OTARo: Once Tuning for All Precisions Toward Robust On-Device LLMs

Shaoyuan Chen[1,2,†,△], **Zhixuan Chen**[1,†], **Dawei Yang**[1,*,†], **Zhihang Yuan**[3], **Qiang Wu**[1]

[1]Houmo AI
[2]Sun Yat-sen University
[3]Peking University
chenshy299@mail2.sysu.edu.cn, zhixuan.chen@houmo.ai, dawei.yang@houmo.ai, yuanzhihang@pku.edu.cn, qiang.wu@houmo.ai

Abstract

Large Language Models (LLMs) fine-tuning techniques not only improve the adaptability to diverse downstream tasks, but also mitigate adverse effects of model quantization. Despite this, conventional quantization suffers from its structural limitation that hinders flexibility during the fine-tuning and deployment stages. Practical on-device tasks demand different quantization precisions (i.e. different bit-widths), e.g., understanding tasks tend to exhibit higher tolerance to reduced precision compared to generation tasks. Conventional quantization, typically relying on scaling factors that are incompatible across bit-widths, fails to support the on-device switching of precisions when confronted with complex real-world scenarios. To overcome the dilemma, we propose OTARo, a novel method that enables on-device LLMs to flexibly switch quantization precisions while maintaining performance robustness through once fine-tuning. OTARo introduces Shared Exponent Floating Point (SEFP), a distinct quantization mechanism, to produce different bit-widths through simple mantissa truncations of a single model. Moreover, to achieve bit-width robustness in downstream applications, OTARo performs a learning process toward losses induced by different bit-widths. The method involves two critical strategies: (1) Exploitation-Exploration Bit-Width Path Search (BPS), which iteratively updates the search path via a designed scoring mechanism; (2) Low-Precision Asynchronous Accumulation (LAA), which performs asynchronous gradient accumulations and delayed updates under low bit-widths. Experiments on popular LLMs, e.g., LLaMA3.2-1B, LLaMA3-8B, demonstrate that OTARo achieves consistently strong and robust performance for all precisions.

Introduction

In recent years, Large Language Models (LLMs) have shown outstanding capabilities in Natural Language Processing (NLP) (Achiam et al. 2023; Zubiaga 2024), and on-device deployment of LLMs has emerged as a frontier research direction to enable real-time responsiveness, privacy, and personalization in intelligent terminal services (Qu et al. 2025; Tang et al. 2025). With advances in processing power, memory bandwidth, and storage capacity of edge devices, on-device deployment of compact LLMs becomes feasible. In parallel, model compression techniques, e.g., quantization (Achiam et al. 2023; Lin et al. 2024), reduce deployment costs while incurring some accuracy degradation.

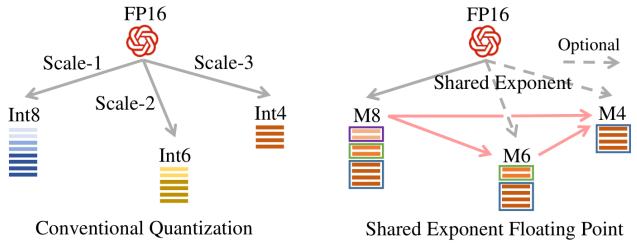


Figure 1: A comparison of conventional quantization and Shared Exponent Floating Point (SEFP) in supporting dynamic precision switching. Gray arrows represent quantization, while red arrows indicate cross-precision conversion. M8, M6, and M4 denote mantissa bit-widths in SEFP.

For a better inference performance of LLMs in devices, fine-tuning techniques play a pivotal role by not only enhancing the adaptability across downstream tasks, but also effectively mitigating the accuracy degradation induced by quantization (Yang et al. 2024b; Lu, Luu, and Buehler 2025). Despite this, conventional quantization still faces challenges in complex real-world scenarios due to its inherent limitation in switching bit-widths.

Practical on-device tasks require different quantization precisions, corresponding to different bit-widths. For instance, generation tasks tolerate increased inference time in exchange for a higher precision, whereas understanding tasks can obtain immediate responses with a lower precision (Manduchi et al. 2024). In addition, different precisions can be employed in the prefill and decoding phases to optimize the inference efficiency (Qiao et al. 2025). Consequently, the practical on-device deployment necessitates support for all precisions, rather than a model with a fixed precision.

In conventional quantization, the original weights are scaled and clipped into different numerical ranges depending on the bit-width, which prevents the reuse of scaling factors (scales) across precisions. Once quantized to a specific bit-width, the model loses the flexibility to switch between

*Corresponding author
†Equal contribution
△Work done as an intern at Houmo AI

precisions. Edge devices can maintain a model zoo containing models of each precision, which substantially increases storage overhead and extends the total fine-tuning time. Alternatively, compressing a full- or half-precision model into various quantization precisions at runtime still requires offline fine-tuning for bit-specific and high-quality scaling factors, thereby incurring similar fine-tuning burdens. It also increases storage demands and introduces non-negligible computational costs during on-device quantization.

To address the challenges, this paper proposes OTARo (**O**nce **T**uning for **A**ll Precisions Toward **Ro**bust On-Device LLMs), which obtains one unified model through once fine-tuning to support all quantization precisions. OTARo introduces Shared Exponent Floating Point (SEFP), a distinct quantization mechanism. Notably, compared to conventional quantization, SEFP eliminates scaling factors and allows flexible switching of precisions, as illustrated in fig. 1. In SEFP, an exponent is shared in each group of parameters, and individual mantissas are maintained for each parameter. After aligning exponents, any expected bit-width can be achieved through simple mantissa truncation, and the higher bit-widths can be easily converted to the lower ones.

