RECAC: RESIDUAL-GUIDED CALIBRATION VIA ACTIVATION COVERAGE FOR LLM QUANTIZATION

Anonymous authors
Paper under double-blind review

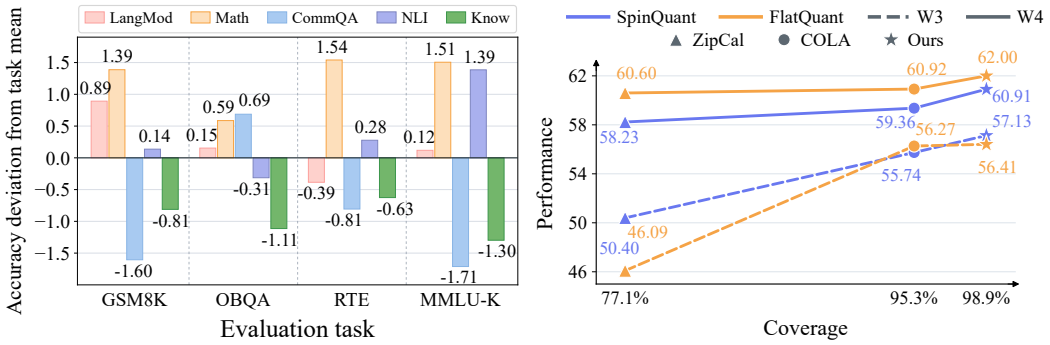
ABSTRACT

As a crucial technique for efficient deployment of large language models (LLMs), post-training quantization (PTQ) performs low-bit quantization for LLM relying on a pre-defined calibration dataset. Logically, the calibration dataset determines quantization parameters and so clearly influences PTQ performance. However, effective calibration construction is still underexplored. Particularly, existing methods typically focus on selecting representative and task-related samples while overlooking the complementarity among selected samples, which hardly guarantees coverage of the calibration dataset and limits the generalization ability of quantized models. Therefore, this paper proposes a Residual-guided Calibration via activation Coverage (ReCaC) method, aiming to construct an effective calibration dataset by maximizing the sample coverage in the activation space under a small fixed budget. Specifically, given a pre-defined benchmark for constructing the calibration dataset, our ReCaC uses residual-guided calibration construction (RGC) to progressively seek and pick the samples with the smallest residual score after being added to the currently selected calibration set, which effectively prioritizes complementary activation information and maximizes calibration coverage in activation space. Furthermore, ReCaC can be flexibly extended to multi-domain scenarios by joint global and domain coverage (JGDC), which selects the samples that provide complementary contributions to the overall activation space while maintaining sufficient coverage for each domain. As such, ReCaC can capture complementary activation patterns among heterogeneous domains and so enlarge the coverage of calibration dataset. Experiments are conducted by using different language models across various post-training quantization methods and settings, and results on a dozen downstream tasks demonstrate ReCaC achieves highly competitive or superior performance compared with existing state-of-the-art calibration methods in both iintra-domain, inter-domain and out-of-domain downstream tasks, particularly under the setting of lower-bit quantization.

1 INTRODUCTION

The growing scale of large language models (LLMs) leads to substantial memory and computational costs, posing significant challenges to their efficient deployment. Post-training quantization (PTQ) (Frantar et al., 2022; Lin et al., 2024; Sun et al., 2025) provides an effective solution to reduce the numerical precision of model weights and activations without costly retraining. To minimize performance degradation induced by low-bit quantization, existing PTQ methods (Xiao et al., 2023; Ashkboos et al., 2024; Liu et al., 2025; Liang et al., 2026) mainly rely on a small calibration dataset to characterize model responses and guide the quantization process through extracting activation statistics, calibrating scale factors, or optimizing layer-wise reconstruction. As one core part of PTQ, prior studies investigate the impact of calibration data from various perspectives, including sample count, sequence length, data source, and domain relationship (Williams & Aletras, 2024; Ji et al., 2025; He et al., 2025; Hai et al., 2026; Monaco et al., 2026), and show that different calibration choices substantially affect post-quantization performance. Analogously, as further illustrated in Fig. 1a, we empirically show that different calibration datasets can lead to noticeable performance variations in downstream tasks. However, how to construct a high-quality calibration dataset for effective and reliable PTQ is still an open and challenging problem.

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.



(a) Influence of Calibration Datasets

(b) Performance vs. Coverage

Figure 1: (a) We compare accuracy deviations of five calibration sets (LangMod, Math, CommQA, NLI, and Know) across four evaluation tasks from different domains using Llama-3.1-8B-Instruct with GPTQ, where it is observed that different calibration datasets can lead to noticeable performance variations under the same quantization setting. (b) Comparison of different methods (i.e., ZipCal, COLA, and our ReCaC) in terms of performance and calibration coverage, where SpinQuant and FlatQuant are used as low-bit quantization under the settings of W3 and W4. We can clearly observe that higher calibration coverage generally leads to better performance.

Recently, several studies have made the attempts to explore how to construct effective calibration datasets by developing different criteria to select informative samples. Specifically, ZipCal (Monaco et al., 2026) selects samples to promote lexical diversity based on Zipfian vocabulary-frequency statistics. COLA (He et al., 2025) first clusters candidates based on their activation representations and then selects samples near the cluster centers, aiming to obtain representative examples from diverse regions in the activation space. ACDM (Hai et al., 2026) favors samples whose activation distributions align with a predefined task distribution, making it dependent on prior knowledge of the target task distribution. However, these methods do not explicitly account for the complementarity among the selected samples with respect to the coverage of calibration. As shown in Fig. 1b, under a limited budget, calibration datasets constructed by existing methods (He et al., 2025; Monaco et al., 2026) still exhibit limited coverage in activation space[1], while calibration datasets with higher coverage consistently achieve better quantization performance.

Based on above observations, we propose a Residual-guided Calibration via activation Coverage (ReCaC) method to maximize calibration coverage in an activation space with a fixed budget, whose core idea is to seek candidate samples by evaluating their complementarity to the preceding selected samples in an activation space rather than solely considering their individual characteristics. Specifically, given a pre-defined benchmark for constructing the calibration dataset, we first select the sample with the highest local density as the initial sample, and then iteratively select candidates with the smallest residual score relative to the subspace spanned by the currently selected calibration samples. In this way, our ReCaC progressively expands calibration coverage in the activation space by considering the greatest complementarity between the selected samples, given the fixed and limited calibration budget. Furthermore, our ReCaC can be flexibly extended to the multi-domain setting by jointly considering global and domain-level coverage in the activation space. Specifically, our multi-domain ReCaC seeks candidate samples by maximizing their complementarity to the overall selected samples in the activation space, simultaneously regularizing by holding the coverage of each individual domain. By jointly optimizing two objectives, our multi-domain ReCaC explores and exploits the complementary patterns across heterogeneous domains, while preventing each individual domain from being insufficiently represented in the constructed calibration dataset. Extensive experiments under both single- and multi-domain calibration settings are conducted on three LLMs, including Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-2-9B-it (Team et al., 2024), and

[1]Coverage is computed as $\text{Coverage} = \|XB^{\mathsf{T}}\|_F^2 / \|X\|_F^2 \times 100\%$, where $\|\cdot\|_F$ is the Frobenius norm, X denotes the LLM activation matrix of all candidate samples from the calibration data source, and B is an orthonormal basis obtained via QR decomposition from the activations of 128 samples selected by each method.

Qwen3.5-9B (Qwen Team, 2026), with four representative PTQ methods, including GPTQ (Frantar et al., 2022), AWQ (Lin et al., 2024), SpinQuant (Liu et al., 2025), and FlatQuant (Sun et al., 2025), across four quantization settings and sixteen downstream tasks spanning five functional domains. The results clearly verify the effectiveness of activation-space coverage as a criterion for calibration dataset construction and good generalization of our ReCaC across LLMs, PTQ methods, and downstream tasks. Our contributions are summarized as follows:

- This paper, to our best knowledge, makes the first attempt to investigate calibration dataset construction based on coverage in activation space and proposes a Residual-guided Calibration via activation Coverage (ReCaC) method for effective calibration of LLM quantization.
- To this end, our ReCaC progressively expands the coverage of calibration in activation space by residual-guided calibration construction (RGC). Furthermore, ReCaC can be flexibly extended to multi-domain by joint global-domain coverage (JGDC) to maximize complementary contributions while maintaining domain-level coverage.
- Extensive experiments across different LLMs with various post-training quantization methods and settings demonstrate that our ReCaC achieves highly competitive or superior performance over state-of-the-art calibration methods on a dozen downstream tasks, particularly under lower-bit quantization.

2 Related Work

Post-Training Quantization and Calibration. Post-training quantization (PTQ) reduces the memory and computational costs of pretrained language models by lowering the numerical precision of their weights and activations without additional training. GPTQ (Frantar et al., 2022) uses calibration activations for layer-wise reconstruction, while AWQ (Lin et al., 2024) identifies salient weights from calibration activations. QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2025) derive orthogonal rotations from calibration data, while FlatQuant (Sun et al., 2025) learns equivalent transformations using calibration data. Despite their different mechanisms, these PTQ methods rely on calibration data to estimate quantization-relevant statistics or transformations. Prior studies show that calibration properties, including sample count, sequence length, data source, and domain alignment, can substantially affect post-quantization performance (Williams & Aletras, 2024; Khanal & Capone, 2024; Ji et al., 2025; He et al., 2025; Hai et al., 2026; Monaco et al., 2026). These factors influence which activation patterns are observed during calibration, while insufficient activation statistics or a mismatch between calibration and deployment activation distributions can lead to inaccurate quantization parameters and degraded performance. This highlights the importance of constructing an effective calibration dataset under a limited budget.

Calibration Data Construction. Recent studies have made initial attempts to explore effective calibration data construction using different strategies. Generation-based methods, such as Self-Gen (Ji et al., 2025) and Self-Calibration (Williams et al., 2025), leverage the knowledge encoded in pretrained models to generate calibration samples. Selection-based methods instead select informative samples from existing data according to different criteria. Specifically, ZipCal (Monaco et al., 2026) relies on vocabulary-frequency statistics to improve lexical coverage, while COLA (He et al., 2025) selects samples near cluster centers in the activation space to improve representativeness and diversity. ACDM (Hai et al., 2026) favors samples whose activation distributions align with a predefined task-specific reference, emphasizing task relevance. Despite their different strategies, existing methods do not explicitly account for the complementarity among calibration samples in the activation space. As a result, the constructed calibration datasets may contain redundant activation patterns under a limited budget, leaving the activation space insufficiently covered.

