Calibration Limits in Mixed-Precision Quantization

Mantis Discovery

Abstract

Mixed-precision allocation is constrained by the information retained from calibration evaluations. We compare scalar loss profiles, reconstructed output vectors, and direct comparisons under matched forward budgets. Exhaustive evaluation of 25 small trained networks separates approximation bias, sampling variance, and omissions from a shortlist. With round-to-nearest quantization, output scoring reduces normalized allocation regret by 61.0% relative to additive scalar scoring. After a GPTQ Cholesky update, the reduction is 6.7%, and its interval includes zero. Fisher controls show that cross-tensor terms can help under rounding yet increase sampling error after reconstruction. Native task metrics, a corrupted-input experiment, and a residual-depth series likewise show that lower teacher discrepancy need not translate into a general accuracy or speed advantage. The accompanying constructions characterize limits of the restricted observation models. These results concern the small networks and quantizers studied here.

1 Introduction

A calibration evaluation returns a model output, but sensitivity tables often retain only a scalar loss. That reduction can discard information: two perturbations with equal individual errors may reinforce or cancel when applied together. Output vectors retain their directions. Scalar interaction profiles offer another route, but they divide a finite evaluation budget among more configurations.

We ask which measurements improve precision allocation. The networks are small enough that every feasible allocation can be evaluated independently of the selector. This separates approximation error, sampling error, and the cost of omitting a candidate. It also separates allocation from quantizer construction: a method that helps with rounding may add little after data-dependent reconstruction has reduced quantization error.

This distinction changes the main empirical conclusion. Output scoring substantially improves allocation under rounding, but its average advantage becomes small after the GPTQ update, as do the associated accuracy differences. Cross-tensor Fisher terms reduce approximation bias in both settings but increase sampling variance enough to hurt selection after reconstruction. Direct comparisons help within a shortlist; richer shortlist proposals do not help uniformly across quantizers or depths. The timing study measures additional selection work, not a CPU speedup.

The theoretical examples separate limitations of an observation model from behavior of a trained network. Networks can agree on every measured profile while having different feasible optima, and an exact mean interaction model can still be difficult to estimate. Conversely, a classifier can have a highest-order scalar KL interaction while singleton outputs reconstruct every allocation exactly.

Relation to prior work. Mixed-precision allocation by integer optimization is established in HAWQ-V3 and the small-calibration study of Hubara et al. [1, 2]. CLADO measures cross-layer interactions with forward evaluations and uses integer quadratic optimization [3]; CoopQ and MixQuant likewise model interactions or upstream quantization effects [4, 5]. BRECQ studies output

reconstruction and Fisher curvature [6], while GPTQ and AWQ construct calibrated weight quantizers [7, 8]. We hold the quantized candidate bank fixed during allocation rather than proposing another quantizer.

Calibration robustness and unlabeled sensitivity measurements have also been studied by Pandey et al., Williams and Aletras, and COVERCAL [9, 10, 21]. Recent work uses task-conditioned output sensitivity, type-aware Fisher metrics, or learned reconstruction preferences [22–24]. Here, scalar, Fisher, and output representations receive identical observations. The anchored decomposition [12, 13] and successive halving [14, 15] are standard ingredients; the contribution is their controlled comparison and the accompanying network examples.

2 Measurements for allocation

Let f be a trained teacher and f_a its quantized version, where $a = (a_1, \ldots, a_G)$ assigns a bit width to each of G weight tensors. Each tensor has a fixed quantized candidate at each permitted precision. This bank may be obtained by rounding weights or by reconstruction on a separate dataset. Selection changes the precision choices, not the bank. For n_g weights in tensor g and fixed overhead H, the storage constraint is

$$\mathcal{A}_M = \left\{ a : \; C(a) = H + \sum_g \lceil n_g a_g / 8 \rceil \le M \right\}. \tag{1}$$

The overhead includes biases, affine quantization parameters, and the file header. Activations are unquantized. The main experiment uses $a_g \in \{2, 3, 4\}$ and a weight-payload limit $\lfloor \bar{b} \sum_g n_g / 8 \rfloor$, with $\bar{b} = 2.5$ or 3. Temporary measurement configurations need not satisfy the final storage limit.