Furthermore, to enhance performance robustness across bit-widths, OTARo introduces an efficient learning scheme that jointly optimizes for mixed quantization losses induced by different SEFP configurations. The learning process incorporates an iterative bit-width selection, and adopts delayed updates to mitigate the effects of low-precision training. Generally, by unifying all precisions within one fine-tuned model and obviating the need for separate fine-tuning per bit-width, OTARo substantially enhances resource efficiency and deployment flexibility in practical applications.

The contributions of our work are summarized as follows:

- We propose OTARo, a novel fine-tuning paradigm for building robust on-device LLMs. OTARo produces a unified model through once fine-tuning to support all quantization precisions. It introduces Shared Exponent Floating Point (SEFP) to enable lightweight and hardware-friendly multi-precision adaptation, and jointly update weights for quantization errors from different bit-widths.
- We identify the commonality on gradient directions under all bit-width settings, and then propose Exploitation-Exploration Bit-Width Path Search (BPS) strategy to determine the optimal sequence of bit-widths.
- We observe that low-precision training induces severe gradient oscillations. To mitigate the oscillations, we conduct high-dimensional projection analysis on the gradients, and propose Low-Precision Asynchronous Accumulation (LAA) strategy, which performs asynchronous gradient accumulations and delayed updates.
- We conduct the comprehensive evaluations on popular LLMs, e.g., LLaMA3.2-1B and LLaMA3-8B. Experimental results show that OTARo consistently outperforms baselines across all tested bit-widths, in both task-specific and zero-shot settings, highlighting its robustness for multi-precision deployment of LLMs.

Related Work

Quantization

Quantization is a critical technique to reduce the memory footprint and inference costs (Wei et al. 2024). Post-Training Quantization (PTQ) methods quantize a pre-trained model without retraining. GPTQ (Frantar et al. 2022) uses arbitrary order quantization, lazy batch updates, and Cholesky reformulation. CFWS (Yang et al. 2024a) introduces coarse & fine weight splitting and an enhanced KL calibration metric. AWQ (Lin et al. 2024) finds the salient weights based on the activation distribution, and scales the salient weight channels through an equivalent transformation. QuaRot (Ashkboos et al. 2024) uses Hadamard transform and computational invariance to mitigate outliers in activations. SpinQuant (Liu et al. 2024) uses a learnable rotation matrix that is fused into weights. Any-Precision LLM (Park et al. 2024) performs incremental upscaling and specialized hardware design to merge multiple integer formats. QUARK (Zhao et al. 2025) proposes a reordering-based group quantization scheme. Quantization-Aware Training (QAT) methods incorporate fake quantization in training, allowing models to lea quantization noises (Liu et al. 2023; Chen et al. 2024).

Shared Exponent Floating Point

Shared Exponent Floating Point (SEFP) quantization shares an exponent in each parameter group while maintaining individual mantissas for each parameter (Gao et al. 2024; Zhou et al. 2025). SEFP quantization procedure consists of two main steps, illustrated in fig. 2. Firstly, the maximum exponent of each parameter group is selected as the shared one before each mantissa is shifted right according to the difference between the shared exponent and its original one. Secondly, shifted mantissas are truncated to a fixed width. SEFP bit-width is typically donated as EeMm, where e and m denote the bit-width of exponents and mantissas, respectively. Accuracy declines arise almost from the forced truncation in Step 2; the impacts of mantissa overflow in step 1 is negligible. (Kalliojarvi and Astola 2002).

Method

Preliminaries

Due to the non-differentiable operations involved in SEFP quantization, such as mantissa right-shifting and truncation, we incorporate the Straight-Through Estimator (STE) (Yin et al. 2019) to approximate the gradients, and introduce SEFP quantization errors into training losses to facilitate the learning process. The approach is formulated as follows:

$$\frac{\partial Q(\omega, b)}{\partial \omega} = 1 \tag{1}$$

$$\mathcal{L} = \text{Loss}(y, Q(\omega, b)x) \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial \omega} = \frac{\partial \mathcal{L}}{\partial Q(\omega, b)} \frac{\partial Q(\omega, b)}{\partial \omega} = \frac{\partial \mathcal{L}}{\partial Q(\omega, b)} \tag{3}$$

where, ω is weight, $Q(\cdot, b)$ is SEFP function targeting bit-width b, x is input, y is label, $\text{Loss}(\cdot)$ is training loss function. In the above manner, OTARo enables models to adapt

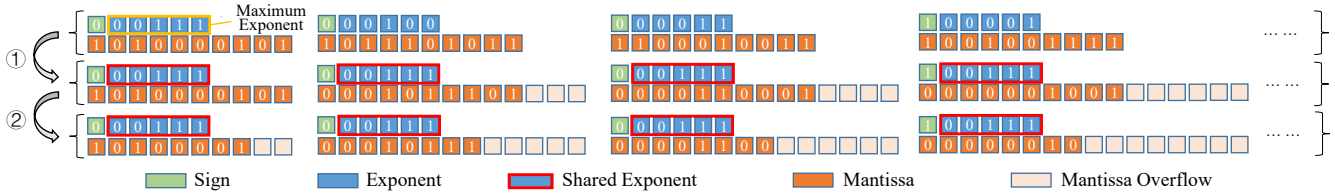


Figure 2: A simple illustration of SEFP quantization (e.g., from FP16 to E5M8) of a group of parameters. Step 1 shows the exponent alignment and mantissa right-shifting for FP16 values. Step 2 shows the forced mantissa truncation.

