preserving correction directions and allocated capacity.
3. **Evaluation across models and quantization settings.** Experiments across Llama and Qwen models and multiple quantization settings show consistent PPL improvements over matched quantized baselines and the benefit of adaptive allocation at equal budgets. Response diagnostics, stagewise comparisons, downstream accuracy, and factor-cost accounting provide complementary evaluations.

2 Background and Motivation

2.1 LLM Post-Training Quantization

PTQ uses calibration data to determine low-precision representations. Reconstruction and activation saliency guide weight quantization (Frantar et al., 2023; Lin et al., 2024b), while scaling, clipping, and equivalent transformations reshape the distributions presented to the quantizers (Xiao et al., 2023; Shao et al., 2024; Ma et al., 2024; Ashkboos et al., 2024; Lin et al., 2024a). ARHQ incorporates activation-residual Hessian information for low-bit W–A quantization (Wang et al., 2026). For $W \in \mathbb{R}^{d_o \times d_i}$ and token-major inputs $X \in \mathbb{R}^{N \times d_i}$, define

$$W_q = Q_w^{(b_w)}(W), \qquad X_q = Q_a^{(b_a)}(X), \qquad Y_q = X_q W_q^\top.$$

Weight-only PTQ leaves $X_q = X$; W4A8 uses 4-bit weights and 8-bit activations. We consider residual correction with these base weights and quantizers fixed.

2.2 Low-Rank Quantization Error Reconstruction

With $E = W - W_q$, QER uses $\widehat{W} = W_q + L$ and $\mathrm{rank}(L) \le r$. Writing $L^\top = FG$, with $F \in \mathbb{R}^{d_i \times r}$ and $G \in \mathbb{R}^{r \times d_o}$, gives

$$X_q \widehat{W}^\top = X_q W_q^\top + (X_q F)G, \tag{1}$$

Both branches receive the same quantized input X_q. Weight-SVD minimizes $\|E - L\|_F^2$ by truncating the singular value decomposition (SVD) of E. QERA distinguishes this weight-error objective from input-aware layer-output reconstruction (Zhang et al., 2025):

$$L_r^{\mathrm{X}} = \arg\min_{\mathrm{rank}(L) \le r} N^{-1} \|X(E - L)^\top\|_F^2 = \arg\min_{\mathrm{rank}(L) \le r} \|(E - L)H_x^{1/2}\|_F^2,$$

where $H_x = N^{-1} X^\top X$ weights the input directions that excite the remaining residual (our convention transposes QERA's weights).

Reconstruction designs differ in both their error metric and their use of capacity. LQER reshapes the residual through activation scaling (Zhang et al., 2024); CALDERA combines low rank with low precision, and SERQ uses saliency-guided single-matrix compensation (Saha et al., 2024; Park et al., 2026). Flexible-rank sketching and splitting capacity between pre-quantization preservation and residual reconstruction provide further choices (Gu et al., 2026; Cho et al., 2026). For a fixed quantized backbone, the question is how to distribute the available correction capacity according to its value to the task.

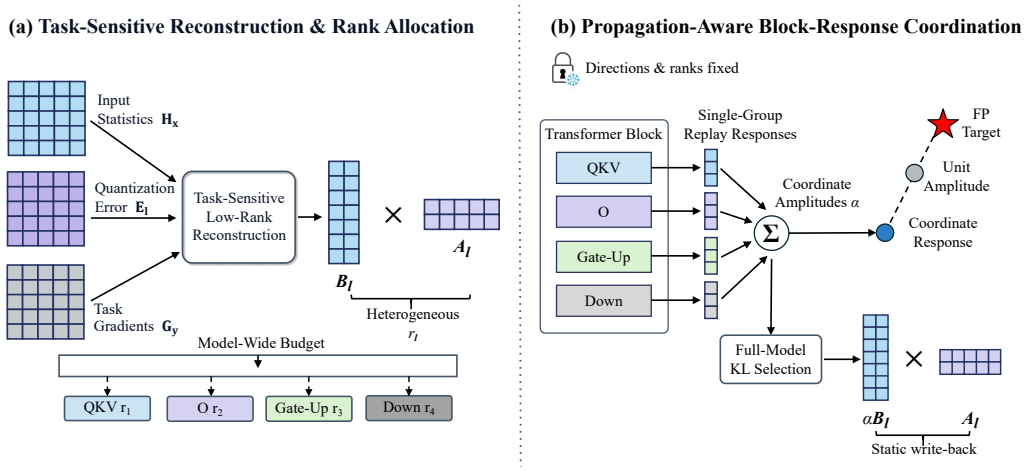


Figure 1: PATR workflow. Stage I constructs F_l, G_l and allocates ranks. Its spectral gains also constrain Stage-II amplitude fitting. Four fixed-prefix replays give each Block's responses; full-model KL selects a shared step, and the amplitudes are written into G_l.

2.3 Task Value and Correction Capacity

Input-aware reconstruction weights the directions excited by the inputs, but measures their output perturbations in the layer's Euclidean metric. Equal-sized output errors can therefore correspond to different task costs. Downstream sensitivity provides a second weighting: YAQA uses Kronecker-factored full-model KL curvature for adaptive rounding (Tseng et al., 2026), while OBD-LLM uses input and output-gradient factors for low-rank reconstruction, including preset-rank quantization-residual compensation (Li et al., 2026). IO-SVD combines predictive-KL whitening with heterogeneous ranks and hybrid low-rank/quantized weight compression (Abbasi et al., 2026).

PATR uses task information to allocate corrections to a fixed quantized backbone. The metric must support two decisions: which residual directions to retain within a projection, and how much capacity to assign across projections. Task-weighted spectral gains and their associated factor costs provide the basis for this allocation. The output gradients carry downstream task information, evaluated on the FP reference trajectory.

2.4 Correction Responses under Quantized States

At deployment, the correction receives inputs from the quantized network rather than the FP reference trajectory used to construct it. Upstream quantization shifts these inputs, and the current Block transforms the correction through attention, MLP nonlinearities, activation quantization, and residual connections. These operations can change the scale and overlap of correction responses. The relevant target therefore includes both incoming state discrepancy and error introduced within the current Block.

Block reconstruction and propagated-error objectives address dependencies beyond an isolated projection. BRECQ develops Block reconstruction for vision PTQ (Li et al., 2021); for LLMs, CBQ jointly optimizes LoRA-Rounding and quantization step sizes across overlapping Blocks, while QEP corrects weights using upstream quantization error and controls its propagation strength (Ding et al., 2025; Arai & Ichikawa, 2025). In low-rank weight compression, LACE-SVD combines loss-aware allocation, factor updates, and cumulative error correction (Liu et al., 2026).

PATR instead retains the selected correction directions and ranks and uses measured quantized-prefix responses to coordinate four group amplitudes per Block. An additive response model accounts for within-Block overlap; full-model KL evaluates their joint deployment. The reconstruction gains constrain amplitude changes, connecting response coordination to the task value used for allocation.

3 METHOD

PATR combines *Task-Sensitive Correction Allocation* and *Quantized-State Response Coordination*. The first component selects correction directions and distributes low-rank capacity according to task-weighted spectral gains. The second holds these directions and ranks fixed and coordinates correction amplitudes using measured responses under quantized states.

3.1 BUDGETED LOW-RANK COMPENSATION

For target projections $l \in \mathcal{L}$, with $W_l \in \mathbb{R}^{d_{o,l} \times d_{i,l}}$, we seek corrections under a common factor-capacity budget $\mathcal{B}$:

$$W_{q,l} = Q_l(W_l), \quad E_l = W_l - W_{q,l}, \quad \widehat{W}_l = W_{q,l} + L_l, \\ \operatorname{rank}(L_l) \le r_l, \qquad \sum_{l \in \mathcal{L}} c_l^{\mathrm{cap}}(r_l) \le \mathcal{B}. \tag{2}$$

The base weights and activation quantizer Q_A remain fixed. Let p_{FP} and $p^{Q}_{\{L_l\}}$ denote the full-precision and corrected predictive distributions. With calibration distribution $\mathcal{D}_{\mathrm{cal}}$ and stop-gradient operator sg, the ideal deployment objective is

$$\min_{\{L_l, r_l\}} \quad \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{cal}}} \mathrm{KL}\Big(\mathrm{sg}[p_{\mathrm{FP}}(\cdot \mid x)] \| p^{Q}_{\{L_l\}}(\cdot \mid x)\Big) \\ \text{s.t.} \quad \sum_{l} c_l^{\mathrm{cap}}(r_l) \le \mathcal{B}, \quad \operatorname{rank}(L_l) \le r_l, \quad r_l \in \mathcal{R}. \tag{3}$$

PATR separates capacity selection from response coordination (Figure 1). Stage I uses observed-token negative log-likelihood (NLL) gradients to choose directions and ranks, yielding the *pre-coordination* model (Pre). Stage II fits weighted Block responses; full-model KL selects the deployment model, or *complete PATR*. The NLL-gradient and response objectives are tractable, complementary surrogates for the discrete deployment problem, not equivalent KL losses.