Write $z_a(x)$ for logits in classification and the scalar prediction in regression. Selection uses no labels. Its loss is teacher KL, $\mathrm{KL}(p_f(x) \| p_a(x))$, or squared teacher discrepancy, $(z_f(x) - z_a(x))^2$, clipped to $[0, L]$ with $L = 5$. Let R_C and R_T be means on the fixed calibration and test pools. A population mean R always refers to an explicitly specified input law; repeated draws from a pool estimate that pool's mean.

A query draws calibration inputs uniformly with replacement and evaluates selected models on them. It pays one unit per student–input evaluation, including repetitions, and one unit per distinct teacher input first used during that run:

$$B(\mathcal{Q}) = N_{\text{student}}(\mathcal{Q}) + N_{\text{distinct teacher inputs}}(\mathcal{Q}). \tag{2}$$

A shared batch of n inputs for m models costs at most $(m + 1)n$. Quantizer construction is charged separately and is common to selectors using the same bank. Forward counts exclude vector storage, score arithmetic, weight copies, and batching effects; the empirical timing study reports their execution cost.

2.1 Scalar and output approximations

Let a^0 be the all-four-bit reference, and let $a^{g,b}$ change only tensor g to b. The pilot measures the reference and all singleton changes on common inputs. Scalar scoring adds their mean loss increments:

$$\widehat{S}_1(a) = \widehat{R}(a^0) + \sum_{g : a_g \neq 4} [\widehat{R}(a^{g,a_g}) - \widehat{R}(a^0)]. \tag{3}$$

Output scoring applies the same first-order anchored approximation before evaluating the loss:

$$\widetilde{z}_a(x) = z_{a^0}(x) + \sum_{g : a_g \neq 4} [z_{a^{g,a_g}}(x) - z_{a^0}(x)]. \tag{4}$$

Both are applications of the anchored decomposition [12, 13]. Equation (4) preserves the directions of the observed perturbations, while (3) retains only scalar losses. The two scores use identical forward evaluations. Neither is generally exact for a serial network.

Two controls distinguish output interactions from the loss's nonlinearity. For teacher probabilities p, the usual output Fisher approximation to KL is

$$q(e) = \tfrac{1}{2} e^\top [\mathrm{diag}(p) - pp^\top] e, \qquad e = \widetilde{z}_a - z_f. \tag{5}$$

For regression, $q(e) = e^2$. Write $e_0 = z_{a^0} - z_f$ and $d_g = z_{a^{g,a_g}} - z_{a^0}$. The full Fisher score uses $q(e_0 + \sum_g d_g)$; its additive control uses $q(e_0) + \sum_g [q(e_0 + d_g) - q(e_0)]$. Their difference is the cross-tensor contribution before clipping and averaging. Output-space curvature and reconstruction are established quantization tools [6, 19]; these controls isolate the contribution of the retained cross terms.

2.2 Selection and comparison

Full-budget scoring evaluates all 11 singleton profiles on $\lfloor B/12 \rfloor$ common draws and returns the best feasible score. Quadratic scalar scoring instead measures all 51 profiles changing at most two tensors. We enumerate the 243 allocations so that differences between selectors cannot be attributed to search error. Integer programming is an established alternative for additive allocation [1, 2]; enumeration of the nonlinear output score is not a scalable search contribution.

A second family uses eight pilot inputs to propose 16 feasible candidates, then compares their actual losses by successive halving [14, 15]. With m survivors, remaining budget B_t, and $r_t = \lceil \log_2 m \rceil$, each round requests $\max\{1, \lfloor B_t / \lceil r_t(m+1) \rceil \rfloor\}$ common draws, subject to the budget check. It retains the best $\lceil m/2 \rceil$ cumulative means. Ties follow a stable order; eliminated candidates do not return. The shortlist is ranked by the scalar score, by the output score, or by alternately taking unused entries from their two rankings. The latter uses no fitted combination weight. Shared observations can reduce comparison variance when losses are positively correlated [16]; this fixed-budget rule has no confidence-based elimination guarantee.