3 Method

To maximize coverage of the constructed calibration dataset under a fixed budget, we propose **ReCaC**, a residual-guided calibration construction method that captures broad and complementary activation patterns in both single-domain and multi-domain settings, as illustrated in Fig. 2. Specifically, Sec. 3.1 first formulates calibration construction as a K-sample activation coverage problem and defines the corresponding objective. Sec. 3.2 develops the residual-guided construction strategy, which initializes

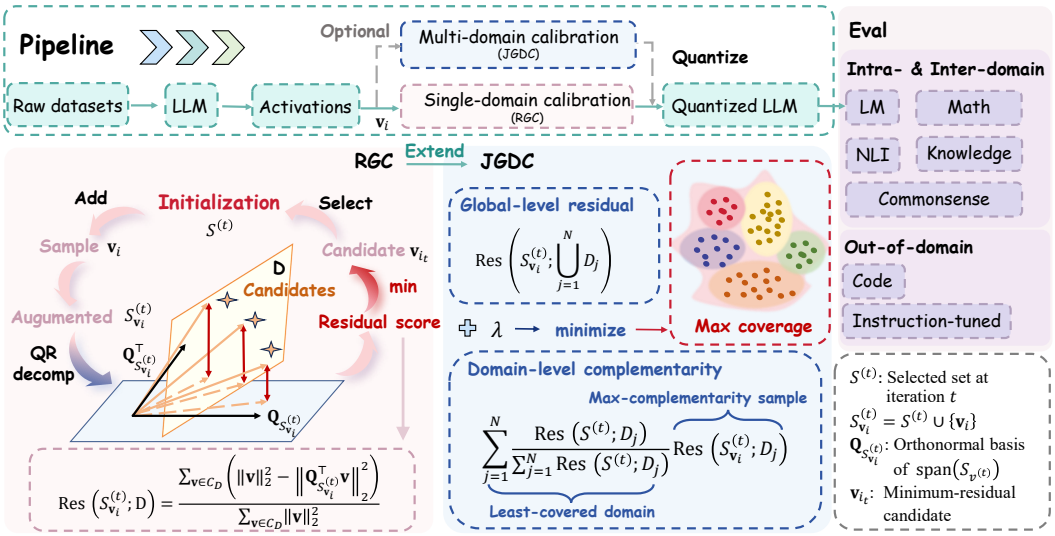


Figure 2: Overview of the proposed ReCaC method. ReCaC first extracts LLM activations, then applies RGC for single-domain calibration selection, and uses JGDC to incorporate domain-level coverage for multi-domain scenarios. The selected calibration dataset is then used for LLM quantization and evaluated across intra-domain, inter-domain, and out-of-domain scenarios.

with a highest-density representative sample and iteratively expands the covered activation subspace based on residual scores. Sec. 3.3 extends our method to multiple domains by jointly considering global and domain-level activation coverage.

3.1 Problem Formulation

To perfrom low-bit quantization for a full-precision model f_θ, PTQ estimates the quantization statistics from a selected calibration dataset $P_C \subseteq P_D$ with the calibration budget of K, where $P_D = \{x_i\}_{i=1}^N$ is a candidate pool of domain D. Let $\widehat{f}_{P_C}$ be the quantized model calibrated on P_C, an ideal calibration construction can be formulated as

$$S_K^*(D) = \arg \min_{\substack{P_C \subseteq P_D \\ |P_C|=K}} \mathbb{E}_{x \sim P_{dep}} \left[\mathcal{L}\Big(f_\theta(x), \widehat{f}_{P_C}(x) \Big) \right], \tag{1}$$

where P_{dep} is the deployment input distribution and $\mathcal{L}$ measures the output error introduced by low-bit quantization. However, the objective in Eq. 1 is difficult to optimize in practice, since P_{dep} is typically unknown and the quantized model is expected to generalize across diverse domains.

We therefore propose a calibration criterion, aiming to directly reflect the information relevant to quantization. Since quantization parameters are estimated based on the activations of calibration samples, criterion in activation space is potentially more essential than one in sample space for calibration construction. More importantly, raw-input properties such as lexical diversity are only indirect proxies, as different inputs can induce similar representations (Liu et al., 2026) while small input changes may substantially alter activations Matthews et al. (2024). In contrast, activation directly characterizes the relevance between representations and quantization. Beyond the importance of individual samples, an effective calibration set should also cover the activation patterns as many as possible. Particularly, larger calibration coverage provides more representative and diverse activation patterns for estimating the statistics required by PTQ and reduces the mismatch between calibration and real-world distributions in the practical deployment. Therefore, under a fixed calibration budget, we seek to select samples that maximize marginal activation coverage, favoring complementary patterns over redundant ones and improving generalization performance of quantization. As such, we reformulate calibration construction as a K-sample activation coverage problem:

$$S_K^*(D) = \arg \min_{\substack{P_C \subseteq P_D \\ |P_C|=K}} \mathrm{Res}(S; D), \tag{2}$$

where P_C denotes the selected set of K calibration samples, and S denotes the corresponding set of activation representations induced by P_C. $\mathrm{Res}(S; D)$ measures the activation space of domain D that remains uncovered by S, as defined in Eq. 7. Note that our formulation in Eq. 2 can be readily extended to the multi-domain setting by replacing the single domain D with the union of M domains, $\bigcup_{j=1}^{M} D_j$. However, solving Eq. 2 exactly requires a combinatorial search over candidate subsets. We therefore develop a coverage-oriented greedy strategy to efficiently construct the calibration set in both single- and multi-domain settings, whose details will be described in the following.

3.2 RGC: Residual-guided Calibration Construction

Based on above analysis in Sec. 3.1, we propose a residual-guided construction (RGC) strategy to obtain a calibration dataset by maximizing coverage in the activation space $S_K^*(D)$. To efficiently achieve this objective, we first initialize the calibration with the highest-density representative sample, and then greedily select the candidate that minimizing the residual score after added to the current set in a greedy manner. As such, RGC ensures that each newly selected sample contributes the largest complementary information to the current calibration in the activation space.

For each candidate sample $x_i \in P_D$, RGC extracts activation representations $\mathbf{h}_i^l$ from L selected layers and concatenate them into a multi-layer activation representation $\mathbf{a}_i = [\mathbf{h}_i^l]_{l \in L}$. We then project $\mathbf{a}_i$ into a lower-dimensional space using a shared random projection matrix $\mathbf{R} \in \mathbb{R}^{r \times d}$ (Bingham & Mannila, 2001), yielding $\mathbf{v}_i = \mathbf{R}\mathbf{a}_i$, where d denotes the dimension of the concatenated activation representation, and r is the projection dimension. All projected $\mathbf{v}_i$ form the candidate set C_D.

Density-max initialization. For a sample $\mathbf{v}_i \in C_D$, $N_k(\mathbf{v}_i; C_D)$ are the k nearest candidates to $\mathbf{v}_i$ in pool C_D, measured by cosine similarity between the same pool representations. Then, the local density $\mathbf{v}_i$ can be defined as

$$\rho(\mathbf{v}_i) = \frac{1}{k} \sum_{j \in N_k(\mathbf{v}_i; C_D)} \mathbf{v}_i^\top \mathbf{v}_j. \tag{3}$$

Based on Eq. 3, we choose the highest-density $\rho(\mathbf{v}_i)$ as the initial anchor $\mathbf{v}^{(0)}$.

Residual-guided selection. Let $S^{(t)}$ denote the samples selected at iteration t, we stack their activation representations row-wise as $\mathbf{V}_{S^{(t)}} = \left[\mathbf{v}_i^\top\right]_{\mathbf{v}_i \in S^{(t)}}$. Then the joint activation subspace can be represented by the orthonormal basis

$$\boldsymbol{Q}_{S^{(t)}} = \mathrm{orth}\left(\mathbf{V}_{S^{(t)}}^\top\right), \qquad \boldsymbol{Q}_{S^{(t)}}^\top \boldsymbol{Q}_{S^{(t)}} = \boldsymbol{I}, \tag{4}$$

where $\mathrm{orth}(\cdot)$ denotes the QR decomposition. For the initialization of single density-max anchor $\mathbf{v}^{(0)}$, we initialize $S^{(0)} = \left\{\mathbf{v}^{(0)}\right\}$ and $\boldsymbol{Q}_{S^{(0)}} = \mathrm{orth}(\mathbf{v}^{(0)})$.

To measure the fraction of a candidate $\mathbf{v}$ that remains unexplained by the subspace spanned by $S^{(t)}$, the residual score is computed as

$$\mathrm{r}\left(\mathbf{v}; S^{(t)}\right) = \|\mathbf{v}\|_2^2 - \|\boldsymbol{Q}_{S^{(t)}} \boldsymbol{Q}_{S^{(t)}}^\top \mathbf{v}\|_2^2 = \|\mathbf{v}\|_2^2 - \left\|\boldsymbol{Q}_{S^{(t)}}^\top \mathbf{v}\right\|_2^2. \tag{5}$$

For the candidate set C_D, we define the coverage of the selected calibration set $S^{(t)}$ over D as

$$\mathrm{Cover}(S^{(t)}; D) = \frac{\sum_{\mathbf{v} \in C_D} (\|\mathbf{v}\|_2^2 - \mathrm{r}\left(\mathbf{v}; S^{(t)}\right))}{E_D} = \frac{\sum_{\mathbf{v} \in C_D} \|\boldsymbol{Q}_{S^{(t)}}^\top \mathbf{v}\|_2^2}{\sum_{\mathbf{v} \in C_D} \|\mathbf{v}\|_2^2}, \tag{6}$$

where all candidates in C_D projected into the subspace spanned by S measures the fraction of their activation representations. It can be explained by the selected subspace and E_D denotes the total activation energy of domain D, which is introduced to account for the differences in activation energy across domains. Accordingly, the uncovered energy of domain D can be calculated by

$$\mathrm{Res}(S^{(t)}; D) = \frac{\sum_{\mathbf{v} \in C_D} \mathrm{r}\left(\mathbf{v}; S^{(t)}\right)}{E_D} = \frac{\sum_{\mathbf{v} \in C_D} (\|\mathbf{v}\|_2^2 - \|\boldsymbol{Q}_{S^{(t)}}^\top \mathbf{v}\|_2^2)}{\sum_{\mathbf{v} \in C_D} \|\mathbf{v}\|_2^2}. \tag{7}$$

For a candidate $\mathbf{v}_i$, we measure its gain by using the resulting residual score of the activation space of domain D achieved by the selected set after incorporating $\mathbf{v}_i$, i.e.,

$$\mathrm{Res}\left(S^{(t)} \cup \{\mathbf{v}_i\}; D\right) = \frac{\sum_{\mathbf{v} \in C_D} (\|\mathbf{v}\|_2^2 - \|\boldsymbol{Q}_{S^{(t)} \cup \{\mathbf{v}_i\}}^\top \mathbf{v}\|_2^2)}{\sum_{\mathbf{v} \in C_D} \|\mathbf{v}\|_2^2}. \tag{8}$$

For simplicity, we denote $S^{(t)} \cup \{\mathbf{v}_i\}$ as $S^{(t)}_{\mathbf{v}_i}$, and then select the candidate with the smallest residual score among the unselected candidates, i.e.:

$$i_t = \arg \min_{\mathbf{v}_i \in C_D \setminus S^{(t)}} \mathrm{Res}\Big(S^{(t)}_{\mathbf{v}_i}; D\Big), \qquad S^{(t+1)} = S^{(t)} \cup \{\mathbf{v}_{i_t}\}. \tag{9}$$

After each selection, we recompute the basis using QR decomposition: $\boldsymbol{Q}_{S^{(t+1)}} = \mathrm{orth}\big(\mathbf{V}^\top_{S^{(t+1)}}\big)$ The procedure is repeated for $K-1$ iterations, yielding the final calibration $S^*_K(D)$.