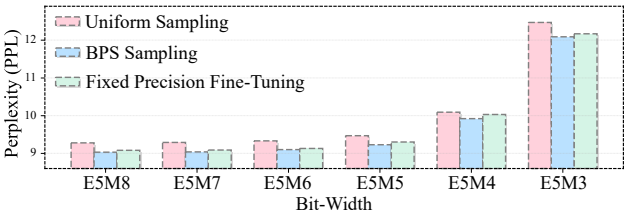


Figure 3: A comparison of uniform sampling, BPS sampling and fixed precision fine-tuning. We illustrate the perplexity results of the three approaches. In the experiments, LLaMA3.2-1B is fine-tuned and evaluated on the WikiText2 train and test set, respectively.

to quantization errors from different SEFP bit-widths, and fixed precision fine-tuning only learns for quantization errors from a specific bit-width.

Problem Statement

In OTARo, once fine-tuning is performed to obtain a model robust to all bit-widths in SEFP. Each bit-width corresponds to a certain level of precision, with higher bit-widths yielding higher precision. The objective can be formulated as:

$$\omega = \arg\min_{\omega} \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathbb{L}(Q(\omega, b)) \tag{4}$$

where, b is the bit-width, $\mathcal{B}$ is the set of bit-widths, $\mathbb{L}(\cdot)$ is test loss function, and $Q(\omega, b)$ is SEFP function under bit-width b. We aim to develop a single model to maintain high performance across all bit-widths, achieving accuracy comparable to the original model before quantization, thus supporting robust and adaptable on-device deployment of LLMs.

Exploitation-Exploration Bit-Width Path Search

To ensure robustness across bit-widths, an intuition is to uniformly sample all target bit-widths, and facilitate models to update based on the gradients induced by quantization errors of different bit-widths. However, our empirical results indicate that this straightforward sampling approach does not align with fixed precision fine-tuning in performance, as shown in fig. 3. This finding prompts us to rethink: whether different bit-widths should be sampled non-uniformly, and scheduled in a deliberate order during fine-tuning.

To assess the feasibility of sampling all bit-widths for robustness, and to inform the design of a practical implementation strategy, we revisit bit-width sampling from a gradient-centric perspective, as illustrated in fig. 4.

It can be observed that the gradients produced at different bit-widths exhibit a certain degree of similarity, which supports the feasibility of enhancing robustness by sampling all bit-widths. Furthermore, for both the higher and lower bit-widths, the gradient direction shows higher consistency with those of the higher ones, but less consistency even with that of the adjacent lower one. This commonality suggests the existence of a larger shared potential subspace between higher bit-widths and others in terms of gradient directions.

Motivated by the observation and analysis, we argue that the bit-width sampling should not only explore all bit-widths to sufficiently learn the loss landscapes across them, but also gradually update toward higher bit-widths that exhibit stronger alignment with the others in gradient directions.

Accordingly, we propose the Exploitation-Exploration Bit-Width Path Search (BPS) strategy, which iteratively selects the bit-width that achieves the highest score in a designed scoring mechanism. The scoring mechanism is formulated as follows:

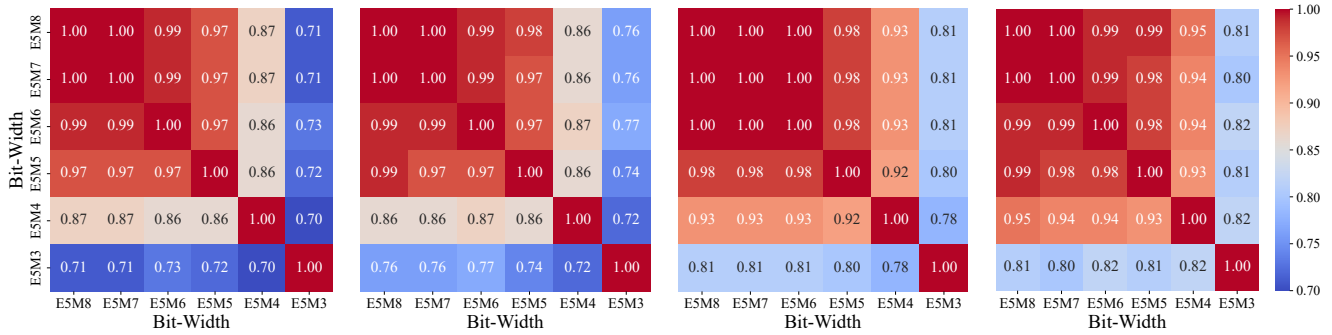
$$\text{Score}(b) = \lambda \sqrt{\frac{\ln t}{t_b}} - \mathcal{L}_b \tag{5}$$

where, the score of bit-width b is denoted as $\text{Score}(b)$, $\mathcal{L}_b$ is the real-time loss for bit-width b, t is the current number of training batches, t_b is the search frequency of b, and exploration coefficient λ is a constant multiplied by the exploration term.