3.2 TASK-SENSITIVE RESIDUAL GEOMETRY

Let $R_l = E_l - L_l$. With other weights fixed, a twice-differentiable full-precision task loss admits the expansion

$$\Delta \mathcal{L}_l \approx -\langle \nabla_{W_l} \mathcal{L}, R_l \rangle_F + \tfrac{1}{2} \operatorname{vec}(R_l)^\top \mathcal{H}_l \operatorname{vec}(R_l), \tag{4}$$

where $\mathcal{H}_l$ is the corresponding Hessian block. Under a near-stationarity approximation, we neglect the linear term (Li et al., 2026) and use the positive-semidefinite, empirical-Fisher/K-FAC-inspired surrogate (Martens & Grosse, 2015; Kunstner et al., 2019)

$$q_l(R_l) = \tfrac{1}{2} \operatorname{vec}(R_l)^\top (H_{x,l} \otimes G_{y,l}) \operatorname{vec}(R_l) \quad = \tfrac{1}{2} \| G_{y,l}^{1/2} R_l H_{x,l}^{1/2} \|_F^2. \tag{5}$$

Here vec is column-wise, $H_{x,l} = \mathbb{E}_{\mathrm{cal}}[x_l x_l^\top]$, and $G_{y,l} = \mathbb{E}_{\mathrm{cal}}[g_{y,l} g_{y,l}^\top]$, with column vectors $x_l \in \mathbb{R}^{d_{i,l}}$, $y_l \in \mathbb{R}^{d_{o,l}}$, and $g_{y,l} = \partial \ell_{\mathrm{task}} / \partial y_l$. Backpropagation of observed-token NLL makes $G_{y,l}$ sensitive to downstream task computation on the full-precision trajectory. The Kronecker factorization and gradient outer products approximate curvature; they are not the exact Hessian or the Fisher of teacher KL. Near-stationarity is also an approximation on the finite calibration set.

The implementation uses a diagonal output metric with a positive floor:

$$\bar{\gamma}_l = \frac{1}{N} \sum_{n=1}^{N} g_{y,l,n} \odot g_{y,l,n}, \quad \Gamma_l = \mathrm{diag}[\max(\bar{\gamma}_l, \tau_l \mathbf{1})], \qquad D_l = \Gamma_l^{1/2}, \tag{6}$$

Here N counts valid tokens, $\odot$ and $\max$ act elementwise, and $\tau_l > 0$. We collect input moments and output gradients on the full-precision trajectory with consistent NLL normalization. The two sides weight input excitation and output task sensitivity, respectively. Under W–A deployment, activation error and shifted inputs also affect the projection output; the exact decomposition is given in Appendix A. Stage II measures the resulting Block responses after the Stage-I corrections have been selected.

3.3 Reconstruction and Global Rank Allocation

For $X_l \in \mathbb{R}^{N \times d_{i,l}}$, decompose

$$H_{x,l} = N^{-1} X_l^\top X_l = U_l \Lambda_l U_l^\top. \tag{7}$$

A fixed numerical threshold retains the active eigenpairs $(U_{A,l}, \Lambda_{A,l})$, with $d_{A,l}$ retained columns. We restrict $L = L U_{A,l} U_{A,l}^\top$ and solve

$$\min_{\substack{\operatorname{rank}(L) \le r \\ L = L U_{A,l} U_{A,l}^\top}} \left\| D_l (E_l - L) H_{x,l}^{1/2} \right\|_F^2. \tag{8}$$

This restriction removes unsupported or weakly supported input directions and avoids inverting negligible eigenvalues. Define

$$M_l = D_l E_l U_{A,l} \Lambda_{A,l}^{1/2} = \widetilde{U}_l \Sigma_l \widetilde{V}_l^\top, \tag{9}$$

and $k_l(r) = \min\{r, \operatorname{rank}(M_l)\}$.

Proposition 1. If $D_l \succ 0$ on the modeled output space and $\Lambda_{A,l}$ is positive definite, an optimal rank-at-most-r solution is

$$L_{l,r} = D_l^{-1} \widetilde{U}_{l,k_l(r)} \Sigma_{l,k_l(r)} \widetilde{V}_{l,k_l(r)}^\top \Lambda_{A,l}^{-1/2} U_{A,l}^\top. \tag{10}$$

For the exact optimum $J_l(r)$, including residual outside the active subspace, the gain over zero correction is

$$\mathcal{G}_l(r) = J_l(0) - J_l(r) = \sum_{k=1}^{k_l(r)} \sigma_{l,k}^2, \tag{11}$$

where $\sigma_{l,k}$ are the exact singular values. The result follows by a rank-preserving change of variables and the Eckart–Young–Mirsky theorem (Appendix A).

The implementation shares one randomized top-$r_{\max}$ decomposition across all candidate ranks and filters numerically invalid components. Let $\kappa_l(r) = \min(r, d_{i,l}, d_{o,l}, d_{A,l})$ and $\widehat{r}_{\text{eff},l}(r)$ be the valid component count. It uses $\widehat{\mathcal{G}}_l(r) = \sum_{k=1}^{\widehat{r}_{\text{eff},l}(r)} \widehat{\sigma}_{l,k}^2$. Hats distinguish these computed proxies from exact-SVD quantities. With balanced factors $L_{l,r}^\top = F_{l,r} G_{l,r}$ and b_F bits per factor element, the assigned capacity and realized payload are

$$c_l^{\text{cap}}(r) = b_F \kappa_l(r)(d_{i,l} + d_{o,l}), \quad c_l^{\text{real}}(r) = b_F \widehat{r}_{\text{eff},l}(r)(d_{i,l} + d_{o,l}). \tag{12}$$

The factor-bit budget links rank choices across all target projections:

$$\max_{\{r_l\}} \quad \sum_l \widehat{\mathcal{G}}_l(r_l) \quad \text{s.t.} \quad r_l \in \mathcal{R}, \qquad \sum_l c_l^{\text{cap}}(r_l) \le \mathcal{B}. \tag{13}$$

Pareto dynamic programming finds the global optimum over these computed discrete gain curves and capacity costs. Larger gains can therefore receive more capacity without increasing the total budget. Algorithm 1 in the appendix summarizes the two stages; all rank candidates use $r_{\max} = \max \mathcal{R}$.

3.4 Propagation-Aware Block-Response Coordination

Stage II measures how fixed corrections act through quantized Block computation. For Block $b \in \{1, \dots, B\}$, let $\mathcal{L}_b$ be its target projections and $\mathcal{F}_b$, $\widehat{\mathcal{F}}_b$ its full-precision and corrected quantized mappings. All Stage-II statistics are collected once under correction-free quantized prefixes, without recollection after final upstream corrections.

Boundary target and measured responses. Let h_{b-1}^* and h_{b-1}^0 be states on the full-precision and uncorrected quantized trajectories. Valid tokens share a token-major layout $z_b \in \mathbb{R}^{N_b \times d_b}$. Each Block

has four groups: QKV, O, Gate/Up, and Down. One amplitude $\alpha_{b,g}$ scales every Stage-I correction in group g. With the same h_{b-1}^0 for all replays, define

$$\begin{aligned} z_b^* &= \mathcal{F}_b(h_{b-1}^*), \quad z_b^0 = \widehat{\mathcal{F}}_b^0(h_{b-1}^0), \\ z_{b,g}^{\text{single}} &= \widehat{\mathcal{F}}_b^{(g,\alpha=1)}(h_{b-1}^0), \\ t_b &= z_b^* - z_b^0, \quad \Delta_{b,g} = z_{b,g}^{\text{single}} - z_b^0. \end{aligned} \tag{14}$$

The target t_b contains incoming state discrepancy and current-Block quantization error (Eq. (25)). Each $\Delta_{b,g} \in \mathbb{R}^{N_b \times d_b}$ is a measured finite difference from amplitude zero to one, including the Block's nonlinear and quantized computation.