3.3 JGDC: Joint Global And Domain Coverage

We further propose a joint global and domain coverage (JGDC) strategy for extension of multi-domain setting, which jointly considers contribution of each candidate sample to global and domain-level activation coverage. At iteration t, rather than evaluating candidate $\mathbf{v}_i$ solely based on its source dataset, we first measure the global coverage of the overall activation space across domains achieved by the selected set after incorporating $\mathbf{v}_i$:

$$\mathrm{Res}\Big(S^{(t)}_{\mathbf{v}_i}; D_M\Big) = \frac{\sum_{\mathbf{v}\in C_{D_M}} (\|\mathbf{v}\|_2^2 - \|\boldsymbol{Q}^\top_{S^{(t)}_{\mathbf{v}_i}}\mathbf{v}\|_2^2)}{\sum_{\mathbf{v}\in C_{D_M}} \|\mathbf{v}\|_2^2}, D_M = \bigcup_{j=1}^{M} D_j, \tag{10}$$

where D_M denote the union of all domains. To ensure that the selected set provides sufficient coverage across different domains, we further incorporate domain-specific information into the global coverage objective. Specifically, we compute the residual score of the candidate-augmented set with respect to each domain D_j as

$$\mathrm{Res}\Big(S^{(t)}_{\mathbf{v}_i}; D_j\Big) = \frac{\sum_{\mathbf{v}\in C_{D_j}} (\|\mathbf{v}\|_2^2 - \|\boldsymbol{Q}^\top_{S^{(t)}_{\mathbf{v}_i}}\mathbf{v}\|_2^2)}{\sum_{\mathbf{v}\in C_{D_j}} \|\mathbf{v}\|_2^2}. \tag{11}$$

However, a small residual score for a particular domain D_j indicates that the candidate-augmented set already provides relatively good coverage of D_j, but it does not indicate that D_j is the least-covered domain among all domains before incorporating the candidate. To prioritize domains that remain insufficiently covered while accounting for the relative coverage across domains, we introduce a weighting factor to balance the coverage requirement of the most underrepresented domains at iteration t against the best relative coverage achieved across different domains after adding $\mathbf{v}_i$:

$$w^{(t)}(D_j) = \frac{\mathrm{Res}\big(S^{(t)}; D_j\big)}{\sum_{j=1}^{M} \mathrm{Res}\big(S^{(t)}; D_j\big)}. \tag{12}$$

Therefore, the overall residual objective consists of two complementary components. The first is the global residual, which quantifies the uncovered portion of the overall activation space after adding candidate $\mathbf{v}_i$. The second is the domain-specific residual, which first identifies the least-covered domain under the current selected set and then evaluates how effectively candidate $\mathbf{v}_i$ reduces the uncovered space of that domain. The overall formulation is given as follows:

$$\mathcal{L}^{(t)}(\mathbf{v}_i) = \mathrm{Res}\Big(S^{(t)}_{\mathbf{v}_i}; D_M\Big) + \lambda \sum_{j=1}^{M} w^{(t)}(D_j)\, \mathrm{Res}\Big(S^{(t)}_{\mathbf{v}_i}; D_j\Big) \tag{13}$$

where λ is a hyperparameter that balances the relative contributions of the global- and domain-level objectives. At each iteration, we select the candidate that minimizes $\mathcal{L}^{(t)}(\mathbf{v}_i)$ and repeat this process for $K-1$ iterations to obtain the final calibration $S^*_K(D_M)$.

4 Experiments

To assess our ReCaC, we compare with two state-of-the-art calibration methods, COLA (He et al., 2025) and ZipCal (Monaco et al., 2026), under both single-domain and multi-domain settings. Experiments are conducted on two LLMs, including Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Gemma-2-9B-it (Team et al., 2024). We employ four PTQ methods, i.e., GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2024) under W3A16 and W4A16 settings, and SpinQuant (Liu et al.,

Table 1: Comparison (%) of different calibration methods in single-domain setting. 'Dense' indicates full-precision models.

Model	Task	Dense	GPTQ						AWQ					
			W4A16			W3A16			W4A16			W3A16		
			COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC
Llama-3.1-8B-Instruct	MMLU-M	24.44	23.19	23.11	24.00	23.06	22.56	25.44	24.44	24.42	24.07	22.35	21.18	23.17
	GSM8k	78.09	67.61	70.00	74.72	4.33	1.85	37.17	68.98	68.88	73.03	26.61	27.76	27.54
	HellaSwag	71.71	71.94	72.05	72.25	40.60	37.51	69.24	71.68	71.68	71.71	68.64	69.03	68.72
	WinoGr.	69.46	66.85	66.96	65.19	53.79	52.57	61.83	66.75	66.74	64.75	60.76	62.49	60.69
	OBQA	47.80	45.60	46.34	46.71	32.43	30.07	43.15	45.90	45.88	45.55	42.32	41.58	41.47
	BoolQ	84.46	84.23	83.27	84.15	55.39	48.07	80.77	84.07	83.96	84.10	78.94	78.42	78.51
	RTE	74.73	74.16	72.20	73.55	54.96	52.67	69.15	73.10	73.14	73.78	66.19	66.26	65.51
	ANLI	58.40	55.30	54.37	55.86	35.15	33.47	46.39	53.36	53.54	51.32	46.66	45.61	45.81
	ARC-C	51.71	50.43	50.47	50.83	29.17	26.99	45.99	50.21	50.24	50.80	45.51	45.54	45.80
	ARC-E	73.86	73.42	73.90	73.37	40.79	36.56	68.33	72.55	72.38	72.77	68.44	67.73	68.77
	MMLU-K	62.28	58.44	57.11	58.09	28.89	26.48	46.29	58.52	58.45	60.60	38.74	40.18	42.23
	Mean	63.36	61.02	60.89	**61.70**	36.23	33.53	**53.98**	60.89	60.85	**61.13**	51.38	51.43	**51.66**
Gemma-2-9B-it	MMLU-M	21.48	22.44	26.27	21.78	21.43	21.95	21.81	21.40	21.39	22.14	21.48	21.25	21.10
	GSM8k	75.44	74.16	73.66	74.18	68.77	64.89	70.85	74.22	74.14	73.62	66.80	67.24	67.44
	HellaSwag	67.24	67.63	67.63	67.94	66.68	65.88	67.13	67.58	67.40	67.99	66.76	66.76	66.63
	WinoGr.	70.48	69.72	69.92	68.15	66.11	67.05	66.47	69.58	69.48	68.05	65.47	66.54	66.00
	OBQA	45.40	44.73	45.03	45.17	43.69	43.43	44.29	45.17	45.12	45.24	44.43	44.32	44.05
	BoolQ	88.59	88.57	88.68	88.51	87.37	87.23	87.35	88.57	88.54	88.39	87.45	87.53	87.70
	RTE	78.34	77.57	77.91	76.94	77.80	77.69	78.01	77.58	77.42	77.55	78.11	77.82	78.22
	ANLI	72.80	71.70	71.98	70.13	67.27	58.41	65.80	71.76	71.70	70.35	64.11	63.71	62.64
	ARC-C	51.79	51.51	52.14	50.78	49.44	47.65	50.08	51.28	51.20	51.22	48.81	48.88	48.23
	ARC-E	66.79	66.88	68.02	66.00	64.47	63.18	66.22	66.66	66.58	67.07	64.92	65.86	64.90
	MMLU-K	33.83	31.26	34.25	46.84	31.36	35.04	39.03	32.00	31.92	43.82	24.87	28.87	25.50
	Mean	61.11	60.56	61.41	**61.49**	58.58	57.49	**59.73**	60.53	60.44	**61.40**	57.56	**58.07**	57.49

2025) and FlatQuant (Sun et al., 2025) under W3A4 and W4A4 settings. Unless otherwise specified, we use a context length of 2048 and a calibration budget of $K = 128$. Calibration and evaluation cover five functional domains: (i) Language Modeling, (ii) Mathematical Reasoning, (iii) Commonsense Reasoning & QA, (iv) Natural Language Inference (NLI), and (v) Knowledge. More experimental details are provided in Appendix A.1.

4.1 Main Results

Results on single-domain. Following ZipCal (Monaco et al., 2026), we first compare different methods in single-domain setting. As shown in Table 1, our ReCaC achieves strong performance across different models and quantization settings. Specifically, for GPTQ with W3A16, our ReCaC outperforms COLA and ZipCal by 17.75 and 20.45 points on Llama-3.1-8B-Instruct and by 1.15 and 2.24 points on Gemma-2-9B-it, respectively. Additionally, our ReCaC also outperforms them under GPTQ with W4A16. For AWQ, the performance of ReCaC is very competitive with its counterparts. These results demonstrate the effectiveness of our ReCaC across LLMs and quantization settings.

Results on multi-domain. We further assess our ReCaC the multi-domain settings following ZipCal (Monaco et al., 2026). Moreover, to better evaluate the generalization of calibration methods to unseen inter-domain tasks, we test on additional datasets that are excluded from calibration dataset construction. As shown in Table 2, our ReCaC outperforms the competing methods on most of unseen tasks, demonstrating strong generalization across models and quantization settings. For GPTQ with W4A16, our ReCaC achieves the highest Mean_5 on both Llama-3.1-8B-Instruct and Gemma-2-9B-it. Particularly, the perfromance gains become larger for more challenging 3-bit GPTQ. For metric of Mean_{16}, ReCaC exhibits substantially greater robustness to reduced GPTQ precision, with its performance on Llama-3.1-8B-Instruct decreasing by only 3.25 points when moving from W4A16 to W3A16, compared with 25.83 and 6.39 points for COLA and ZipCal, respectively. Overall, these results demonstrate the strong generalization and effectiveness of our ReCaC across LLMs and

Table 2: Comparison (%) of different calibration methods in multi-domain setting.