This design encourages the bit-width search path to explore underutilized bit-widths. As fine-tuning progresses, the exploration term of frequently used bit-widths gradually decreases, naturally leading to sampling underutilized ones. The dynamic balance ensures diversity in bit-width selection. This design also ensures that, as the number of batches increases, higher bit-widths are selected more frequently than lower ones. To validate this convergence property, we examine whether the score discrepancy between the higher and lower bit-widths consistently converges toward a positive value as t increases. Suppose h and l denote two distinct bit-widths, and h is the higher one. The score difference is defined as:

$$\Delta = (\lambda \sqrt{\frac{\ln t}{t_h}} - \mathcal{L}_h) - (\lambda \sqrt{\frac{\ln t}{t_l}} - \mathcal{L}_l) \tag{6}$$

where, T is the total number of batches, t_h, t_l are respectively the current counts of h and l being searched (approximately increasing linearly), $\mathcal{L}_h$, $\mathcal{L}_l$ are the corresponding



(a) q proj (b) k proj (c) v proj (d) down proj

Figure 4: Cosine similarities between the gradients produced by different bit-widths of the LLaMA3.2-1B layer-15 q/k/v/down projector. There exists a certain degree of similarity between gradients under different bit-widths. Furthermore, the gradient at each bit-width tends to exhibit stronger similarity with that of its higher bit-width. For example, for q projector, the cosine similarity of gradients between E5M5 and E5M8/E5M7/E5M6 is 0.97, whereas the cosine similarity with E5M4 and E5M3 decreases to 0.86 and 0.72, respectively.

losses at bit-widths h and l, with $\mathcal{L}_l$ generally being larger due to the lower precision of l. We reorganize the terms in the above equation:

$$\Delta = (\mathcal{L}_l - \mathcal{L}_h) + \lambda(\sqrt{\frac{t}{t_h}} - \sqrt{\frac{t}{t_l}})\sqrt{\frac{\ln t}{t}} \tag{7}$$

As t increases to a high value, Δ approaches $\mathcal{L}_l - \mathcal{L}_h$, which is increasingly likely to be positive:

$$\lim_{t \to T} \sqrt{\frac{\ln t}{t}} \to 0 \tag{8}$$

$$\lim_{t \to T} \Delta \to \mathcal{L}_l - \mathcal{L}_h \tag{9}$$

Therefore, the search path in the BPS strategy gradually converges toward the higher bit-widths with smaller losses and more robust gradient directions. As show in fig. 3, the sampling based on the BPS strategy can match or even outperform fixed-precision fine-tuning in each bit-width.

Low-Precision Asynchronous Accumulation

Through the BPS strategy, we obtain a search path that sufficiently explores multiple bit-widths while gradually favoring higher ones. However, the robustness of fine-tuned models suffers from the impacts of low bit-widths. To address this issue, we propose the Low-Precision Asynchronous Accumulation (LAA) strategy, which utilizes a distinctive weight update scheme under continuous low-bit width sampling conditions. To demonstrate the problem, we perform a detailed analysis of errors introduced by SEFP quantization:

$$\mathcal{L}_q = \frac{1}{2}||xQ(\omega, m) - x\omega||^2 \tag{10}$$

$$= \frac{1}{2}||\frac{x[\omega 2^m]}{2^m} - x\omega||^2 \tag{11}$$

$$\frac{\partial \mathcal{L}_q}{\partial \omega} = \frac{x\left(\omega 2^m - [\omega 2^m]\right)}{2^m} \cdot \frac{\partial Q(\omega, m)}{\partial \omega} \tag{12}$$

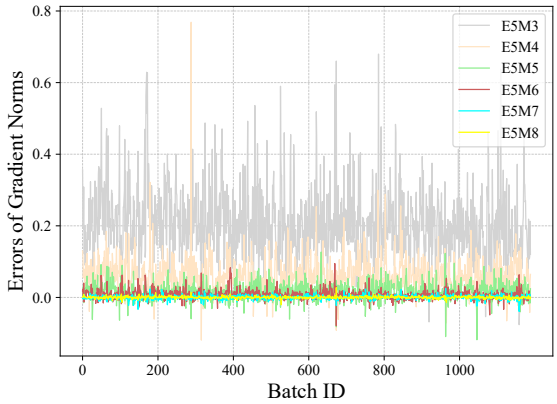


Figure 5: The errors of gradient norms $||\nabla_{\text{sefp}}|| - ||\nabla_{\text{fp}}||$ under different SEFP bit-widths. Gradients are calculated on LLaMA3.2-1B layer-15 down projector.

where, $\mathcal{L}_q$ denotes an approximation of the SEFP error, $\partial \mathcal{L}_q / \partial \omega$ is the derivative of the error with respect to the weight., m denotes the number of mantissa bits in SEFP, and $[\cdot]$ is rounding function. Suppose ω_0 is an arbitrary parameter in ω, we define the function $\epsilon(\omega_0)$ as:

$$\epsilon(\omega_0) = \frac{\omega_0 2^m - [\omega_0 2^m]}{2^m} \tag{13}$$

Function $\epsilon(\omega_0)$ describes a sawtooth wave with both period and amplitude equal to $1/2^m$. Due to the discontinuities introduced by the rounding operation, the function linearly increases from 0 within each period, then abruptly drops to a negative value before jumping back to 0, forming a periodic oscillatory pattern. When m is small, the wave exhibits larger amplitudes, leading to more pronounced oscillations. In low bit-width settings, changes in parameters can induce significant and periodic variations in SEFP quantization errors, resulting in periodic and intense oscillations in both

