OBEQuant: Outlier-aware Batch-norm Equivalent Quantization for Large Language Models

Ye Zhong, Bin Ji*, Xiaodong Liu*, Shasha Li, Jun Ma, Jie Yu*
College of Computer Science and Technology, National University of Defense Technology, Changsha, China
Email: {zhongye25, jibin, liuxiaodong, shashali, majun, yj}@nudt.edu.cn

***Abstract*—Post-training quantization (PTQ) is pivotal for deploying Large Language Models (LLMs) in resource-constrained environments, yet aggressive low-bit quantization (*e.g.* W4A4) leads to severe accuracy degradation due to channel-wise mean bias and dynamic residual errors. To address this issue, we propose OBEQuant, a collaborative error-compensation framework featuring a cascaded dual-module design: a Bias Mean Correction (BMC) module that first eliminates systematic per-channel mean bias after rotation, and a Quantization-Aware Efficient Fine-Tuning (EFT) module that then compensates for task-dependent dynamic residuals via an ultra-low-rank adapter. These modules synergize to form an error-suppression loop, enabling layer-adaptive error decoupling and compensation. Extensive experiments on LLaMA-2 and LLaMA-3 show that under the challenging W4A4KV4 configuration, OBEQuant significantly outperforms previous state-of-the-art baselines in terms of zero-shot average accuracy, while maintaining competitive perplexity. Quantitative and qualitative analyses further validate the effectiveness of OBEQuant.**

***Index Terms*—large language models; post-training quantization; low-bit quantization; error compensation; bias correction; parameter-efficient fine-tuning**

I. Introduction

Large Language Models (LLMs) [1] have demonstrated exceptional performance across various natural language processing tasks, driving profound innovation in both academic research and industrial applications. As LLMs gradually integrate into daily life and industrial practices [2], they play an irreplaceable role in critical domains such as intelligent interaction, content generation, and decision support. However, the exponential growth in models scale (*i.e.*, parameters ranging from billions to trillions) incurs substantial memory consumption and computational overhead, posing severe challenges for LLMs' deployment. These challenges not only hinder applying LLMs to resource-limited edge-side devices but also impose significant burdens on resource-rich cloud devices, leading to high operational costs and inference latency.

Post-Training Quantization (PTQ), as mainstream techniques to quantize and accelerate LLMs, which enables LLMs lightweighting without relying on their original training datasets, thus facilitates LLMs' practical deployment. Among various PTQ schemes, low-bit quantization (*e.g.*, 4-bit or lower) is particularly crucial for tackling the aforementioned challenges. However, this aggressive reduction in LLMs' bit-width inevitably leads to significant LLMs' ability degradation, which can be attributed to the prevalent presence of notable outliers in LLMs' activations and weights [3]–[5]. To be more precise, these outliers passively force the range of quantization values to expand, thereby harming the precision of these values that are normally distributed, As a result, they amplify quantization errors and lead to the reduction of LLMs' ability.

To tackle the above issues, existing studies have tried to propose solutions that can be broadly categorized into linear transformation and rotation transformation [6]. Linear transformation studies aim to optimize the data distribution in specific regions of the quantization space through smooth scaling. For example, SmoothQuant [7] adjusts scaling factors between weights and activations, and transfers the dynamic value range of activations to the weight side, thereby reducing the interference of outliers on quantization accuracy. Rotation transformation methods suppress outlier influence through spatial rotation. For instance, QuaRot [8] employs rotation matrices to restructure the distributions of quantization values and compress dynamic ranges. Furthermore, subsequent studies, *e.g.*, SpinQuant [9] and OSTQuant [10], propose to use fixed rotation-based strategies such as the Hadamard transform. Although these studies have achieved promising performance, they still suffer from inherent limitations. Firstly, the fixed rotation transformation mode is difficult to adapt to the uneven distribution of quantized values in different layers of large language models [11]–[13]. Secondly, these methods have limited effectiveness in dealing with the mean deviation issues between channels after rotation, and fail to effectively eliminate the systematic rounding errors caused by them.

To tackle the above issue, we propose OBEQuant, an **O**utlier-aware **B**atch-norm **E**quivalent **Quant**ization method, with the following core contributions:

1) To address the systematic quantization error caused by channel mean deviation, we proposed a learnable bias mean correction (BMC) module. This module explicitly aligns the activation distribution means of each channel before quantization, and more efficiently utilizes the quantization intervals to represent the variance of the real data. This preprocessing step effectively relieves the residual channel mean deviation after rotation and significantly reduces the base rounding error.
2) To address the problem that fixed quantization strategies are difficult to adapt to task-specific distributions and suf-

*Corresponding author.

fer from information loss, we introduced a quantization-aware efficient fine-tuning (EFT) module. In this stage, based on the BMC correction, by integrating a low-rank adapter to learn the output differences between the quantization model and the full-precision model on a specific task, it compensates for the task-critical information lost during quantization with minimal parameter overhead, achieving precise adaptation to downstream tasks.

3) We conduct extensive experiments on LLaMA models to validate the effectiveness of OBEQuant. Experimental results demonstrate that under the aggressive W4A4KV4 setting, LLaMA-3-8B achieves 64.46% zero-shot average accuracy, outperforming the previous best OSTQuant by a large margin, while WikiText-2 perplexity falls to 7.36, matching or even surpassing the full-precision LLMs. Consistent gains are observed on LLaMA-2-7B across multiple tasks. Whereas 4-bit recipes often incur severe degradation, our OBEQuant delivers stable and reproducible recovery, extending the practical frontier of PTQ techniques.

