Journal of Science and Technology – Data Science and Artificial Intelligence
Volume 36, Issue 1, October 2026, 001–008

Escaping Rounding Barriers in Large Language Model Quantization via Barrier-Aware Ising Correction

Anh Nguyen Duc[1], Linh Ngo Van[1], Sang Dinh Viet[1], Hoang Ta Duy[1], Tung Nguyen Son[1], Canh Dinh The[2], Tu Vu Van[3*]

[1]*Hanoi University of Science and Technology, Ha Noi, Vietnam*
[2]*The University of Sydney, Sydney, Australia*
[3]*Australian National University, Canberra, Australia*
[*]*Corresponding author email: tu.vu@anu.edu.au*

Abstract

Weight-only post-training quantization ultimately commits each scalar weight to a discrete grid point, turning the final rounding stage into a combinatorial optimization problem. We study this stage under a fixed quantization grid and show that the layer-wise calibration reconstruction loss is an exact quadratic objective over binary rounding variables, equivalently a classical Ising energy whose pairwise interactions are induced by the activation Gram matrix. This representation exposes rounding barriers*: configurations in which every considered single flip is non-improving while a coordinated multi-spin flip decreases the same empirical objective. We give an explicit positive-definite, round-to-nearest-consistent two-spin barrier and characterize the role of frustration between local fields and pairwise couplings. Motivated by this analysis, we propose Barrier-Aware Ising Correction (BAIC), a post-processing routine evaluated here only on 4-bit AWQ. BAIC first performs thresholded exact coordinate descent on a boundary-selected pool and then examines small correlation-based clusters with exhaustive local enumeration. Candidate generation is heuristic, but every accepted transition is evaluated with the exact empirical reconstruction objective; therefore each accepted move strictly decreases, and the sequence of accepted moves makes the per-layer calibration objective non-increasing. This guarantee does not imply population-loss or end-to-end perplexity improvement. On Mistral-7B-v0.3 and Llama-3-8B, the complete BAIC pipeline improves WikiText-2 and C4 perplexity over the tested 4-bit AWQ baseline without changing inference-time operators, kernels, or the storage format. The experiments establish proof-of-concept effectiveness of the complete corrector, not the prevalence of barriers or the isolated contribution of cluster moves; those questions are explicitly left for future work.*

Keywords: Ising optimization, discrete optimization, post-training quantization, weight-only quantization.

1. Introduction

Large language models are expensive to deploy because inference is dominated by parameter storage and memory bandwidth. Weight-only post-training quantization (PTQ) reduces these costs by mapping pretrained weights to low-bit grids without retraining. Representative methods include GPTQ [1], AWQ [2], SmoothQuant [3], and more recent transformation-based approaches such as FlatQuant [4].

Once a grid and its scales are fixed, however, each scalar weight must still be assigned to a representable level. Restricting a weight to its two adjacent levels makes this final decision binary. For a linear layer, independent nearest rounding is optimal only when the relevant reconstruction metric—the activation Gram matrix—is diagonal. Real calibration activations have nonzero cross-channel correlations, so the layer reconstruction loss contains pairwise interactions between rounding decisions in the same output row.

Discrete coordinate methods such as QuantEase [5] and CDQuant [6] optimize related layer-wise objectives directly on the quantization grid. Their success motivates a more specific structural question: *can a state be stable to all single-coordinate moves considered at that state while still admitting an improving coordinated move?* We show that the answer is yes. Under a fixed grid, the exact calibration reconstruction loss is a classical Ising energy, and this energy can contain multi-spin barriers. Importantly, our result is a structural possibility result; it does not establish that such barriers are frequent in real LLM layers.

Motivated by this observation, we introduce *Barrier-Aware Ising Correction* (BAIC), a post-hoc discrete corrector. The method uses a computationally cheap boundary score to restrict the mobile spins, performs thresholded exact coordinate descent, and then searches small clusters constructed from activation correlations. The current implementation does *not* directly implement the frustration/synergy diagnostic derived by the theory: boundary selection and $|G_{jk}|$ are

p-ISSN 3093-3331
e-ISSN 3093-3323
https://doi.org/10.51316/jst.194.dsai.2026.36.1.1
Received: Aug 19, 2026; Revised: Oct 8, 2026;
Accepted: Oct 9, 2026; Online: Oct 15, 2026

proposal heuristics, whereas exact energy is used for acceptance. This distinction is central to the revised scope of the paper.