Model	Task	Dense	GPTQ						AWQ					
			W4A16			W3A16			W4A16			W3A16		
			COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC
Llama-3.1-8B-Instruct	MMLU-M	24.44	27.41	23.01	23.70	21.85	23.70	32.59	24.07	24.44	22.59	22.22	26.67	22.59
	GSM8k	78.09	73.24	74.07	73.69	0.23	43.90	53.53	73.69	70.50	72.78	24.64	28.66	32.60
	HellaSwag	71.71	72.17	71.73	72.59	26.68	69.09	71.34	71.57	71.95	71.01	69.25	68.46	68.86
	WinoGr.	69.46	64.25	67.95	64.72	48.54	63.30	63.54	64.33	68.10	63.69	60.14	60.30	60.38
	OBQA	47.80	45.60	44.40	47.80	30.40	43.20	44.00	45.60	45.90	45.60	42.80	41.80	42.40
	BoolQ	84.46	84.74	85.38	83.73	45.54	81.31	83.85	83.79	84.60	84.28	77.65	78.90	78.72
	RTE	74.73	74.73	74.72	72.56	52.71	67.15	74.73	73.65	74.50	73.65	66.43	58.12	67.15
	ANLI	58.40	58.40	56.90	58.30	33.40	40.20	49.20	51.60	55.00	50.50	49.30	40.10	48.60
	ARC-C	51.71	49.57	49.31	51.02	23.38	46.67	47.35	50.09	50.30	50.68	45.90	45.90	46.16
	ARC-E	73.86	71.93	72.97	73.36	28.54	70.41	71.97	73.11	72.80	72.56	69.65	69.49	70.58
	MMLU-K	62.28	59.69	55.86	57.41	24.57	44.49	54.44	61.15	58.60	59.94	36.36	39.56	39.96
	Mean$_{11}$	63.36	**61.98**	61.48	61.72	30.53	53.95	**58.78**	61.15	**61.52**	60.66	51.30	50.72	**52.55**
	MathQA	26.83	26.40	26.16	25.90	20.97	26.06	24.99	27.24	26.37	26.30	25.80	25.13	24.46
	SIQA	50.05	48.57	48.82	50.10	34.29	47.80	50.46	49.18	48.52	48.82	44.73	44.42	45.96
	PIQA	78.56	77.97	79.05	77.31	53.05	76.39	78.02	78.78	78.94	78.24	77.86	76.82	76.66
	MNLI	57.89	57.32	57.14	64.61	35.39	43.14	55.64	55.22	56.35	55.93	43.02	36.94	42.52
	GPQA	27.23	26.56	27.01	35.27	25.67	25.45	24.33	26.79	25.89	26.34	25.22	22.32	25.22
	Mean$_5$	53.33	47.36	47.64	**50.64**	33.87	43.77	**46.69**	**47.44**	47.21	47.13	**43.33**	41.13	42.96
	Mean$_{16}$	60.68	57.41	57.16	**58.25**	31.58	50.77	**55.00**	56.87	**57.05**	56.43	48.81	47.72	**49.55**
Gemma-2-9B-it	MMLU-M	21.48	21.11	23.29	21.48	21.11	21.48	20.74	21.48	21.30	21.85	21.11	21.11	21.11
	GSM8k	75.44	75.28	73.61	74.68	69.83	70.43	72.63	73.31	74.10	73.84	65.43	66.94	66.94
	HellaSwag	67.24	67.70	67.55	68.00	67.19	66.59	66.90	68.21	67.58	68.16	67.06	66.17	66.54
	WinoGr.	70.48	68.59	69.61	68.03	66.54	67.80	66.22	67.40	69.58	67.88	65.43	65.04	65.98
	OBQA	45.40	45.20	45.40	45.00	43.60	44.20	44.60	44.00	45.17	45.60	44.80	44.00	44.00
	BoolQ	88.59	88.96	88.04	88.47	88.41	87.37	88.26	88.17	88.57	88.26	87.52	87.46	87.27
	RTE	78.34	77.26	79.06	77.26	77.62	78.34	77.98	77.62	77.58	77.98	79.42	78.70	77.61
	ANLI	72.80	70.10	72.90	69.40	69.00	70.30	71.10	71.50	71.76	69.40	60.50	57.40	63.40
	ARC-C	51.79	51.19	51.11	50.77	50.34	48.21	48.98	51.45	51.28	51.28	48.89	50.77	48.12
	ARC-E	66.79	65.15	68.22	66.67	64.94	63.34	64.39	67.63	66.66	67.26	64.65	66.92	64.10
	MMLU-K	33.93	31.44	35.19	41.98	25.96	24.03	31.99	33.71	32.00	42.17	25.92	25.17	26.78
	Mean$_{11}$	61.11	60.18	**61.27**	61.07	58.59	58.37	**59.44**	60.41	60.51	**61.24**	57.34	57.24	**57.44**
	MathQA	34.57	34.00	33.96	34.10	32.16	31.49	32.53	34.17	33.96	34.14	30.79	31.72	31.56
	SIQA	49.69	49.74	49.54	50.10	49.28	48.52	48.77	49.39	49.08	49.44	48.46	49.18	48.16
	PIQA	79.00	77.26	78.18	77.31	78.40	77.69	78.18	78.29	78.18	78.13	77.31	77.69	76.55
	MNLI	62.33	61.63	62.59	64.61	56.84	55.60	54.22	62.46	63.09	63.07	41.52	42.56	47.88
	GPQA	35.49	35.71	33.93	35.27	32.81	30.58	34.38	35.49	34.38	36.16	33.71	32.14	32.24
	Mean$_5$	52.22	51.67	51.64	**52.28**	**49.90**	48.78	49.62	51.96	51.74	**52.19**	46.36	46.66	**47.28**
	Mean$_{16}$	58.34	57.52	58.26	**58.32**	55.88	55.37	**56.37**	57.77	57.77	**58.41**	53.91	53.94	**54.27**

quantization settings, particularly under lower-bit quantization. Due to page limitations, we report the results with Qwen3.5-9B in Table 7 of the appendix.

4.2 Results on advanced quantization methods

As shown in Table 3, we evaluate different calibration construction strategies with SpinQuant (Liu et al., 2025) and FlatQuant (Sun et al., 2025) under W4A4 and W3A4 quantization settings. Overall, ReCaC demonstrates strong competitiveness across all four settings. Under SpinQuant-W4A4, ReCaC consistently outperforms both COLA and ZipCal across the evaluated metrics. This advantage is further observed with SpinQuant-W3A4, where the corresponding improvements reach 1.39/6.73, 1.26/3.57, and 1.36/5.75 points. The same trend holds with FlatQuant-W4A4, where ReCaC surpasses COLA and ZipCal by 1.08/1.40, 0.18/0.92, and 0.80/1.25 points on the three metrics. Under FlatQuant-W3A4, ReCaC remains competitive. These results demonstrate the consistent effectiveness of ReCaC across advanced PTQ methods, including settings with both low-bit weights and activations.

Table 3: Comparisons (%) of calibration methods with SpinQuant and FlatQuant.

Task	SpinQuant						FlatQuant					
	W4A4			W3A4			W4A4			W3A4		
	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC	COLA	ZipCal	ReCaC
MMLU-M	21.48	23.70	29.26	24.81	23.70	25.93	29.26	22.59	26.30	25.19	25.19	30.74
GSM8k	69.60	63.23	68.01	59.67	31.77	58.00	72.71	72.86	72.71	51.55	0.30	48.52
HellaSwag	71.02	69.89	71.78	68.19	66.11	68.77	71.46	71.42	72.19	67.67	65.07	67.72
WinoGr.	61.40	63.85	63.06	63.06	56.67	61.33	64.56	62.35	64.33	63.54	58.17	60.54
OBQA	44.80	43.80	44.60	43.60	39.60	41.80	45.00	44.20	47.00	42.00	39.40	42.00
BoolQ	82.97	83.18	81.56	76.67	80.12	81.04	80.92	81.93	84.22	79.14	72.14	81.07
RTE	72.20	73.65	75.09	66.06	61.01	71.48	70.04	70.04	71.84	67.51	57.40	69.68
ANLI-R1	53.50	48.80	54.80	45.60	40.90	48.00	50.00	59.10	57.60	50.80	40.60	51.50
ARC-C	48.98	45.39	48.89	46.67	41.98	48.04	51.02	49.83	50.68	48.29	41.47	46.59
ARC-E	71.59	70.41	73.11	68.81	66.62	71.46	74.24	73.82	74.03	71.13	65.70	70.08
MMLU-K	55.39	54.60	59.84	49.98	45.87	52.58	60.95	58.50	61.10	52.15	41.51	52.03
Mean$_{11}$	59.36	58.23	**60.91**	55.74	50.40	**57.13**	60.92	60.60	**62.00**	56.27	46.09	**56.41**
MathQA	25.96	26.13	25.93	26.43	25.70	26.57	26.50	26.80	26.53	25.80	25.13	26.67
SIQA	47.03	47.59	48.72	46.16	43.96	48.36	48.62	47.59	49.18	46.93	44.32	47.39
PIQA	76.50	76.88	78.40	76.44	74.59	76.61	78.35	78.13	79.16	78.13	75.52	78.56
MNLI	52.82	51.02	50.95	46.23	39.48	48.47	54.64	56.16	54.18	51.23	40.60	46.26
GPQA	24.11	26.34	27.23	27.68	27.68	29.24	28.57	24.33	28.57	25.67	27.01	27.01
Mean$_5$	45.28	45.59	**46.25**	44.59	42.28	**45.85**	47.34	46.60	**47.52**	**45.55**	42.52	45.18
Mean$_{16}$	54.96	54.28	**56.33**	52.25	47.86	**53.61**	56.68	56.23	**57.48**	**52.92**	44.97	52.90

Table 4: Out-of-domain generalization (%) of SpinQuant.

Bit	Method	Alpaca	Alpaca LC	HumanEval	MBPP
W4A4	COLA	18.32	18.57	**60.37**	47.00
	ZipCal	16.52	17.15	52.44	45.00
	ReCaC	**19.63**	**20.09**	56.71	**49.80**
W3A4	COLA	12.30	12.94	40.24	39.40
	ZipCal	10.00	11.07	29.27	26.00
	ReCaC	**12.48**	**13.92**	**42.68**	**40.40**

Table 5: Initialization comparison (%).

Method	Setting	W4A16	W3A16
GPTQ	Density-max	**59.57**	**55.87**
	Centroid	59.25	55.14
	Random	59.38	55.48
AWQ	Density-max	58.44	51.18
	Centroid	**58.90**	50.29
	Random	58.60	**51.24**

4.3 Ablation Study

Generalization to out-of-domain datasets. Table 4 further evaluates the generalization of calibration construction methods to instruction-tuned and code-generation tasks. Overall, when evaluated within the SpinQuant framework, ReCaC remains competitive with the baseline across both precision settings, with more pronounced advantages under W3 quantization.