To sum up, OBEQuant systematically addresses the limitations existing in current advanced methods through a three-level collaborative mechanism of global distribution optimization (*i.e.*, orthogonal scaling), local deviation correction (*i.e.*, MBC), and parameter efficient adaptation (*i.e.*, EFT), supplemented by a robust optimization objective (*i.e.*, KL-Top Loss). We demonstrate that OBEQuant provides a new idea for achieving higher-precision and more robust low-bit LLM quantization.

II. RELATED WORK

Post-training quantization (PTQ) for large language models (LLMs) compresses and accelerates inference by mapping high-precision weights to low-bit formats without full retraining. Research has followed three tracks: weight-only, activation-only, and joint weight-activation quantization. Each generation tried to shrink the error source found by the previous one.

Early works set the core templates. GPTQ [14] extends optimal brain damage to billion-parameter models; it compensates second-order error layer-wise with inverse-Hessian information and first enables accurate W4A16 weight-only quantization. LLM.int8() [1] shows that outliers in activations are the main barrier to low-bit activation quantization; it keeps outlier feature dimensions in FP16 and quantizes the rest to INT8, giving later studies a clear target.

As research deepened, scholars realized that not all weight channels contribute equally to the final output in weighted quantization. A series of works, such as AWQ [15] and OWQ [16], proposed the idea of activation-aware weighted quantization, which reduces errors by protecting significant weight channels. This idea evolved into parameterized calibration, such as OmniQuant [17] and SignRound [18], which achieves breakthroughs in lower bit depths within the PTQ framework through learnable weight pruning and quantization parameter optimization.

In the more challenging area of weight-activation joint quantization, research focuses on smoothing the activation distribution to suppress outliers. SmoothQuant [7] transfers some of the activation quantization difficulty to the weights through transferable mathematical equivalent transformations, achieving high-precision W8A8 quantization. Subsequent works, such as Outlier Suppression+ [19], introduce channel offsets, and RPTQ [20] uses reordering group quantization to address differences in channel ranges. Another important path is rotation-based quantization, such as QuaRot [8] and SpinQuant [9], which disperse outliers through random or learned rotations. Recent works such as DuQuant [21], ParoQuant [12] and WUSH [22] further systematically handle outliers through rotation and permutation, paired rotation or optimized linear transformation.

Yet rotation and orthogonal methods still face two challenges. First, they realign the global distribution but leave per-channel means mis-aligned, creating systematic bias that turns into rounding error at low bit-width. Second, PTQ uses only a few calibration samples; optimizing with cross-entropy over the full vocabulary injects tail-distribution noise and yields sub-optimal, less generalizable transforms.

III. ERROR ANALYSES

This section conducts a theoretical error analysis of the rotation quantization method, aiming to reveal its inherent limitations and provide a theoretical basis for the proposed method. We first establish a general analytical framework for rotation quantization errors, then decompose and quantize different error sources, focusing on clarifying the causes of channel mean deviation and its fundamental constraint on quantization performance.

A. Rotational Quantization Error Decomposition

Consider the forward propagation process of a linear layer, whose full-precision computation is $\mathbf{Y} = \mathbf{X}\mathbf{W}^T$, where $\mathbf{X} \in \mathbb{R}^{k\times n}$ are the input activations and $\mathbf{W} \in \mathbb{R}^{m\times n}$ are the weights. After applying the orthogonal transformation $\mathbf{R} \in \mathbb{R}^{n\times n}$, the computation process is equivalently transformed to $\mathbf{Y} = (\mathbf{X}\mathbf{R})(\mathbf{R}^T\mathbf{W}^T)$. Quantizing the transformed activations $\mathbf{X}' = \mathbf{X}\mathbf{R}$ and weights $\mathbf{W}' = \mathbf{W}\mathbf{R}$ yields the quantized output:

$$\hat{\mathbf{Y}} = Q_a(\mathbf{X}')Q_w(\mathbf{W}')^T, \tag{1}$$

where $Q_a(\cdot)$ and $Q_w(\cdot)$ represent the quantization functions for the activation and weights, respectively. The quantization error ε is defined as the difference between the quantized output and the full-precision output:

$$\varepsilon = \hat{\mathbf{Y}} - \mathbf{Y} = Q_a(\mathbf{X}\mathbf{R})Q_w(\mathbf{W}\mathbf{R})^T - (\mathbf{X}\mathbf{R})(\mathbf{W}\mathbf{R})^T. \tag{2}$$

This error can be further decomposed into two main parts: rounding error $\varepsilon_{\text{round}}$ and clipping error $\varepsilon_{\text{clip}}$. Rounding error arises from the process of mapping continuous values to a discrete quantization grid, while clipping error arises from values outside the quantization range being truncated to the boundary.

B. Channel Mean Deviation and Inherent Rounding Error

The core assumption of existing rotation quantization methods is that an orthogonal transformation $\mathbf{R}$ can smooth the distribution of activation $\mathbf{X}$ and suppress outliers. However, our analysis shows that even after the optimal rotation transformation, mean bias of the activation tensor still exists across different channels.