The experimental evidence is also deliberately narrow. We evaluate the complete BAIC pipeline on two 7B–8B models with a 4-bit AWQ baseline and report perplexity on WikiText-2 and C4. These experiments show that the *complete corrector* improves the tested AWQ models. They do not isolate the cluster phase, quantify barrier prevalence, establish robustness across seeds or calibration draws, compare against optimized-rounding baselines under matched budgets, or validate 3-bit and 70B-scale operation. We therefore make no such claims.

Our contributions are as follows:

- **Exact fixed-grid Ising formulation.** The calibration reconstruction objective is a quadratic energy over binary rounding decisions, with couplings determined by the scaled activation Gram matrix.
- **RTN-consistent barrier construction.** We provide a positive-definite two-spin example initialized exactly by round-to-nearest in which both single flips increase the objective while the pair flip decreases it.
- **Frustration characterization.** The pair-flip identity separates local single-spin costs from the cross term and yields a signed synergy quantity that diagnoses when joint flipping is more favorable than isolated flipping.
- **Exact-acceptance discrete correction.** BAIC combines heuristic proposals with exact empirical-energy acceptance. Every accepted move strictly decreases the empirical layer objective; no population or perplexity guarantee is claimed.
- **Proof-of-concept 4-bit AWQ evaluation.** On the two tested models and two perplexity datasets, the complete BAIC pipeline improves over the corresponding AWQ baseline without changing inference kernels or storage format.

2. Related Work

Post-training quantization. GPTQ uses approximate second-order information to quantize weights sequentially [1]; AWQ protects salient weights through activation-aware scaling [2]; SmoothQuant redistributes activation difficulty through equivalent transformations [3]; and FlatQuant learns affine transformations to flatten weights and activations [4]. Our experiments use AWQ only. Although the fixed-grid formulation applies whenever a quantized tensor and its adjacent grid levels are available, effectiveness on GPTQ, FlatQuant, or other base quantizers is not established here.

Discrete coordinate optimization. QuantEase formulates layer quantization as a discrete structured optimization problem and uses coordinate-descent updates [5]. CDQuant similarly applies greedy coordinate descent to layer-wise reconstruction objectives [6], while COMQ uses coordinate-wise minimization without backpropagation in other PTQ settings [7]. These methods motivate our structural comparison with single-coordinate dynamics. We do not claim empirical superiority over them because no matched-budget comparison is included in the present evaluation.

Optimized rounding. AdaRound relaxes discrete rounding choices to continuous variables and optimizes a regularized reconstruction objective before projection [8]; BRECQ extends block reconstruction [9]; and SignRound optimizes signed rounding offsets by gradient descent [10]. Successive-rounding methods can additionally use bounded look-ahead search [11]. Continuous or bounded-search algorithms are not generally restricted to strictly improving single-coordinate trajectories, so our barrier theorem should not be read as a statement that all optimized-rounding methods must terminate in the same states.

Ising optimization. Mapping binary optimization to Ising energies is classical [12]. Quantum annealing and simulated-bifurcation methods provide broader solver families [13, 14]. The present algorithm, however, uses only a *classical* Ising energy, coordinate updates, and finite cluster enumeration. For this reason we use the name Barrier-Aware Ising Correction rather than "transverse-field" terminology.

3. Methodology

3.1. Scaled Reconstruction Objective

Consider a linear layer with full-precision weights $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and calibration activations $X \in \mathbb{R}^{n \times d_{\text{in}}}$. Following the AWQ reparameterization, let $C = \operatorname{diag}(c)$ be the per-channel scaling matrix,

$$W_{\text{sc}} = WC, \qquad X_{\text{sc}} = XC^{-1} \tag{1}$$

so that $XW^\top = X_{\text{sc}} W_{\text{sc}}^\top$. The calibration metric used by the correction is therefore

$$G = \frac{X_{\text{sc}}^\top X_{\text{sc}}}{n} = C^{-1} \frac{X^\top X}{n} C^{-1} \succeq 0 \tag{2}$$

For each scaled weight, let q_{ij}^- and q_{ij}^+ be its two adjacent grid levels and define

$$M = \frac{q^- + q^+}{2}, \qquad H = \frac{q^+ - q^-}{2}, \qquad D = W_{\text{sc}} - M \tag{3}$$

A binary spin $s_{ij} \in \{-1, +1\}$ selects

$$W_{q,\text{sc}}(s) = M + H \odot s \tag{4}$$

with residual

$$R(s) = H \odot s - D \tag{5}$$

The empirical layer reconstruction objective is

$$E(s) = \frac{1}{n} \|X_{\text{sc}} R(s)^\top\|_F^2 = \operatorname{Tr}\Big(R(s) G R(s)^\top\Big) \tag{6}$$

All monotonicity statements in this paper refer to the finite-calibration objective in (6).