Suffix-sensitive amplitude fitting. For Block b, use a hybrid model that is quantized and uncorrected through b and full precision thereafter. Its teacher-KL gradient gives

$$\begin{aligned} \ell_b^{\text{pre}} &= \mathrm{KL}\big(\mathrm{sg}[p_{\text{FP}}] \| p_{Q_{\le b}, W_{>b}}\big)\,, \\ \Omega_b &= \mathrm{diag}\left[\max\left(\frac{1}{N_b}\sum_{n=1}^{N_b}\left(\frac{\partial \ell_b^{\text{pre}}}{\partial z_{b,n}}\right)^{\odot 2}, \tau_b \mathbf{1}\right)\right]. \end{aligned} \tag{15}$$

The diagonal $\Omega_b \in \mathbb{R}^{d_b \times d_b}$ weights boundary channels by suffix sensitivity, with floor $\tau_b > 0$. It is a gradient-square proxy, not an explicit propagation Jacobian. For $\boldsymbol{\alpha}_b \in \mathbb{R}^4$, the response objective is

$$\begin{aligned} J_b^{\text{prop}}(\boldsymbol{\alpha}_b) &= \frac{1}{N_b}\left\| \left(t_b - \sum_{g=1}^{4} \alpha_{b,g}\Delta_{b,g}\right)\Omega_b^{1/2} \right\|_F^2 \\ &= c_b - 2\boldsymbol{h}_b^\top \boldsymbol{\alpha}_b + \boldsymbol{\alpha}_b^\top \boldsymbol{R}_b \boldsymbol{\alpha}_b, \end{aligned} \tag{16}$$

Right multiplication by $\Omega_b^{1/2}$ weights channels in the token-major layout. The weighted inner products defining the coefficients are given in Eq. (26). The 4×4 positive-semidefinite Gram $\boldsymbol{R}_b$ measures within-Block overlap; $\boldsymbol{h}_b \in \mathbb{R}^4$ measures target alignment. The unconstrained continuous stationary equation $\boldsymbol{R}_b \boldsymbol{\alpha}_b = \boldsymbol{h}_b$ explains how these quantities determine amplitudes after directions and ranks are fixed. The implementation instead searches the fixed amplitude grid.

This additive secant model differs from an actual multi-group forward pass. $\mathbf{0}$ and $\mathbf{1}$ denote zero and unit coefficients; coefficients above one extrapolate the model. Full-model selection evaluates actual joint deployment.

Link to Stage-I task value. Under Proposition 1's exact-SVD conditions, scaling a correction gives $J_l(\alpha) = J_l(r_l) + (1-\alpha)^2 \mathcal{G}_l(r_l)$. Unit amplitude is the local optimum, while changing it incurs a cost weighted by the selected spectral gain. The implementation defines $\widehat{J}_l(r) = J_l(0) - \widehat{\mathcal{G}}_l(r)$ from its randomized decomposition and constrains

$$\begin{aligned} \widehat{J}_b^{\text{local}}(\boldsymbol{\alpha}_b) &= \sum_{l \in \mathcal{L}_b}\left[\widehat{J}_l(r_l) + (1-\alpha_{b,g(l)})^2 \widehat{\mathcal{G}}_l(r_l)\right], \\ \widehat{J}_b^{\text{local}}(\boldsymbol{\alpha}_b) &\le (1+\rho)\widehat{J}_b^{\text{local}}(\mathbf{1}). \end{aligned} \tag{17}$$

Here $g(l)$ maps projections to groups. With a nonnegative unit residual proxy, higher Stage-I gains penalize amplitude departures more strongly. For approximate factors this is a spectral-proxy constraint, not a Block KL/NLL bound (Appendix A). Grid search minimizes Eq. (16) subject to Eq. (17) to obtain $\boldsymbol{\alpha}_b^*$.

Full-model selection and factor write-back. All Blocks share an interpolation step from a finite $\Lambda \subseteq [0,1]$ containing zero:

$$\begin{aligned} \boldsymbol{\alpha}_b(\lambda) &= \mathbf{1} + \lambda(\boldsymbol{\alpha}_b^* - \mathbf{1}), \quad \lambda \in \Lambda, \quad 0 \in \Lambda, \\ C^* &= \arg \min_{C \in \{\widehat{W}(\lambda):\lambda\in\Lambda\} \cup \{W_q\}} \widehat{\mathrm{KL}}_{\text{sel}}(p_{\text{FP}} \| p_C^Q). \end{aligned} \tag{18}$$

Here $\widehat{W}(\lambda)$ is the complete corrected model and p_C^Q its distribution under the fixed activation quantizer. The candidates include Pre ($\lambda = 0$) and the Quant-only fallback W_q. Actual full-model selection KL is therefore no higher than Pre's, up to numerical accuracy; this ordering does not extend to independent-test NLL. For a corrected candidate, write

$$G_l^{\text{final}} = \alpha_{b,g(l)}(\lambda^*)G_l, \qquad G_l = G_{l,r_l}, \tag{19}$$

for every $l \in \mathcal{L}_b$. If Quant-only is selected, disable the branches. Otherwise, write-back preserves Pre's allocated ranks, factor shapes, payload, and nominal branch multiply–accumulate (MAC) counts. A zero amplitude can lower algebraic rank without changing allocated shape.

4 Experiments

4.1 Experimental Setup

Models and quantization. We evaluate Llama-2-7B/13B (Touvron et al., 2023), Llama-3-8B (Grattafiori et al., 2024), Llama-3.2-1B/3B (Meta AI, 2024), Qwen2.5-3B (Yang et al., 2024), and Qwen3-4B-Base (Yang et al., 2025) at W4A8, W4A4, and W3A4. GPTQ uses symmetric group-128 weights, no zero point, damping 0.01, block size 128, and activation ordering. RTN uses symmetric group-128 weights. Both use dynamic asymmetric strict group-128 activation fake quantization. Internal pairs share the model, tokenizer, base weights, quantizers, calibration stream, and evaluator.

Calibration and budget. Statistics, rank allocation, and Block fitting use 128 WikiText2 training sequences of length 2048 (Merity et al., 2017). Full-model selection uses 32 additional windows with distinct starting positions; token overlap is not excluded. Test data do not select ranks, amplitudes, or λ. We use $\mathcal{R} = \{0, 16, \ldots, 128\}$, the uniform-rank-64 FP32 factor budget, $\mathcal{A} = \{0, 0.5, 0.75, 1, 1.25\}^4$, $\Lambda = \{0, 0.25, 0.5, 0.75, 1\}$, and $\rho = 0.05$. Tables report frozen canonical runs. Representative configurations repeated with three calibration seeds retained their qualitative ordering (Appendix B).

Metrics and references. We report WikiText2 PPL, $\mathrm{NLL} = \log(\mathrm{PPL})$, and paired relative changes. Avg_8 is the unweighted zero-shot mean over ARC-C/E, BoolQ, HellaSwag, OpenBookQA, PIQA, Social-IQA, and WinoGrande; MMLU is question-weighted over 57 subjects (Hendrycks et al., 2021). W4 LLM.int4(), $\mathrm{L}^2\mathrm{QER}$, and SERQ values are published references from SERQ Table 1 (Zhang et al., 2024; Park et al., 2026); their factor precision, budgets, and protocols differ from ours. W3A4 $\mathrm{L}^2\mathrm{QER}$/SERQ values are separate runs from our evaluation. Appendix C.5 records provenance and precision.

4.2 Overall Quantization Performance

Complete PATR lowers PPL in all 42 matched configurations: 21 GPTQ and 21 RTN cases (Tables 1 and 2). Median paired reductions are 2.33% for GPTQ and 12.41% for RTN, and 2.56%, 3.84%, and 32.21% for W4A8, W4A4, and W3A4. For Qwen2.5-3B RTN W3A4, PPL falls from 32,925.4 to 14.557 (NLL 10.402 to 2.678). Full-precision and complete paired results for all seven models appear in Tables 6 and 7. At W3A4, PATR(GPTQ) has the lowest PPL among completed entries on all seven models. The PATR runs also avoid the severe Qwen degradation in the additional baselines. Qwen3 improves over its own Quant-only backbone at all three precisions; it is absent from the published SERQ W4 comparison. On ten shared W4 model–precision pairs, both PATR backends are below the reported LLM.int4() and $\mathrm{L}^2\mathrm{QER}$ PPL. Against backend-labeled SERQ, PATR is lower in 16/20 settings (RTN 9/10, GPTQ 7/10), with median relative differences of -1.95% and -0.79%. These are cross-protocol accuracy comparisons; Table 9 also reports their costs.