Let μ'_j be the mean of the j-th column of the transformed activation $\mathbf{X}'$. The quantization process typically uses a scaling factor s_j per channel. The expected sum of squares of the rounding error can be approximated as:

$$\hat{X}'_{ij} = \text{clamp}\left(\left\lfloor \frac{X'_{ij}}{s_j} \right\rceil, q_{\min}, q_{\max}\right) \cdot s_j, \tag{3}$$

where $\lfloor \cdot \rceil$ represents rounding, clamp$(\cdot)$ is the clipping function, and $q_{\min}$ and $q_{\max}$ are the integer boundaries of the quantization grid. The expected sum of squares of rounding errors can be approximated as:

$$\mathbb{E}[\|\varepsilon_{\text{round}}\|_F^2] \approx \frac{s_j^2}{12} \cdot k \cdot n + \sum_{j=1}^{n} k \cdot (\mu'_j - \bar{\mu}')^2, \tag{4}$$

where $\bar{\mu}' = \frac{1}{n}\sum_{j=1}^{n} \mu'_j$ is the global mean of all channels. The first term $\frac{s_j^2}{12} \cdot k \cdot n$ is the noise term, which is proportional to the square of the scaling factor s_j. The second term, $\sum_{j=1}^{n} k \cdot (\mu'_j - \bar{\mu}')^2 = k \cdot n \cdot \text{Var}(\mu'_j)$, directly quantifies the contribution of the channel mean variance $\text{Var}(\mu'_j)$ to the total error.

Analysis reveals that the orthogonal transformation $\mathbf{R}$ is essentially a zero-mean operation and cannot change or eliminate the differences between the original means μ_j of each channel. The variance $Var(\mu'_j)$ of the transformed channel mean $\mu'_j = \sum_i R_{ji}\mu_i$ may still be significant. This variance directly contributes to rounding error and is independent of the number of quantization bits, constituting an inherent bias in the quantization process. This means that a portion of the quantization interval must be used to represent these static channel mean differences, leading to reduced effective utilization of the quantization interval and introducing systematic errors that cannot be eliminated by optimizing the rotation itself.

C. Gaussianization of Distribution and Pruning Energy Loss

Another effect of rotation is to make the activation distribution closer to a Gaussian distribution. For a Gaussian distribution, more probability mass is concentrated around the mean. When a fixed pruning threshold τ is used, the rotated distribution $p_{\mathbf{X}'}(x)$, which is closer to Gaussian, will have more effective information pruned away. Let the total energy (second moment) clipped be:

$$E_{\text{clip}}^{\text{orig}} = \int_{|x|>\tau} x^2 p_{\mathbf{X}}(x) dx. \tag{5}$$

Although the absolute tail value of $p_{\mathbf{X}'}(x)$ may be smaller, the change in its distribution shape increases the proportion of activations clipped at a fixed threshold. This energy loss due to clipping accumulates layer by layer, violating the computational equivalence established by rotation, and ultimately leading to a degraded model performance.

IV. METHODOLOGY

This section introduces the OBEQuant method. The core idea of this method is to optimize the distribution of weights and activations during quantization by synergistically employing orthogonal transformation, bias mean correction, and quantization-aware efficient fine-tuning, thereby minimizing quantization error. The overall framework is shown in Fig.1, comprising three key modules: distribution reshaping based on orthogonal and scaling transformations, explicit bias mean correction, and quantization-aware efficient fine-tuning.

A. Orthogonal and Scaling Transformations

We first introduce a learnable equivalent transformation pair for each fully connected layer, consisting of an orthogonal transformation $\mathbf{R} \in \mathbb{R}^{n\times n}$ and a channel-wise scaling transformation $\mathbf{S} = \text{diag}(s_1, s_2, ..., s_n)$, where $s_i > 0$. For the linear layer operation $\mathbf{Y} = \mathbf{X}\mathbf{W}^T$, we transform it equivalently as:

$$\mathbf{Y} = (\mathbf{X}\mathbf{R}\mathbf{S}^{-1})(\mathbf{S}\mathbf{R}^T\mathbf{W}^T), \tag{6}$$

$\mathbf{S}\mathbf{R}^T\mathbf{W}^T$ can be pre-computed offline and fused into the weights, while $\mathbf{R}\mathbf{S}^{-1}$ acts on the activation during inference. The orthogonal transformation $\mathbf{R}$ aims to smooth outliers through inter-channel mixing; the scaling transformation $\mathbf{S}$ is used to adjust the amplitude of each channel, collectively optimizing the overall data distribution in the quantization space. This design inherits and enhances the ideas of OSTQuant [10]: the orthogonal transformation preserves the vector norm and rotates the distribution without introducing numerical bias; the scaling transformation provides direct control over the distribution amplitude.

B. Explicit Bias Mean Correction

As BASE-Q [3] points out, while orthogonal transformations can suppress outliers, they cannot eliminate mean bias between channels. Let the activation tensor $\mathbf{X} \in \mathbb{R}^{k\times n}$ have a mean of μ_j in the j-th column. After the transformation, the variance $Var(\mu'_j)$ of the channel means μ'_j of $\mathbf{X}' = \mathbf{X}\mathbf{R}$ may still be significant. According to quantization error analysis, the rounding error energy is proportional to the square of the quantization scale s, and s is usually related to the activation standard deviation σ. Channel mean misalignment causes some quantization intervals to be used to represent static bias rather than the true data variance, thus reducing the effective utilization of the quantization intervals and introducing inherent errors:

$$\mathbb{E}[\|\varepsilon_{\text{round}}\|_2^2] \propto s^2 \propto \sigma^2 + \text{Var}(\mu'_j). \tag{7}$$