3.2. Classical Ising form

For a fixed output row i, expansion of (6) gives, up to an additive constant,

$$E_i(s_i) = s_i^\top J^{(i)} s_i + h_i^\top s_i + \text{const} \tag{7}$$

where

$$J_{jk}^{(i)} = H_{ij} H_{ik} G_{jk}, \qquad (h_i)_j = -2H_{ij}(GD_i^\top)_j \tag{8}$$

Thus each output row defines a classical Ising optimization problem over the input-channel rounding spins. The full layer is row-separable for fixed G.

3.3. Exact Flip Gains and Rounding Barriers

For a spin (i, j), define the residual change caused by flipping it as

$$\delta_j = -2 s_{ij} H_{ij} \tag{9}$$

The exact single-flip gain is

$$\Delta E_j = 2\delta_j (RG)_{ij} + \delta_j^2 G_{jj} \tag{10}$$

For two spins $j \neq k$ in the same output row, simultaneous flipping gives the following identity.

Lemma 1 (Pair-flip identity)**.**

$$\Delta E_{\{j,k\}} = \Delta E_j + \Delta E_k + 2\delta_j \delta_k G_{jk} \tag{11}$$

The final term in (11) is absent from an evaluation that treats the two flips independently at the same state.

Definition 1 (Pair rounding barrier)**.** *At configuration s, the pair $\{j,k\}$ is a rounding barrier if*

$$\Delta E_j \geq 0, \qquad \Delta E_k \geq 0, \qquad \Delta E_{\{j,k\}} < 0 \tag{12}$$

Proposition 1 (RTN-consistent barrier existence)**.** *There exist a positive-definite Gram matrix G, a valid two-level quantization geometry, and an RTN-initialized state that is stable to both individual flips but is strictly improved by flipping the pair. In the resulting two-spin problem, no strictly decreasing single-spin path connects the RTN state to the jointly flipped state.*

A concrete instance is

$$H = (1,1), \quad D = (0.9, 0.1), \quad s_{\text{RTN}} = (+1, +1) \tag{13}$$

with

$$G = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix} \succ 0 \tag{14}$$

Here $R = (0.1, 0.9)$, and direct substitution into (10)–(11) gives

$$\Delta E_1 = 9.0, \qquad \Delta E_2 = 2.2, \qquad \Delta E_{\{1,2\}} = -0.8 \tag{15}$$

The complete proof is in Appendix B. This example answers an existence question only; it does not establish barrier frequency in trained LLMs.

Corollary 1 (Frustration is necessary for a pair barrier)**.** *Because $G \succeq 0$,*

$$\delta_j^2 G_{jj} + \delta_k^2 G_{kk} + 2\delta_j \delta_k G_{jk} \geq 0 \tag{16}$$

Therefore a pair barrier requires the aggregate linear field contribution to favor the joint move strongly enough to overcome the nonnegative quadratic cost. Large $|G_{jk}|$ alone is insufficient.

A useful signed diagnostic is the *barrier synergy*

$$S_{jk} := \Delta E_j + \Delta E_k - \Delta E_{\{j,k\}} = -2\delta_j \delta_k G_{jk} \tag{17}$$

Positive S_{jk} means that the joint move is better than the sum of the two isolated moves at the same state. Equation (17) is a theoretical diagnostic. The current implementation uses $|G_{jk}|$ for cluster proposals and therefore does not fully exploit this signed criterion.

3.4. Boundary-Selected Mobile Pool

The implemented corrector restricts optimization to weights close to a rounding midpoint. Define

$$U_{ij}^{\text{bnd}} = \left(1 - \frac{|D_{ij}|}{H_{ij}}\right)_+ \tag{18}$$

with zero score for degenerate $H_{ij} = 0$. The mobile pool is

$$\mathscr{U} = \{(i,j) : U_{ij}^{\text{bnd}} > \gamma_{\text{th}},\ H_{ij} > 0\} \tag{19}$$

We use $\gamma_{\text{th}} = 0.7$. This is a computational heuristic: a weight far from a rounding midpoint can still participate in a profitable correlated move. The theory therefore motivates, but the present experiments do not evaluate, a richer boundary-plus-frustration pool score.

3.5. Group-Flip Gain

For a cluster T within one output row, let $f \in \{0,1\}^{|T|}$ denote a flip pattern and

$$\delta_T = (-2f \odot s_{i,T}) \odot H_{i,T} \tag{20}$$

The exact gain is

$$\Delta E_T = 2\langle \delta_T, (RG)_{i,T} \rangle + \delta_T^\top G_{T,T} \delta_T \tag{21}$$

For $|T| \leq 6$, exhaustive enumeration is inexpensive relative to the dense layer operations.