Two controls require no allocation measurements. Uniform precision uses the highest strictly feasible common width. Balanced precision minimizes $\sum_g n_g (a_g - 3)^2$ over the same feasible set, with stable ties. The second control avoids disadvantaging nearly uniform three-bit weights when byte alignment alone excludes the exactly uniform configuration.

3 Empirical evaluation

3.1 Design and outcomes

The main panel contains 25 checkpoints: digit classification with a plain multilayer perceptron (MLP), a residual MLP, and a convolutional network (CNN); breast-cancer benchmark classification with an MLP; and diabetes regression with an MLP. Each setting uses five training/split seeds and the corresponding scikit-learn dataset [20]. The networks have five quantized tensors and 7,633–10,666 parameters. Approximately 40% of examples train the teacher, 10% determine early stopping, 25% form the allocation-calibration pool, and 25% form the test pool. Preprocessing uses training data only. No quantized model is fine-tuned.

We compare per-channel and per-tensor grids with either round-to-nearest weights or a GPTQ Cholesky update [7]. The latter builds every candidate from 32 separate training inputs before allocation, with that construction shared by the selectors. Each checkpoint therefore supplies four

Method	Normalized allocation regret ↓ Round-to-nearest	GPTQ update	Mean allocation evaluations
Uniform precision	1.2175	1.3181	0
Balanced precision	0.4229	0.4848	0
Scalar score	0.2523	0.1497	585.6
Additive Fisher score	0.2137	0.1396	585.6
Full Fisher score	0.1368	0.1545	585.6
Output score	0.0983	0.1397	585.6
Quadratic scalar score	0.3539	0.4418	554.3
Scalar shortlist, halving	0.1042	0.1670	581.0
Combined shortlist, halving	0.0883	0.1681	581.0

Table 1: Matched allocation budgets on the 25-checkpoint panel. The three caps average 597.3 evaluations. GPTQ-update banks additionally require 32 teacher evaluations for construction, common to all methods in that column. Regret is normalized separately within each bank; native task values are in Table 2.

fixed candidate banks. This implementation uses teacher layer inputs throughout and does not propagate reconstructed activations downstream. Full implementation details and its distinction from GPTQ appear in the supplementary material.

All 243 allocations are evaluated offline on the calibration and test pools. The selector receives only charged calibration entries, never the exhaustive table or test outcomes. There are 27–125 feasible allocations per storage constraint. Calibration caps are 256, 512, and 1,024, with 30 matched input-resampling seeds per condition. The comparison records 162,000 policy outcomes; resampling and repeated control choices are not additional training experiments. The reconstruction, Fisher, and mechanism analyses are exploratory.

We report native task metrics and retrospective normalized allocation regret,

$$\mathrm{NReg}(\widehat{a}) = \frac{R_T(\widehat{a}) - \min_{a \in \mathcal{A}_M} R_T(a)}{|\mathcal{A}_M|^{-1} \sum_{a \in \mathcal{A}_M} R_T(a) - \min_{a \in \mathcal{A}_M} R_T(a)}. \tag{6}$$

Every denominator is positive in the study. This normalization expresses regret relative to uniform random feasible selection, so values can exceed one. Panel summaries give equal weight to settings, seeds, granularities, storage constraints, and budgets. Intervals resample five joint seed groups, preserving shared digit splits and matched methods. They describe the benchmark panel; the resampling procedure is given in the supplementary material.

3.2 The quantizer changes which measurements help

With rounding, output scoring reduces mean regret from 0.2523 to 0.0983, a 61.0% reduction with joint-seed interval 44.3–71.9% (Table 1). After the GPTQ update, the reduction is 6.7%, with interval −0.4–13.3%. Combining scalar and output shortlists improves paired halving by 15.3% under rounding, but changes its mean regret from 0.1670 to 0.1681 after reconstruction. The large rounding gain therefore does not persist unchanged under this calibrated quantizer.

The Fisher controls separate two effects of retaining outputs. Under rounding, retaining cross-tensor terms improves regret from 0.2137 to 0.1368, and applying the teacher loss to the reconstructed output further improves it to 0.0983. After the GPTQ update, full Fisher scoring is worse than its additive control, although output scoring remains better than full Fisher scoring. The additional terms can therefore be less reliable when estimated from the same small sample.