To eliminate this error, we add a learnable channel-wise bias term $\mathbf{B}^c \in \mathbb{R}^{1\times n}$ to the quantization-forward activation:

$$\hat{\mathbf{X}} = \mathbf{X}' - \mathbf{B}^c. \tag{8}$$

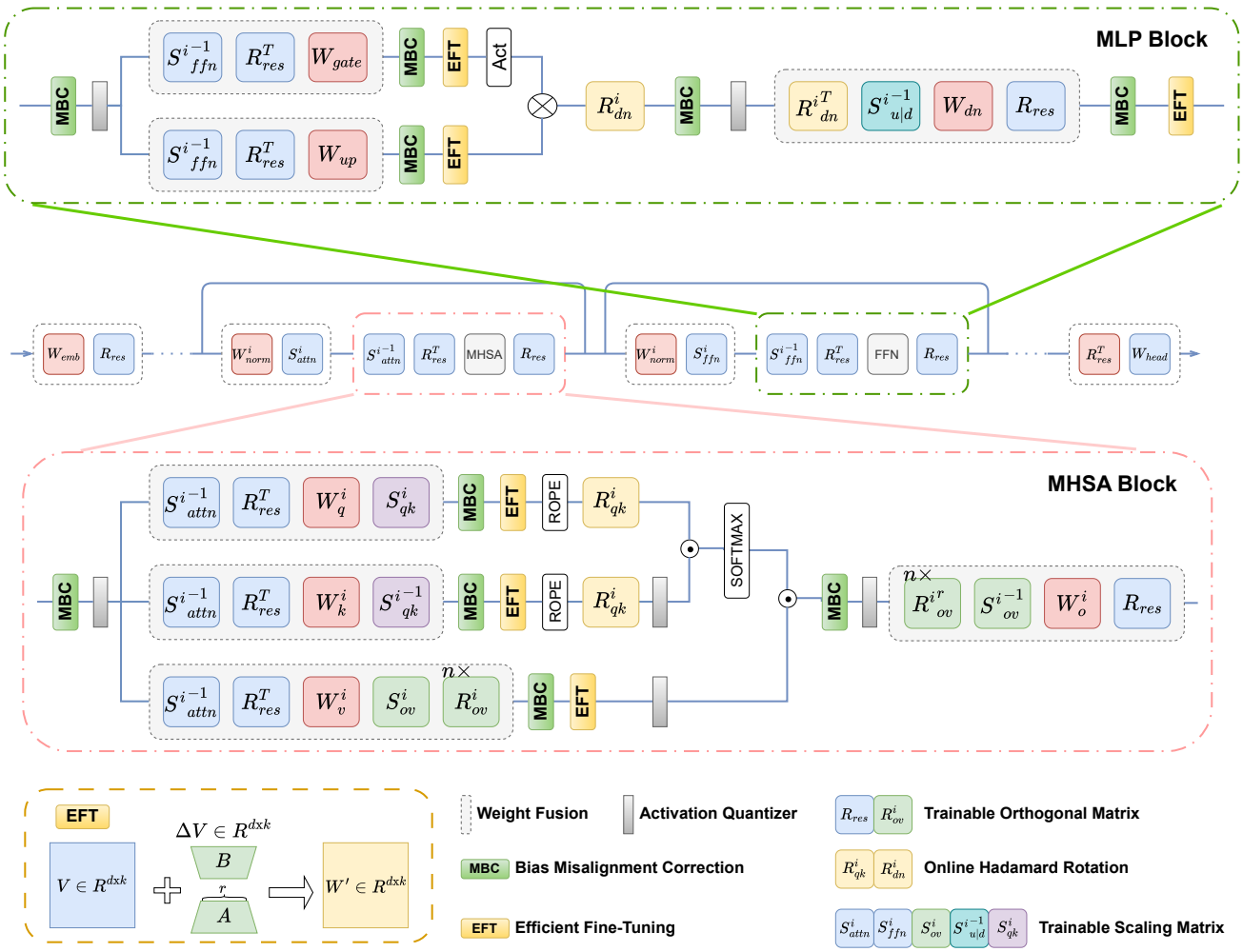

Fig. 1. Overall architecture of the OBEQuant method. The input activation X and weights W are first reshaped through a learnable orthogonal transformation R and a scaling transformation S. Subsequently, the activations enter the explicit bias mean correction module, where the channel mean is aligned by subtracting the learnable channel bias B^c. The corrected activations $\hat{X}$ and the transformed weights W' are quantized (Q_a and Q_w) and matrix multiplication is performed. Simultaneously, the quantization-aware efficient fine-tuning module introduces a low-rank adapter $\Delta W = AB$, whose output is added to the quantization result to produce the final output Y_Q. All transformation parameters (R, S, B^c) and adapter parameters (A, B) can be fused with the weights, achieving zero additional inference overhead.

The quantization calculation process is as follows:

$$\mathbf{Y}_Q = Q_w(\mathbf{W}\mathbf{R}\mathbf{S}^{-1}) \cdot Q_a(\hat{\mathbf{X}}) + \underbrace{(\mathbf{W}\mathbf{R}\mathbf{S}^{-1})\mathbf{B}^c + \mathbf{b}}_{\text{fusion bias}}, \quad (9)$$

where $Q_w(\cdot)$ and $Q_a(\cdot)$ are the quantization functions for the weights and activations, respectively. The fusion bias term can be merged with the layer bias $\mathbf{b}$, resulting in zero overhead during inference. The bias correction term $\mathbf{B}^c$ has very few parameters (less than 0.1%) and is learned efficiently through block-level optimization.

C. Quantization-Aware Efficient Fine-Tuning

While bias mean correction addresses static bias at the distribution level, rounding and pruning operations during quantization still introduce task-related dynamic errors. To compensate for these errors and improve the model's adaptability to downstream tasks, we introduce a quantization-aware efficient fine-tuning module based on low-rank adaptation.