3.6. Gain-Scale Acceptance Margin

The original manuscript described the following quantity as a statistical "noise floor." That interpretation is too strong. We instead define the robust within-layer gain scale

$$\tau = \text{median}_{(i,j)\in\mathscr{U}} |\Delta E_{ij}^{(0)}|, \tag{22}$$

computed once from initial single-flip gains. It summarizes the typical magnitude of candidate gains in the selected pool; it is *not* an estimator of calibration-sampling variance.

Phase 1 accepts a single flip only if

$$\Delta E < -\kappa\tau, \qquad \kappa = 2 \tag{23}$$

Thus a genuinely negative move in $[-\kappa\tau, 0)$ may be rejected. The resulting state is not necessarily a standard single-spin local minimum. If a complete sweep accepts no moves, then the state is threshold-stationary with respect to (23); if the fixed sweep cap is reached first, no convergence claim is made. Cluster moves use the strict empirical criterion $\Delta E_T < 0$.

The role of (23) is heuristic conservatism, not statistical certification. Estimating population-level uncertainty would require independent calibration draws, bootstrap/subsample analysis, or another explicit variance model, none of which is part of the present evaluation.

3.7. Two-Phase Correction Procedure

BAIC uses two phases.

Phase 1: thresholded coordinate descent. Columns are processed sequentially. Within a column, output rows are independent and may be evaluated in parallel. For each accepted row flip in column j, the corresponding spin is updated explicitly as $s_{ij} \leftarrow -s_{ij}$. The residual and maintained product are then updated exactly:

$$R_{:,j} \leftarrow R_{:,j} + \Delta R_{:,j}, \qquad RG \leftarrow RG + \Delta R_{:,j} \otimes G_{j,:} \tag{24}$$

We use at most $T = 3$ sweeps and stop early after a sweep with zero accepted moves.

Phase 2: small-cluster correction. For each row, current exact single-flip gains are recomputed. Candidate spins satisfy

$$0 \le \Delta E_j \le c_{\text{cand}}\tau, \qquad c_{\text{cand}} = 8 \tag{25}$$

For each candidate, we inspect the $m = 32$ strongest neighbors according to $|G_{jk}|$. Clusters are seeded by strong candidate pairs and grown greedily by the same magnitude heuristic up to $g_{\max} = 6$. All $2^{|T|}$ flip patterns are evaluated using (21), and the best strictly decreasing pattern is accepted.

The implementation evaluates clusters *sequentially with fresh state*. If a pattern on T is accepted, we apply

$$s_{i,T} \leftarrow s_{i,T} \odot (1 - 2f) \tag{26}$$

$$R_{i,T} \leftarrow R_{i,T} + \delta_T \tag{27}$$

$$(RG)_{i,:} \leftarrow (RG)_{i,:} + \delta_T^\top G_{T,:} \tag{28}$$

We then recompute the row's single-flip gains, candidate set, and subsequent cluster proposals. Overlapping clusters are therefore permitted, but no proposal computed before an accepted state change is reused with a stale energy. If no state change occurs, the next remaining proposal may be evaluated against the same current state.

Because Phase 1 is thresholded and capped, an accepted Phase-2 move is evidence of a profitable joint search move, but it is *not automatically evidence that a strict rounding barrier was crossed*. Establishing such attribution requires logging that all admissible single-spin gains are nonnegative at the accepted cluster state; this was not recorded in the reported experiments.

3.8. Exact Acceptance Guarantee

Proposition 2 (Empirical descent of accepted moves)**.** *Every move accepted by BAIC strictly decreases the empirical calibration objective* (6)*. Consequently, the objective after the complete correction is no larger than at initialization.*

Phase 1 accepts only gains satisfying (23), hence $\Delta E < 0$. Phase 2 evaluates the exact group gain (21) at the current state and accepts only if $\Delta E_T < 0$. The Phase-1 update (24) and the complete Phase-2 state updates (26)–(28) are algebraically exact, so summing accepted negative gains proves the result.

This guarantee follows from exact acceptance, not from the gain-scale margin. It concerns only the empirical layer reconstruction objective. No theorem in this paper connects this monotonicity to population reconstruction error, downstream accuracy, or perplexity.

3.9. Initialization

RTN initialization must reproduce the base quantizer's tie-breaking rule. Rather than setting $s = \text{sign}(D)$, which can disagree at exact midpoints under round-half-to-even, we compute the actual nearest grid level and set

$$s_{ij}^{(0)} = +1 \quad \Longleftrightarrow \quad q_{ij}^{\text{nearest}} = q_{ij}^{+} \tag{29}$$

A reconstruction check verifies that $M + H \odot s^{(0)}$ exactly matches the baseline quantized tensor before correction.