Setting	Quantizer	Balanced	Scalar	Output	Combined
Digits CNN	Rounding	78.01	82.10	84.35	84.75
	GPTQ update	94.60	95.04	95.17	95.20
Digits MLP	Rounding	85.62	87.81	88.29	88.37
	GPTQ update	91.20	91.89	91.87	91.77
Digits residual	Rounding	91.37	93.29	93.58	93.52
	GPTQ update	93.74	93.87	93.86	93.84
Breast cancer	Rounding	96.92	96.81	96.80	96.79
	GPTQ update	96.75	96.75	96.72	96.71
Diabetes	Rounding	0.54137	0.53940	0.53787	0.53795
	GPTQ update	0.53825	0.53780	0.53699	0.53715

Table 2: Accuracy in percent, except diabetes standardized MSE, where lower is better. Combined denotes combined-shortlist halving. Full-precision values are 95.78, 92.71, 94.36, 96.78%, and 0.53777 in row order. Means cover both storage constraints, granularities, and all three allocation budgets.

Exact calibration-pool calculations help explain this reversal. For the GPTQ banks, full Fisher scoring has lower normalized squared comparison bias than its additive version, 0.1465 versus 0.2766. At budget 512, however, its sampling variance is 1.1602 rather than 0.8336, giving total comparison MSE 1.3067 rather than 1.1102. Output scoring has bias squared 0.0341 and variance 0.7522. These are fixed-pool diagnostics against the calibration-optimal allocation, not quantities available to selection. They instantiate the bias–variance identity in Section 4 and show why ranking the approximations by bias alone is insufficient.

3.3 Native task performance

Table 2 distinguishes allocation quality from the quantizer's effect. For the digit CNN, output scoring adds 2.24 accuracy points over scalar scoring under rounding (interval 0.81–4.06). After the GPTQ update, the difference is 0.13 points (-0.01–0.28). The reconstruction itself raises both methods to about 95% accuracy, close to the 95.78% teacher.

The remaining GPTQ accuracy differences are -0.02 points for the plain digit MLP, -0.01 for the residual MLP, and -0.03 for breast cancer; each interval includes zero. Diabetes MSE decreases by 0.00081, with interval -0.00006–0.00190 for the reduction. Allocation regret and native task performance need not agree: the breast-cancer benchmark has nearly unchanged accuracy under several simpler choices.

A separate residual depth series contains 12 further checkpoints, three seeds at each of $G = 3, 5, 7, 9$ tensors, with up to 72,394 parameters. It uses the digit dataset and binary two/four-bit rounding. Scalar and output scoring have mean regrets 0.0423 and 0.0449; combined-shortlist halving has 0.0440. At the two smaller depths every feasible allocation fits in the 16-candidate shortlist. These results do not support a general superiority claim for output scoring as depth increases.

3.4 Calibration under an input corruption

For the 15 digit checkpoints with channel quantization, we add independent Gaussian noise of standard deviation 0.15 to inputs scaled to $[0, 1]$, then clip them to that range. One fixed corruption is generated for each example. The same noisy test inputs evaluate selection from either clean calibration data or a disjoint, similarly corrupted calibration pool. Weight banks remain constructed from clean training

Quantizer	Calibration	Noisy-test regret ↓ Scalar	Output	Combined	Output-score accuracy (%)
Rounding	Clean	0.2194	0.0890	0.0789	84.57
	Noisy	0.2057	0.0695	0.0670	84.72
GPTQ update	Clean	0.1477	0.1669	0.2253	89.16
	Noisy	0.1260	0.1308	0.1790	89.20

Table 3: Same noisy test examples, different allocation-calibration laws. The full-precision teacher averages 88.70% accuracy on these examples. Quantized models can differ from the teacher in ways that improve task accuracy while worsening teacher fidelity.

data. This produces 30 new exhaustive domain tables and 21,600 policy records at budget 512.