Initialization and the choice of λ. Table 5 compares three initialization strategies on Llama-3.1-8B-Instruct, with density-max achieving better or competitive performance across the evaluated settings. Fig. 3 in Appendix A.5 further studies the sensitivity to λ on Llama-3.1-8B-Instruct, where $\lambda = 1.5$ achieves the best overall performance and is therefore used throughout the paper.

5 Conclusion

This paper investigates calibration dataset construction for post-training quantization from the perspective of activation-space coverage. We propose **ReCaC**, a residual-guided method that progressively captures complementary activation patterns under a fixed budget, with an extension to multi-domain settings that considers both global and domain-level coverage. Experiments across diverse LLMs, quantization methods, and bit-widths demonstrate the effectiveness of ReCaC, particularly under low-bit quantization and on held-out tasks. As future work, we plan to extend ReCaC to multimodal LLMs by jointly considering the activation characteristics of language and visual components.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript language polishing and LaTeX formatting. We have not used generative AI tools for experimental design, data collection, data analysis, model training, or result interpretation. We have reviewed all AI-assisted work by manually checking the revised text and LaTeX structure for accuracy and consistency. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. We are not aware of any ethical concerns associated with the methodology or experiments presented in this work.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results throughout the paper. The main text describes the proposed method, datasets, baselines, and evaluation protocols, while the appendix provides implementation details, hyperparameters, and additional experimental results.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245/.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 245–250, 2001.

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, pp. 7432–7439, 2020.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 2924–2936, 2019.

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 preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

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.

Benben Hai, Anzheng Wang, Xiangdong Liu, Lingwei Meng, Tingting Wang, Shuli Zheng, Yixin Zhou, Haibiao Chen, et al. Understanding and selecting calibration data for llm quantization: From sensitivity analysis to activation-based curation. In *The 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models (SELVA) at ACL 2026*, 2026.

Bowei He, Lihao Yin, Hui-Ling Zhen, Shuqi Liu, Han Wu, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. Preserving llm capabilities through calibration data curation: From analysis to optimization. *Advances in Neural Information Processing Systems*, 38:58531–58572, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Beware of calibration data for pruning large language models. In *International Conference on Learning Representations*, volume 2025, pp. 18604–18620, 2025.

Bishwash Khanal and Jeffery M Capone. Evaluating the impact of compression techniques on task-specific performance of large language models. *arXiv preprint arXiv:2409.11233*, 2024.

Yesheng Liang, Haisheng Chen, Song Han, and Zhijian Liu. Paroquant: Pairwise rotation quantization for efficient reasoning llm inference. In *International Conference on Learning Representations*, volume 2026, pp. 77000–77017, 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Xuyuan Liu, Shengyu Chen, Xinshuai Dong, Yanchi Liu, Xujiang Zhao, Haoyu Wang, Yujun Yan, Haifeng Chen, and Zhengzhang Chen. Representation interventions enable lifelong knowledge memory control in llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5414–5436, 2026.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025.

Jacob A Matthews, John R Starr, and Marten van Schijndel. Semantics or spelling? probing contextual word embeddings with orthographic noise. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4495–4504, 2024.

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 *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Francesco Pio Monaco, Elia Cunegatti, Flavio Vella, and Giovanni Iacca. Frequency matters: Fast model-agnostic data curation for pruning and quantization. *arXiv preprint arXiv:2603.16105*, 2026.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 2080–2094, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

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 *International conference on machine learning*, 2025.

Lintang Sutawika, Leo Gao, Hailey Schoelkopf, Stella Biderman, Jonathan Tow, Baber Abbasi, Ben Fattori, Charles Lovering, Jason Phang, Anish Thite, Niklas Muennighoff, Thomas Wang, Eric Tang, et al. EleutherAI/lm-evaluation-harness: Major refactor, 2023. URL https://doi.org/10.5281/zenodo.10256836. Computer software.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

Miles Williams and Nikolaos 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 (Volume 1: Long Papers)*, pp. 10100–10118, 2024.

Miles Williams, George Chrysostomou, and Nikolaos Aletras. Self-calibration for language model quantization and pruning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10149–10167, 2025.

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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

A APPENDIX

A.1 DETAILED EXPERIMENTS

Baselines. For calibration, we compare two representative strategies: COLA (He et al., 2025) and ZipCal (Monaco et al., 2026). For post-training quantization, we evaluate our method under two representative quantization frameworks, including GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2024) with both W3A16 and W4A16 quantization settings and SpinQuant (Liu et al., 2025), and FlatQuant (Sun et al., 2025), with W3A4 and W4A4 quantization settings

Implementation. We evaluate downstream task performance using Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-2-9B-it (Team et al., 2024), and Qwen3.5-9B (Qwen Team, 2026). Unless otherwise specified, we use a context length of 2048 and a calibration budget of $K = 128$ samples. For Llama and Qwen, we extract activations from layers $\mathcal{L} = \{-1, -8, -16\}$ and concatenate their representations, where negative indices count backward from the final layer. Since Gemma has a different model depth, we use only its final-layer representation, i.e., $\mathcal{L} = \{-1\}$. For all models, we map the resulting representations to $r = 384$ dimensions using a shared random projection. Local density is estimated using the $k = 20$ nearest neighbors.

Calibration. For single- and multi-domain, we use datasets form five functional domains as calibration. **(i) Language Modeling**: WikiText (Merity et al., 2016), C4 (Raffel et al., 2020), and Pile (Gao et al., 2020), which have commonly been used for model quantization. **(ii) Mathematical Reasoning**: GSM8k (Cobbe et al., 2021) and SVAMP (Patel et al., 2021). **(iii) Commonsense Reasoning & QA**: WinoGrande (Sakaguchi et al., 2021) and OpenBookQA Mihaylov et al. (2018). **(iv) Natural Language Inference(NLI)**: RTE Wang et al. (2018) and ANLI Nie et al. (2020). **(v) Knowledge**: MMLU (Hendrycks et al., 2020) and WMT14 (Bojar et al., 2014).

Evaluation. To assess downstream performance post-quantization, we use the LM-Evaluation-Harness (Sutawika et al., 2023) framework across five functional domains. Results are reported on the subset of datasets that support standardized zero- or few-shot metrics. We categorize the datasets in functional domains as follows. **(i) Language Modeling**: zero-shot perplexity is measured on WikiText (Merity et al., 2016), C4 (Raffel et al., 2020), and Pile (Gao et al., 2020), which are general datasets that have commonly been used for model quantization. **(ii) Mathematical Reasoning**: evaluated using GSM8K (5-shot) (Cobbe et al., 2021), MathQA (Amini et al., 2019), and MMLU-M, the subset of mathematical tasks in MMLU (Hendrycks et al., 2020). **(iii) Commonsense Reasoning & QA**: zero-shot assessed using HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), OpenBookQA Mihaylov et al. (2018), BoolQ (Clark et al., 2019), SIQA (Sap et al., 2019) and PIQA (Bisk et al., 2020). **(iv) Natural Language Inferenec(NLI)**: tested on RTE Wang et al. (2018), ANLI Nie et al. (2020) and MNLI Williams et al. (2018) benchmarks. **(v) Knowledge**: general world knowledge and reasoning are measured via MMLU-K the MMLU corpus with math tasks excluded, ARC-Easy, ARC-Challenge (Clark et al., 2018) and GPQA (Rein et al., 2023). Due to time constraints, MathQA is not evaluated for some experiments. However, the results on the remaining 15 datasets provide a comprehensive comparison of different calibration methods, and therefore we report their average as Mean_{15}.

A.2 LANGUAGE MODELING RESULTS

We evaluate the perplexity (PPL) of multi-domain calibration on Llama-3.1-8B-Instruct. The final PPL is computed as the average of the PPLs on C4, Pile, and WikiText. The results are reported in Table 6.

Table 6: Average perplexity on WikiText, Pile, and C4 under different quantization settings. Lower is better. Δ denotes Ours minus the lower perplexity of COLA and ZipCal; negative values indicate improvements.

Model	Method	Setting	COLA	ZipCal	ReCaC	Δ
Llama-3.1-8B	Dense			8.04		–
	GPTQ	W3A16	18.83	13.00	10.59	−2.41
		W4A16	8.76	8.76	8.50	−0.26
	AWQ	W3A16	12.27	12.41	12.35	+0.08
		W4A16	8.71	8.72	8.77	+0.06

A.3 Multi-Domain Results on Qwen3.5-9B

We evaluate multi-domain calibration on Qwen3.5-9B under GPTQ and AWQ with W4A16 and W3A16 quantization. The results are reported in Table 7.

Table 7: Multi-domain comparison of calibration methods on Qwen3.5-9B.

Task	Dense	GPTQ						AWQ					
		W4A16			W3A16			W4A16			W3A16		
		COLA	ZipCal	Ours	COLA	ZipCal	Ours	COLA	ZipCal	Ours	COLA	ZipCal	Ours
MMLU-M	40.74	47.04	47.04	42.96	27.78	27.78	32.22	44.81	33.70	40.00	28.89	27.41	28.15
GSM8k	84.91	83.78	84.31	79.83	54.06	33.43	61.41	82.71	81.73	82.34	24.72	32.68	34.80
HellaSwag	68.94	68.89	68.78	68.71	65.36	67.63	67.54	68.04	68.52	68.38	64.40	63.98	63.44
WinoGr.	63.46	63.38	63.30	63.06	60.69	60.62	61.56	62.90	62.67	62.90	58.64	58.17	58.48
OBQA	44.60	44.00	43.40	42.80	39.60	40.40	41.60	42.80	44.20	44.00	37.60	39.00	38.40
BoolQ	89.30	89.02	89.42	89.14	86.54	86.42	87.89	88.78	88.87	88.78	75.81	80.58	82.35
RTE	81.95	81.59	79.78	80.87	77.98	79.06	79.42	80.51	80.87	80.14	59.21	64.62	70.04
ANLI	69.30	69.00	66.50	67.70	57.40	61.30	56.90	69.20	68.70	68.60	40.40	42.60	44.80
ARC-C	55.97	55.72	58.62	53.84	48.21	48.29	54.10	55.63	56.14	57.17	43.09	44.71	46.33
ARC-E	70.83	70.71	74.58	67.00	63.72	65.61	69.02	70.75	73.53	74.03	55.72	61.15	62.21
MMLU-K	66.36	73.71	73.71	75.12	64.55	64.55	71.37	73.14	65.72	66.89	40.36	49.66	47.15
Mean$_{11}$	66.94	67.89	**68.13**	66.46	58.72	57.74	**62.09**	**67.21**	65.88	66.66	48.08	51.32	**52.38**
MathQA	41.27	40.60	42.41	37.86	35.14	33.57	35.85	38.73	41.14	40.17	28.61	28.58	29.31
SIQA	51.79	50.56	51.79	50.92	49.39	48.82	49.08	50.51	50.26	51.64	44.37	43.14	45.60
PIQA	76.99	76.99	77.58	76.77	76.06	76.93	76.88	77.37	77.80	76.55	73.23	74.27	74.92
MNLI	72.12	67.81	69.16	68.27	56.13	60.92	62.24	67.04	63.67	69.88	42.47	41.82	47.85
GPQA	38.17	35.71	37.05	39.29	29.02	26.79	32.59	39.73	38.62	37.28	27.23	24.55	26.79
Mean$_5$	56.07	54.33	**55.60**	54.62	49.15	49.41	**51.33**	54.68	54.30	**55.10**	43.18	42.47	**44.89**
Mean$_{16}$	63.54	63.66	**64.21**	62.76	55.73	55.13	**58.73**	**63.29**	62.26	63.05	46.55	48.56	**50.04**

A.4 Ablation Study on Calibration Set Initialization Strategies

This section provides detailed results for the ablation study on calibration set initialization strategies. The complete task-level results under GPTQ and AWQ are reported in Table 8.