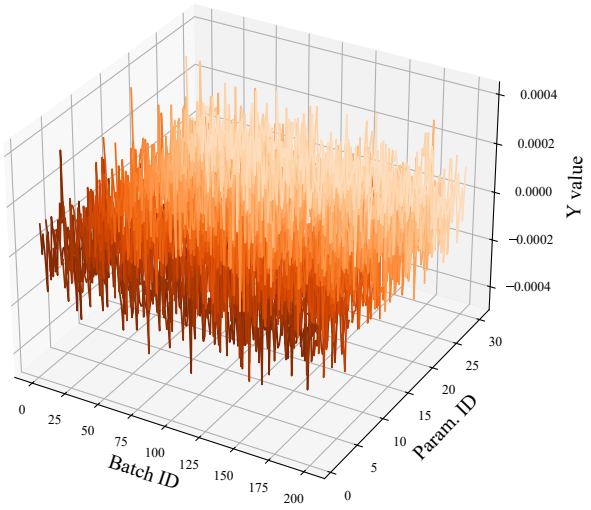


Figure 6: Y values. For model LLaMA3.2-1B and dataset WikiText2, the figure shows the Y values of 30 gradient values in 200 batches, and indicates that $\mathbb{E}[Y] \approx 0$.

loss and gradient during fine-tuning.

Suppose that the gradient in FP16, SEFP fine-tuning are respectively ∇_{fp} and ∇_{sefp}, the error of gradient norms induced by SEFP is $||\nabla_{\text{sefp}}|| - ||\nabla_{\text{fp}}||$. As shown in fig. 5, we observe that the oscillations in $||\nabla_{\text{sefp}}|| - ||\nabla_{\text{fp}}||$ become more intense as the bit-width decreases, and the oscillation exhibits an overall periodic pattern, providing indirect support for the above inference.

Based on these analysis, the error introduced by SEFP quantization is a critical issue to be tackled under low bit-widths. To measure the space distance between gradients with and without introducing SEFP quantization errors (i.e., ∇_{fp} and ∇_{sefp}, respectively), we model the relationship of them through a high-dimensional linear mapping:

$$\nabla_{\text{sefp}} = X \cdot \nabla_{\text{fp}} + Y \tag{14}$$

where X denotes the global linear mapping matrix from ∇_{fp} to ∇_{sefp}, and Y represents the residual perturbation term introduced by quantization, which corresponds to the non-linear and unexplained quantization noise. Estimation of X,Y is achieved by Least Squares Method (LSM).

To empirically validate the statistical properties of the perturbation Y, we conduct experiments under the low bit-width setting (e.g., E5M3). As shown in fig. 6, the results indicate that while Y exhibits considerable fluctuation between batches, its mean remains close to zero, that is,

$$\mathbb{E}[Y] \approx 0 \tag{15}$$

In the LAA strategy, gradients are asynchronously accumulated over N batches to generate a gradient ∇^N_{sefp} to update models. ∇^N_{sefp} can be expressed as:

$$\nabla^N_{\text{sefp}} = \sum_{i=1}^{N} \nabla_{\text{sefp},i} = X \sum_{i=1}^{N} \nabla_{\text{fp},i} + \sum_{i=1}^{N} Y_i. \tag{16}$$

Algorithm 1: Overall Pipeline of OTARo.

Require: Batch id t, total number of batches T, bit-width b, training loss function Loss(), weight ω, quantized ω under a bit-width b $Q(\omega, b)$, exploration coefficient λ, search count of a bit-width b t_b, inputs x, labels y, batch id and also a flag for gradient accumulations i (starts from 0), delayed interval N, learning rate η.

for each batch $t = 1$ to T **do**
  Calculate scores: $\text{Score}(b) = \lambda\sqrt{(\ln t)/t_b} - \mathcal{L}_b$
  Select optimal bit-width $b^* = \arg\max_b \text{Score}(b)$
  Calculate the training loss: $\mathcal{L} = \text{Loss}(y, Q(\omega, b^*)x)$
  Calculate the gradient: $\nabla_{\text{sefp}} = \partial\mathcal{L}/\partial Q(\omega, b)$
  if bit-width is ultra-low **then**
    if i is 0 **then**
      Initialize the general gradient: $\nabla^N_{\text{sefp}} \leftarrow \nabla_{\text{sefp}}$
    else
      Accumulate gradients: $\nabla^N_{\text{sefp}} \leftarrow \nabla^N_{\text{sefp}} + \nabla_{\text{sefp}}$
    end if
    $i \leftarrow i + 1$
    if i is N **then**
      Update weights: $\omega \leftarrow \omega - \eta\nabla^N_{\text{sefp}}$
      $i \leftarrow 0$
    end if
  else
    Standard gradient updates: $\omega \leftarrow \omega - \eta\nabla_{\text{sefp}}$
  end if
end for

Suppose independence or weak correlation among the Y_i, the relative influence of the perturbation with N rising diminishes as:

$$\frac{\left\|\sum_{i=1}^{N} Y_i\right\|}{\left\|\sum_{i=1}^{N} \nabla_{\text{fp},i}\right\|} \propto \frac{1}{\sqrt{N}} \to 0 \tag{17}$$

Accordingly, the parameter update rule under the LAA strategy becomes:

$$\omega \leftarrow \omega - \eta \sum_{i=1}^{N} \nabla_{\text{sefp},i} = \omega - \eta X \sum_{i=1}^{N} \nabla_{\text{fp},i} - \eta \sum_{i=1}^{N} Y_i \tag{18}$$

where η is the learning rate. For the cumulative perturbation tends toward zero expectation, the high-frequency oscillations introduced by low bit-widths are effectively mitigated. This stabilizes the training process and enhances convergence. In addition, asynchronous accumulation avoids the memory growth in the computing device.