Matching the calibration corruption reduces output-scoring regret from 0.0890 to 0.0695 under rounding and from 0.1669 to 0.1308 after the GPTQ update. The corresponding accuracy changes are only 0.16 and 0.05 points. Scalar scoring remains slightly better in the reconstructed noisy-input comparison. The experiment separates allocation sampling from a changed input law; it tests one synthetic corruption and does not establish robustness to arbitrary distribution shift.

3.5 Why more scalar profiles can be worse

The rounding experiment measures every order-r scalar profile. Exact quadratic profiles have less approximation error than additive profiles: selecting and assessing on the same test pool gives regrets 0.0193 and 0.1750. At budget 512, calibration-based selection reverses them, giving 0.3364 and 0.2500 (the supplementary interaction-order table).

At that budget, additive profiling has 42 shared inputs, normalized squared bias 0.4442, and variance 0.3804. Quadratic profiling has nine inputs, squared bias 0.0553, and variance 2.8462. Its 7.48-fold variance increase reflects both a 1.60-fold per-input increase and the 42/9 reduction in sample count. Removing cross-profile covariance at fixed sample count raises the quadratic variance a further 7.80-fold. Shared inputs help, although quadratic profiling remains worse. Across 1,500 order/budget conditions, 99.3% of predicted MSEs fall within three estimated Monte Carlo standard errors of 4,096-batch resampling estimates.

Canceling shared coefficients saves measurements but leaves approximation bias. In 10,980 pairs selected without inspecting losses, quadratic comparisons at budget 256 have wrong-order rates 27.74% using all 51 profiles, 19.04% using only nonzero coefficients (16.43 profiles on average), and 8.34% using two directly evaluated models. At budget 1,024 the rates are 17.86, 11.79, and 3.95%. Direct comparisons are more accurate for these pairs; the cancellation identity is not evidence for a better global search algorithm.

The local output approximation has a narrow empirical domain. On fixed test subsets preserving every relevant activation and pooling choice for $\tau \leq 1/8$, median measured slopes are 2.001 for output remainder and 3.001 for loss discrepancy. Five of 30 conditions have empty stable subsets and no slope estimate. Only 0.12% of test inputs preserve all choices at the low-bit endpoints, averaged across conditions. The small-perturbation result in the supplementary material therefore does not explain endpoint performance by itself.

Method	Median live selection time (ms) Round-to-nearest	GPTQ update
Scalar score	2.29	2.32
Full Fisher score	4.14	3.70
Output score	7.32	7.01
Combined shortlist, halving	8.39	8.60

Table 4: Twenty timed repetitions for each seed-1 task, backend, and method, after two warmups; 800 runs in total. Method order is randomized. Budget 512 and weight-bit limit 2.5 are fixed. Timings include queried forwards, caching, weight copies, and score computation, but exclude object setup and shared bank construction.

3.6 Execution cost and storage

Table 4 uses live neural evaluations rather than table lookup. Output scoring and halving are slower than scalar scoring on these CPUs. Shared bank construction has median time 1.70 ms for rounding and 50.35 ms for the GPTQ update across the five seed-1 tasks. These small-model timings do not predict large-model throughput.

For 100 selected GPTQ models, one per checkpoint, granularity, and storage constraint at budget 512, the recorded packing checks recover the expected weights and test predictions and meet the payload constraint. The format stores integer codes and float32 affine parameters, then decodes to floating-point tensors for inference. Together with 100 rounding files, these checks address storage accounting rather than integer-kernel acceleration.

4 What the measurements omit

4.1 Interaction approximations and comparison error

For an allocation changing coordinates J from a^0, let $m = |J|$ and let a_T retain only changes in $T \subseteq J$. An order-r profile changes at most r tensors. For a scalar or vector function h, its anchored approximation for $r < m$ is

$$S_r[h](a) = \sum_{\substack{T \subseteq J \\ |T| \le r}} (-1)^{r-|T|} \binom{m - |T| - 1}{r - |T|} h(a_T). \tag{7}$$

For $r \ge m$, it is $h(a)$. Equation (7) is the dimension-reduction formula of Xu and Rahman [11], Kuo et al. [12], Rahman [13]. The number of measured profiles is $P_r = \sum_{k=0}^{\min(r,G)} \binom{G}{k}(q-1)^k$ for a q-element precision menu.