Table 8: Comparison of calibration set initialization strategies under W4 and W3 quantization.

Task	W4						W3					
	GPTQ			AWQ			GPTQ			AWQ		
	Density	Centroid	Random	Density	Centroid	Random	Density	Centroid	Random	Density	Centroid	Random
MMLU-M	23.70	22.96	24.07	22.59	23.70	23.33	21.85	24.44	21.48	24.44	21.85	24.44
GSM8k	73.69	75.59	77.33	72.78	73.24	72.18	58.45	57.24	54.97	30.55	27.60	29.26
HellaSwag	72.59	72.10	72.32	71.01	71.49	72.13	70.82	70.26	70.86	68.77	69.12	69.02
WinoGr.	64.72	64.80	65.90	63.69	64.96	64.96	63.14	63.14	62.59	60.93	59.43	60.14
OBQA	47.80	45.40	45.00	45.60	43.60	45.60	43.00	42.40	43.00	42.00	43.20	41.00
BoolQ	83.73	83.73	84.10	84.28	83.94	84.07	82.57	81.83	84.04	78.81	78.13	79.27
RTE	72.56	73.29	72.92	73.65	72.56	73.29	72.92	69.31	72.56	66.06	66.43	66.43
ANLI	58.30	58.00	53.70	50.50	50.70	49.10	56.50	47.50	53.50	48.20	49.00	47.60
ARC-C	51.02	49.66	50.09	50.68	51.88	51.62	48.12	48.63	46.84	45.73	45.39	46.08
ARC-E	73.36	72.35	73.23	72.56	73.70	73.53	71.84	71.17	71.63	68.73	68.48	69.36
MMLU-K	57.41	56.06	59.35	59.94	62.01	59.41	49.47	52.43	51.23	39.67	37.84	45.27
Mean$_{11}$	**61.72**	61.27	61.64	60.66	**61.07**	60.84	**58.06**	57.13	57.52	52.17	51.50	**52.53**
SIQA	50.36	48.77	50.31	48.82	49.18	49.59	49.74	46.93	47.49	45.70	45.65	46.01
PIQA	79.05	78.67	78.94	78.24	78.94	78.78	77.58	76.99	78.45	77.69	75.46	76.44
MNLI	56.43	57.29	55.71	55.93	56.44	55.26	45.51	46.43	48.16	40.97	39.82	42.64
GPQA	28.79	30.13	27.68	26.34	27.23	26.12	26.56	28.35	25.45	29.46	27.01	25.67
Mean$_{4}$	**45.94**	43.51	43.20	42.03	42.44	**42.65**	42.67	40.81	**42.97**	**39.68**	37.95	38.58
Mean$_{15}$	**59.57**	59.25	59.38	58.44	**58.90**	58.60	**55.87**	55.14	55.48	51.18	50.29	**51.24**

A.5 Sensitivity Analysis of λ

This section presents detailed task-level results for the sensitivity analysis of λ under GPTQ-W3A16 and AWQ-W3A16, covering $\lambda \in \{0, 0.5, 1, 1.5, 2\}$. The complete results are reported in Table 9, while the corresponding trend visualization is shown in Fig. 3.

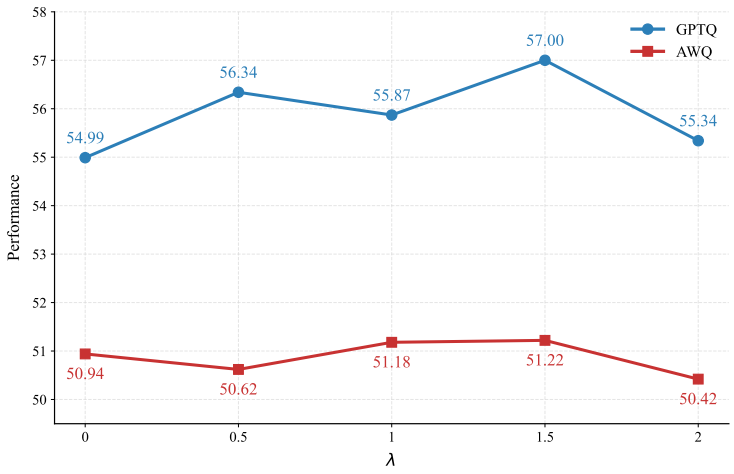


Figure 3: Performance under W3A16 with different values of λ.

Table 9: Ablation study of λ under GPTQ and AWQ W3 quantization.

Task	Setting of λ									
	GPTQ-W3A16					AWQ-W3A16				
	0	0.5	1	1.5	2	0	0.5	1	1.5	2
MMLU-M	21.11	22.59	21.85	32.59	21.85	21.48	22.59	24.44	22.59	21.85
GSM8k	54.59	60.35	58.45	53.53	60.42	29.72	28.73	30.55	32.60	28.28
HellaSwag	69.88	71.36	70.82	71.34	71.01	68.88	69.52	68.77	68.86	69.18
WinoGr.	64.25	62.83	63.14	63.54	62.12	59.91	59.59	60.93	60.38	58.41
OBQA	45.80	44.80	43.00	44.00	43.60	40.00	41.60	42.00	42.40	42.00
BoolQ	83.73	82.29	82.57	83.85	81.47	79.51	78.69	78.81	78.72	78.69
RTE	76.53	71.12	72.92	74.73	72.92	65.70	65.70	66.06	67.15	67.51
ANLI	50.20	55.20	56.50	49.20	48.10	48.40	41.90	48.20	48.60	49.00
ARC-C	47.18	47.44	48.12	47.35	48.29	46.16	46.67	45.73	46.16	43.86
ARC-E	67.72	70.79	71.84	71.97	72.01	69.44	69.44	68.73	70.58	67.00
MMLU-K	46.99	51.57	49.47	54.44	49.22	41.76	46.80	39.67	39.96	35.72
Mean$_{11}$	57.09	58.21	58.06	**58.78**	57.37	51.91	51.93	52.17	**52.55**	51.04
SIQA	47.03	48.62	49.74	50.46	47.75	45.80	46.11	45.70	45.96	45.04
PIQA	78.45	77.86	77.58	78.02	76.22	76.28	77.15	77.69	76.66	77.15
MNLI	45.95	51.88	45.51	55.64	48.77	44.48	42.25	40.97	42.52	44.66
GPQA	25.45	26.34	26.56	24.33	26.34	26.56	22.54	29.46	25.22	27.90
Mean$_{4}$	42.27	**44.00**	42.67	42.78	41.98	39.15	38.46	**39.68**	38.90	39.34
Mean$_{15}$	54.99	56.34	55.87	**57.00**	55.34	50.94	50.62	51.18	**51.22**	50.42

A.6 Task-Wise Single-Domain Results for ReCaC on Llama-3.1-8B-Instruct

Tables 10–13 present detailed task-wise results for ReCaC using single-domain calibration data on Llama-3.1-8B-Instruct across 11 evaluation tasks. The tables cover AWQ-W3A16, AWQ-W4A16, GPTQ-W3A16, and GPTQ-W4A16, respectively. The five calibration domains are LangMod, Math, CommQA, NLI, and Know. The task-label colors follow the domain grouping used in Table 2.

Table 10: Task-wise results of ReCaC single-domain calibration on Llama-3.1-8B-Instruct under AWQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	22.35	23.15	24.63	23.52	22.22	23.17
GSM8k	28.58	24.98	29.08	27.63	27.45	27.54
HellaSwag	68.54	68.15	69.05	69.23	68.60	68.72
WinoGr.	60.09	61.52	61.25	60.46	60.14	60.69
OBQA	43.33	39.00	42.80	41.30	40.90	41.47
BoolQ	78.30	77.91	78.35	78.91	79.07	78.51
RTE	66.55	66.79	64.80	66.61	62.82	65.51
ANLI	47.27	44.10	44.15	48.25	45.30	45.81
ARC-C	45.88	45.56	46.25	45.44	45.90	45.80
ARC-E	67.61	67.13	70.73	68.64	69.76	68.77
MMLU-K	39.43	41.72	44.62	41.78	43.60	42.23
Mean$_{11}$	51.63	50.91	52.34	51.98	51.43	51.66

Table 11: Task-wise results of ReCaC single-domain calibration on Llama-3.1-8B-Instruct under AWQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	24.81	24.07	23.15	23.89	24.44	24.07
GSM8k	72.48	72.86	72.90	73.24	73.69	73.03
HellaSwag	71.71	71.58	71.40	71.97	71.88	71.71
WinoGr.	65.04	64.72	64.96	64.52	64.52	64.75
OBQA	45.13	46.60	45.10	45.30	45.60	45.55
BoolQ	84.25	84.14	84.19	84.01	83.93	84.10
RTE	73.41	74.91	74.01	72.74	73.83	73.78
ANLI	50.23	51.95	51.35	51.40	51.65	51.32
ARC-C	50.63	50.94	50.73	50.43	51.28	50.80
ARC-E	72.53	72.12	72.69	73.06	73.44	72.77
MMLU-K	61.69	60.22	60.13	60.62	60.37	60.60
Mean$_{11}$	61.08	61.28	60.96	61.02	61.33	61.13

Table 12: Task-wise results of ReCaC single-domain calibration on Llama-3.1-8B-Instruct under GPTQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	24.81	26.11	27.59	25.93	22.78	25.44
GSM8k	51.93	44.66	25.28	28.54	35.44	37.17
HellaSwag	70.19	68.87	68.80	69.47	68.86	69.24
WinoGr.	62.35	60.54	62.43	61.13	62.71	61.83
OBQA	43.67	41.20	44.10	43.70	43.10	43.15
BoolQ	82.19	80.31	78.35	81.21	81.77	80.77
RTE	70.28	67.69	63.36	73.83	70.58	69.15
ANLI	51.23	43.95	41.20	49.00	46.55	46.39
ARC-C	47.58	43.98	47.35	46.54	44.50	45.99
ARC-E	71.14	63.68	70.75	69.28	66.81	68.33
MMLU-K	48.02	44.15	44.52	45.45	49.29	46.29
Mean$_{11}$	56.67	53.19	52.16	54.01	53.85	53.98