3.10. Cost and Implementation Scope

The Gram matrix is accumulated in streaming batches from scaled activations, so the full activation matrix need not remain in memory. The dense Gram matrix requires $O(d_{\text{in}}^2)$ storage.

For a square $d \times d$ layer, one thresholded descent sweep with rank-one RG maintenance costs $O(d_{\text{out}} d_{\text{in}}^2)$. Cluster enumeration is bounded by the small cluster cap and the number of rows/clusters considered. In the archived evaluated setup, end-to-end correction of one 7B–8B model took approximately 2–3 hours on a single NVIDIA RTX 4090; because per-layer cost grows with d_{in}^2, the wide MLP layers dominate this total. Measured peak GPU memory was not recorded in the reported experiment set, and 70B-scale behavior was not evaluated, so we do not claim 70B-scale feasibility from these results.

4. Algorithm Summary

For clarity, the complete state-update logic is summarized below.

Algorithm 1: Barrier-Aware Ising Correction (BAIC)

1) Compute M, H, D and initialize s from the base quantizer's exact RTN decisions.
2) Stream scaled calibration activations to form G; initialize $R = H \odot s - D$ and RG.
3) Build boundary pool $\mathscr{U}$ and compute τ from (22).
4) For at most T sweeps, visit columns sequentially. Evaluate fresh single-flip gains using (10); accept pool moves satisfying (23); flip each accepted spin via $s_{ij} \leftarrow -s_{ij}$, and update R and RG immediately using (24). Stop if a complete sweep accepts no moves.
5) For each selected output row, recompute current single-flip gains and candidate spins. Build one or more $|G_{jk}|$-based clusters, capped at $g_{\max}$.
6) Enumerate every pattern in a proposed cluster with the exact gain (21). If the best pattern has negative gain, accept it, apply the complete state updates (26)–(28), and rebuild the row's candidates/clusters from the fresh state before evaluating further overlapping proposals.
7) Return $W_q = (M + H \odot s)C^{-1}$.

5. Experiments

5.1. Scope and Setup

We evaluate only the setting for which experimental results are available: Mistral-7B-v0.3 [15] and Llama-3-8B [16], using 4-bit weight-only asymmetric AWQ with group size 128. BAIC is applied after the AWQ scale search and initial grid quantization. Hyperparameters are fixed to $\gamma_{\text{th}} = 0.7$, $g_{\max} = 6$, $T = 3$, $\kappa = 2$, $m = 32$, and $c_{\text{cand}} = 8$. The calibration set contains 128 C4 sequences of length 2048 (approximately 262k tokens). The Gram matrix for each corrected layer is accumulated over the layer's calibration activations. The `lm_head` is not corrected.

The available experimental record contains one reported configuration per model; it does not support across-seed or across-calibration-draw variance estimates. It also does not contain the broader 3-bit, downstream-task, matched-baseline, column-permutation robustness, or peak-memory measurements requested during review. We therefore do not infer those quantities from the present results.

5.2. Perplexity Results

Table 1 reports the same evaluated model–dataset pairs as the original submission. Gap Reduction is defined as

$$\text{GapRed} = \frac{\text{PPL}_{\text{AWQ}} - \text{PPL}_{\text{AWQ+BAIC}}}{\text{PPL}_{\text{AWQ}} - \text{PPL}_{\text{FP16}}} \times 100\% \quad (30)$$

Across these four tested model–dataset pairs, the complete BAIC correction reduces perplexity relative to the corresponding AWQ baseline. The correction changes only the stored quantized weights, so it introduces no additional inference-time operators or model parameters.

5.3. What the Experiment does and does not Establish

The empirical result supports the narrow statement that the complete correction procedure improves the tested 4-bit AWQ checkpoints. It does not determine whether the gain comes primarily from thresholded single-spin descent, cluster search, or their interaction. In particular, because the original runs did not separately log the number of accepted single-spin moves, proposed/accepted cluster moves, or whether accepted clusters occurred at states with all single gains nonnegative, these results cannot be used to estimate barrier prevalence or to claim that the measured perplexity gain was caused specifically by barrier crossing.

Likewise, the exact acceptance guarantee and the perplexity measurements are distinct pieces of evidence. Exact acceptance proves non-increase of the finite-calibration reconstruction objective. Perplexity is evaluated end-to-end and happens to improve in the tested cases. No theorem is supplied that maps a layer-wise decrease in (6) to a guaranteed perplexity decrease.