Finally, the pseudo-code (algorithm 1) is presented to more vividly demonstrate the overall pipeline.

Experiments

In this section, we present a comprehensive evaluation of the proposed OTARo method. We report the experimental setup, i.e., models, datasets, methods and implementation details, and show the overall results. In addition, we perform ablation studies, and calculate the memory reduction and speedup of SEFP quantization.

Models	Methods	E5M8	E5M7	E5M6	E5M5	E5M4	E5M3
LLaMA2-7B	Before Fine-Tuning	57.64%	57.65%	57.64%	57.42%	57.79%	56.99%
	FP16 Fine-Tuning	58.24%	58.27%	58.14%	57.96%	57.82%	56.46%
	Fixed Precision Fine-Tuning	58.31%	58.32%	58.27%	58.20%	58.14%	57.09%
	Ours	**59.15%**	**59.09%**	**58.93%**	**58.75%**	**58.70%**	**57.72%**
LLaMA2-13B	Before Fine-Tuning	60.87%	60.75%	60.76%	60.87%	60.40%	60.41%
	FP16 Fine-Tuning	61.43%	61.42%	61.43%	61.13%	60.84%	60.01%
	Fixed Precision Fine-Tuning	61.54%	61.43%	61.50%	61.17%	61.23%	60.63%
	Ours	**62.33%**	**62.05%**	**62.11%**	**61.83%**	**61.89%**	**61.10%**
LLaMA3-8B	Before Fine-Tuning	62.44%	62.35%	62.49%	62.31%	62.01%	59.81%
	FP16 Fine-Tuning	63.42%	63.50%	63.44%	63.29%	61.23%	59.62%
	Fixed Precision Fine-Tuning	63.89%	63.64%	63.55%	63.34%	62.30%	59.84%
	Ours	**64.09%**	**64.12%**	**63.99%**	**63.84%**	**63.34%**	**60.67%**
LLaMA3.2-3B	Before Fine-Tuning	57.21%	57.10%	57.39%	57.29%	56.68%	54.52%
	FP16 Fine-Tuning	57.80%	57.75%	57.92%	57.74%	56.78%	54.57%
	Fixed Precision Fine-Tuning	57.89%	58.08%	58.14%	57.85%	57.00%	54.78%
	Ours	**58.48%**	**58.46%**	**58.64%**	**58.21%**	**57.82%**	**56.03%**
Qwen3-8B	Before Fine-Tuning	64.45%	64.38%	64.72%	63.29%	63.13%	58.65%
	FP16 Fine-Tuning	65.82%	65.62%	65.80%	64.95%	63.62%	58.60%
	Fixed Precision Fine-Tuning	65.93%	66.10%	65.91%	65.14%	64.82%	61.05%
	Ours	**66.74%**	**66.64%**	**66.58%**	**66.19%**	**66.13%**	**61.51%**

Table 1: Zero-shot results of LLaMA2-7B, LLaMA2-13B, LLaMA3-8B, LLaMA3.2-3B, Qwen3-8B. We report average accuracies of all zero-shot tasks, i.e., Arc-Challenge, Arc-Easy, BoolQ, HellaSwag, MATHQA, OpenBookQA, PIQA, WinoGrande.

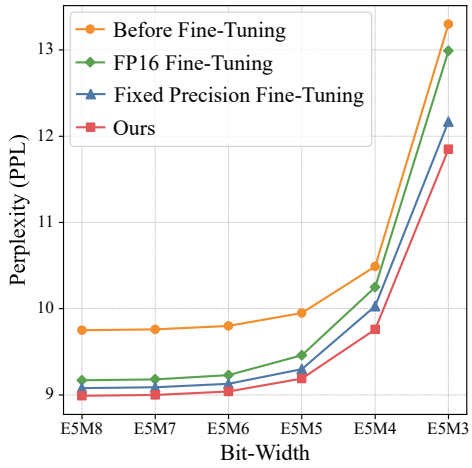


Figure 7: Task-specific fine-tuning results of LLaMA3.2-1B.

Setup

Models In the current development of LLMs, LLaMA (Touvron et al. 2023; Grattafiori et al. 2024) and Qwen(Yang et al. 2025) families have gained popularity due to strong performance and broad community support. Accordingly, in zero-shot experiments, we use LLaMA2-7B, LLaMA2-13B, LLaMA3-8B, LLaMA3.2-3B and Qwen3-8B, and in task-specific fine-tuning experiments, we use LLaMA3.2-1B.