Bits	Method	Llama-2-7B	Llama-3-8B	Llama-3.2-1B	Llama-3.2-3B	Qwen2.5-3B
W4A8	FP16	5.47	6.13	9.75	7.81	8.03
	LLM.int4()	7.28	10.82	32.39	12.47	9.98×10^5
	L^2QER	5.83	7.16	12.06	8.78	12.25
	Quant(RTN)	5.78	7.03	12.70	8.81	11.47
	SERQ(RTN)	5.64	6.71	11.14	8.38	8.67
	PATR(RTN)	**5.62**	**6.59**	**10.59**	**8.22**	**8.38**
	Quant(GPTQ)	5.63	6.57	10.55	8.90	8.39
	SERQ(GPTQ)	**5.59**	6.52	10.45	8.18	8.36
	PATR(GPTQ)	5.60	**6.50**	**10.30**	**8.11**	**8.27**
W4A4	FP16	5.47	6.13	9.75	7.81	8.03
	LLM.int4()	6.32×10^2	4.87×10^2	1.7×10^3	4.37×10^2	8.51×10^5
	L^2QER	7.37	11.44	30.83	14.11	221.5
	Quant(RTN)	6.13	7.97	16.61	10.09	12.88
	SERQ(RTN)	6.03	8.07	13.57	9.43	**9.66**
	PATR(RTN)	**5.95**	**7.43**	**13.10**	**9.23**	10.35
	Quant(GPTQ)	5.94	7.61	13.03	9.59	10.42
	SERQ(GPTQ)	5.97	7.75	**12.52**	9.15	**9.35**
	PATR(GPTQ)	**5.91**	**7.29**	12.57	**9.08**	10.18

Table 1: WikiText2 test PPL (↓) under W4A8 and W4A4. Within each backend, bold and underlining mark the lowest and second-lowest results among Quant, SERQ, and PATR.

Method	Llama-2-7B	Llama-2-13B	Llama-3-8B	Llama-3.2-1B	Llama-3.2-3B	Qwen2.5-3B	Qwen3-4B
L^2QER	7.83	6.66	14.59	93.01	19.95	125.47	3.00×10^5
Quant(RTN)	9.70	6.64	43.36	538.59	47.04	3.29×10^4	305.81
PATR(RTN)	7.17	5.98	10.79	27.97	13.70	14.55	13.06
Quant(GPTQ)	7.111	5.72	14.28	26.15	18.00	13.96	12.17
SERQ(GPTQ)	6.93	5.86	11.74	22.73	13.20	3.54×10^4	1.30×10^6
PATR(GPTQ)	**6.80**	**5.63**	**9.21**	**18.58**	**11.55**	**12.26**	**11.22**

Table 2: WikiText2 test PPL (↓) under W3A4. L^2QER and SERQ are separately completed runs; Quant/PATR use the parenthesized backend. Bold marks the lowest value.

4.3 Task-Sensitive Reconstruction and Rank Allocation

Stage I selects directions using input moments and the diagonal output-gradient metric, then assigns their ranks. Its pre-coordination output improves over Quant-only in all 42 configurations (Table 4). This evaluates the combined reconstruction-and-allocation procedure.

Table 3: Llama-3.2-1B/3B rank ablation. Only the rank map changes under a uniform-rank-64 FP32 budget. Reductions are means of paired relative PPL reductions.

Bits	Cells	Adaptive wins	Mean PPL red. (%)
W4A8	4	4/4	0.184
W4A4	4	4/4	0.479
W3A4	4	4/4	3.034
All	12	12/12	1.233

The controlled allocator comparison fixes the base quantizer, top-128 basis, data, evaluator, and FP32 factor budget, changing only the rank map. Adaptive allocation wins all 12 pairs (Table 3). Mean relative PPL reductions are 0.184%, 0.479%, and 3.034% at W4A8, W4A4, and W3A4, respectively, and 1.233% overall. The largest mean benefit is at W3A4. Tables 11 and 12 contain the full values from these separate allocator runs.

4.4 Quantized-State Response Mismatch and Coordination

Figure 2 evaluates the additive Block-response surrogate at unit amplitude against no correction, $J_{\text{unit}} > J_{\text{ref}}$. The reported micro-pooled fractions are 1.01%, 92.57%, and 79.39% at W4A8, W4A4, and W3A4. Both backends show the W4A8-to-W4A4 transition. This diagnostic concerns the sum of measured single-group responses: unit scale in the Stage-I geometry does not fix the preferred scale in the quantized-state response model.

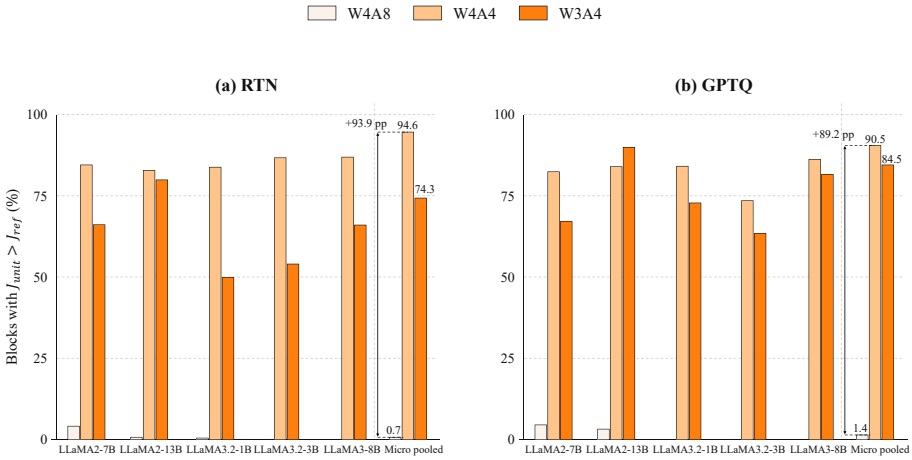


Figure 2: Fraction of Blocks with $J_{\text{unit}} > J_{\text{ref}}$ in the additive response surrogate. Micro denotes the source diagnostic's reported aggregates across five models; pp denotes percentage points.

Coordination changes amplitudes without increasing the allocated capacity. Full-model selection uses a nonzero update in 20 configurations and retains Pre in 22. Among the updates, independent-test PPL decreases in 9 and increases in 11; the reported relative changes span -0.2909% to $+0.2023\%$. Complete PATR remains below Quant-only in all 42 cases. Table 4 gives the stagewise medians and counts; Tables 6–8 retain every result

Table 4: Stagewise PPL summary across seven models and two backends. Q, Pre, and Full are Quant-only, pre-coordination, and complete PATR. Red. is the median of paired relative reductions. I/R/H counts lower/retained/higher independent-test PPL relative to Pre.

Bits	Cells	$Q \to$ Pre Red. (%)	Pre $\to$ Full I/R/H	$Q \to$ Full Red. (%)
W4A8	14	2.55	3/9/2	2.56
W4A4	14	3.84	4/4/6	3.84
W3A4	14	32.30	2/9/3	32.21
All	42	6.73	9/22/11	6.73

and selected step. These test-PPL outcomes are evaluated separately from the calibration-selection KL used to choose deployment.

4.5 Downstream Performance and Cost

PATR(RTN) exceeds the SERQ reference on all four W4A8 aggregates in Table 5. At W4A4, SERQ is higher on three, while PATR is higher on Llama-3.2-3B MMLU. These absolute accuracies complement the PPL results; the table does not contain a matched Quant-only downstream ablation. Tables 15 and 14 include W3A4 and the remaining completed evaluations.

Table 5: Selected Llama-3.2 downstream accuracies (%, ↑), copied from Table 15. Bold compares the two quantized methods within a precision; FP16 supplies context. The methods use different protocols and factor precisions.

		Llama-3.2-1B		Llama-3.2-3B	
Bits	Method	Avg_8	MMLU	Avg_8	MMLU
FP16	Baseline	54.82	36.76	62.66	54.06
W4A8	SERQ(RTN)	52.76	32.77	61.27	51.45
	PATR(RTN)	**53.26**	**33.11**	**61.46**	**52.56**
W4A4	SERQ(RTN)	**50.15**	**29.11**	**58.40**	46.36
	PATR(RTN)	49.40	27.83	57.71	**46.87**

FP32 factors add 0.63–1.48 bits per target base weight (Table 13). This nominal accounting excludes scales, packing metadata, non-target parameters, and workspace. Stage-II write-back preserves Pre's factor shapes, payload, and nominal branch MACs. The low-rank branch itself still requires storage and computation; these counts are not measured packed-kernel latency or peak memory.

5 Conclusion

PATR allocates low-rank correction capacity by task value and coordinates the selected corrections through measured quantized-state responses. The Stage-I spectral gains determine rank allocation and constrain Stage-II amplitude changes, while full-model KL selects deployment. Across the evaluated W–A settings, complete PATR lowers PPL in all 42 matched configurations, and adaptive allocation improves all 12 equal-budget pairs. Response coordination retains the allocated factor capacity.

AI use statement

Generative AI assisted manuscript restructuring and drafting, language editing, literature lookup, mathematical exposition, interpretation of the reported results, and preparation of LaTeX and conceptual figures. The revision used existing experimental records; no new experimental measurements were generated by AI. The authors are responsible for the final text, claims, citations, and artifacts.

Reproducibility statement

Algorithm 1 specifies the workflow. Section 4 and Appendix B document models, quantizers, calibration and selection windows, candidate grids, and metrics. Appendix A gives the reconstruction proof, numerical conditions, response construction, and cost accounting. The remaining appendix tables retain the complete stagewise and allocator results.