For a fixed pair, let c be the difference of the two coefficient vectors in (7), Y the shared-input profile losses, and $\Delta = R(a) - R(b)$. The ordinary bias–variance identity gives

$$\mathbb{E}[(c^\top \overline{Y} - \Delta)^2] = (c^\top \mathbb{E}Y - \Delta)^2 + \frac{c^\top \Sigma c}{n}, \qquad \Sigma = \mathrm{Cov}(Y). \tag{8}$$

Only the second term decreases with repeated measurements. Coefficients shared by the two allocations cancel. In particular, if they differ only in tensor i, then $S_r[R](a) - S_r[R](b) = S_{r-1}[g_i](u)$, where u contains the common other choices and $g_i(u) = R(a_i, u) - R(b_i, u)$. This is a direct application of the

projection form of the anchored decomposition [12, Theorem 2.1]. The supplementary material gives the coefficient counts used in the experiment.

Writing $A = \|w\|_1$ for one extrapolated score, bounded profile losses imply variance at most $L^2A^2/(4n)$ and the usual bounded-mean sample complexity [17, 18]. The supplementary material gives network examples attaining the extremal profile laws while their mean risk is exactly $S_r[R]$. Correct mean structure therefore does not preclude difficult estimation. Equation (8), rather than the worst-case norm alone, is the diagnostic used on trained networks.

4.2 Two limits within quantized networks

Equal singleton squared losses can conceal opposite signed errors. Three parallel branches with errors $\pm 2/15$ have the same individual loss, while their joint errors can cancel or reinforce. Observing their outputs resolves that example. A stronger construction shows why output access still does not justify a global optimum from an arbitrary restricted profile set.

Theorem 1. *For integers $G \geq 2$, $0 \leq r < G$, and $0 < \lambda \leq L$, affine quantization with menu $\{2, 4\}$ admits a family of ReLU networks with G three-weight branches and a three-unit unquantized readout. Their teacher outputs and every order-r profile output coincide. A common storage bound leaves $M_r = \sum_{j=r+1}^{G} \binom{G}{j}$ feasible allocations; each network has one of them at loss zero and all others at loss λ. For integer query and list limits $T \geq 0$ and $K \geq 1$, rules observing only those profiles, costs, and T distinct complete allocations have minimax expected regret*

$$\lambda \left(1 - \frac{T+1}{M_r}\right)_+ .$$

The largest worst-case probability that a returned list of at most K candidates contains the optimum is $\min\{1, (T+K)/M_r\}$.

The supplementary material realizes an unstructured search within this network family. The result excludes access to hidden weights or other unmeasured information and uses real-arithmetic quantization. It concerns this observation model, not typical trained-network behavior.

Proposition 2. *For every $G \geq 2$, there is an affine-quantized binary classifier with G three-weight branches and menu $\{2, 4\}$ whose teacher-KL loss has a nonzero order-G anchored interaction, although its complete outputs are exactly reconstructed by* (4).

The supplementary material gives the construction. Thus high scalar interaction order alone does not imply that high-order output measurements are necessary. Conversely, a serial network need not have additive outputs. Standard smoothness gives

$$|\ell(a, x) - \widetilde{\ell}(a, x)| \leq \min\{L, \tfrac{\beta}{2}(\|e\| + \|\widetilde{e}\|)\|e - \widetilde{e}\|\}, \tag{9}$$

where $e = z_a - z_f$, $\widetilde{e} = \widetilde{z}_a - z_f$, and $\beta = 1/2$ for teacher KL or 2 for squared loss [19]. If the anchor is the teacher and scaled perturbations τ preserve activation and pooling choices, the omitted product terms are $O(\tau^2)$, giving $O(\tau^3)$ loss discrepancy. The supplementary material specifies the norm bound and examples attaining the cubic order or violating it at a ReLU boundary. The teacher anchor differs from the all-four-bit reference used in selection.

5 Discussion

Calibration quality depends on what is measured, how it is represented, and how much error remains after quantization. Output vectors can recover interactions hidden by scalar losses without extra forward evaluations, but reconstruction can be biased and its sampling variance can be substantial. Scoring every candidate also has an arithmetic cost. The GPTQ and depth-series results qualify the large gain observed with rounding.