Table 13: Task-wise results of ReCaC single-domain calibration on Llama-3.1-8B-Instruct under GPTQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	22.59	25.56	23.70	24.44	23.70	24.00
GSM8k	75.56	76.12	73.12	74.87	73.92	74.72
HellaSwag	72.21	72.56	72.08	72.52	71.89	72.25
WinoGr.	65.09	64.72	65.39	64.72	66.02	65.19
OBQA	46.87	47.30	47.40	46.40	45.60	46.71
BoolQ	83.96	84.10	84.08	84.28	84.36	84.15
RTE	73.16	75.09	72.74	73.83	72.92	73.55
ANLI	55.70	52.55	54.55	58.05	58.45	55.86
ARC-C	50.97	51.41	50.34	51.15	50.30	50.83
ARC-E	73.89	73.02	73.65	73.61	72.69	73.37
MMLU-K	58.21	59.59	56.38	59.47	56.79	58.09
Mean$_{11}$	61.66	62.00	61.22	62.12	61.51	61.70

A.7 Task-Wise Single-Domain Results for COLA on Llama-3.1-8B-Instruct

Tables 14–17 present detailed task-wise results for COLA using single-domain calibration data on Llama-3.1-8B-Instruct across 11 evaluation tasks. The tables cover AWQ-W3A16, AWQ-W4A16, GPTQ-W3A16, and GPTQ-W4A16, respectively. The five calibration domains are LangMod, Math, CommQA, NLI, and Know. The task-label colors follow the domain grouping used in Table 2.

Table 14: Task-wise results of COLA single-domain calibration on Llama-3.1-8B-Instruct under AWQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	21.73	22.59	22.22	22.04	23.15	22.35
GSM8k	28.78	22.25	23.69	26.46	31.88	26.61
HellaSwag	69.03	68.67	67.51	69.08	68.89	68.64
WinoGr.	60.91	60.10	61.52	60.73	60.54	60.76
OBQA	43.20	41.30	40.50	43.50	43.10	42.32
BoolQ	78.85	79.43	77.97	78.69	79.79	78.94
RTE	64.14	66.25	67.51	65.88	67.15	66.19
ANLI	47.77	45.15	47.10	47.15	46.15	46.66
ARC-C	45.71	44.58	45.69	45.09	46.50	45.51
ARC-E	69.19	66.73	68.16	68.18	69.93	68.44
MMLU-K	39.41	39.89	33.97	38.05	42.38	38.74
Mean$_{11}$	51.70	50.63	50.53	51.35	52.68	51.38

Table 15: Task-wise results of COLA single-domain calibration on Llama-3.1-8B-Instruct under AWQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	24.40	24.60	24.30	24.40	24.50	24.44
GSM8k	72.35	62.51	69.67	69.60	70.77	68.98
HellaSwag	72.20	71.38	71.28	72.10	71.45	71.68
WinoGr.	67.19	67.05	67.52	65.98	66.02	66.75
OBQA	46.60	44.90	46.50	45.20	46.30	45.90
BoolQ	83.92	84.28	84.20	83.70	84.24	84.07
RTE	74.01	72.92	73.47	72.02	73.10	73.10
ANLI	56.33	50.95	53.40	53.30	53.80	53.56
ARC-C	50.03	51.41	49.62	50.04	49.96	50.21
ARC-E	73.36	71.72	72.14	72.69	72.85	72.55
MMLU-K	58.52	58.52	58.31	58.11	59.12	58.52
Mean$_{11}$	60.81	59.97	60.84	60.72	61.13	60.89

Table 16: Task-wise results of COLA single-domain calibration on Llama-3.1-8B-Instruct under GPTQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	24.94	24.81	21.30	21.48	22.78	23.06
GSM8k	0.88	0.57	0.53	2.08	17.59	4.33
HellaSwag	49.16	26.29	34.94	44.57	48.02	40.60
WinoGr.	54.46	51.42	52.01	54.97	56.08	53.79
OBQA	33.73	29.70	29.50	34.40	34.80	32.43
BoolQ	58.74	43.81	48.99	65.76	59.65	55.39
RTE	54.03	52.53	51.99	55.42	60.83	54.96
ANLI	34.50	33.85	31.40	34.75	41.25	35.15
ARC-C	31.48	24.02	25.13	31.31	33.92	29.17
ARC-E	47.97	28.24	35.29	43.64	48.82	40.79
MMLU-K	31.70	23.73	24.22	29.85	34.93	28.89
Mean$_{11}$	38.33	30.82	32.30	38.02	41.70	36.23

Table 17: Task-wise results of COLA single-domain calibration on Llama-3.1-8B-Instruct under GPTQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	23.33	24.44	23.01	21.85	23.33	23.19
GSM8k	71.75	68.46	68.84	63.99	65.01	67.61
HellaSwag	72.07	72.14	71.91	72.06	71.54	71.94
WinoGr.	66.75	68.15	66.34	66.30	66.73	66.85
OBQA	45.60	46.40	46.20	43.90	45.90	45.60
BoolQ	84.70	84.08	84.08	84.74	83.53	84.23
RTE	74.61	75.63	71.30	74.91	74.37	74.16
ANLI	57.47	52.85	53.55	54.40	58.25	55.30
ARC-C	50.14	50.38	50.21	50.77	50.64	50.43
ARC-E	73.55	73.76	72.62	73.30	73.88	73.42
MMLU-K	57.46	60.45	57.48	58.45	58.35	58.44
Mean$_{11}$	61.58	61.52	60.50	60.42	61.05	61.02

A.8 TASK-WISE SINGLE-DOMAIN RESULTS FOR ZIPCAL ON LLAMA-3.1-8B-INSTRUCT

Tables 18 and 19 present detailed task-wise results for ZipCal using single-domain calibration data on Llama-3.1-8B-Instruct across 11 evaluation tasks. The tables cover AWQ-W4A16 and GPTQ-W4A16, respectively. The five calibration domains are LangMod, Math, CommQA, NLI, and Know. The task-label colors follow the domain grouping used in Table 2.

Table 18: Task-wise results of ZipCal single-domain calibration on Llama-3.1-8B-Instruct under AWQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	24.50	24.30	24.40	24.20	24.70	24.42
GSM8k	71.80	63.10	68.90	70.10	70.50	68.88
HellaSwag	72.40	71.10	71.50	71.80	71.60	71.68
WinoGr.	66.90	67.30	67.20	66.50	65.80	66.74
OBQA	46.80	44.50	46.10	45.50	46.50	45.88
BoolQ	83.50	84.40	83.80	84.10	84.00	83.96
RTE	73.80	73.10	73.00	72.50	73.30	73.14
ANLI	55.80	51.50	53.10	53.90	53.40	53.54
ARC-C	50.50	50.80	49.90	49.70	50.30	50.24
ARC-E	72.90	71.90	71.80	72.80	72.50	72.38
MMLU-K	58.10	58.80	58.15	58.30	58.90	58.45
Mean$_{11}$	60.64	60.07	60.67	60.85	61.17	60.85

Table 19: Task-wise results of ZipCal single-domain calibration on Llama-3.1-8B-Instruct under GPTQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	23.01	26.29	21.24	22.02	23.01	23.11
GSM8k	71.99	74.60	65.13	70.96	67.32	70.00
HellaSwag	71.84	72.71	71.46	71.89	72.37	72.05
WinoGr.	66.34	67.40	66.69	67.05	67.32	66.96
OBQA	45.20	47.20	45.60	47.50	46.20	46.34
BoolQ	82.63	82.68	83.79	83.91	83.35	83.27
RTE	72.38	70.76	71.84	73.10	72.92	72.20
ANLI	53.45	51.90	55.20	56.25	55.05	54.37
ARC-C	50.21	50.34	50.17	49.66	51.96	50.47
ARC-E	72.73	74.07	74.33	72.90	75.48	73.90
MMLU-K	54.42	58.64	57.40	56.96	58.12	57.11
Mean$_{11}$	60.38	61.51	60.26	61.11	61.19	60.89

A.9 TASK-WISE SINGLE-DOMAIN RESULTS FOR RECAC ON GEMMA-2-9B-IT

Tables 20–23 present detailed task-wise results for ReCaC using single-domain calibration data on Gemma-2-9B-it across 11 evaluation tasks. The tables cover AWQ-W3A16, AWQ-W4A16, GPTQ-W3A16, and GPTQ-W4A16, respectively. The five calibration domains are LangMod, Math, CommQA, NLI, and Know. The task-label colors follow the domain grouping used in Table 2.

Table 20: Task-wise results of ReCaC single-domain calibration on Gemma-2-9B-it under AWQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	20.86	21.11	21.67	20.93	20.93	21.10
GSM8k	66.94	67.02	68.65	66.38	68.20	67.44
HellaSwag	66.75	66.58	66.77	66.58	66.47	66.63
WinoGr.	66.40	66.77	65.39	65.63	65.79	66.00
OBQA	44.33	43.90	44.30	43.50	44.20	44.05
BoolQ	87.57	87.60	87.83	87.77	87.71	87.70
RTE	78.46	77.80	77.44	79.06	78.34	78.22
ANLI	61.20	65.80	64.30	59.75	62.15	62.64
ARC-C	48.58	48.17	47.40	49.23	47.78	48.23
ARC-E	65.01	65.76	65.05	64.54	64.14	64.90
MMLU-K	25.65	26.30	25.59	25.21	24.74	25.50
Mean$_{11}$	57.43	57.89	57.67	57.14	57.31	57.49

Table 21: Task-wise results of ReCaC single-domain calibration on Gemma-2-9B-it under AWQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	22.72	21.48	21.67	22.59	22.22	22.14
GSM8k	73.77	73.96	73.54	73.58	73.28	73.62
HellaSwag	68.04	68.15	67.79	68.02	67.96	67.99
WinoGr.	67.98	68.47	67.60	68.03	68.15	68.05
OBQA	45.00	45.30	45.60	45.30	45.00	45.24
BoolQ	88.39	88.35	88.73	88.17	88.30	88.39
RTE	77.62	77.98	77.26	77.44	77.44	77.55
ANLI	69.97	70.75	71.10	69.60	70.35	70.35
ARC-C	51.51	51.45	50.98	51.07	51.11	51.22
ARC-E	67.17	66.90	66.54	67.45	67.30	67.07
MMLU-K	47.49	38.17	43.09	45.96	44.38	43.82
Mean$_{11}$	61.79	61.00	61.26	61.56	61.41	61.40

