DISTRIQUANT: BEYOND RECONSTRUCTION FOR ROBUST BINARY LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) for large language models (LLMs) is commonly formulated as a reconstruction problem, where quantized weights are optimized to reproduce full-precision behavior on a small calibration set. We show that, in the 1-bit regime, reconstruction plays a more consequential role: it implicitly determines *which representation directions survive quantization*. Since reconstruction error places greater emphasis on directions with high variance on the calibration distribution, binary PTQ preferentially preserves calibration-dominant structure while weak directions receive substantially less protection. Crucially, directions that are weak on the calibration distribution are not necessarily unimportant elsewhere. We find that many become relatively more prominent on unseen domains, yet are disproportionately suppressed after binarization. To quantify this effect, we introduce a directional survivor rate that measures how much of each full-precision representation direction remains after quantization. Motivated by this finding, we propose **DistriQuant**, a spectrum-balanced approach to salient-mask selection for binary PTQ. DistriQuant uses *Covariance-Balanced Alignment* to remove the variance imbalance among calibration directions before comparing full-precision and quantized representations, preventing dominant calibration directions from overwhelming the allocation of limited binary capacity. The method requires only the standard calibration set, leaves the underlying binary representation unchanged, and introduces no additional inference-time computation. Extensive experiments across different architectures and datasets demonstrate the effectiveness of our proposed method.

1 INTRODUCTION

Large language models (LLMs) achieve strong performance across a wide range of tasks, but their scale makes efficient inference increasingly expensive. Post-training quantization (PTQ) reduces this cost by compressing pretrained models using only a small calibration set and limited computation (Frantar et al., 2023; Ding et al., 2023; Sun et al., 2025). While modern PTQ methods can preserve model quality at moderate precision, pushing weights toward one bit remains considerably more challenging. Recent evidence further suggests that PTQ can be sensitive to the domain of its calibration data (Williams & Aletras, 2024; Liu et al., 2025). For example, (Liu et al., 2025) show that, under 3-bit GPTQ, replacing generic WikiText2 calibration by math-oriented calibration improves average reasoning accuracy by nearly 10%, with substantially larger gains on individual mathematical benchmarks. Prior work has primarily characterized this sensitivity or mitigated it through calibration-data selection and curation (Williams & Aletras, 2024). Here we ask a complementary question: *with the calibration distribution fixed, through what mechanism does reconstruction determine which representation directions survive extreme quantization?*

Most PTQ methods formulate quantization as a reconstruction problem. Specifically, each quantized layer is optimized to approximate its full-precision (FP) counterpart on calibration data. At moderate precision, this is an effective proxy for preserving model behavior. At one bit, however, reconstruction takes on an additional role. Binary weights have insufficient capacity to reproduce every FP representation direction equally well, so minimizing reconstruction error implicitly determines *which information receives the limited binary capacity*. Directions with large variance on the calibration distribution contribute strongly to reconstruction error and are therefore heavily protected, whereas errors along weaker directions incur a much smaller penalty.

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.

This preference would be harmless if directions that are weak on the calibration distribution were also unimportant elsewhere. We find that this is not the case. Although low-variance calibration directions often contribute less in absolute magnitude, many become *relatively more important* when the input distribution changes, including on mathematical reasoning and code data. We further observe that binary quantization preferentially suppresses these directions. To measure this effect, we introduce a *directional survivor rate*, which compares the contribution of each FP representation direction before and after quantization. Across transformer blocks, directions whose relative importance increases outside the calibration distribution exhibit systematically lower survivor rates, with the imbalance becoming particularly pronounced in deeper layers.

These observations reveal a limitation of reconstruction-based binary PTQ that is hidden by aggregate reconstruction error. A binary model may reproduce the dominant calibration structure well while discarding weaker directions inherited from pretraining. The resulting representation can therefore remain accurate on the calibration distribution yet lose information that becomes useful when the model encounters a different domain. Viewed from this perspective, the central problem is not simply that binarization introduces larger reconstruction error, but that the calibration distribution determines *where that error is allowed to occur*.

Motivated by this finding, we introduce **DistriQuant**, which changes how salient binary structure is selected. Our key idea is to prevent high-variance calibration directions from dominating the criterion used to allocate binary capacity. We propose *Covariance-Balanced Alignment* (CBA), which independently standardizes the FP and quantized representations before measuring their agreement. After this normalization, representation directions contribute on a comparable scale, allowing mask selection to account for activation structure that would otherwise be overshadowed by dominant calibration directions. DistriQuant uses CBA to select the salient masks while retaining the underlying binary quantization procedure. It requires no out-of-domain data and introduces no additional parameters or inference-time computation.

Extensive experiments across OPT, LLaMA-2/3, and Qwen models show that DistriQuant consistently improves perplexity and downstream accuracy over strong 1-bit PTQ baselines. Beyond aggregate performance, our analysis shows that the improvement is concentrated in the representation directions predicted by our diagnosis: compared with reconstruction-driven selection, DistriQuant better preserves directions whose relative importance increases beyond the calibration distribution. This provides a direct connection between the failure mode identified in our analysis, the design of the proposed method, and the resulting performance gains.

In summary, our contributions are:

- We show that reconstruction-based 1-bit PTQ induces a directional bias: as the error is weighted by calibration variance, the quantized model preferentially preserves calibration-dominant directions while weak directions receive less protection. We introduce a directional survivor rate to quantify which FP directions are retained after quantization.
- We uncover a systematic mismatch between calibration importance and cross-domain importance. Directions with small calibration variance can become relatively more prominent on unseen domains, yet these directions are disproportionately suppressed by binary quantization, especially in deeper transformer blocks.
- Motivated by this failure mode, we introduce **DistriQuant** with *Covariance-Balanced Alignment* for salient-mask selection. By removing the variance imbalance before comparing FP and quantized representations, DistriQuant allocates binary capacity without letting dominant calibration directions overwhelm the selection criterion. It uses only standard calibration data, incurs no additional inference-time cost, and consistently improves 1-bit quantization across multiple LLM families.