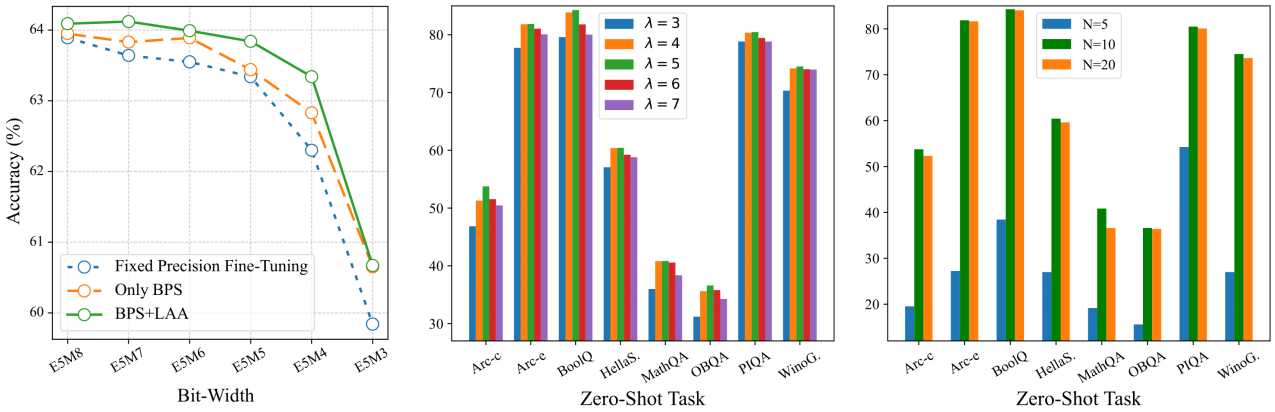
Datasets For zero-shot experiments, we adopt the Alpaca (Taori et al. 2023) dataset. Each model is fine-tuned on Alpaca, and evaluated on a suite of benchmarks spanning different reasoning and understanding skills: ARC-Easy, ARC-Challenge (Clark et al. 2018), BoolQ (Clark et al. 2019), HellaSwag (Zellers et al. 2019), MATHQA (Amini et al. 2019), OpenBookQA (Mihaylov et al. 2018), PIQA (Bisk et al. 2020), WinoGrande (Sakaguchi et al. 2021). For task-specific fine-tuning experiments, we adopt the WikiText2 (Merity et al. 2016) dataset. The model is fine-tuned and evaluated using its train and test set, respectively.

Methods We evaluate pre-trained models to show the optimizations from fine-tuning. FP16 fine-tuning and fixed precision fine-tuning are also included as baselines to highlight the bit-width robustness achieved by OTARo. In fixed precision fine-tuning, the backpropagation-based weight update mechanism aligns with that employed in OTARo. Fixed precision fine-tuning make the models adapt to quantization errors from each fixed bit-width, while multiplying the total fine-tuning time.

Implementation Details In the experiments, we set the learning rate to 1e-5, and use SGD optimizer. SEFP group size is set to 64, and the bit-widths include {E5M8, E5M7, E5M6, E5M5, E5M4, E5M3}. In the BPS strategy, exploration coefficient λ is set to 5, and in the LAA strategy, delayed updates are performed with a delay step N=10. For zero-shot experiments, each model is fine-tuned in a single epoch to avoid excessive adaptation, and evaluated with accuracy (%). For task-specific fine-tuning experiments, the model is fine-tuned in 20 epochs for a well-converged state, and evaluated with perplexity, i.e. PPL. In ablation studies, strategies comparison, λ, and N are included. All experiments are conducted on NVIDIA A6000 GPUs.

Zero-Shot Results

We report overall results of zero-shot experiments in table 1, where OTARo consistently achieves the highest average ac-



(a) strategies comparison (b) exploration coefficient λ (c) delay step N

Figure 8: Ablation results. We show the ablation results for strategies, exploration coefficient λ, delay step N. For strategies comparison, we report average accuracies of all zero-shot tasks, and for the others, we report E5M8 accuracies in each task.

curacy across all bit-widths (E5M8 to E5M3) for all models.

For challenging low-bit settings (E5M4, E5M3), OTARo also obtains high accuracies and outperforms the baselines. For example, in LLaMA3-8B, OTARo achieves 63.34% at E5M4 and 60.67% at E5M3, compared with 62.30% and 59.84% from fixed precision fine-tuning, and 61.23% and 59.62% from FP16 fine-tuning. And, in Qwen3-8B, OTARo reaches 66.13%, 61.51% respectively at E5M4, E5M3, exceeding both fixed precision fine-tuning results (64.82%, 61.05%) and FP16 fine-tuning results (63.62%, 58.60%).

These experimental results highlight the ability of OTARo to maintain robust performance in zero-shot tasks.

Task-Specific Fine-Tuning Performances

As shown in fig. 7, in the task-specific fine-tuning task, OTARo outperforms all baselines, achieving the lowest perplexity in all bit-widths. Then, specific numerical values are provided to show the optimization. Compared to fixed precision fine-tuning, OTARo obtains an average reduction of 0.16 PPL, and at lower bit-widths, the benefits are more pronounced, with reductions of 0.27 and 0.32 PPL respectively at E5M4 and E5M3. The experimental results confirm the effectiveness of OTARo in task-specific fine-tuning.

Ablation Studies

We perform ablation studies to evaluate the impacts of the BPS and LAA strategies. BPS alone can achieve robustness by exploring multiple bit-widths rather than relying on a fixed one. However, BPS-only fine-tuned models still suffer from low-precision gradient oscillations. By incorporating LAA, the model obtains further performance improvements.

Moreover, we ablate exploration coefficient λ in BPS. A value too large causes an under-use of high precisions, whereas a value too small leads to an insufficient exploration of low precisions. Experiments with $\lambda \in \{3, 4, 5, 6, 7\}$ confirm that λ=5 offers the best balance.

We also explore different values of delay step N in LAA.

Precisions	Mem. (GB)	Dec. Thpt. (token/s)
FP16	15.20	39.00
SEFP-E5M4	4.77 (69% ↓)	95.65 ($\times$2.45)

Table 2: The memory consumption (Mem.) and decoding throughput (Dec. Thpt.) of FP16 and SEFP LLaMA-8B (E5M4 taken as an example for SEFP).