For linear layers requiring fine-tuning, we add a low-rank adapter $\Delta\mathbf{W} = \mathbf{AB}$ in its bypass, where $\mathbf{A} \in \mathbb{R}^{m\times r}, \mathbf{B} \in \mathbb{R}^{r\times n}$, and the rank $r \ll \min(m, n)$. The quantized forward propagation becomes:

$$\begin{aligned}\mathbf{Y} &= Q_a(\mathbf{X})Q_w(\mathbf{W} + \Delta\mathbf{W})^T \\ &= Q_a(\mathbf{X})[Q_w(\mathbf{W})^T + (\mathbf{AB})^T].\end{aligned} \quad (10)$$

The low-rank adapter parameters $\mathbf{A}, \mathbf{B}$ are trained during the fine-tuning phase, while the quantization weights $Q_w(\mathbf{W})$ remain frozen. This design originates from the QLoRA concept, requiring only a very small number of parameters to be fine-tuned.

The low-rank adapter and the bias correction module have a synergistic effect. Bias correction globally aligns the input distribution, while the low-rank adapter learns to compensate for the complex task-related residual errors in subsequent quantization calculations. The adapter learns the residual mapping

$\mathbf{AB}$, which not only compensates for general quantization noise but also performs secondary calibration for task-related distribution mismatches that may remain after bias correction.

D. Theoretical Analysis

OBEQuant's performance stems from its collaborative error sensing and cascaded compensation design. The BMC module explicitly corrects static distribution bias, providing a high-quality baseline for quantization; the EFT module introduces extremely low-rank increments to dynamically compensate for residual task bias. These two modules form a cascaded optimization. The BMC reduces activation variance and outliers, making it easier for the EFT to learn smooth error patterns; the secondary calibration of the EFT further improves the input quality of the next-layer BMC, constructing a positive error suppression loop and significantly enhancing the model's robustness under extreme quantization conditions.

V. EXPERIMENTS

A. Baselines, LLMs, and Datasets.

Baselines. To comprehensively evaluate the effectiveness and advancement of the proposed OBEQuant method, we systematically compare it with current mainstream PTQ methods. The baseline methods encompass various technical approaches, which can be categorized into three representative advances: weight-only quantization methods, represented by RTN [23] and GPTQ [14]; weight-activation joint quantization methods, represented by SmoothQuant [7], AWQ [15], and OmniQuant [17], which optimize quantization difficulty through activation smoothing or mixed-precision strategies; and rotation-based quantization methods, represented by QuaRot [8], SpinQuant [9], and OSTQuant [10], which aim to improve quantization adaptability through parameter space transformation. These methods collectively constitute a rigorous performance benchmark.

LLMs and Datasets. For evaluation, we conduct experiments on the LLaMA series models, focusing on two medium-sized architectures, LLaMA-2-7B [24] and LLaMA-3-8B [25], to achieve a balance between computational feasibility and model complexity. We employ two core metrics for comprehensive evaluation: first, we use perplexity on the WikiText2 [26] test set to measure the fidelity of quantization to the model's basic language modeling capabilities. Secondly, the performance of downstream tasks is evaluated using the lm-evaluation-harness framework on nine zero-shot commonsense reasoning tasks. These tasks cover multiple cognitive dimensions, including BoolQ [27], HellaSwag [28], LAMBADA [29], OpenBookQA [30], PIQA [31], SIQA [32], WinoGrande [33], ARC-Easy and ARC-Challenge [34], comprehensively reflecting the model's ability to maintain its capabilities in real-world scenarios. This dual evaluation strategy ensures rigorous testing from statistical likelihood to actual performance.

B. Main Results

This paper compares OBEQuant with several baseline methods, and the results are shown in Table I. Experiments were conducted under three quantization configurations, and the specific analysis is as follows:

1) Under the W4A16KV16 configuration: OBEQuant achieved an average accuracy of 64.34% on nine downstream tasks, outperforming all baseline methods, with the best performance on five tasks including ARC-Challenge and BoolQ. Its WikiText2 perplexity was 5.62, comparable to the best baseline.
2) Under the W4A4KV16 configuration: Most baseline methods, such as RTN, SmoothQuant, and GPTQ, failed, while OBEQuant performed robustly, achieving an average accuracy of 62.63%, outperforming the second-best OSTQuant method by 0.16%. It achieved accuracies of 74.37% and 71.30% on the BoolQ and LAMBADA tasks, respectively, with recovery accuracy percentages as high as 95.7% and 96.8%. Its perplexity of 5.89 was also the lowest under this configuration.
3) Under the W4A4KV4 configuration: Even under this most extreme configuration, OBEQuant still achieves an average accuracy of 62.43%, surpassing all similar methods. It achieves 44.62% accuracy on the SIQA task, a 1.17% improvement over OSTQuant. Notably, this performance even exceeds QuaRot's 61.87% accuracy in the more relaxed W4A16 configuration.

In summary, OBEQuant demonstrates SOTA or highly competitive performance across all test configurations, particularly exhibiting stable effectiveness in the highly challenging low-bit activation quantization scenario, proving the method's state-of-the-art nature and robustness.