2 Related Work

Post-training quantization of LLMs. Post-training quantization (PTQ) compresses pretrained LLMs using a small calibration set, avoiding the cost of retraining or quantization-aware training. GPTQ (Frantar et al., 2023) uses approximate second-order information to compensate for errors during sequential weight quantization. AWQ (Lin et al., 2023) protects activation-salient weights, while SmoothQuant (Xiao et al., 2023) redistributes quantization difficulty between weights and activations. QuIP# and QuaRot (Tseng et al., 2024; Ashkboos et al., 2024) instead transform the

representation into a more quantization-friendly space. OAC (Edalati et al., 2025) incorporates model-output sensitivity into calibration and saliency estimation. Despite their different formulations, these approaches primarily seek to improve quantization fidelity on calibration data. Our work studies a complementary question that becomes critical under extreme compression: *which representation directions are retained when calibration fidelity cannot be preserved uniformly?*

Binary post-training quantization. Binary PTQ imposes a substantially tighter representation budget and has therefore motivated specialized quantization structures and saliency mechanisms. PB-LLM (Shang et al., 2023) preserves a small fraction of salient weights at higher precision while binarizing the remainder. BiLLM (Huang et al., 2024) combines salient-weight separation with residual binary approximation, while DBellQuant (Ye et al., 2025) reshapes weight distributions to reduce binarization error and activation outliers. ARB-LLM (Li et al., 2024) further introduces alternating refinement of binary values and scaling parameters and uses calibration activations to optimize structured binary representations. DistriQuant addresses a different source of error. Rather than increasing the expressiveness of the binary representation, we study how reconstruction allocates a *fixed* binary capacity. We show that its calibration-weighted objective preferentially protects dominant calibration directions while under-preserving directions that can become relatively more important on other domains. This observation motivates covariance-balanced salient-mask selection, which changes how the available binary capacity is allocated while leaving the underlying binary representation unchanged.

3 Diagnosing Calibration-Spectrum Bias

Binary PTQ is typically optimized by reconstructing the full-precision representation on a small calibration set. At moderate precision, the quantizer may retain sufficient capacity to approximate most representation directions simultaneously. At one bit, however, this is no longer possible: the reconstruction objective implicitly determines which directions receive the limited representational capacity of the binary model. We first expose this implicit allocation rule, then examine whether the directions weakly protected by reconstruction remain unimportant when the input distribution changes. For this analysis, NuminaMath(LI et al., 2024) and MBPP(Austin et al., 2021) are used only as diagnostic datasets but not for quantization.

3.1 Reconstruction Favors Dominant Calibration directions

Consider the FP activation matrix $X \in \mathbb{R}^{N\times d}$ collected on the calibration set, with centered form $X_c = X - \mathbf{1}\mu_X^\top$, where $\mu_X = X^\top \mathbf{1}/N$. Let its covariance eigendecomposition be: $C_{\text{cal}} = \frac{1}{N} X_c^\top X_c = V\Lambda V^\top$, where $V = [v_1, \ldots, v_d]$ is a matrix whose columns are eigenvectors and $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix containing the corresponding eigenvalues on the diagonal. For a full-precision weight W and its quantized counterpart $\widehat{W}$, define $\Delta W = W - \widehat{W}$. The average reconstruction error can be written as follows:

$$\mathcal{L}_{\text{rec}} = \tfrac{1}{N}\|X\Delta W^\top\|_F^2 = \|\mu_X^\top \Delta W^\top\|_2^2 + \textstyle\sum_{i=1}^d \lambda_i \|\Delta W v_i\|_2^2. \tag{1}$$

Eq. (1) reveals an implicit spectral weighting: the same weight error incurs a larger reconstruction penalty when it acts along a high-variance calibration direction. Consequently, under the severe capacity constraint of binary weights, dominant calibration directions exert substantially greater pressure on the quantized solution, while low-variance directions can be distorted at comparatively little reconstruction cost. This bias is not specific to a particular binary parameterization or optimization algorithm. It follows directly from the reconstruction objective. The relevant question is therefore whether weak calibration directions are genuinely dispensable.

3.2 Weak Calibration directions Can Be Important Under Distribution Shift

We next examine how the importance of the calibration eigenvectors v_i changes when the input distribution shifts, using V from Section 3.1 as a common reference basis. In our experiments, the calibration distribution is C4. For another input distribution D, let C_D denote the FP covariance of the same activations on D, and define the contribution score of direction v_i as

$$p_D(v_i) = {}^{v_i^\top C_D^{\text{FP}} v_i}/_{\text{Tr}(C_D^{\text{FP}})}. \tag{2}$$

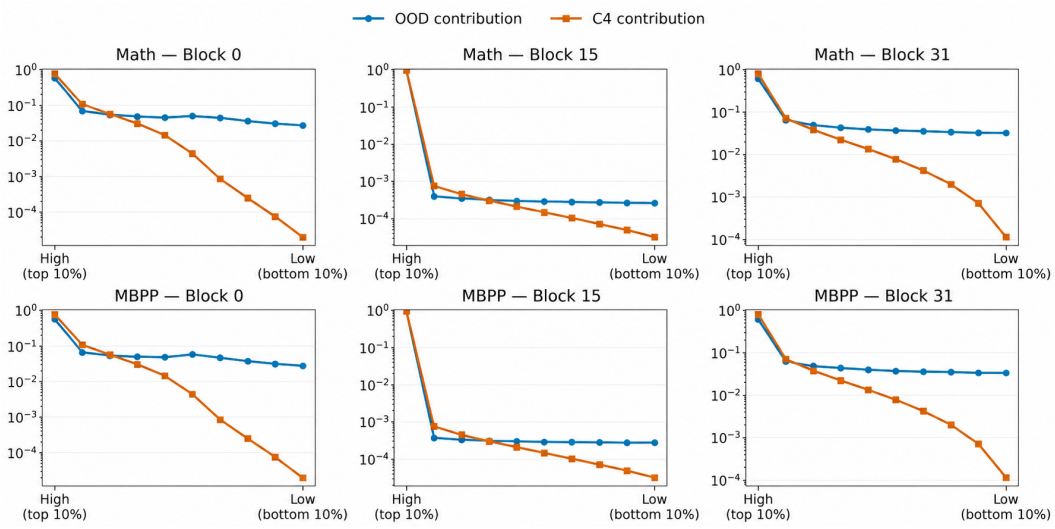


Figure 1: *Weak calibration directions are enriched under distribution shift.* Eigenvectors are grouped by their eigenvalues on the C4 calibration distribution. We report the contribution score defined in Eq. (2) on C4, NuminaMath and MBPP data. Directions toward the tail of the calibration spectrum exhibit substantially greater relative enrichment under distribution shift.