6. Discussion

6.1. Theory Versus Implemented Proposals

The barrier analysis gives a signed, state-dependent quantity S_{jk} in (17), whereas the present implementation proposes clusters using $|G_{jk}|$ and selects the mobile pool using boundary distance. This is intentionally presented as a gap between diagnosis and the first practical implementation, not as an implementation of the theoretically most targeted score. Exact acceptance protects the empirical objective from poor proposals, but a better proposal score could improve search efficiency or barrier recall. Testing synergy-based clustering and frustration-aware pool expansion is future work.

6.2. Relation to Coordinate Descent

Proposition 1 shows that strictly improving single-coordinate dynamics can be obstructed in a valid fixed-grid instance. It does not imply that every real layer contains such a barrier, nor that every coordinate-descent quantizer reaches one. It also does not apply unchanged to continuous relaxations or bounded look-ahead algorithms that can traverse different trajectories. The contribution is therefore a structural limitation of a class of single-flip trajectories, not a general empirical ranking of quantization algorithms.

6.3. Calibration Objective Versus Downstream Quality

The empirical reconstruction objective is a local surrogate for preserving layer outputs on calibration activations. Exact acceptance gives a useful invariant: BAIC never accepts a move that worsens that measured objective. However, the calibration objective is estimated

Table 1. 4-bit AWQ results in the evaluated setting. BAIC denotes the complete two-phase correction; no phase ablation is available

		WikiText-2		C4	
Model	**Method**	**PPL ↓**	**Gap Red. ↑**	**PPL ↓**	**Gap Red. ↑**
Mistral-7B-v0.3	FP16 (Ref.)	4.8454	–	7.6040	–
	AWQ	4.9778	0.00%	7.7892	0.00%
	AWQ + BAIC	4.9612	12.54%	7.7740	8.21%
Llama-3-8B	FP16 (Ref.)	5.4425	–	8.6383	–
	AWQ	6.7386	0.00%	10.4595	0.00%
	AWQ + BAIC	6.6731	5.05%	10.2826	9.71%

on a finite distribution and is not identical to the model's end-to-end language-modeling loss. The paper therefore uses "strictly decreasing accepted move" and "non-increasing empirical calibration objective" rather than the broader phrase "monotonic improvement." Population reconstruction and perplexity remain empirical questions.

6.4. Scalability

The $O(d_{\text{in}}^2)$ dense Gram matrix is manageable for the dimensions evaluated here but grows quickly in wider feed-forward layers. Blocking, CPU offload, sparse proposal graphs, and factorized exact residual updates are possible engineering directions. Since we do not report measured peak memory or 70B-scale runs, we make no empirical scalability claim beyond the evaluated 7B–8B setting.

7. Limitations

The present study is intentionally limited in several ways. First, the experimental evaluation covers two model families, one 4-bit AWQ configuration, and perplexity on WikiText-2 and C4. There is no 3-bit evaluation, no MMLU/PIQA/HellaSwag/ARC evaluation, no matched comparison with QuantEase/CDQuant/AdaRound/SignRound, and no evaluation on a second base quantizer. Consequently, the empirical claims are restricted to the reported 4-bit AWQ setting.

Second, the experiments do not isolate Phase 1 from Phase 2 and do not log barrier-specific events. The paper therefore cannot quantify how common strict barriers are in real LLM layers or how much of the final perplexity change is attributable to multi-spin moves. The theoretical result establishes existence, including an RTN-consistent construction, but not prevalence.

Third, τ is a robust gain-scale statistic rather than a calibrated estimate of sampling noise, and κ is a fixed heuristic multiplier. No bootstrap, independent calibration draw, or formal population-generalization analysis is included. Accordingly, the monotonicity guarantee is limited to the observed calibration objective.

Fourth, the hyperparameters γ_{th}, $g_{\max}$, κ, m, c_{cand}, and T are fixed implementation choices in the reported runs; sensitivity to them is not established. Robustness to alternative or randomized column permutations is also not evaluated across repeated runs.

Fifth, the dense Gram matrix scales quadratically in input width. End-to-end correction time was recorded (approximately 2–3 hours per model on a single RTX 4090), but measured peak GPU memory and 70B-scale behavior are not. The `lm_head` is left unchanged.

Finally, the present revision is based on the archived experimental outputs available from the original study. Some reproducibility metadata requested during review, such as complete package-version and run-seed records, are not available in those outputs; we therefore avoid inventing them or claiming version-level reproducibility that the record cannot support.

8. Future Work

The highest-priority empirical extension is a phase ablation that reports AWQ, AWQ plus thresholded single-spin descent, cluster-only correction from the same initialization, and the complete method, together with accepted-move counts, per-layer reconstruction changes, and phase runtimes. Logging whether a cluster is accepted when all current single gains are nonnegative would directly measure barrier-crossing events and estimate their prevalence by layer.