N=10 strikes a balance between smoothing gradient oscillations and ensuring sufficient updates, achieving the best performance compared to N=5 and N=20.

Ablation results are shown in fig. 8.

Memory and Speed

We benchmark the memory consumption (including storage spaces for weights and KV cache) and decoding throughput of LLaMA3-8B respectively under FP16 and SEFP data formats, suppose an input of 2000 tokens, to show the efficiency gains from SEFP quantization. As shown in table 2, from FP16 to SEFP, the memory consumption is reduced by 69%, and the speedup of decoding throughput reaches $\times$2.45. These results highlight the practical advantages of SEFP in enabling light-weight and high-throughput on-device LLMs.

Conclusion

In this paper, we have proposed OTARo, a novel fine-tuning method that enables robust, flexible, and hardware-friendly LLMs at edge. OTARo can obtain one unified model to support all precisions after once fine-tuning. It introduces SEFP quantization to switch precisions flexibly, and performs a complete end-to-end learning workflow that stabilizes the performances across precisions. Experiments demonstrate its effectiveness in multiple LLMs and domains. In summary, this work advances the multi-precision adaptation of on-device LLMs, and provides a crucial methodological support for the real-world edge intelligence.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amini, A.; Gabriel, S.; Lin, P.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Ashkboos, S.; Mohtashami, A.; Croci, M. L.; Li, B.; Cameron, P.; Jaggi, M.; Alistarh, D.; Hoefler, T.; and Hensman, J. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37: 100213–100240.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.

Chen, M.; Shao, W.; Xu, P.; Wang, J.; Gao, P.; Zhang, K.; and Luo, P. 2024. Efficientqat: Efficient quantization-aware training for large language models. *arXiv preprint arXiv:2407.11062*.

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

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Gao, J.; Shen, J.; Zhang, Y.; Ji, W.; and Huang, H. 2024. Precision-Aware Iterative Algorithms Based on Group-Shared Exponents of Floating-Point Numbers. *arXiv preprint arXiv:2411.04686*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kalliojarvi, K.; and Astola, J. 2002. Roundoff errors in block-floating-point systems. *IEEE transactions on signal processing*, 44(4): 783–790.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6: 87–100.

Liu, Z.; Oguz, B.; Zhao, C.; Chang, E.; Stock, P.; Mehdad, Y.; Shi, Y.; Krishnamoorthi, R.; and Chandra, V. 2023. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*.

Liu, Z.; Zhao, C.; Fedorov, I.; Soran, B.; Choudhary, D.; Krishnamoorthi, R.; Chandra, V.; Tian, Y.; and Blankevoort, T. 2024. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

Lu, W.; Luu, R. K.; and Buehler, M. J. 2025. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(1): 84.

Manduchi, L.; Pandey, K.; Meister, C.; Bamler, R.; Cotterell, R.; Däubener, S.; Fellenz, S.; Fischer, A.; Gärtner, T.; Kirchler, M.; et al. 2024. On the challenges and opportunities in generative ai. *arXiv preprint arXiv:2403.00025*.

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

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

Park, Y.; Hyun, J.; Cho, S.; Sim, B.; and Lee, J. W. 2024. Any-precision llm: Low-cost deployment of multiple, different-sized llms. *arXiv preprint arXiv:2402.10517*.

Qiao, Y.; Chen, Z.; Zhang, Y.; Wang, Y.; and Huang, S. 2025. TeLLMe: An Energy-Efficient Ternary LLM Accelerator for Prefilling and Decoding on Edge FPGAs. *arXiv preprint arXiv:2504.16266*.

Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; and Huang, K. 2025. Mobile edge intelligence for large language models: A contemporary survey. *IEEE Communications Surveys & Tutorials*.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Tang, J.; Sorokin, R.; Ignasheva, E.; Jensen, G.; Chen, L.; Lee, J.; Kulik, A.; and Grundman, M. 2025. Scaling On-Device GPU Inference for Large Generative Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6355–6364.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei, L.; Ma, Z.; Yang, C.; and Yao, Q. 2024. Advances in the neural network quantization: A comprehensive review. *Applied Sciences*, 14(17): 7445.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, D.; He, N.; Hu, X.; Yuan, Z.; Yu, J.; Xu, C.; and Jiang, Z. 2024a. Post-training quantization for re-parameterization via coarse & fine weight splitting. *Journal of Systems Architecture*, 147: 103065.

Yang, H.; Zhang, Y.; Xu, J.; Lu, H.; Heng, P. A.; and Lam, W. 2024b. Unveiling the generalization power of fine-tuned large language models. *arXiv preprint arXiv:2403.09162*.

Yin, P.; Lyu, J.; Zhang, S.; Osher, S.; Qi, Y.; and Xin, J. 2019. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*.

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

Zhao, Z.; Li, H.; Liu, F.; Lu, Y.; Wang, Z.; Yang, T.; Jiang, L.; and Guan, H. 2025. QUARK: Quantization-Enabled Circuit Sharing for Transformer Acceleration by Exploiting Common Patterns in Nonlinear Operations. arXiv:2511.06767.

Zhou, S.; Wang, S.; Yuan, Z.; Shi, M.; Shang, Y.; and Yang, D. 2025. GSQ-Tuning: Group-Shared Exponents Integer in Fully Quantized Training for LLMs On-Device Fine-tuning. *arXiv preprint arXiv:2502.12913*.

Zubiaga, A. 2024. Natural language processing in the era of large language models. *Front Artif Intell*.