This quantity measures the fraction of the representation variance associated with v_i on distribution D. Absolute contribution alone, however, does not indicate whether a direction becomes more or less important after distribution shift. We therefore define its *cross-domain enrichment*

$$r_D(v_i) = {}^{(p_D(v_i)+\epsilon)}\!/_{(p_{\text{cal}}(v_i)+\epsilon)}, \tag{3}$$

where $\epsilon > 0$ is a small numerical constant. $r_D(v_i) > 1$ means that v_i occupies a larger fraction of the representation on D than on the calibration distribution.

Fig. 1 groups the calibration eigenvectors according to their eigenvalues and reports their contribution scores on shift domain distributions. As expected, high-eigenvalue calibration directions remain important in absolute magnitude. Their relative contribution, however, changes comparatively little across domains. In contrast, directions toward the tail of the calibration spectrum exhibit substantially stronger cross-domain enrichment.

Importantly, this observation does *not* imply that low-variance directions dominate the shifted representation in absolute magnitude. Rather, calibration variance systematically underestimates the relative importance of a subset of weak directions once the input distribution changes. A quantizer that allocates capacity solely according to calibration variance can therefore under-protect precisely the directions whose contribution grows outside the calibration distribution.

3.3 Binary Quantization Under-Preserves Enriched Directions

The preceding analysis identifies a mismatch between the importance weighting induced by reconstruction and the directions that can gain relative importance under distribution shift. We next ask whether this mismatch is actually reflected in the representation produced by a binary model. For an FP calibration eigenvector v_i, define its normalized contribution after quantization on distribution D as $p_D^Q(v_i) = v_i^\top C_D^Q v_i / \operatorname{Tr}(C_D^Q)$, where C_D^Q is the covariance of the same activations in the quantized model. We then define its *directional survivor rate*

$$s_D(v_i) = {}^{p_D^Q(v_i)}\!/_{p_D(v_i)}. \tag{4}$$

The normalization in Eq. 4 factors out global changes in activation scale. Thus, $s_D(v_i) < 1$ indicates that direction v_i occupies a smaller fraction of the quantized representation than of its FP counterpart, whereas $s_D(v_i) > 1$ indicates relative amplification.

Fig. 2 groups directions according to their cross-domain enrichment from Eq. 3. In early Transformer blocks, different groups are retained at relatively similar rates. As quantization error prop-

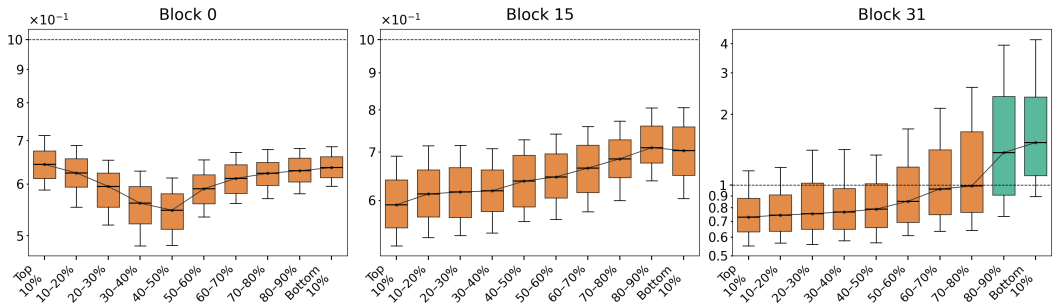


Figure 2: *Cross-domain-enriched directions survive binary quantization less reliably.* Directions are grouped according to their cross-domain enrichment in Eq. 3, and we report the directional survivor rate in Eq. 4 across Transformer blocks. Binary quantization increasingly favors calibration-dominant directions in deeper blocks, while weaker directions that are relatively more important for mathematical reasoning are suppressed more strongly.

agates through the network, however, a systematic imbalance emerges: directions with high cross-domain enrichment exhibit lower survivor rates, whereas calibration-dominant directions are preserved more strongly. The separation becomes particularly pronounced in later blocks and appears consistently on both mathematical-reasoning and code distributions.

This failure mode is more specific than a global reduction in activation variance. Binary quantization does not merely contract the representation uniformly; it changes its directional allocation. In particular, the directions weakly protected by calibration reconstruction are also the directions most likely to be suppressed when their relative importance increases under distribution shift.

Taken together, these results expose a structural weakness of reconstruction-based binary PTQ. Reconstruction allocates error according to the calibration covariance spectrum, yet weak directions can become relatively more important under distribution shift and are disproportionately suppressed by the resulting binary representation. This motivates a calibration-only allocation criterion that retains reconstruction fidelity while preventing the quantization process from being governed solely by dominant calibration directions.

4 METHOD

Let X_Q and X_{FP} respectively be the input of quantized model and full-precision model, $\widehat{W}$ and W_{FP} are respectively the quantized and full-precision weights. We can parameterize $\widehat{W}$ as

$$\widehat{W} = \mathrm{Diag}(\alpha_r)\, B\, \mathrm{Diag}(\alpha_c), \qquad B \in \{-1, +1\}^{d_{\text{out}} \times d_{\text{in}}}, \tag{5}$$

where α_r and α_c denote row- and column-wise scaling factors. Within each weight block, the parameters are partitioned into disjoint masks: salient subsets are represented using two binary components, whereas the remaining subsets use one. Mask selection therefore determines where the additional binary capacity is allocated. Our goal is to reduce the bias of this allocation toward high-variance calibration directions and better balance capacity across the representation spectrum.

4.1 COVARIANCE-BALANCED MASK SELECTION

Let $S_{\text{FP}} = X_{\text{FP}} W_{\text{FP}}^\top P$, and $S_Q = X_Q \widehat{W}^\top P$, where $P \in \mathbb{R}^{d_{\text{out}} \times r}$ has orthonormal columns and $r \ll d_{\text{out}}$. The projection P is used only to reduce the dimensionality of the representation statistics. We construct P by sampling a Gaussian random matrix and orthonormalizing its columns using reduced QR decomposition. The same fixed projection is used for all candidate structures.