References

Ali Abbasi, Chayne Thrash, Haoran Qin, Hamed Pirsiavash, and Soheil Kolouri. IO-SVD: Input-output whitened SVD for adaptive-rank LLM compression, 2026. URL https://arxiv.org/abs/2605.15626.

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=a3l3K9khbL.

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 *NeurIPS*, volume 37, pp. 100213–100240, 2024.

Yoonjun Cho, Dongjae Jeon, Soeun Kim, Moongyu Jeon, and Albert No. Preserve-then-quantize: Balancing rank budgets for quantization error reconstruction in LLMs. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02001.

Xin Ding, Xiaoyu Liu, Zhijun Tu, Yun Zhang, Wei Li, Jie Hu, Hanting Chen, Yehui Tang, Zhiwei Xiong, Baoqun Yin, and Yunhe Wang. CBQ: Cross-block quantization for large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eW4yh6HKz4.

E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *ICLR*, 2023.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783.

Hongyaoxing Gu, Lijuan Hu, Shuzi Niu, and Fangfang Liu. FLRQ: Faster LLM quantization with flexible low-rank matrix sketching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21369–21377, 2026. doi: 10.1609/aaai.v40i26.39283. URL https://ojs.aaai.org/index.php/AAAI/article/view/39283.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical Fisher approximation for natural gradient descent, 2019. URL https://arxiv.org/abs/1905.12558.

Y. Li, R. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. Yu, W. Wang, and S. Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *ICLR*, 2021.

Yuhang Li, Donghyun Lee, Ruokai Yin, and Priyadarshini Panda. Optimal brain decomposition for accurate LLM low-rank approximation, 2026. URL https://arxiv.org/abs/2604.00821.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *NeurIPS*, volume 37, 2024a.

J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In *MLSys*, 2024b.

Zhuowen Liu, Longkun Hao, Shiyu Feng, Xiaowen Chang, Ruiqun Li, and Changqun Li. LACE-SVD: Loss-aware SVD with cumulative error correction for LLM compression, 2026. URL https://arxiv.org/abs/2607.03057v2.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. AffineQuant: Affine transformation quantization for large language models. In *ICLR*, 2024. URL https://openreview.net/forum?id=of2rhALq8l.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *ICLR*, 2017.

Meta AI. Llama 3.2 model card. https://huggingface.co/meta-llama/Llama-3.2-1B, 2024. Released September 25, 2024.

Yeonsik Park, Hyeonseong Kim, and Seungkyu Choi. SERQ: Saliency-aware low-rank error reconstruction for LLM quantization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nFjj8NEBqv.

R. Saha, N. Sagan, V. Srivastava, A. J. Goldsmith, and M. Pilanci. Compressing large language models using low rank and low precision decomposition. In *NeurIPS*, 2024.

W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, K. Zhang, P. Gao, Y. Qiao, and P. Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *ICLR*, 2024.

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

Albert Tseng, Zhaofeng Sun, and Christopher De Sa. Model-preserving adaptive rounding. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2505.22988v3.

YiFeng Wang, Zhun Sun, and Keisuke Sakaguchi. Technical report: Activation residual hessian quantization (ARHQ) for low-bit LLM quantization, 2026. URL https://arxiv.org/abs/2605.00140.

G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *ICML*, 2023.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388.

Cheng Zhang, Jianyi Cheng, George Anthony Constantinides, and Yiren Zhao. LQER: Low-rank quantization error reconstruction for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58763–58779, 2024. URL https://proceedings.mlr.press/v235/zhang24j.html.

Cheng Zhang, Jeffrey T. H. Wong, Can Xiao, George Constantinides, and Yiren Zhao. QERA: an analytical framework for quantization error reconstruction. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 12531–12560, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/21718991f6acf19a42376b5c7a8668c5-Paper-Conference.pdf.

A Additional Method Details

A.1 Notation and Proof Sketch under the Stated Assumptions

For target projection l, let $E_l = W_l - W_{q,l}$ be its weight-quantization residual. Let D_l be the positive diagonal output-gradient factor, and let $(U_{A,l}, \Lambda_{A,l})$ contain the retained eigenpairs of the active input subspace. Write $U_A = U_{A,l}$ and $\Pi_A = U_A U_A^\top$. The candidate is explicitly restricted to $L = L\Pi_A$; hence $L = (LU_A)U_A^\top$ and $\mathrm{rank}(L) = \mathrm{rank}(LU_A)$. Since D_l and $\Lambda_{A,l}^{1/2}$ are invertible, the mapping

$$L \longmapsto D_l L U_{A,l} \Lambda_{A,l}^{1/2} \tag{20}$$

preserves rank when $D_l \succ 0$ on the modeled output space and $\Lambda_{A,l}$ is positive definite. The part of the local objective that depends on L is therefore equivalent to approximating

$$M_l = D_l E_l U_{A,l} \Lambda_{A,l}^{1/2} \tag{21}$$

with a matrix of rank at most r. In particular, for the discarded eigenpairs $(U_{\perp,l}, \Lambda_{\perp,l})$, orthogonality gives

$$\|D_l(E_l - L)H_{x,l}^{1/2}\|_F^2 = \|M_l - D_l L U_{A,l}\Lambda_{A,l}^{1/2}\|_F^2 + C_{\perp,l},$$

where $C_{\perp,l} = \|D_l E_l U_{\perp,l}\Lambda_{\perp,l}^{1/2}\|_F^2$ is independent of L and may be nonzero when small positive eigenvalues are discarded. Every active-space matrix Z has the inverse $L = D_l^{-1} Z \Lambda_{A,l}^{-1/2} U_{A,l}^\top$, satisfying $L = L\Pi_A$. Thus the mapping is rank preserving on this constrained domain, not for arbitrary L. Define

$$k_l(r) = \min\{r, \mathrm{rank}(M_l)\}.$$

The Eckart–Young–Mirsky theorem gives an optimal approximation of rank at most r: the truncated singular value decomposition (SVD) of M_l using its first $k_l(r)$ components. Inverting the mapping in Eq. (20) yields an optimal solution in the main paper. Orthogonality of the retained exact singular components and the spectral tail gives

$$\mathcal{G}_l^\star(r) = \sum_{k=1}^{k_l(r)} \sigma_{l,k}^2, \tag{22}$$

where $\sigma_{l,k}$ is the kth exact singular value. This argument is a proof sketch for the fixed-active-subspace, exact-SVD case in which D_l is strictly positive on the modeled output space and $\Lambda_{A,l}$ is positive definite. These conditions specify the domain of Proposition 1.

The implementation separately uses randomized SVD and a numerical-validity rule. Let $\widehat{r}_{\mathrm{eff},l}(r)$ be the number of retained computed components and let

$$\widehat{\mathcal{G}}_l(r) = \sum_{k=1}^{\widehat{r}_{\mathrm{eff},l}(r)} \widehat{\sigma}_{l,k}^2.$$

These hatted quantities define the numerical proxy curves used for allocation and amplitude constraints. Proposition 1 describes the corresponding exact-SVD problem.

A.2 Balanced Deployment Factors

Let $k = k_l(r)$ and take the first k exact singular components of M_l. (The implementation instead uses the valid computed component count.) A balanced factorization is

$$\begin{aligned} F_{l,r} &= U_{A,l}\Lambda_{A,l}^{-1/2}\widetilde{V}_{l,k}\Sigma_{l,k}^{1/2}, \\ G_{l,r} &= \Sigma_{l,k}^{1/2}\widetilde{U}_{l,k}^\top D_l^{-1}, \end{aligned} \tag{23}$$

Here $F_{l,r} \in \mathbb{R}^{d_{i,l}\times k}$ and $G_{l,r} \in \mathbb{R}^{k\times d_{o,l}}$, so $L_{l,r}^\top = F_{l,r}G_{l,r}$. The factor $G_{l,r}$ is distinct from the output metric $G_{y,l} \in \mathbb{R}^{d_{o,l}\times d_{o,l}}$. At runtime,

$$X_{q,l}\widehat{W}_l^\top = X_{q,l}W_{q,l}^\top + (X_{q,l}F_{l,r})G_{l,r}, \tag{24}$$

where $X_{q,l}$ is the quantized layer input. Both branches receive the same $X_{q,l}$.

A.3 Boundary Target Decomposition

Adding and subtracting the FP Block evaluated at the quantized-prefix state gives the exact decomposition

$$\begin{aligned} u_b &= \mathcal{F}_b(h^*_{b-1}) - \mathcal{F}_b(h^0_{b-1}), \\ v_b &= \mathcal{F}_b(h^0_{b-1}) - \widehat{\mathcal{F}}^0_b(h^0_{b-1}), \qquad t_b = u_b + v_b. \end{aligned} \tag{25}$$

The first term propagates the incoming state discrepancy through the FP Block. The second measures the current Block's quantization discrepancy at the same input. The implemented fit uses their sum t_b; these terms are an algebraic explanation, not separately measured curves.

