REP-Q: REPLACEABILITY-AWARE MIXED-PRECISION QUANTIZATION FOR MOE LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

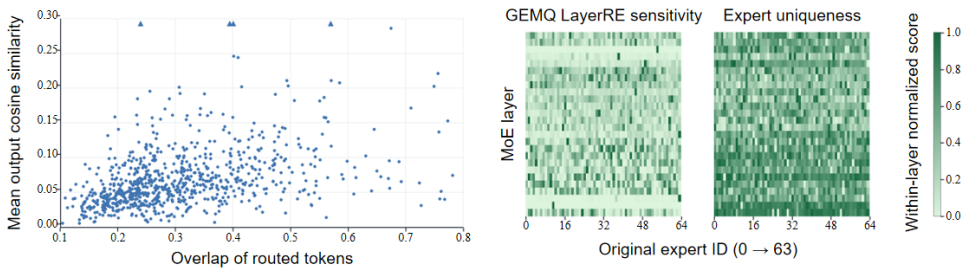
Mixture-of-Experts Large Language Models (MoE-LLMs) can expand model capacity while controlling computational costs, yet their massive expert parameters still introduce substantial memory overhead. Mixed-precision quantization mitigates this issue by assigning differentiated bit widths to distinct experts. Existing approaches mostly assess expert importance based on activation frequency or loss sensitivity but largely overlook expert replaceability, *i.e.*, whether the functionality of one expert can be fulfilled by others. Highly replaceable experts encode knowledge already covered by other experts, but experts with low replaceability require higher precision. Accordingly, this paper proposes Rep-Q, a mixed-precision quantization framework that jointly models expert replaceability and importance for quantization bit-width adjustment. (1) To translate the directly unmeasurable replaceability among experts into a computable metric, we design a similarity score to quantify similarities between expert output representations, establishing a quantitative foundation for modeling information dependencies across experts. (2) Leveraging these similarity scores, we construct a unified expert uniqueness metric by aggregating representation similarities between each expert and all remaining experts. This metric identifies single-expert replaceability from pairwise expert correlations and quantifies the irreplaceability of each expert's information, enabling the transition from relationship-level measurements to individual-level evaluations. (3) Subject to quantization budget constraints, we jointly model expert quantization sensitivity and uniqueness scores. This allocation ensures experts with high importance and low replaceability are assigned higher precision. Evaluations across multiple datasets on Mixtral 8×7B (pure routing experts) and DeepSeek-V2-Lite (with shared experts) demonstrate that under an identical quantization budget, Rep-Q yields consistent improvements in perplexity and downstream task accuracy over existing expert-level mixed-precision quantization baselines, and exhibits strong cross-architecture generalization.

1 INTRODUCTION

Large Language Models (LLMs) have achieved excellent performance in natural language processing tasks (Minaee et al., 2024). In recent years, Mixture-of-Experts Large Language Models (MoE-LLMs) have gradually become an important paradigm for model scale expansion (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Dai et al., 2024). MoE-LLMs replace the dense feedforward network with multiple parallel experts, and activate only a few experts for each token through the routing mechanism, so as to realize the expansion of model capacity with low computational overhead. However, the huge memory overhead is still the main obstacle to the actual deployment of MoE-LLMs.

Kim et al. (2022) propose a low-bit quantization of expert weights in MoE-LLMs, which can reduce the memory occupation of model storage and inference by representing expert weights with low bit widths. Li et al. (2025) pioneer expert-level mixed-precision quantization by assigning bit widths according to expert activation frequency. Many methods incorporate routing weights, reconstruction error, or theoretical sensitivity (Huang et al., 2025; Chowdhury et al., 2026; Deng et al., 2026). The core assumption of these methods is that more important experts should achieve higher quantification accuracy. However, *none of them has considered the replaceability among experts*, that is, whether

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) Experts processing more of the same tokens produce more similar outputs

(b) GEMQ LayerRE sensitivity and uniqueness highlight different experts

Figure 1: **Motivation for expert uniqueness.** (a) Correlation between token-overlap and output similarity across expert pairs. Each point represents one expert pair from 26 MoE layers. The x-axis is the Jaccard overlap of tokens processed by the two experts, and the y-axis is the cosine similarity of their output representations on shared tokens, confirming systematic functional overlap among experts. (b) Layer-wise comparison of GEMQ LayerRE sensitivity (left) and expert uniqueness (right) across 26 MoE layers with 64 routed experts each. Both heatmaps are row-wise min–max normalized, and darker colors indicate higher relative scores within each layer. The high-score regions do not coincide, showing that importance and replaceability capture distinct expert properties.