For $a \in \{\text{FP}, Q\}$, define $\bar{S}_a = S_a - \mathbf{1}\mu_a^\top$, and $C_a = \bar{S}_a^\top \bar{S}_a / N$, where $\mu_a = S_a^\top \mathbf{1}/N$. Let $C_{\text{FP}} = U \Sigma U^\top$ with $U = [u_1, \ldots, u_r]$ and $\Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$. We assume $C_{\text{FP}} \succ 0$, i.e., $\sigma_i > 0$, so that $C_{\text{FP}}^{-1/2} = U \Sigma^{-1/2} U^\top$ is well defined. In practice, inverse square roots are regularized by a small ϵI. For an FP covariance eigenvector u_i, applying Eq. (4) to the projected outputs gives its

directional survivor rate as follows:

$$s_i = \frac{\mathrm{Tr}(C_{\mathrm{FP}})}{\mathrm{Tr}(C_Q)} \frac{u_i^\top C_Q u_i}{\sigma_i}. \tag{6}$$

Thus, $s_i = 1$ indicates that quantization preserves the relative contribution of FP direction u_i. The survivor rate, however, does not by itself compare the two models' outputs. To connect it to such a comparison, let $\alpha = \sqrt{\mathrm{Tr}(C_{\mathrm{FP}})/\,\mathrm{Tr}(C_Q)}$ and project both outputs onto u_i, rescaled by $\sqrt{\sigma_i}$. Then $\frac{1}{N}\|\bar{S}_{\mathrm{FP}} u_i/\sqrt{\sigma_i}\|_2^2 = 1$ and $\frac{1}{N}\|\alpha \bar{S}_Q u_i/\sqrt{\sigma_i}\|_2^2 = s_i$. In other words, s_i is the mean-square of the rescaled quantized projection, while the FP counterpart has the unit mean-square. Measuring the deviation from $s_i = 1$ by $(\sqrt{s_i} - 1)^2$, the squared gap between their root-mean-square magnitudes, yields the bound in Proposition 1.

Proposition 1 (Survival-rate control)**.** *The survivor rates specified in Eq.* (6) *satisfy*

$$\tfrac{1}{r}\textstyle\sum_{i=1}^{r}\left(\sqrt{s_i} - 1\right)^2 \le \tfrac{1}{Nr}\sum_{i=1}^{r}\left\|\frac{\alpha \bar{S}_Q u_i - \bar{S}_{\mathrm{FP}} u_i}{\sqrt{\sigma_i}}\right\|_2^2 = \tfrac{1}{Nr}\left\|\left(\alpha \bar{S}_Q - \bar{S}_{\mathrm{FP}}\right) C_{\mathrm{FP}}^{-1/2}\right\|_F^2. \tag{7}$$

The middle term in (7) compares the FP outputs projected onto u_i, rescaled to unit variance, with the quantized outputs projected onto the same direction. Matching these projections token by token, not only their variances, therefore controls survival rate. Moreover, since every direction is rescaled to unit variance, high-variance calibration directions carry no more weight than tail directions, unlike the reconstruction objective in Eq. (1). The proof is provided in Appendix A.

Eq. 7 motivates balancing the contribution of different representation directions. In block-wise quantization, however, preceding quantization changes the current-layer covariance, so generally $C_Q \neq C_{\mathrm{FP}}$. We therefore normalize the FP and quantized representations using their respective covariances: $Z_{\mathrm{FP}} = \bar{S}_{\mathrm{FP}}(C_{\mathrm{FP}} + \epsilon I)^{-1/2}$, $\quad Z_{\mathrm{Q}} = \bar{S}_{\mathrm{Q}}(C_{\mathrm{Q}} + \epsilon I)^{-1/2}$. This gives a practical surrogate for the exact directional-preservation objective rather than an equivalent reformulation. It recovers the exact directional target when the FP and quantized representations have matching normalized covariance:

$$C_Q^{-1/2} = \sqrt{\frac{\mathrm{Tr}(C_{\mathrm{FP}})}{\mathrm{Tr}(C_Q)}} C_{\mathrm{FP}}^{-1/2} = \alpha C_{\mathrm{FP}}^{-1/2}. \tag{8}$$

Proposition 1 motivates driving every s_i toward 1 by reducing the output mismatch in Eq. (7). We therefore use the surrogate mismatch $\frac{1}{Nr}\|Z_Q - Z_{\mathrm{FP}}\|_F^2$, which yields the *covariance-balanced alignment* loss:

$$\mathcal{L}_{\mathrm{CBA}} = \tfrac{1}{2Nr}\|Z_Q - Z_{\mathrm{FP}}\|_F^2 = 1 - \tfrac{1}{r}\textstyle\sum_{j=1}^{r}\cos(Z_{Q,:,j}, Z_{\mathrm{FP},:,j}). \tag{9}$$

Using CBA for mask selection.

When identifying salient parameters within a weight block, we represent a selection by a binary mask $M \in \mathcal{M}$, where $\mathcal{M}$ denotes the set of masks considered by the search procedure. Let $H = X^\top X$ denote the calibration Hessian, and let U be the upper-triangular Cholesky factor of its regularized inverse, $U^\top U = (H + \lambda I)^{-1}$. For each input column j, the aggregated saliency score is defined as

$$s_j = \frac{\sum_i W_{ij}^2}{U_{jj}^2}. \tag{10}$$

We rank the input columns in descending order of s_j. Each considered prefix of this ranking defines a candidate mask: the top-t ranked columns form one subset and the remaining columns form its complement, with the two subsets pseudo-quantized independently. The resulting candidate masks constitute $\mathcal{M}$. We then select the candidate mask with the best covariance-balanced alignment,

$$M^\star = \arg\min_{M \in \mathcal{M}} \mathcal{L}_{\mathrm{CBA}}(M). \tag{11}$$

Thus, the GPTQ-inspired column saliency score determines the ordering of candidate columns, while CBA determines the preferred partition among the candidate prefixes. All candidate masks quantize the same set of parameters using the same binary parameterization and differ only in their partition of the columns.

Hierarchical mask selection. For each weight block, we first search for a salient mask M_{sal} and its complement using the procedure above. Each of the two resulting parameter sets is then independently ranked and further partitioned using the same prefix-based search, producing four disjoint

Table 1: Perplexity (↓) results of OPT and LLaMA models on C4, WikiText2, and PTB. Separate weight-bit columns are reported for each model family. Green values in parentheses denote the absolute perplexity reduction of each enhanced method over its corresponding base method. Bold values indicate the best quantized result in each column.