C. Analyses

The overall design of the OBEQuant method demonstrates systematic innovation at both the theoretical and practical levels, specifically in the following aspects:

1) **Theoretical innovation perspective.** This method is the first to explicitly decompose quantization error into two orthogonal components: static distribution bias and dynamic task bias, and designs a cascaded compensation path for this purpose. This transcends the limitations of traditional methods that attempt to solve all problems with a single unified transformation, providing a more refined perspective on error control. Distribution reshaping based on orthogonality and scaling inherits and optimizes existing rotational quantization ideas, while the explicit bias mean correction module innovatively solves the generally neglected static error source of channel mean offset, theoretically reducing the invalid occupancy in the quantization interval.
2) **Quantization accuracy and model capability preservation.** The main experimental data fully validates the excellence of this framework. Under the most challenging W4A4 configuration, OBEQuant achieves an average accuracy of 62.43% on multiple commonsense reasoning tasks, surpassing not only all similar solutions but also outperforming the performance of some methods

TABLE I

A COMPLETE COMPARISON OF THE PERPLEXITY SCORE OF THE LLAMA2-7B MODEL ON WIKITEXT2 WITH THE AVERAGE ACCURACY OF NINE COMMONLY USED ZERO-SHOT COMMONSENSE REASONING TASKS. THE BASELINES INCLUDE REPRESENTATIVE GENERAL METHODS SUCH AS RTN, SMOOTHQUANT, GPTQ, OMNIQUANT, AND AWQ, AS WELL AS STATE-OF-THE-ART SOLUTIONS FOCUSED ON ROTATION QUANTIZATION, INCLUDING QUAROT, SPINQUANT, AND OSTQUANT. ALL BASELINE EXPERIMENTAL RESULTS ARE REPRODUCED BASED ON THEIR OFFICIAL OPEN-SOURCE IMPLEMENTATIONS. ↑ DENOTES THAT THE LARGER THE VALUE, THE BETTER THE PERFORMANCE.

Bits #W-A-KV	Method	Wiki2 (↓)	ARC-c (↑)	ARC-e (↑)	BoolQ (↑)	Hella. (↑)	Lam. (↑)	OBQA (↑)	PIQA (↑)	SIQA (↑)	WinoG. (↑)	Avg. (↑)
16-16-16	Full Precision	5.47	46.42	74.33	77.71	75.94	73.69	44.2	79.16	45.91	69.53	65.21
4-16-16	RTN	7.02	42.15	67.59	73.06	72.34	67.18	41.8	76.50	44.11	66.69	61.27
	SmoothQuant	8.03	39.59	65.19	69.82	68.84	62.27	40.2	75.95	44.17	63.85	58.88
	GPTQ	9.84	42.49	69.53	61.31	73.83	67.61	42.4	77.64	44.52	68.43	60.86
	Omniquant	5.74	42.49	71.0	74.34	73.85	70.7	44.2	78.40	44.93	68.82	63.19
	AWQ	5.83	44.11	70.75	78.07	74.98	70.68	43.8	78.13	45.14	69.38	63.89
	QuaRot	5.62	43.94	73.15	76.97	74.87	73.06	44.0	78.24	45.09	69.38	64.30
	SpinQuant	5.58	43.34	72.69	73.36	75.10	73.8	43.0	77.86	45.6	67.56	63.59
	OSTQuant	5.63	44.03	74.54	74.89	74.75	73.47	43.4	78.94	45.39	68.43	64.20
	OBEQuant	**5.62**	**44.78**	74.03	**75.95**	**74.90**	73.15	43.36	78.55	**45.49**	**68.86**	**64.34**
4-4-16	RTN	NAN	25.34	28.03	50.52	27.71	1.01	26.2	50.82	33.93	48.38	32.44
	SmoothQuant	NAN	28.33	26.39	49.39	27.28	1.18	23.4	48.8	33.62	50.75	32.13
	GPTQ	NAN	24.40	28.70	51.62	28.66	1.36	24.6	51.14	34.49	49.49	32.72
	QuaRot	6.05	42.32	69.65	74.77	72.91	70.81	39.8	77.20	43.55	65.82	61.87
	SpinQuant	6.78	37.54	62.58	71.16	70.48	67.16	34.8	75.46	39.76	60.62	57.37
	OSTQuant	5.90	42.75	71.21	74.07	74.06	70.91	41.4	77.48	44.22	66.14	62.47
	OBEQuant	**5.89**	**42.99**	71.18	**74.37**	73.58	**71.30**	**41.76**	77.26	**44.28**	**66.91**	**62.63**
4-4-4	RTN	NAN	27.22	27.06	50.83	27.34	0.93	25.8	49.51	34.85	50.51	32.67
	SmoothQuant	NAN	26.37	25.63	47.71	27.05	1.11	26.4	51.90	34.49	48.38	32.12
	GPTQ	NAN	26.96	27.65	52.84	28.83	1.63	29.2	49.62	35.11	49.80	33.52
	Omniquant	14.26	31.40	53.75	63.79	55.06	35.63	34.4	56.59	40.28	54.70	48.40
	QuaRot	6.11	41.43	69.32	74.19	72.50	70.66	39.8	77.42	43.35	64.64	61.48
	SpinQuant	5.96	40.44	71.08	74.40	73.51	70.66	41.8	76.88	43.50	65.82	62.01
	OSTQuant	5.94	42.15	70.37	74.95	73.53	70.72	42.0	76.99	43.45	65.43	62.18
	OBEQuant	**5.94**	**42.44**	**70.95**	74.42	73.12	**70.97**	41.66	**77.20**	**44.62**	**66.50**	**62.43**

under more relaxed configurations. The BMC module significantly improves the effective utilization of the quantization interval by correcting the mean; while the EFT module accurately corrects task-related dynamic errors through extremely low-rank adaptation. The cascaded framework of coarse correction and fine correction formed by their collaboration constitutes the key to the performance breakthrough.

3) **Deployment efficiency and practicality.** This design is highly advantageous. The channel-by-channel bias correction introduced by the BMC module allows for offline parameter fusion, achieving zero inference overhead. The extremely low-rank increment of the EFT module also supports pre-fusion with weights, manifesting as only a tiny set of weight updates during inference, with negligible impact on computational latency and memory usage. Overall, OBEQuant achieves the optimal balance between accuracy and efficiency without introducing significant additional computational costs, providing a reliable technical path for high-fidelity, low-loss deployment of large models.