the function of an expert can be completed by other experts. Some experts may be frequently activated or have high loss sensitivity, but their functions may have been covered by other experts; On the contrary, **some experts have a low activation frequency but undertake irreplaceable specialized functions**. Assigning high-precision bit widths to highly replaceable experts not only wastes limited quantization budget, but may also result in irreversible performance loss due to truly irreplaceable experts being assigned to low bit widths.

We have analyzed the input-output relationship of expert pairs on C4 (Raffel et al., 2020) with DeepSeek-V2-Lite. As shown in Figure 1 (a), for each pair of experts, the horizontal axis measures the overlap of their received token sets, and the vertical axis measures the cosine similarity of their output representations on the jointly processed tokens. Based on the statistical results of the 32 pairs of experts with the most frequent co-routing in each layer of 26 MoE layers (a total of 832 pairs), experts who handle similar tokens usually produce more similar outputs (overall Spearman correlation coefficient +0.461, with 24 out of 26 layers positively correlated). This result confirms that there is indeed a systematic overlapping structure of expert functions in the MoE model. If two experts frequently receive the same input, their output representations tend to be consistent, indicating that at least one expert's function can be covered by another part.

However, the existing importance measures cannot identify which experts are more irreplaceable. As shown in Figure 1 (b), we normalized and visualized the GEMQ LayerRE sensitivity and expert uniqueness scores of all 26 layers and 64 routing experts in each layer. It can be clearly observed that the high scoring regions of the two indicators do not overlap: experts with high sensitivity to LayerRE may not necessarily have low uniqueness (*i.e.* high replaceability), and experts with high uniqueness (strong irreplaceability) may not necessarily rank high in LayerRE. This indicates that importance and replaceability capture different attributes of experts. Relying solely on importance for accuracy allocation will allocate high-precision bit widths to important but replaceable experts, while ignoring unimportant but irreplaceable experts. Based on this observation, this paper proposes Rep-Q, a mixed-precision quantization framework that takes expert replaceability into consideration. There are three main challenges to realizing the above ideas.

(1) It is difficult to directly observe and quantify the degree of functional overlap between experts from expert outputs. To evaluate the replaceability among experts, it is necessary to establish clear and computable evaluation criteria, so as to reliably measure the similarity of representations between different expert pairs.

(2) The representation similarity describes the relationship between two experts, while the precision allocation needs to evaluate the replaceability of a single expert. MoE-LLMs usually contain multiple MoE layers and a large number of experts, and the organization and functional roles of experts differ across different layers and architectures. Therefore, it is hard to directly use the scattered pairwise similarity to make individual-level replaceability judgments. It is necessary to combine the

complex MoE structure to aggregate the relationships between experts and transform them into a comparable expert-level replaceability index.

(3) Mixed-precision quantization requires not only evaluating experts, but also specific bit-width allocation strategies based on the evaluation results. The replaceability of experts reflects information redundancy and function compensation ability, whereas existing evaluation indicators measure the importance or quantization sensitivity of experts themselves. These two dimensions may conflict with each other, and any single dimension is difficult to fully describe the precision requirements of experts. Therefore, we need to explicitly model the trade-off between replaceability and importance, and transform the joint evaluation results into reasonable expert precision configurations under the given average bit-width constraint.

To address the above issues, this paper proposes the following three methods:

(1) To transform the functional overlap that is difficult to directly quantify among experts into a computable measure, a *similarity score* is designed based on linear CKA (Kornblith et al., 2019) to convert abstract representation dependencies into measurable similarity scores, providing a quantitative basis for subsequent replaceability analysis. CKA has been widely used to measure the alignment of inter-layer representations in neural networks and is suitable as an indicator for quantifying functional overlap between experts. By calculating the linear CKA values between different expert output representations, a computable expert-level similarity matrix is obtained.