Dataset	Method	OPT	OPT			LLaMA	LLaMA		
		Bits	1.3B	2.7B	6.7B	Bits	LLaMA-2 7B	LLaMA-2 13B	LLaMA-3 8B
C4	Full Precision	16	16.07	14.34	12.71	16	7.26	6.73	9.45
	PB-LLM	1.7	168.12	222.15	104.78	1.7	80.69	184.67	104.15
	BiLLM	1.11	64.14	44.77	42.13	1.08	39.38	25.87	61.04
	OAC	1.11	–	–	–	1.08	21.64	15.25	–
	DBellQuant	1.11	42.86	33.19	21.75	1.08	21.83	15.14	–
	ARB-X	1.11	47.60	34.97	22.54	1.08	28.02	19.82	41.86
	ARB-X + Ours	1.11	35.39(↓12.21)	29.18(↓5.79)	20.19(↓2.35)	1.08	13.39(↓14.63)	11.37(↓8.45)	25.24(↓16.62)
	ARB-RC	1.11	27.70	21.46	16.97	1.08	20.40	14.77	36.04
	ARB-RC + Ours	1.11	**22.99**(↓4.71)	**18.82**(↓2.64)	**15.50**(↓1.47)	1.08	**12.14**(↓8.26)	**10.07**(↓4.70)	**21.22**(↓14.82)
WikiText2	Full Precision	16	14.62	12.47	10.86	16	5.47	4.88	6.14
	PB-LLM	1.7	239.81	278.27	144.25	1.7	66.41	236.40	73.08
	BiLLM	1.11	69.05	48.61	47.65	1.08	32.31	21.35	55.80
	OAC	1.11	–	–	–	1.08	19.50	13.15	–
	DBellQuant	1.11	43.42	31.47	18.89	1.08	17.91	12.79	–
	ARB-X	1.11	45.40	34.37	20.07	1.08	21.61	14.86	31.98
	ARB-X + Ours	1.11	32.65(↓12.75)	26.63(↓7.74)	17.53(↓2.54)	1.08	10.21(↓11.40)	8.42(↓6.44)	18.03(↓13.95)
	ARB-RC	1.11	26.40	19.84	14.92	1.08	16.25	12.47	27.42
	ARB-RC + Ours	1.11	**21.67**(↓4.73)	**16.85**(↓2.99)	**13.38**(↓1.54)	1.08	**9.68**(↓6.57)	**7.67**(↓4.80)	**15.68**(↓11.74)
PTB	Full Precision	16	20.29	17.97	15.77	16	37.91	50.93	11.18
	PB-LLM	1.7	324.62	183.97	169.49	1.7	657.24	816.31	106.25
	BiLLM	1.11	115.94	88.52	69.41	1.08	5243.01	309.12	87.25
	ARB-X	1.11	71.96	54.28	31.23	1.08	783.66	208.06	62.69
	ARB-X + Ours	1.11	54.32(↓17.64)	43.66(↓10.62)	26.56(↓4.67)	1.08	**158.16**(↓625.50)	**99.14**(↓108.92)	26.47(↓36.22)
	ARB-RC	1.11	43.03	31.77	22.31	1.08	763.19	197.70	47.88
	ARB-RC + Ours	1.11	**34.40**(↓8.63)	**26.22**(↓5.55)	**19.85**(↓2.46)	1.08	308.26(↓454.93)	138.43(↓59.27)	**24.45**(↓23.43)

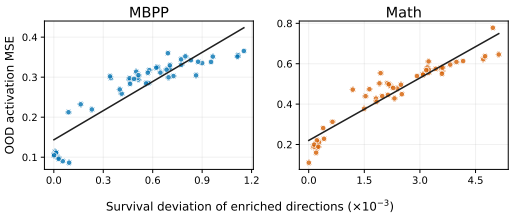
masks in total. At each mask-selection step, we retain the K candidate prefixes and select the one with the lowest CBA loss according to Eq. 11. Once the four masks are determined, the binary values and scaling factors within each mask are optimized group-wise.

5 EXPERIMENTS

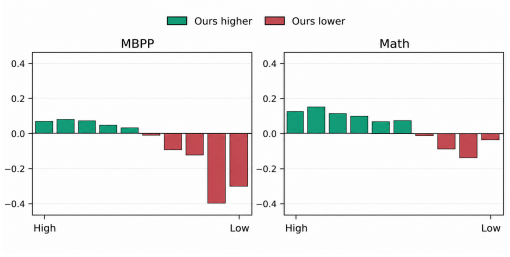
Models and datasets. We evaluate DISTRIQUANT across multiple model families and generations, including OPT-1.3B/2.7B/6.7B (Zhang et al., 2022), LLaMA-2-7B/13B (Touvron et al., 2023), LLaMA-3-8B (Grattafiori et al., 2024), and Qwen2-7B, Qwen2.5-7B, and Qwen3-8B. Language-modeling quality is evaluated by perplexity on C4 (Raffel et al., 2020), WikiText2 (Merity et al., 2016), and PTB (Marcus et al., 1993). We additionally report zero-shot performance on ARC-Easy and ARC-Challenge (Clark et al., 2018), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), OpenBookQA (Mihaylov et al., 2018), COPA(Gordon et al., 2012), and LAMBADA (Paperno et al., 2016). To further assess capabilities beyond conventional language modeling, we evaluate recent Qwen models on challenging reasoning and knowledge benchmarks, including GSM8K(Cobbe et al., 2021) and MMLU(Hendrycks et al., 2021). Importantly, all models are calibrated exclusively on C4; no samples from the evaluation benchmarks are used during quantization.

Baselines. We compare against representative 1-bit PTQ methods, including PB-LLM (Shang et al., 2023), BiLLM (Huang et al., 2024), ARB-LLM (Li et al., 2024), OAC, and DBellQuant (Ye et al., 2025). We follow their reported calibration and GPTQ-style error-compensation settings as closely as possible. ARB-RC serves as our primary reconstruction-based baseline since it provides the strongest overall performance among the ARB variants. For controlled comparison, we primarily consider 1-bit PTQ methods that retain the standard binary-weight representation without introducing additional representation transforms .

Experimental setting. Unless otherwise specified, all experiments are implemented in PyTorch and run on a single NVIDIA A100 GPU. Following the standard calibration protocol adopted by prior PTQ methods such as GPTQ (Frantar et al., 2023) and BiLLM (Huang et al., 2024), we randomly sample 128 sequences of length 2048 from C4 as calibration data. We use a quantization block size of 128 throughout all experiments. The quantization variables are updated for 15 iterations,



(a) *Preserving cross-domain-enriched directions is associated with lower shifted-domain reconstruction error.* Each point denotes a partially quantized model with similar quantization budget but a different structural allocation. The x-axis reports the mean survival-rate deviation from Eq. 7, computed over five bins spanning the top 50% most cross-domain-enriched eigenvectors, while the y-axis reports output error on the corresponding shifted domain.