A second direction is broader evaluation: 3-bit quantization, downstream zero-shot tasks, multiple calibration draws and random seeds, matched-budget comparisons against discrete/optimized-rounding baselines, and additional base quantizers. These experiments are necessary before making claims about robustness or general superiority.

A dedicated hyperparameter-sensitivity study should vary γ_{th}, $g_{\max}$, κ, m, c_{cand}, and T, and report perplexity, calibration reconstruction error, runtime, mobile-pool size, and accepted-move counts. In particular, varying κ would test the effect of the heuristic gain-scale margin rather than treating the reported value as a validated default.

On the algorithmic side, the theory suggests replacing boundary-only mobility and raw $|G_{jk}|$ clustering with frustration-aware pool expansion and synergy-based proposals from (17). On the statistical side, the gain-scale threshold could be replaced by uncertainty estimates from calibration resampling or independent draws. On the systems side, blocked Gram construction, CPU offloading, and wide-layer scheduling should be measured on 70B-scale models rather than inferred from tensor-size arithmetic.

9. Conclusion

We formulated fixed-grid rounding correction as an exact classical Ising optimization problem and showed, by an RTN-consistent positive-definite example, that strictly improving single-spin dynamics can encounter a genuine multi-spin barrier. This provides a precise structural reason to consider coordinated discrete moves. BAIC is a practical corrector motivated by that analysis: it combines thresholded coordinate descent with small-cluster enumeration while using exact empirical-energy acceptance for every accepted transition. The present implementation uses the theoretical analysis as motivation, but its proposal heuristic is based on boundary proximity and $|G_{jk}|$ rather than the signed synergy criterion in (17).

The guarantee is deliberately narrow: accepted moves strictly decrease the finite-calibration reconstruction objective. The reported experiments are likewise narrow: on two 7B–8B models with a 4-bit AWQ baseline, the complete BAIC pipeline improves WikiText-2 and C4 perplexity without changing inference-time computation. The present evidence does not establish barrier prevalence, isolate the cluster contribution, compare against optimized-rounding baselines, or demonstrate broader bit-width and scale robustness. Those questions define the next empirical stage of this work.

A. Proof of the Single-Spin and Group-Flip Formulas

For one output row, a single flip changes R by δe_j. Expanding the quadratic form gives

$$E(R+\delta e_j) - E(R) = 2\delta\, e_j^\top G R^\top + \delta^2 e_j^\top G e_j \quad (31)$$

$$= 2\delta (RG)_j + \delta^2 G_{jj} \quad (32)$$

which is (10). For a group T, replacing δe_j by the sparse vector δ_T yields (21); taking $T = \{j,k\}$ yields (11).

B. Proof of Proposition 1

Take $H = (1,1)$ and $D = (0.9, 0.1)$. Both offsets are positive, so nearest rounding chooses $s = (+1,+1)$. Let

$$G = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix} \quad (33)$$

Its leading principal minors are $1 > 0$ and $4 - 2.25 = 1.75 > 0$, hence $G \succ 0$. The residual is $R = (0.1, 0.9)$ and both flips have $\delta = -2$. Therefore

$$(RG)_1 = 0.1 - 1.35 = -1.25 \quad (34)$$

$$(RG)_2 = -0.15 + 3.6 = 3.45 \quad (35)$$

Using (10),

$$\Delta E_1 = 2(-2)(-1.25) + 4(1) = 9.0 \quad (36)$$

$$\Delta E_2 = 2(-2)(3.45) + 4(4) = 2.2 \quad (37)$$

The pair identity then gives

$$\Delta E_{\{1,2\}} = 9.0 + 2.2 + 2(4)(-1.5) = -0.8 \quad (38)$$

Thus the RTN state is single-flip stable but the pair flip is strictly improving. Because the system contains only two spins, any single-spin path from $(+1,+1)$ to $(-1,-1)$ must begin with one of those two individually non-improving flips. Hence no strictly decreasing single-spin path connects the two states. More generally, a barrier on a subset T blocks a strictly decreasing path that is restricted to changing spins in T from the barrier state; we do not claim that unrelated auxiliary spins in a larger system can never alter the landscape first.

C. Banker's-Rounding-Safe Initialization

If the base implementation uses round-half-to-even, $\mathrm{sign}(D)$ disagrees at exact midpoints. BAIC therefore records the actual nearest grid level and maps that decision to the spin variable. This guarantees that the pre-correction state exactly matches the stored baseline quantized tensor.