A fixed shortlist creates a separate limitation. For $\widehat{a} \in S$, regret splits into the gap between the best member of S and the global optimum, plus the gap from $\widehat{a}$ to the best member of S. Additional comparisons within S can affect only the latter. Under rounding, combining scalar and output proposals reduces the mean normalized omission term from 0.0276 to 0.0130; after the GPTQ update it changes from 0.0287 to 0.0270.

The study covers three datasets, 37 trained networks below 73,000 parameters, and the specified weight quantizers. The reconstruction bank uses 32 training inputs. Exhaustive allocation removes search error but does not scale to many tensors, and equal forward counts do not imply equal time. Larger pretrained networks, scalable output-based search, and integer inference kernels remain outside this evidence.

References

[1] Z. Yao, Z. Dong, Z. Zheng, A. Gholami, J. Yu, E. Tan, L. Wang, Q. Huang, Y. Wang, M. Mahoney, and K. Keutzer. HAWQ-V3: Dyadic neural network quantization. In *International Conference on Machine Learning*, PMLR 139:11875–11886, 2021.

[2] I. Hubara, Y. Nahshan, Y. Hanani, R. Banner, and D. Soudry. Accurate post training quantization with small calibration sets. In *International Conference on Machine Learning*, PMLR 139:4466–4475, 2021.

[3] Z. Deng, S. Sharify, X. Wang, and M. Orshansky. Mixed-precision quantization for deep vision models with integer quadratic programming. In *62nd ACM/IEEE Design Automation Conference*, 2025.

[4] J. Zhao, A. Derakhshan, J. Hyman, J. Dong, S. Abdu Jyothi, and I. Harris. CoopQ: Cooperative game inspired layerwise mixed precision quantization for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7566–7578, 2026.

[5] A. Misra, M. Agrawal, A. Jain, and T. Abdelzaher. MixQuant: Adaptive mixed-precision quantization for large language models. *arXiv:2607.23047*, version 1, 2026.

[6] 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*, 2021.

[7] 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*, 2023.

[8] 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*, vol. 6, 2024.

[9] M. Williams and N. Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 10100–10118, 2024.

[10] I. F. Shihab, S. Akter, and A. Sharma. Coverage-based calibration for post-training quantization via weighted set cover over outlier channels. *arXiv:2604.24008*, version 1, 2026.

[11] H. Xu and S. Rahman. A generalized dimension-reduction method for multidimensional integration in stochastic mechanics. *International Journal for Numerical Methods in Engineering*, 61(12):1992–2019, 2004.

[12] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79(270):953–966, 2010.

[13] S. Rahman. Approximation errors in truncated dimensional decompositions. *Mathematics of Computation*, 83(290):2799–2819, 2014.

[14] Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, PMLR 28(3):1238–1246, 2013.

[15] K. Jamieson and A. Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *International Conference on Artificial Intelligence and Statistics*, PMLR 51:240–248, 2016.

[16] E. M. Saad, G. Blanchard, and N. Verzelen. Covariance-adaptive best arm identification. In *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[18] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

[19] D. Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] N. P. Pandey, M. Nagel, M. van Baalen, Y. Huang, C. Patel, and T. Blankevoort. A practical mixed precision algorithm for post-training quantization. In *BMVC Workshop on Computational Aspects of Deep Learning*, 2023.

[22] A. LeVi, R. Lapid, R. Himelstein, C. Baskin, R. Shwartz Ziv, and A. Mendelson. You had one job: Per-task quantization using LLMs' hidden representations. In *ICML Workshop on Resource-Adaptive Foundation Model Inference*, 2026. arXiv:2511.06516, version 4.

[23] Z. Yao, H. Zhao, H. Wang, and X. Han. GAMMA: Global bit allocation for mixed-precision models under arbitrary budgets. *arXiv:2605.18475*, version 2, 2026.

[24] M. Kim, J. Lee, J. Kim, J. Yun, Y. Kwon, and U Kang. LampQ: Towards accurate layer-wise mixed precision quantization for vision transformers. *arXiv:2511.10004*, version 2, 2025. Author manuscript for AAAI 2026.