A.4 Response Regression Coefficients

With $t_b, \Delta_{b,g} \in \mathbb{R}^{N_b \times d_b}$ and the channel metric $\Omega_b \in \mathbb{R}^{d_b \times d_b}$, the additive response fit expands as

$$\begin{aligned} c_b &= N_b^{-1} \|t_b \Omega_b^{1/2}\|_F^2, \\ h_{b,g} &= N_b^{-1} \langle \Delta_{b,g} \Omega_b^{1/2}, t_b \Omega_b^{1/2} \rangle_F, \\ R_{b,gg'} &= N_b^{-1} \langle \Delta_{b,g} \Omega_b^{1/2}, \Delta_{b,g'} \Omega_b^{1/2} \rangle_F. \end{aligned} \tag{26}$$

Thus c_b is scalar, $\boldsymbol{h}_b \in \mathbb{R}^4$, and $\boldsymbol{R}_b \in \mathbb{R}^{4\times 4}$ is a positive-semidefinite Gram. The products act on columns because rows index valid tokens.

A.5 Linearized Cross-Block Motivation

Consider a network with B Transformer Blocks. For each $b \in \{1, \dots, B\}$, let $\mathcal{F}_b$ and $\widehat{\mathcal{F}}_b$ denote the full-precision and corrected quantized mappings of Block b. Let h^*_b and $\widehat{h}_b$ be their states at the output boundary, vectorized consistently over tokens and channels. Define $e_b = h^*_b - \widehat{h}_b$ with $e_0 = 0$, and evaluate along one fixed realized quantized trajectory. Set

$$A_b = \left.\frac{\partial \mathcal{F}_b(h)}{\partial h}\right|_{h=\widehat{h}_{b-1}}, \qquad \varepsilon_b = \mathcal{F}_b(\widehat{h}_{b-1}) - \widehat{\mathcal{F}}_b(\widehat{h}_{b-1}).$$

Define the remainder exactly by $\xi_b = \mathcal{F}_b(h^*_{b-1}) - \mathcal{F}_b(\widehat{h}_{b-1}) - A_b e_{b-1}$. Then

$$\begin{aligned} e_b &= A_b e_{b-1} + \varepsilon_b + \xi_b, \\ e_B &= \sum_{b=1}^{B} P_b \varepsilon_b + \sum_{b=1}^{B} P_b \xi_b, \\ P_b &= A_B A_{B-1} \cdots A_{b+1}, \qquad P_B = I. \end{aligned} \tag{27}$$

This definition assigns quantization, including state switching, to ε_b; ξ_b is the nonlinear remainder of the smooth full-precision mapping. Under local smoothness it is second order in the state error, but aggressive quantization need not make it small. Reusing responses at a different trajectory introduces additional state-switching and modeling errors. Only after neglecting the accumulated remainder $\sum_b P_b \xi_b$ do we obtain $e_B \approx \sum_b P_b \varepsilon_b$. With terminal metric $H_{\text{out}} \succeq 0$ and $\|v\|^2_{H_{\text{out}}} = v^\top H_{\text{out}} v$,

$$\begin{aligned} \|e_B\|^2_{H_{\text{out}}} \approx \sum_b \varepsilon_b^\top P_b^\top H_{\text{out}} P_b \varepsilon_b \\ + 2 \sum_{b<c} \varepsilon_b^\top P_b^\top H_{\text{out}} P_c \varepsilon_c. \end{aligned} \tag{28}$$

The off-diagonal terms explain how local discrepancies can interact after propagation; they do not require error norms to increase at every Block. PATR uses measured, fixed-prefix single-group responses within each Block and full-model KL interpolation for joint deployment. This implementation avoids forming the propagation operators P_b or a cross-Block Gram.

A.6 Scaling Identity and Numerical Trust Proxy

Write $Z = D_l L_{l,r} U_{A,l} \Lambda_{A,l}^{1/2}$. For arbitrary Z,

$$J_l(\alpha) - J_l(1) = (1-\alpha)^2 \|Z\|_F^2 + 2(1-\alpha)\langle M_l - Z, Z\rangle_F.$$

Exact truncated SVD makes the cross term zero and $\|Z\|_F^2 = \mathcal{G}_l(r)$, which proves the scaling identity used in the main text. Randomized decomposition alone does not establish this orthogonality for every numerical implementation. The hatted trust constraint is therefore defined on spectral proxy curves, not claimed as an exact bound on the realized local residual. Its excess over unit amplitude is $\sum_l (1-\alpha_{b,g(l)})^2 \widehat{\mathcal{G}}_l(r_l)$. With nonnegative gains and a nonnegative unit spectral-residual proxy, this defines a convex ellipsoidal sublevel set (possibly degenerate or unbounded) centered at $\mathbf{1}$. The unit candidate is feasible; for $0 \le \lambda \le 1$, interpolation toward a feasible proposal scales its excess by λ^2. This preserves the proxy constraint, not actual Block loss or independent-test PPL. The separate Quant-only deployment fallback need not lie in that trust region. Likewise, selecting from a set containing Pre guarantees only non-increasing measured selection KL (up to numerical accuracy), not independent-test NLL.

A.7 Approximation Hierarchy and Cost

PATR separates task value from quantized-state response. Stage I uses input moments and observed-token NLL output-gradient statistics collected on the full-precision trajectory. Its active-subspace objective reconstructs the fixed weight residual. For W–A quantization, the corrected projection also has an activation-error contribution:

$$Y_l^{\mathrm{FP}} - Y_l^{\mathrm{corr}} = (X_l - X_{q,l})W_l^\top + X_{q,l}(E_l - L_l)^\top,$$

where $Y_l^{\mathrm{FP}} = X_l W_l^\top$ and $Y_l^{\mathrm{corr}} = X_{q,l}(W_{q,l} + L_l)^\top$. The Stage-I geometry values residual directions through downstream NLL sensitivity on the reference trajectory. Its empirical gradient moments are not the output-space Hessian or Fisher of the forward-KL objective.

Stage II holds the selected directions and ranks fixed. It fits four unit-amplitude single-group finite differences to the boundary discrepancy in a suffix-sensitive metric. The boundary target includes both incoming state discrepancy and current-Block quantization error, as decomposed in Eq. (25). The resulting additive model retains within-Block response overlap. Its stationary equation $\boldsymbol{R}_b \boldsymbol{\alpha}_b = \boldsymbol{h}_b$ describes the unconstrained continuous quadratic model; the actual solver searches the fixed amplitude grid with the numerical spectral constraint. Stage-I gains therefore inform both rank allocation and admissible amplitude changes.

All Stage-II statistics are collected once under correction-free quantized prefixes. The selected upstream corrections do not alter these collection trajectories. Actual joint deployment is evaluated by the full-model KL search over the shared interpolation step. This distinction separates the measured single-group response model from the multi-group corrected network.

Stage I eigendecomposes the input second moments and computes one top-$r_{\max}$ SVD per layer, where $r_{\max}$ is the maximum candidate rank. All rank candidates share that decomposition. With an exact SVD, the active-space result is exact under its stated conditions; randomized SVD and numerical filtering produce the hatted proxy curves defined above. Pareto dynamic programming exactly solves the allocation problem over those computed discrete proxy curves. Stage II uses one suffix-sensitive gradient-statistics collection, four single-group replays per Block, and a small number of full-model step evaluations. Relative to the pre-coordination model, it introduces no inference operator, parameter, factor bit, or nominal low-rank multiply–accumulate.

A.8 COMPLETE WORKFLOW

Input: Full-precision model; fixed quantizers; fitting and selection windows $\mathcal{D}_{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{sel}}$; rank, amplitude, and interpolation grids $\mathcal{R}$, $\mathcal{A}$, and Λ; factor-bit budget $\mathcal{B}$ and trust tolerance ρ.
Output: Selected deployment model C^*.

Quantize and freeze the base weights $\{W_{q,l}\}_{l\in\mathcal{L}}$.
for each layer $l \in \mathcal{L}$ **do**
Collect FP-trajectory $H_{x,l}$ and Γ_l; decompose M_l once.
Tabulate $\widehat{J}_l$, $\widehat{\mathcal{G}}_l$, and c_l^{cap} over $\mathcal{R}$.
end for
Solve Eq. (13); materialize $\{F_l, G_l\}$.
Freeze the correction directions and ranks; use correction-free quantized prefixes.
for each Transformer Block $b = 1, \ldots, B$ **do**
Collect Ω_b and t_b; obtain $\{\Delta_{b,g}\}_{g=1}^{4}$ from four fixed-prefix single-group replays.
Grid-search $\boldsymbol{\alpha}_b^*$ by Eq. (16) subject to Eq. (17).
end for
Select C^* from $\{\widehat{W}(\lambda) : \lambda \in \Lambda\} \cup \{W_q\}$ on $\mathcal{D}_{\mathrm{sel}}$ by Eq. (18).
if $C^* = W_q$ **then**
Disable all low-rank branches.
else
Absorb the selected amplitudes into $\{G_l\}$ by Eq. (19).
end if
return C^*.