A few outlier data points in the experiment, such as the OBQA task accuracy being slightly lower than some baselines under the W4A4KV4 configuration, are analyzed to be likely due to the sensitivity of open-domain question answering tasks to the breadth of knowledge coverage and subtle semantic differences. Their composite error patterns may have exceeded the current modeling boundaries of low-rank compensation to some extent. However, OBEQuant maintained its leading position in the vast majority of tasks and average accuracy. Furthermore, compared to the complete failure of traditional methods under extreme quantization, its BMC module effectively prevented numerical collapse through stable distribution, ensuring the robustness and reliability of the overall solution.

D. Ablation Study

To deeply explore the specific contributions of each innovative module in the OBEQuant framework and verify the effectiveness of its design, we conducted a systematic ablation experiment on the Llama-2-7B model. The experimental setup remained consistent with the main experiment, and evaluation was performed under the W4A4KV4 quantization configuration. The experimental results are shown in Table II.

Based on the data results from the ablation experiment, the analysis is as follows:

1) The BMC module brings fundamental improvements. When only the BMC module is ablated, the model

TABLE II
ABLATION STUDY ON THE INFLUENCE OF DIFFERENT TRANSFORMATION MATRICES ON WIKI PPL AND ZERO-SAMPLE SCORE OF LLAMA-2-7B UNDER DIFFERENT QUANTIZATION CONFIGURATIONS.

	W4A16KV16		W4A4KV16		W4A4KV4	
	Wiki2 (↓)	**Avg. (↑)**	**Wiki2 (↓)**	**Avg. (↑)**	**Wiki2 (↓)**	**Avg. (↑)**
OBEQuant	5.617	64.34	5.89	62.67	5.94	62.38
-BMC-EFT	5.626(+0.09)	64.2(-0.14)	5.9(+0.01)	62.47(-0.2)	5.94(+0)	62.18(-0.2)
-BMC	5.615(-0.02)	64.24(-0.1)	5.89(+0)	62.55(-0.12)	5.93(+0.01)	62.36(-0.02)
-EFT	5.617(+0)	64.4(+0.06)	5.9(+0.01)	62.4(-0.27)	5.94(+0)	62.3(-0.08)

performs differently under different configurations. Under the relatively relaxed W4A16KV16 configuration, the perplexity is 5.615, a slight improvement of 0.02; however, the average accuracy of 64.24% is slightly impaired, decreasing by 0.1%. Under the more stringent quantization configurations W4A4KV16 and W4A4KV4, the average accuracy of 62.55% and 62.36% respectively is better than the case where both BMC and EFT modules are ablated, but significantly lower than the full OBEQuant. This demonstrates that the BMC module is the foundation for maintaining the overall performance of the model under stringent quantization, especially accuracy, and ensuring stability.

2) The EFT module provides efficient directional parameter correction. When only the EFT module is ablated, the average accuracy under W4A16KV16 is even slightly better than the full model's 64.4%, indicating that BMC alone can achieve good results under this configuration. However, under configurations with lower bit widths, the average accuracy drops significantly, especially under W4A4KV16, where it is 0.27% lower than the full model. This proves that the EFT module is a key compensator for accuracy under extremely low bit width quantization, and its directional correction capability is indispensable under stress scenarios.
3) The synergistic effect of the module combination is excellent. Simultaneously ablating both the BMC and EFT modules resulted in the lowest average accuracy across all configurations, with most perplexity metrics deteriorating. This contrasts sharply with the complete OBEQuant. This directly verifies that both the BMC and EFT modules are indispensable and exhibit a significant synergistic effect. Their combination is not a simple additive process but rather constitutes a more robust error compensation system than a single module or none at all, addressing quantization losses from various sources and levels.

This ablation experiment confirms that the BMC module is the fundamental component ensuring the robustness of the quantization model, while the EFT module is a key innovation for achieving breakthroughs in accuracy under high-stress quantization. Their synergistic work enables OBEQuant to achieve optimal overall performance, surpassing the baseline and any single module, across different quantization configurations, especially in the most demanding W4A4 scenario.

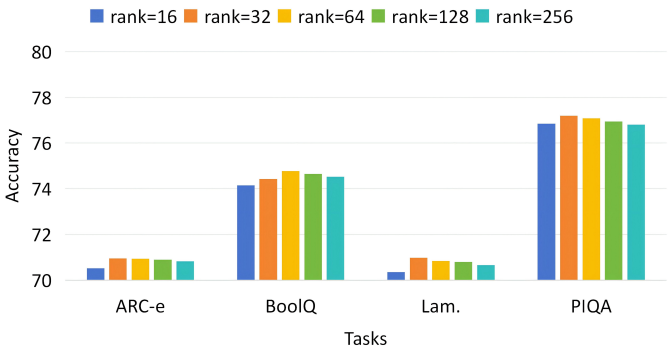

Fig. 2. Comparison of the impact of different low-rank configurations on the accuracy of the OBEQuant method across various tasks.

EFT suffers from a saturation problem. As shown in Fig.2, although the EFT module provides targeted refinement capabilities, it suffers from saturation. While all metrics show significant increases when the rank is increased from 16 to 32, further increases to 64 do not improve performance, and most metrics show a slight downward trend as the rank increases to 128 and 256. This indicates that the error compensation capability of the low-rank adapter is close to optimal at a rank of 32. Blindly increasing the rank not only fails to bring effective accuracy gains but may also introduce optimization noise due to over-parameterization, while increasing deployment overhead. Therefore, OBEQuant adopts a rank of 32 to achieve the best balance between efficiency and performance.