(b) *Covariance-balanced allocation preferentially improves the survival of cross-domain-enriched directions.* We report the change in directional survivor rate relative to ARB-RC. Positive values indicate stronger preservation by our method. Across MBPP and NuminaMath, the gain is concentrated in directions with high cross-domain enrichment, while reduced retention occurs primarily in calibration-dominant directions.

Figure 3: *Analysis of cross-domain-enriched directions.* (a) Preserving cross-domain-enriched directions is associated with lower shifted-domain reconstruction error. (b) DistriQuant preferentially improves the survival of these directions relative to ARB-RC.

following the setting of ARB-LLM (Li et al., 2024). Consistent with the compared baselines, quantization is applied only to model weights, while activations remain in full precision. For the CBA mask selection procedure, we use $K = 20$ across all settings. Since the selected prefix can vary across groups, K bounds the candidate range and is chosen such that the resulting average bit-width remains comparable to the underlying baseline. For DistriQuant, we use a projection dimension of $r = 128$ throughout all experiments.

5.1 Main Results

Results on OPT and LLaMA. Table 1 reports perplexity results across OPT and LLaMA models on C4, WikiText2, and PTB. Our method achieves the **best performance across all 18 configurations** among the compared 1-bit PTQ methods, demonstrating consistent gains across model families, parameter scales, and evaluation distributions. The gains are particularly pronounced on the LLaMA family. On C4 and WikiText2, DistriQuant improves the strongest competing results from $20.40/16.25$ to $12.14/9.68$ on LLaMA-2-7B, and from $36.04/27.42$ to $21.22/15.68$ on LLaMA-3-8B. For PTB, the perplexity of LLaMA-2-7B drops from 763.19 to 308.26, while LLaMA-3-8B improves from 47.88 to 24.45, nearly halving perplexity. The advantage of DistriQuant extends beyond perplexity. As shown in Table 2, our method achieves the **highest accuracy on every one of the 24 configurations** across the eight zero-shot QA benchmarks. At the same precision, the average accuracy improves over the strongest competing baseline by **7.29**, **7.30**, and **11.17** percentage points on LLaMA-2-7B, LLaMA-2-13B, and LLaMA-3-8B, respectively.

Generalization to Qwen architectures and reasoning tasks. Table 3 shows that DistriQuant generalizes consistently across recent Qwen architectures and evaluation regimes. Across Qwen2-7B, Qwen2.5-7B, and Qwen3-8B, DistriQuant improves over ARB-RC on every reported metric, simultaneously reducing perplexity on C4, WikiText2, and PTB and improving zero-shot QA, GSM8K, and MMLU accuracy. The gains are particularly pronounced on Qwen3-8B: C4/PTB perplexity decreases from 32.71/44.56 to 24.38/30.22, AveQA improves from 53.56 to 58.53 (+4.97 points), and GSM8K increases from 14.18 to 39.65 (+25.47 points). Improvements on MMLU are also consistent across all three Qwen generations. These results suggest that DistriQuant does not merely improve reconstruction-oriented metrics, but better preserves capabilities that transfer across architectures, domains, and reasoning-intensive tasks under binary quantization. Detailed category-level MMLU results are provided in Appendix B.

5.2 Ablation and Mechanistic Analysis

Does spectrum-balanced mask selection preserve cross-domain-enriched directions? Our analysis in Sec. 3 shows that reconstruction-based binary PTQ preferentially suppresses directions whose

Table 2: Zero-shot accuracy (%) on commonsense reasoning benchmarks. Green values in parentheses denote the absolute improvement of our method over ARB-RC in percentage points. †The average is computed over the eight reported tasks.

Model	Method	Bits	Lambada	PIQA	OBQA	WinoGrande	ARC-E	ARC-C	HellaSwag	COPA	Avg.†
LLaMA-2-7B	ARB-RC	1.08	51.66	64.85	19.40	59.51	50.97	24.32	36.79	71.00	47.31
	DBELL	1.08	41.63	64.85	17.60	56.91	50.38	23.98	35.65	63.00	44.25
	ARB-RC + Ours	1.08	**66.76** (+15.10)	**70.78** (+5.93)	**21.40** (+2.00)	**60.85** (+1.34)	**64.94** (+13.97)	**30.63** (+6.31)	**44.43** (+7.64)	**77.00** (+6.00)	**54.60** (+7.29)
LLaMA-2-13B	ARB-RC	1.08	66.02	69.59	23.20	61.48	61.28	29.18	38.73	81.00	53.81
	DBELL	1.08	55.75	68.44	20.80	61.64	60.10	28.16	39.08	78.00	51.50
	ARB-RC + Ours	1.08	**75.08** (+9.06)	**74.86** (+5.27)	**28.40** (+5.20)	**68.03** (+6.55)	**69.82** (+8.54)	**36.35** (+7.17)	**48.31** (+9.58)	**88.00** (+7.00)	**61.11** (+7.30)
LLaMA-3-8B	ARB-RC	1.08	48.85	61.43	20.20	57.46	45.83	22.01	35.39	76.00	45.90
	DBELL	1.08	34.78	59.30	17.80	54.85	36.78	18.86	33.40	73.00	41.10
	ARB-RC + Ours	1.08	**69.32** (+20.47)	**72.69** (+11.26)	**25.60** (+5.40)	**64.80** (+7.34)	**66.04** (+20.21)	**31.48** (+9.47)	**44.64** (+9.25)	**82.00** (+6.00)	**57.07** (+11.17)

Table 3: Perplexity (↓) on C4, WikiText2, and PTB, average zero-shot QA accuracy, GSM8K accuracy, and MMLU accuracy (↑) for Qwen2-7B, Qwen2.5-7B, and Qwen3-8B. Green annotations report the absolute improvement of our method over ARB-RC.