Algorithm 1: PATR

B EXPERIMENTAL PROTOCOL

B.1 MODELS, QUANTIZERS, AND SPLITS

The target-precision evaluation covers Llama-2-7B/13B, Llama-3-8B, Llama-3.2-1B/3B, Qwen2.5-3B, and Qwen3-4B-Base. GPTQ is complete for all seven models, and RTN is complete for the same seven models. W4A8, W4A4, and W3A4 are evaluated separately. GPTQ uses symmetric group-128 weights, no zero point, damping 0.01, block size 128, and activation ordering. RTN uses symmetric group-128 weights without a zero point. Activations use dynamic asymmetric strict group-128 fake quantization.

All paired internal comparisons preserve the same base W_q, activation quantizer, model revision, tokenizer, calibration stream, and test evaluator. Statistics and ranks use 128 deterministic WikiText2 training sequences of length 2048. Block responses use the same fitting set. Full-model selection uses 32 additional windows with starting positions distinct from the fitting set. The WikiText2 test set is excluded from all rank, amplitude, and step-size decisions. Representative configurations were repeated with three calibration seeds and retained their qualitative ordering; the tables below report the frozen canonical run.

B.2 STAGES, CANDIDATE SETS, AND METRICS

The pre-coordination model consists of the two-sided task-sensitive reconstruction and global adaptive-rank allocation. PATR denotes the selected complete model after propagation-aware Block-response coordination. The fixed candidate sets are

$$\begin{aligned} \mathcal{R} &= \{0, 16, 32, 48, 64, 80, 96, 112, 128\}, \\ \mathcal{A} &= \{0, 0.5, 0.75, 1, 1.25\}^4, \\ \Lambda &= \{0, 0.25, 0.5, 0.75, 1\}. \end{aligned} \tag{29}$$

Here $r_{\max} = 128$, factor capacity matches uniform-rank-64 FP32 factors, and the local trust ratio is $\rho = 0.05$.

For a standard causal-language-model evaluation,

$$\mathrm{PPL} = \exp(\mathrm{NLL}), \qquad \mathrm{NLL} = \log(\mathrm{PPL}), \tag{30}$$

where NLL is averaged over prediction tokens. For candidate M, NLL-gap recovery is

$$R_{\mathrm{NLL}}(M) = \frac{\mathrm{NLL}_Q - \mathrm{NLL}_M}{\mathrm{NLL}_Q - \mathrm{NLL}_{\mathrm{FP}}}, \tag{31}$$

when the denominator is positive. Because raw PPL can be dominated by severe low-bit failures, aggregate statements use paired relative changes, win counts, and NLL-gap recovery rather than a cross-model arithmetic mean of PPL.

C Additional Results

C.1 Main-Text Diagnostics and Summaries

The unit-amplitude diagnostic and W3A4 stress test are presented in Figure 2 and Table 2. Tables 3 and 4 summarize allocation and coordination; the complete numerical records follow.

C.2 Complete GPTQ Target-Precision Results

Model	Bits	FP	Quant-only	Pre-coordination	λ	PATR	Δ_{BC} (%)
Llama-2-7B	W4A8	5.4751	5.6287	5.6048	0	5.6048	0
	W4A4	5.4751	5.9402	5.9235	.50	**5.9125**	−0.1844
	W3A4	5.4751	7.1109	6.8254	.25	**6.8055**	−0.2909
Llama-2-13B	W4A8	4.8917	4.9822	4.9727	.25	**4.9723**	−0.0076
	W4A4	4.8917	5.2194	5.2034	.75	**5.2030**	−0.0082
	W3A4	4.8917	5.7214	5.6371	.25	**5.6368**	−0.0039
Llama-3-8B	W4A8	6.1684	6.5764	6.5070	0	6.5070	0
	W4A4	6.1684	7.6125	7.2967	0	7.2967	0
	W3A4	6.1684	14.2850	**9.2067**	.25	9.2175	+0.1175
Llama-3.2-1B	W4A8	9.7640	10.5494	10.3043	.25	**10.3033**	−0.0094
	W4A4	9.7640	13.0358	12.5760	.25	**12.5756**	−0.0029
	W3A4	9.7640	26.1516	**18.5546**	.25	18.5839	+0.1578
Llama-3.2-3B	W4A8	7.8214	8.9036	8.1107	0	8.1107	0
	W4A4	7.8214	9.5896	**9.0704**	.25	9.0826	+0.1347
	W3A4	7.8214	18.0062	11.5514	0	11.5514	0
Qwen2.5-3B	W4A8	8.0307	8.3948	**8.2684**	.25	8.2708	+0.0291
	W4A4	8.0307	10.4258	**10.1815**	.50	10.1879	+0.0633
	W3A4	8.0307	13.9602	12.2669	0	12.2669	0
Qwen3-4B-Base	W4A8	7.8997	8.2279	8.1264	0	8.1264	0
	W4A4	7.8997	9.6521	**9.5123**	.25	9.5155	+0.0329
	W3A4	7.8997	12.1767	11.2233	0	11.2233	0

Table 6: Complete GPTQ WikiText2 PPL. The pre-coordination and PATR columns share W_q, correction directions, allocated ranks, and factor bits. $\Delta_{\mathrm{BC}} = 100(\mathrm{PPL}_{\mathrm{PATR}}/\mathrm{PPL}_{\mathrm{pre}} - 1)$; negative is better. Bold distinguishes the lower of the two low-rank stages when they differ.

C.3 Complete RTN Target-Precision Results

Model	Bits	FP	Quant-only	Pre-coordination	λ	PATR	Δ_{BC} (%)
Llama-2-7B	W4A8	5.4751	5.7875	5.6265	0	5.6265	0
	W4A4	5.4751	6.1319	5.9541	0	5.9541	0
	W3A4	5.4751	9.7030	7.1741	0	7.1741	0
Llama-2-13B	W4A8	4.8917	5.0511	4.9934	0	4.9934	0
	W4A4	4.8917	5.3049	5.2349	.50	**5.2301**	-0.0920
	W3A4	4.8917	6.6430	**5.9816**	.25	5.9854	$+0.0635$
Llama-3-8B	W4A8	6.1684	7.0341	**6.5956**	.25	6.5958	$+0.0034$
	W4A4	6.1684	7.9738	7.4341	0	7.4341	0
	W3A4	6.1684	43.3609	10.7931	0	10.7931	0
Llama-3.2-1B	W4A8	9.7631	12.7006	10.5971	0	10.5971	0
	W4A4	9.7631	16.6173	**13.0724**	.75	13.0989	$+0.2023$
	W3A4	9.7631	538.5922	27.9702	0	27.9702	0
Llama-3.2-3B	W4A8	7.8216	8.8148	8.2252	0	8.2252	0
	W4A4	7.8216	10.0907	**9.2269**	.25	9.2366	$+0.1055$
	W3A4	7.8216	47.0395	13.7021	0	13.7021	0
Qwen2.5-3B	W4A8	8.0307	11.4739	8.3859	0	8.3859	0
	W4A4	8.0307	12.8860	10.3576	0	10.3576	0
	W3A4	8.0307	32925.4065	14.5570	0	14.5570	0
Qwen3-4B-Base	W4A8	7.8997	9.0734	8.2238	.50	**8.2208**	-0.0374
	W4A4	7.8997	11.0585	**9.6707**	.25	9.6866	$+0.1640$
	W3A4	7.8997	305.8093	13.0640	0	13.0640	0

Table 7: Complete RTN WikiText2 PPL under the same notation as Table 6.

C.4 Coordination Selection Behavior

Backend	Cells	Improve	Retain	Higher	Nonzero	Nonzero win
GPTQ	21	7	8	6	13	7/13
RTN	21	2	14	5	7	2/7
All	42	9	22	11	20	9/20

Table 8: Independent-test outcome of PATR relative to pre-coordination. "Retain" means selection of the pre-coordination model; "Higher" means a selected nonzero candidate has higher test PPL.

GPTQ selects $\lambda = 0, .25, .50, .75$ in 8, 10, 2, and 1 configurations; RTN selects them in 14, 4, 2, and 1. No configuration selects $\lambda = 1$. Thus all nonzero deployments interpolate between unit amplitudes and the Block proposals. The reported Δ_{BC} values range from -0.2909% to $+0.2023\%$. Selection uses KL on the additional calibration windows, while the tables report independent-test PPL. Among GPTQ Llama-2 runs, five of six 7B/13B settings have lower PPL and one retains Pre.