D. Why Low-Rank Gram Approximations Are Unsafe for Acceptance

A low-rank Gram matrix may be useful for proposing moves, but it cannot generally replace G in the acceptance test.

Proposition 3 (Low-rank acceptance can reverse the sign of a flip)**.** *Let $\widetilde{G}_k$ be a rank-k truncation of $G \succeq 0$. There exist a valid residual state and a single flip such that the approximate gain is negative while the exact gain is positive.*

Proof. A complete two-dimensional example suffices. Let

$$G = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix} \quad (39)$$

with eigenvalues 2 and 1. Its rank-one truncation is

$$\widetilde{G}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad (40)$$

Choose $H = (1,1)$, $s = (+1,+1)$, and $D = (1, -0.2)$, giving the valid residual $R = (0, 1.2)$. Flip the first spin, so $\delta = -2$. Under the rank-one approximation,

$$(R\widetilde{G}_1)_1 = 1.2, \qquad \Delta\widetilde{E} = -4(1.2) + 4(1) = -0.8 < 0 \tag{41}$$

Under the exact Gram matrix,

$$(RG)_1 = 0.6, \qquad \Delta E = -4(0.6) + 4(1.5) = 3.6 > 0 \tag{42}$$

Thus approximate acceptance would move uphill in the exact objective. Low-rank or other approximate matrices may still be used to nominate candidates, provided the final acceptance test uses the current exact energy.

Acknowledgment

This work is supported by the Ministry of Education and Training of Vietnam under project number CT2025.EA.BKA.09.

References

[1] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, GPTQ: Accurate post-training quantization for generative pre-trained transformers, in International Conference on Learning Representations (ICLR), 2023.

[2] 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 on-device LLM compression and acceleration, in Proceedings of Machine Learning and Systems, 2024.

[3] 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 International Conference on Machine Learning (ICML), 2023, pp. 38087–38099.

[4] Y. Sun, R. Liu, H. Bai, H. Bao, K. Zhao, Y. Li, J. Hu, X. Yu, L. Hou, C. Yuan, X. Jiang, W. Liu, and J. Yao, FlatQuant: Flatness matters for LLM quantization, in International Conference on Machine Learning (ICML), 2025, pp. 57587–57613.

[5] K. Behdin, A. Acharya, A. Gupta, Q. Song, S. Zhu, S. Keerthi, and R. Mazumder, QuantEase: Optimization-based quantization for language models, arXiv preprint arXiv:2309.01885, 2023. https://doi.org/10.48550/arXiv.2309.01885.

[6] P. A. Nair and A. S. Suggala, CDQuant: Greedy coordinate descent for accurate LLM quantization, arXiv preprint arXiv:2406.17542, 2024. https://doi.org/10.48550/arXiv.2406.17542.

[7] A. Zhang, Z. Yang, N. Wang, Y. Qi, J. Xin, X. Li, and P. Yin, COMQ: A backpropagation-free algorithm for post-training quantization, IEEE Access, vol. 13, pp. 99879–99891, 2025. https://doi.org/10.1109/ACCESS.2025.3576737.

[8] M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort, Up or down? Adaptive rounding for post-training quantization, in International Conference on Machine Learning (ICML), 2020, pp. 7197–7206.

[9] 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 International Conference on Learning Representations (ICLR), 2021.

[10] W. Cheng, W. Zhang, H. Shen, Y. Cai, X. He, L. Kaokao, and Y. Liu, Optimize weight rounding via signed gradient descent for the quantization of LLMs, in Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 11332–11350.

[11] S. Cha, H. Chen, D. Kim, H. Zhang, K. Chan, G. de Veciana, and H. Vikalo, CoreQ: Learning-free mismatch correction and successive rounding for quantization, arXiv preprint arXiv:2602.05902, 2026. https://doi.org/10.48550/arXiv.2602.05902.

[12] A. Lucas, Ising formulations of many NP problems, Frontiers in Physics, vol. 2, pp. 5, 2014.

[13] T. Kadowaki and H. Nishimori, Quantum annealing in the transverse Ising model, Physical Review E, vol. 58, no. 5, pp. 5355–5363, 1998.

[14] H. Goto, K. Tatsumura, and A. R. Dixon, Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems, Science Advances, vol. 5, no. 4, pp. eaav2372, 2019.

[15] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, Mistral 7B, arXiv preprint arXiv:2310.06825, 2023. https://doi.org/10.48550/arXiv.2310.06825.

[16] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, *et al.*, The Llama 3 herd of models, arXiv preprint arXiv:2407.21783, 2024. https://doi.org/10.48550/arXiv.2407.21783.