Model	Method	Perplexity (↓)			Accuracy (↑)		
		C4	WikiText2	PTB	AveQA	GSM8K	MMLU
Qwen2-7B	ARB-RC	22.63	16.99	28.33	53.95	7.43	40.96
	DBellQuant	29.03	28.52	47.27	47.01	2.35	26.27
	ARB-RC + Ours	**18.15** (↓4.48)	**13.51** (↓3.48)	**22.86** (↓5.47)	**56.89** (+2.94)	**17.66** (+10.23)	**46.47** (+5.51)
Qwen2.5-7B	ARB-RC	24.32	17.08	29.47	54.79	11.45	43.36
	DBellQuant	29.52	25.10	43.99	49.02	2.58	29.26
	ARB-RC + Ours	**19.65** (↓4.67)	**14.12** (↓2.96)	**25.98** (↓3.49)	**57.14** (+2.35)	**21.53** (+10.08)	**45.90** (+2.54)
Qwen3-8B	ARB-RC	32.71	26.03	44.56	53.56	14.18	49.68
	DBellQuant	50.88	58.16	86.66	43.84	1.44	39.18
	ARB-RC + Ours	**24.38** (↓8.33)	**22.67** (↓3.36)	**30.22** (↓14.34)	**58.53** (+4.97)	**39.65** (+25.47)	**55.85** (+6.17)

relative importance increases under distribution shift. We first examine whether the proposed allocation directly mitigates this failure mode. Figure 3b reports the change in directional survivor rate relative to ARB-RC at the final Transformer block. Across mathematical reasoning and code distributions, the improvement is concentrated in directions with high cross-domain enrichment. Conversely, the small reduction in retention occurs predominantly among directions whose importance is concentrated on the calibration distribution. Thus, covariance-balanced allocation does not simply increase representation magnitude uniformly; it redistributes the limited binary capacity toward directions that are systematically under-protected by reconstruction.

Does preserving enriched directions matter for shifted-domain fidelity? The previous result shows that our method changes *which* representation directions survive quantization, but does not establish whether this change is related to fidelity under distribution shift. We therefore construct controlled partially quantized models with same quantization budgets but different structural allocations. For a fixed layer, each sampled model quantizes 80% of the weight columns while leaving the remainder in full precision. For each shifted domain (NuminaMath and MBPP), we use two disjoint subsets: one to estimate the survival of the most cross-domain-enriched directions, and a held-out subset to measure output reconstruction error. Figure 3a shows a consistent correlation: models with smaller survival-rate deviation for enriched directions also exhibit lower reconstruction error on both mathematical and code data. This indicates that *where* quantization error is allocated across representation directions is closely associated with shifted-domain fidelity. Together with Fig. 3b, this connects the directional failure mode identified in Sec. 3 to the behavior induced by our method.

6 Conclusion

In this paper, we identified a directional bias in reconstruction-based binary PTQ: calibration-dominant directions are preferentially preserved, while weaker directions can be disproportionately suppressed under distribution shift. Motivated by this observation, we proposed DistriQuant, which combines reconstruction-constrained mask selection with covariance-balanced alignment to better preserve under-protected representation directions. Experimental results demonstrate the effectiveness of our method across different architectures and datasets.

AI USE STATEMENT

In this work, we used generative AI tools to assist with aspects of implementation and language editing, including improving the clarity and presentation of the manuscript and suggesting improvements to scientific visualizations and figure presentation. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text and visualization suggestions were reviewed and revised to ensure that they accurately reflect the underlying methods and experimental results. We did not rely on generative AI to generate experimental results or make final methodological decisions. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *NeurIPS*, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34(05), pp. 7432–7439, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Xin Ding, Xiaoyu Liu, Zhijun Tu, Yun Zhang, Wei Li, Jie Hu, Hanting Chen, Yehui Tang, Zhiwei Xiong, Baoqun Yin, et al. Cbq: Cross-block quantization for large language models. *arXiv preprint arXiv:2312.07950*, 2023.

Ali Edalati, Alireza Ghaffari, Mahsa Ghazvini Nejad, Lu Hou, Boxing Chen, Masoud Asgharian, and Vahid Partovi Nia. Oac: Output-adaptive calibration for accurate post-training quantization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16453–16461, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i16.33807. URL http://dx.doi.org/10.1609/aaai.v39i16.33807.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *ICLR*, 2023.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret (eds.), **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*, 2024.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath, 2024.

Zhiteng Li, Xianglong Yan, Tianao Zhang, Haotong Qin, Dong Xie, Jiang Tian, Zhongchao Shi, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Arb-llm: Alternating refined binarizations for large language models. *ArXiv*, abs/2410.03129, 2024. URL https://api.semanticscholar.org/CorpusID:273163233.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. Quantization hurts reasoning? an empirical study on quantized reasoning models, 2025. URL https://arxiv.org/abs/2504.04823.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Yuzhang Shang, Zhihang Yuan, Qiang Wu, and Zhen Dong. Pb-llm: Partially binarized large language models. *arXiv*, 2023.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, et al. Flatquant: Flatness matters for llm quantization. In *ICML*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *ICML*, 2024.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10100–10118, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.544. URL https://aclanthology.org/2024.acl-long.544/.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *ICML*, 2023.

Zijian Ye, Wei Huang, Yifei Yu, Tianhe Ren, Zhongrui Wang, and Xiaojuan Qi. Dbellquant: Breaking the bell with double-bell transformation for llms post training binarization. *ArXiv*, abs/2507.01027, 2025. URL https://api.semanticscholar.org/CorpusID:280080895.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv*, 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Proof of Proposition 1.

Proof. Let $g_i^{\mathrm{FP}} = \bar{S}_{\mathrm{FP}} u_i / \sqrt{\sigma_i}$ and $g_i^{Q} = \alpha\, \bar{S}_Q u_i / \sqrt{\sigma_i}$ denote the i-th columns of $\bar{S}_{\mathrm{FP}} U \Sigma^{-1/2}$ and $\alpha \bar{S}_Q U \Sigma^{-1/2}$, respectively. Since $\frac{1}{N}\|\bar{S}_{\mathrm{FP}} u_i\|_2^2 = u_i^\top C_{\mathrm{FP}} u_i = \sigma_i$ and $\frac{1}{N}\|\bar{S}_Q u_i\|_2^2 = u_i^\top C_Q u_i$, Eq. (6) gives

$$\|g_i^{\mathrm{FP}}\|_2 = \sqrt{N}, \qquad \|g_i^{Q}\|_2 = \sqrt{N s_i}.$$

By the reverse triangle inequality,

$$\left\| g_i^{Q} - g_i^{\mathrm{FP}} \right\|_2 \geq \left| \|g_i^{Q}\|_2 - \|g_i^{\mathrm{FP}}\|_2 \right| = \sqrt{N}\,|\sqrt{s_i} - 1|\,.$$

Squaring both sides gives

$$\frac{1}{N} \left\| g_i^{Q} - g_i^{\mathrm{FP}} \right\|_2^2 \geq (\sqrt{s_i} - 1)^2\,.$$