C.5 Cross-Protocol External References

PATR is numerically lower in 9/10 RTN and 7/10 GPTQ configurations. The median relative differences are -1.95% and -0.79%, respectively (-1.16% over all 20). These are comparisons under the named protocols, with the factor costs and precision differences reported alongside them. Qwen3 is intentionally absent because SERQ Table 1 provides no Qwen3 result.

The main paper also includes the LLM.int4() and $\mathrm{L}^2\mathrm{QER}$ values from SERQ Table 1; neither method is labeled by backend. We independently reproduced the corresponding public-code configurations to the reported precision. LLM.int4() has dynamically varying effective precision; the reported effective weight precisions are 4.35/4.24 bits for $\mathrm{L}^2\mathrm{QER}$ at W4A8/W4A4 and 4.24 bits for SERQ. Our model-dependent FP32 factors correspond to 4.77–5.48 modeled bits per target weight on these

Model	RTN W4A8	RTN W4A4	GPTQ W4A8	GPTQ W4A4	PATR W4 eBPP
Llama-2-7B	5.640/**5.626**	6.030/**5.954**	**5.590**/5.605	5.970/**5.913**	4.7892
Llama-3-8B	6.710/**6.596**	8.070/**7.434**	6.520/**6.507**	7.750/**7.297**	4.7692
Llama-3.2-1B	11.140/**10.597**	13.570/**13.099**	10.450/**10.303**	**12.520**/12.576	5.4828
Llama-3.2-3B	8.380/**8.225**	9.430/**9.237**	8.180/**8.111**	9.150/**9.083**	5.1042
Qwen2.5-3B	8.670/**8.386**	**9.660**/10.358	8.360/**8.271**	**9.350**/10.188	5.3810

Table 9: SERQ/PATR WikiText2 PPL on the 20 nominally overlapping configurations; bold is lower. SERQ values are taken from Table 1 Park et al. (2026) and use about 4.24 effective weight bits. PATR eBPP counts the base W4 weights plus FP32 factors under the manuscript's target-weight denominator. Factor precision, decomposition/GPTQ ordering, exact calibration tokens, and evaluation implementations differ, so this is a reported-reference accuracy comparison rather than a matched-cost head-to-head.

five models. Table 10 separately retains the stronger MXFP4 variants so that they are not mixed into the integer W4A4 comparison.

Model	L^2QER INT	L^2QER MXFP4	SERQ-MXFP4	PATR RTN INT	PATR GPTQ INT
Llama-2-7B	7.37	6.30	6.22	5.954	5.913
Llama-3-8B	11.44	7.83	7.63	7.434	7.297
Llama-3.2-1B	30.83	13.78	13.71	13.099	12.576
Llama-3.2-3B	14.11	9.59	8.74	9.237	9.083
Qwen2.5-3B	221.5	10.46	9.79	10.358	10.188

Table 10: W4A4 WikiText2 PPL across arithmetic-format variants reported in SERQ Table 1 and our integer PATR runs. L^2QER-MXFP4 and SERQ-MXFP4 use MXFP4 with the reported effective weight precision of 4.37 bits and default group size 32; our activation path uses asymmetric group-128 integer quantization and model-dependent FP32 factors. The columns therefore document the omitted stronger MXFP4 references but do not form an identical-format or equal-storage comparison.

C.6 Complete Equal-Budget Rank-Allocation Ablation

Backend	Bits	Uniform $r = 64$	Adaptive	Change
GPTQ	W4A8	10.340350	**10.304302**	-0.3486%
GPTQ	W4A4	12.690972	**12.575960**	-0.9063%
GPTQ	W3A4	19.278538	**18.554644**	-3.7549%
RTN	W4A8	10.624074	**10.598924**	-0.2367%
RTN	W4A4	13.095809	**13.038082**	-0.4408%
RTN	W3A4	29.312141	**27.979878**	-4.5451%

Table 11: Llama-3.2-1B equal-budget rank-allocation PPL. Pairs share W_q, activation quantization, data, evaluator, a top-128 basis, and the uniform-rank-64 FP32 factor budget; only the rank map differs. Change is relative to uniform rank 64.

All 12 model–backend–precision pairs favor adaptive allocation. Mean relative PPL reductions are 0.184%, 0.479%, and 3.034% for W4A8, W4A4, and W3A4, and 1.233% overall. This matched allocator ablation remains separate from the frozen seven-model stagewise results.

Backend	Bits	Uniform $r = 64$	Adaptive	Change
GPTQ	W4A8	8.119438	**8.110723**	-0.1073%
GPTQ	W4A4	9.094173	**9.070395**	-0.2615%
GPTQ	W3A4	11.733588	**11.551356**	-1.5531%
RTN	W4A8	8.229928	**8.226212**	-0.0452%
RTN	W4A4	9.252145	**9.223764**	-0.3068%
RTN	W3A4	14.015481	**13.695261**	-2.2848%

Table 12: Llama-3.2-3B equal-budget rank-allocation PPL. The matched protocol and change definition follow Table 11.

C.7 Factor Accounting

Model	Factor	Ovhd.	W4	W3
Llama-2-7B	5.11086 to 5.11115	0.78920 to 0.78925	4.78920 to 4.78925	3.78920 to 3.78925
Llama-2-13B	8.01060 to 8.01112	0.63136 to 0.63140	4.63136 to 4.63140	3.63136 to 3.63140
Llama-3-8B	5.36871	0.76923	4.76923	3.76923
Llama-3.2-1B	1.44284	1.48276	5.48276	4.48276
Llama-3.2-3B	3.11217	1.10417	5.10417	4.10417
Qwen2.5-3B	3.83150	1.38095	5.38095	4.38095
Qwen3-4B	4.22786	1.16364	5.16364	4.16364

Table 13: Nominal target-weight accounting. "Factor" is the FP32-factor payload in 10^9 bits; "Ovhd." is the overhead in bits per target base weight, and W4/W3 are the resulting eBPP values after adding the base-weight precision. The accounting excludes scales, packing metadata, non-target parameters, and runtime workspace, and is not a measured packed-checkpoint size. Small Llama-2-7B ranges reflect cell-specific adaptive rank plans. The Llama-2-13B range reflects the backend-specific frozen rank plans. Block coordination leaves every factor payload unchanged.

C.8 Additional Downstream Results

Model	Bits	Zero-shot 8 (%)	MMLU 0-shot (%)
GPTQ			
Llama-2-7B	W4A8	63.339	40.187
Llama-2-7B	W4A4	62.105	36.861
Llama-2-7B	W3A4	58.183	30.302
Llama-2-13B	W4A8	65.798	51.681
Llama-2-13B	W4A4	64.861	48.953
Llama-2-13B	W3A4	62.743	44.310
Llama-3-8B	W4A8	66.894	60.575
Llama-3-8B	W4A4	64.431	56.303
Llama-3-8B	W3A4	57.651	43.384
RTN			
Llama-2-7B	W4A8	63.813	39.610
Llama-2-7B	W4A4	62.317	36.640
Llama-2-7B	W3A4	58.861	30.452

Table 14: Additional downstream results. Zero-shot 8 is the macro-average over eight tasks, and MMLU is evaluated with zero in-context examples. All values are percentages.

C.9 COMPLETE LLAMA-3.2 EXTERNAL REFERENCES

		Llama-3.2-1B		Llama-3.2-3B	
Bits	Method	Avg_8	MMLU	Avg_8	MMLU
FP16	Baseline	54.82	36.76	62.66	54.06
W4A8	LLM.int4()	44.56	24.01	56.06	41.48
	$\mathrm{L}^2\mathrm{QER}$	50.97	31.23	60.36	51.24
	SERQ(RTN)	<u>52.76</u>	<u>32.77</u>	<u>61.27</u>	<u>51.45</u>
	PATR(RTN)	**53.26**	**33.11**	**61.46**	**52.56**
W4A4	LLM.int4()	35.19	24.11	35.77	23.50
	$\mathrm{L}^2\mathrm{QER}$	42.17	25.67	50.92	32.25
	SERQ(RTN)	**50.15**	**29.11**	**58.40**	<u>46.36</u>
	PATR(RTN)	<u>49.40</u>	<u>27.83</u>	<u>57.71</u>	**46.87**
W3A4	$\mathrm{L}^2\mathrm{QER}$	<u>37.51</u>	**23.96**	<u>45.89</u>	<u>29.37</u>
	PATR(RTN)	**42.14**	<u>23.66</u>	**50.15**	**32.56**

Table 15: Llama-3.2 zero-shot accuracy (%, ↑). Avg_8 is the unweighted mean of the eight tasks defined in the main paper; MMLU is separate. Bold/underlining indicate the best/second-best completed quantized results.