VI. CONCLUSION

This study proposes an efficient quantization framework named OBEQuant that employs collaborative error compensation. The bias correction module is used to eliminate the systematic error caused by channel mean deviation, and the quantization-aware low-rank fine-tuning module is combined to compensate for the dynamic residual error. These two modules form a cascaded optimized error suppression loop, achieving zero additional overhead in precision restoration at extremely low bit rates. Experiments on models such as LLaMA demonstrate that OBEQuant outperforms existing methods under the W4A4 configuration, providing a practical solution for high-fidelity deployment of large models in resource-constrained scenarios.

ACKNOWLEDGMENT

This work was supported by the National Key Laboratory Foundation Project under Grant 2024-KJWPDL-14.

REFERENCES

[1] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale," *Advances in neural information processing systems*, vol. 35, pp. 30 318–30 332, 2022.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] L. He, S. Zheng, K. Sun, Y. Liu, Y. Zhao, C. Tan, H. Yang, Y. Du, and L. Du, "Base-q: Bias and asymmetric scaling enhanced rotational quantization for large language models," *arXiv preprint arXiv:2506.15689*, 2025.

[4] P. Czakó, G. Kertész, and S. Szénási, "Smoothrot: Combining channel-wise scaling and rotation for quantization-friendly llms," *arXiv preprint arXiv:2506.05413*, 2025.

[5] Y. Sun, R. Liu, H. Bai, H. Bao, K. Zhao *et al.*, "Flatquant: Flatness matters for llm quantization," *arXiv preprint arXiv:2410.09426*, 2024.

[6] P. Czakó, G. Kertész, and S. Szénási, "Addressing activation outliers in llms: A systematic review of post-training quantization techniques," *IEEE Access*, 2025.

[7] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *International conference on machine learning*. PMLR, 2023, pp. 38 087–38 099.

[8] S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, P. Cameron *et al.*, "Quarot: Outlier-free 4-bit inference in rotated llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 100 213–100 240, 2024.

[9] Z. Liu, C. Zhao, I. Fedorov, B. Soran, D. Choudhary *et al.*, "Spinquant: Llm quantization with learned rotations," *arXiv preprint arXiv:2405.16406*, 2024.

[10] X. Hu, Y. Cheng, D. Yang, Z. Xu, Z. Yuan *et al.*, "Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting," *arXiv preprint arXiv:2501.13987*, 2025.

[11] J. Xiao, B. Ji, S. Li, X. Liu, M. Jun, Y. Zhong, W. Li, X. Xie, Q. Wu, and J. Yu, "Singlequant: Efficient quantization of large language models in a single pass," *arXiv preprint arXiv:2511.22316*, 2025.

[12] Y. Liang, H. Chen, S. Han, and Z. Liu, "Paroquant: Pairwise rotation quantization for efficient reasoning llm inference," *arXiv preprint arXiv:2511.10645*, 2025.

[13] B. Zeng, B. Ji, X. Liu, J. Yu, S. Li *et al.*, "Lsaq: Layer-specific adaptive quantization for large language model deployment," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[14] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," *arXiv preprint arXiv:2210.17323*, 2022.

[15] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen *et al.*, "Awq: Activation-aware weight quantization for on-device llm compression and acceleration," *Proceedings of machine learning and systems*, vol. 6, pp. 87–100, 2024.

[16] C. Lee, J. Jin, T. Kim, H. Kim, and E. Park, "Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, 2024, pp. 13 355–13 364.

[17] W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao *et al.*, "Omniquant: Omnidirectionally calibrated quantization for large language models," *arXiv preprint arXiv:2308.13137*, 2023.

[18] W. Cheng, W. Zhang, H. Shen, Y. Cai, X. He, L. Kaokao, and Y. Liu, "Optimize weight rounding via signed gradient descent for the quantization of llms," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 11 332–11 350.

[19] X. Wei, Y. Zhang, Y. Li, X. Zhang, R. Gong, J. Guo, and X. Liu, "Outlier suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling," *arXiv preprint arXiv:2304.09145*, 2023.

[20] Z. Yuan, L. Niu, J. Liu, W. Liu, X. Wang *et al.*, "Rptq: Reorder-based post-training quantization for large language models," *arXiv preprint arXiv:2304.01089*, 2023.

[21] H. Lin, H. Xu, Y. Wu, J. Cui, Y. Zhang *et al.*, "Duquant: Distributing outliers via dual transformation makes stronger quantized llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 87 766–87 800, 2024.

[22] J. Chen, V. Egiazarian, T. Hoefler, and D. Alistarh, "Wush: Near-optimal adaptive transforms for llm quantization," *arXiv preprint arXiv:2512.00956*, 2025.

[23] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[25] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[26] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," *arXiv preprint arXiv:1609.07843*, 2016.

[27] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *arXiv preprint arXiv:1905.10044*, 2019.

[28] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" *arXiv preprint arXiv:1905.07830*, 2019.

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[30] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," *arXiv preprint arXiv:1809.02789*, 2018.

[31] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.

[32] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi, "Socialiqa: Commonsense reasoning about social interactions," *arXiv preprint arXiv:1904.09728*, 2019.

[33] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," *Communications of the ACM*, vol. 64, no. 9, pp. 99–106, 2021.

[34] M. Boratko, H. Padigela, D. Mikkilineni, P. Yuvraj, R. Das *et al.*, "A systematic classification of knowledge, reasoning, and context within the arc dataset," in *Proceedings of the Workshop on Machine Reading for Question Answering*, 2018, pp. 60–70.