Averaging over the r covariance modes yields

$$\frac{1}{r}\sum_{i=1}^{r} (\sqrt{s_i} - 1)^2 \leq \frac{1}{Nr}\sum_{i=1}^{r} \left\| g_i^{Q} - g_i^{\mathrm{FP}} \right\|_2^2 = \frac{1}{Nr} \left\| \left(\alpha \bar{S}_Q - \bar{S}_{\mathrm{FP}}\right) U \Sigma^{-1/2} \right\|_F^2,$$

since the sum of squared column norms is the squared Frobenius norm. Finally, $C_{\mathrm{FP}}^{-1/2} = U \Sigma^{-1/2} U^\top$ and right multiplication by the orthogonal matrix $U^\top$ preserves the Frobenius norm, so $\|(\alpha \bar{S}_Q - \bar{S}_{\mathrm{FP}}) U \Sigma^{-1/2}\|_F = \|(\alpha \bar{S}_Q - \bar{S}_{\mathrm{FP}}) C_{\mathrm{FP}}^{-1/2}\|_F$, which proves Proposition 1. □

Ablation of covariance normalization. Table 4 examines the covariance normalization used during mask selection. The reconstruction-based baseline does not use directional alignment, while ***FP covariance*** applies the derived FP covariance normalization to both branches. CBA instead normalizes the quantized branch using its current covariance. Across the evaluated models and datasets, CBA consistently achieves the best performance, showing that covariance-balanced normalization provides a more effective criterion for binary mask selection. The FP covariance variant also verifies that directly enforcing the fixed FP normalization can be overly restrictive once quantization error has accumulated through preceding blocks.

Table 4: Ablation of covariance-balanced mask selection. We compare the reconstruction-based baseline, alignment using the FP covariance for both branches (*FP covariance*), and the proposed CBA using the current covariance for the quantized branch. Accuracy (%) is reported on eight commonsense reasoning benchmarks. The average is computed over all eight tasks.

Method	Bits	Lambada	PIQA	OBQA	WinoGrande	ARC-E	ARC-C	HellaSwag	COPA	Avg.
Baseline	1.08	51.66	64.85	19.40	59.51	50.97	24.32	36.79	71.00	47.31
FP covariance	1.08	63.89	70.62	20.40	60.93	59.09	27.99	43.79	65.00	51.46
CBA	1.08	**66.76**	**70.78**	**21.40**	**60.85**	**64.94**	**30.63**	**44.43**	**77.00**	**54.60**

Computational overhead. DistriQuant incurs approximately 50% additional offline quantization time relative to ARB-RC (approximately 1.5 versus 1.0 hours for LLaMA-2-7B). Since PTQ is performed only once before deployment, this additional cost remains modest in absolute terms. More importantly, DistriQuant introduces no inference-time overhead: the measured prefill and decoding throughput in Table 5 show no degradation in inference throughput.

B Additional Ablations

Additional results on MMLU

Robustness to the calibration distribution. The main experiments use C4 for calibration. To examine whether the benefit of DistriQuant depends on this particular calibration distribution, we additionally quantize OPT-1.3B using WikiText2 calibration data while keeping all other settings unchanged. As shown in Table 7, DistriQuant consistently improves over ARB-RC under both calibration distributions, indicating that its benefit is not specific to C4 calibration.

Table 5: Inference throughput (tokens/s) on the Intel Xeon Gold 6548Y+ CPU. Higher is better.

Model	Prefill	Decode
FP16	2.83	0.87
ARB-RC	11.44	14.15
Ours	12.96	14.96

Table 6: Cross-domain and mathematical reasoning performance (%) of binary-quantized Qwen models. GSM8K evaluates multi-step mathematical reasoning, while MMLU evaluates performance across diverse knowledge domains. Green values in parentheses denote absolute improvements of our method over ARB-RC in percentage points.

Model	Method	Bits	GSM8K	MMLU				
				Overall	STEM	Humanities	Social Sci.	Other
Qwen2-7B	ARB-RC	1.08	7.43	40.96	36.31	38.41	45.92	44.61
	DBELL	1.08	2.35	26.27	27.34	24.27	27.53	26.97
	ARB-RC + Ours	1.08	**17.66** (+10.23)	**46.47** (+5.51)	**39.96** (+3.65)	**42.74** (+4.33)	**53.43** (+7.51)	**51.82** (+7.21)
Qwen2.5-7B	ARB-RC	1.08	11.45	43.36	39.49	38.87	49.85	47.67
	DBELL	1.08	2.58	29.26	30.61	26.78	30.32	30.61
	ARB-RC + Ours	1.08	**21.53** (+10.08)	**45.90** (+2.54)	**42.97** (+3.48)	**41.62** (+2.75)	**51.38** (+1.53)	**49.92** (+2.25)
Qwen3-8B	ARB-RC	1.08	14.18	49.68	47.35	44.12	58.14	52.08
	DBELL	1.08	1.44	39.18	35.59	35.69	45.63	41.71
	ARB-RC + Ours	1.08	**39.65** (+25.47)	**55.85** (+6.17)	**51.25** (+3.90)	**49.48** (+5.36)	**66.66** (+8.52)	**59.45** (+7.37)

Sensitivity to the projection dimension. We further study the sensitivity of CBA to the projection dimension r on OPT-1.3B. We vary r while keeping all remaining quantization settings fixed. Table 8 shows that the method remains effective across a broad range of projection dimensions. Performance is relatively insensitive over r. We use $r = 128$ as the default throughout the main experiments.

Table 7: Robustness to the calibration distribution on OPT-1.3B. Both methods use the same calibration samples and quantization settings within each calibration domain. Perplexity is reported on C4, WikiText2, and PTB.

Calibration	Method	C4	WikiText2	PTB
C4	ARB-RC	27.70	26.40	43.03
	ARB-RC + Ours	**22.99**	**21.67**	**34.40**
WikiText2	ARB-RC	31.53	32.92	44.95
	ARB-RC + Ours	**26.78**	**20.78**	**34.56**

Table 8: Sensitivity to the projection dimension r on OPT-1.3B. All models are calibrated on C4, with all other quantization settings fixed. Perplexity is reported on C4, WikiText2, and PTB.

r	C4	WikiText2	PTB
32	22.94	21.45	33.31
64	22.98	21.63	32.58
128	22.99	21.67	34.40
256	23.73	22.12	34.36