Table 22: Task-wise results of ReCaC single-domain calibration on Gemma-2-9B-it under GPTQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	22.22	21.48	21.48	22.22	21.67	21.81
GSM8k	70.41	72.29	70.43	69.64	71.49	70.85
HellaSwag	67.30	66.89	67.32	67.21	66.92	67.13
WinoGr.	65.98	66.61	66.93	65.75	67.09	66.47
OBQA	43.47	44.90	44.50	44.30	44.30	44.29
BoolQ	87.75	86.33	87.83	87.06	87.78	87.35
RTE	78.34	78.16	77.98	77.26	78.34	78.01
ANLI	63.40	66.75	68.05	66.65	64.15	65.80
ARC-C	49.69	50.26	50.68	49.66	50.13	50.08
ARC-E	65.59	66.81	67.78	64.60	66.33	66.22
MMLU-K	37.77	43.24	37.37	44.49	32.28	39.03
Mean$_{11}$	59.26	60.34	60.03	59.89	59.13	59.73

Table 23: Task-wise results of ReCaC single-domain calibration on Gemma-2-9B-it under GPTQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	22.22	21.30	21.85	22.04	21.48	21.78
GSM8k	74.10	74.94	74.07	73.92	73.88	74.18
HellaSwag	68.02	68.02	67.93	67.76	67.96	67.94
WinoGr.	68.11	67.84	68.43	68.11	68.27	68.15
OBQA	46.07	45.10	45.00	45.10	44.60	45.17
BoolQ	88.40	88.58	88.38	88.69	88.52	88.51
RTE	77.50	75.99	76.71	76.35	78.16	76.94
ANLI	69.37	70.50	70.60	69.15	71.05	70.13
ARC-C	51.17	51.02	50.17	50.73	50.81	50.78
ARC-E	66.71	65.93	65.43	65.76	66.18	66.00
MMLU-K	51.77	37.67	49.35	51.82	43.60	46.84
Mean$_{11}$	62.13	60.63	61.63	61.77	61.32	61.49

A.10 TASK-WISE SINGLE-DOMAIN RESULTS FOR COLA ON GEMMA-2-9B-IT

Tables 24–27 present detailed task-wise results for COLA using single-domain calibration data on Gemma-2-9B-it across 11 evaluation tasks. The tables cover AWQ-W3A16, AWQ-W4A16, GPTQ-W3A16, and GPTQ-W4A16, respectively. The five calibration domains are LangMod, Math, CommQA, NLI, and Know. The task-label colors follow the domain grouping used in Table 2.

Table 24: Task-wise results of COLA single-domain calibration on Gemma-2-9B-it under AWQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	21.48	22.22	21.30	21.30	21.11	21.48
GSM8k	66.24	67.59	66.41	66.98	66.76	66.80
HellaSwag	66.85	66.57	66.72	66.70	66.95	66.76
WinoGr.	65.46	65.75	65.67	65.04	65.43	65.47
OBQA	43.73	44.40	44.50	44.20	45.30	44.43
BoolQ	87.55	87.43	87.16	87.60	87.51	87.45
RTE	78.82	78.34	77.80	77.26	78.34	78.11
ANLI	64.30	65.75	59.30	66.20	65.00	64.11
ARC-C	48.75	48.46	49.23	48.89	48.72	48.81
ARC-E	64.10	64.71	64.73	64.96	66.12	64.92
MMLU-K	24.76	25.42	25.13	24.38	24.66	24.87
Mean$_{11}$	57.46	57.88	57.09	57.59	57.81	57.56

Table 25: Task-wise results of COLA single-domain calibration on Gemma-2-9B-it under AWQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	21.40	21.40	21.35	21.45	21.40	21.40
GSM8k	74.17	74.91	74.07	73.96	74.00	74.22
HellaSwag	67.25	67.81	68.04	67.39	67.42	67.58
WinoGr.	69.53	69.65	69.73	69.34	69.65	69.58
OBQA	45.47	44.70	45.20	45.50	45.00	45.17
BoolQ	88.53	88.53	88.61	88.58	88.59	88.57
RTE	79.06	77.26	76.90	77.26	77.44	77.58
ANLI	71.77	72.10	71.90	70.70	72.35	71.76
ARC-C	51.11	51.49	51.41	51.58	50.81	51.28
ARC-E	66.40	66.35	67.09	66.88	66.58	66.66
MMLU-K	35.73	28.90	30.12	31.79	33.46	32.00
Mean$_{11}$	60.95	60.26	60.22	60.40	60.61	60.53

Table 26: Task-wise results of COLA single-domain calibration on Gemma-2-9B-it under GPTQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	20.86	22.41	21.11	21.30	21.48	21.43
GSM8k	70.41	72.18	64.18	67.74	69.33	68.77
HellaSwag	67.33	66.48	66.84	65.82	66.93	66.68
WinoGr.	66.80	64.88	67.09	65.04	66.73	66.11
OBQA	43.73	44.40	43.30	43.00	44.00	43.69
BoolQ	87.87	87.17	86.73	87.02	88.06	87.37
RTE	78.34	75.63	77.98	78.34	78.70	77.80
ANLI	68.07	64.80	67.00	68.35	68.15	67.27
ARC-C	49.69	48.89	49.83	48.98	49.83	49.44
ARC-E	65.36	65.07	64.54	62.67	64.71	64.47
MMLU-K	27.23	36.28	28.95	32.83	31.49	31.36
Mean$_{11}$	58.70	58.93	57.96	58.28	59.04	58.58

Table 27: Task-wise results of COLA single-domain calibration on Gemma-2-9B-it under GPTQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	24.44	23.01	22.02	21.24	21.85	22.44
GSM8k	74.12	74.30	74.41	74.18	73.81	74.16
HellaSwag	67.43	67.56	68.23	67.43	67.48	67.63
WinoGr.	69.93	70.01	69.69	69.34	69.61	69.72
OBQA	45.47	44.50	44.20	44.10	45.40	44.73
BoolQ	88.66	88.62	88.32	88.50	88.76	88.57
RTE	78.10	76.53	78.34	77.26	77.62	77.57
ANLI	71.07	71.75	71.80	71.90	72.00	71.70
ARC-C	52.28	50.77	51.79	51.32	51.41	51.51
ARC-E	67.31	66.50	66.73	66.22	67.66	66.88
MMLU-K	33.23	26.74	29.17	29.70	37.46	31.26
Mean$_{11}$	61.09	60.03	60.43	60.11	61.19	60.56

A.11 TASK-WISE SINGLE-DOMAIN RESULTS FOR ZIPCAL ON GEMMA-2-9B-IT

Tables 28–31 present detailed task-wise results for ZipCal using single-domain calibration data on Gemma-2-9B-it across 11 evaluation tasks. The tables cover AWQ-W3A16, AWQ-W4A16, GPTQ-W3A16, and GPTQ-W4A16, respectively. The five calibration domains are LangMod, Math, CommQA, NLI, and Know. The task-label colors follow the domain grouping used in Table 2.

Table 28: Task-wise results of ZipCal single-domain calibration on Gemma-2-9B-it under AWQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	21.60	21.30	21.11	21.48	20.74	21.25
GSM8k	66.57	67.66	67.59	67.06	67.32	67.24
HellaSwag	66.92	66.94	66.33	66.96	66.64	66.76
WinoGr.	66.35	67.64	66.65	66.34	65.71	66.54
OBQA	44.00	44.40	44.90	44.00	44.30	44.32
BoolQ	87.64	87.25	87.74	87.31	87.71	87.53
RTE	78.10	78.16	77.80	77.26	77.80	77.82
ANLI	66.33	62.45	66.40	61.60	61.75	63.71
ARC-C	48.63	49.15	48.68	50.04	47.91	48.88
ARC-E	66.57	64.41	65.95	67.07	65.30	65.86
MMLU-K	29.15	27.44	26.84	32.08	28.86	28.87
Mean$_{11}$	58.35	57.89	58.18	58.29	57.64	58.07

Table 29: Task-wise results of ZipCal single-domain calibration on Gemma-2-9B-it under AWQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	21.40	21.35	21.30	21.50	21.45	21.39
GSM8k	74.50	74.50	73.80	74.20	73.70	74.14
HellaSwag	66.90	67.90	67.50	67.50	67.20	67.40
WinoGr.	69.80	69.20	69.90	69.10	69.40	69.48
OBQA	45.10	45.00	44.90	45.80	44.80	45.12
BoolQ	88.70	88.40	88.50	88.70	88.40	88.54
RTE	78.50	77.50	76.50	77.00	77.60	77.42
ANLI	72.00	71.80	72.10	70.50	72.10	71.70
ARC-C	50.80	51.70	51.20	51.80	50.50	51.20
ARC-E	66.60	66.10	66.80	67.10	66.30	66.58
MMLU-K	35.10	29.50	29.80	32.10	33.10	31.92
Mean$_{11}$	61.03	60.27	60.21	60.48	60.41	60.44

Table 30: Task-wise results of ZipCal single-domain calibration on Gemma-2-9B-it under GPTQ-W3A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	21.60	22.59	21.85	22.04	21.67	21.95
GSM8k	63.00	68.20	66.79	61.11	65.35	64.89
HellaSwag	65.74	65.99	66.09	65.27	66.31	65.88
WinoGr.	66.25	68.51	67.05	66.30	67.17	67.05
OBQA	42.47	44.20	43.40	42.80	44.30	43.43
BoolQ	87.37	87.66	86.97	87.26	86.87	87.23
RTE	76.53	78.34	77.62	77.44	78.52	77.69
ANLI	52.03	63.85	66.10	59.60	50.45	58.41
ARC-C	47.81	47.87	46.63	48.29	47.65	47.65
ARC-E	63.33	63.43	63.26	62.69	63.19	63.18
MMLU-K	36.01	42.39	34.46	33.66	28.66	35.04
Mean$_{11}$	56.56	59.37	58.20	56.95	56.38	57.49

Table 31: Task-wise results of ZipCal single-domain calibration on Gemma-2-9B-it under GPTQ-W4A16.

Task	LangMod	Math	CommQA	NLI	Know	Mean
MMLU-M	27.40	24.44	26.29	28.79	24.44	26.27
GSM8k	73.67	73.92	73.39	73.81	73.54	73.66
HellaSwag	68.02	67.61	67.81	67.36	67.37	67.63
WinoGr.	69.69	70.60	70.38	70.01	69.02	69.92
OBQA	45.07	44.90	45.20	45.00	45.00	45.03
BoolQ	88.64	88.62	88.58	89.07	88.49	88.68
RTE	77.26	77.80	77.80	78.16	78.52	77.91
ANLI	72.20	72.90	70.30	72.55	71.95	71.98
ARC-C	52.22	52.09	52.43	51.45	52.52	52.14
ARC-E	68.83	67.82	68.81	66.41	68.22	68.02
MMLU-K	36.76	30.64	34.03	28.79	41.04	34.25
Mean$_{11}$	61.80	61.03	61.37	61.04	61.83	61.41