(2) To identify the *uniqueness metric* for a single expert from pairwise functional similarity among experts, the similarity between a certain expert and all other experts (the lower the similarity, the higher the uniqueness) is aggregated based on the negative correlation between functional similarity and replaceability. The reference benchmark is adaptively determined according to the organization of experts in the model. For architectures without shared experts, other routing experts within the layer are used as the reference set. For architectures that include shared experts, additional shared experts are included in the reference benchmark. Thus, the pairwise similarity between experts is transformed into a comparable expert-level uniqueness index, quantifying the replaceability of each expert.

(3) A joint *experts rank* for quantization is constructed by considering both expert importance and uniqueness, and all experts are ranked in descending order based on this score. Given the constraint of average bit width, bit widths are allocated to experts in descending order of ranking results, so that experts with high importance and low replaceability can obtain higher precision, while experts with high replaceability use lower bit widths, thereby achieving a better balance between model compression rate and task performance.

The main contributions of this paper are as follows:

- An expert correlation analysis method is designed based on representation similarity to measure the overlap of expert functions. To address the issue of difficulty in directly measuring functional overlap among experts, a linear-CKA-based similarity score is used to provide a reliable quantitative basis for expert replaceability analysis in the MoE model.
- By aggregating and characterizing similarity and adapting to different MoE architectures (including or without shared experts), relationship-level metrics are transformed into comparable expert-level replaceability indicators in response to the issue that relationship-level similarity cannot directly reflect the replaceability of a single expert.
- In the precision allocation strategy, a joint evaluation and precision allocation strategy based on expert importance and uniqueness is proposed. We jointly consider expert importance and expert replaceability, generate expert-level bit-width allocation schemes under quantization budget constraints, and enable quantitative decision-making to simultaneously consider the dimensions of expert importance and expert replaceability.
- The proposed Rep-Q has been verified under various MoE architectures and experimental settings. These models are evaluated with significant structural differences, such as Mixtral 8×7B (pure routing experts) and DeepSeek-V2-Lite (including shared experts), across multiple datasets and quantization frameworks. The experimental results show that under the same quantization budget, Rep-Q can stably improve model performance and has good architecture generalization ability compared with existing mixed-precision quantization baselines.

2 Related Work

MoE-LLMs include multiple experts in the feed-forward sublayer and use a router to select a small subset of experts for each token. Although sparse activation lowers the computation of a single inference pass, the large number of expert parameters must still be loaded into memory, resulting in substantial storage and deployment costs. Existing work therefore commonly adopts weight-only quantization to represent expert weights at low bit-widths and reduce memory use and deployment cost. Depending on the bit-width, existing methods can be divided into uniform-precision quantization and expert-level mixed-precision quantization.

2.1 Uniform-Precision Quantization for MoE Models

Uniform-precision quantization methods assign the same bit-width to all experts and improve quantization through representation, calibration data, and error-mitigation techniques. Representative work includes the system-oriented expert-weight quantization of (Kim et al., 2022), MoQE (Kim et al., 2023), QMoE (Frantar & Alistarh, 2024), MoEQuant (Chen et al., 2025b), and EAQuant (Fu et al., 2025). MoQE (Kim et al., 2023) studies the relationship between low-bit expert quantization and model robustness. QMoE (Frantar & Alistarh, 2024) explores sub-1-bit compression for extremely large models. MoEQuant (Chen et al., 2025b) improves quantization calibration through expert-balanced sampling and affinity guidance, and EAQuant (Fu et al., 2025) reduces quantization error through expert-aware optimization and outlier handling. Other MoE quantization methods have also been explored (Chen et al., 2025a; Yin et al., 2026; Fang & Huang, 2026).

2.2 Mixed-Precision Quantization for MoE Models

Mixed-precision quantization for MoE-LLMs assigns different bit-widths to different experts. Its key differences lie in the criterion used to evaluate expert importance and the scope over which precision is allocated. QuantMoE-Bench (Li et al., 2025) studies bit allocation using expert activation statistics. The pre-loading mixed-precision quantization component of MC-MoE (Huang et al., 2025) jointly considers expert activation frequency, routing weights, and quantization error, and uses linear programming to determine the expert precision configuration. MxMoE (Duanmu et al., 2025) further studies sub-expert-granularity bit allocation and co-designs precision with hardware performance. Chowdhury et al. (2026) establish expert priorities using router changes and expert-weight statistics, and then perform multi-level bit-width allocation. GEMQ (Deng et al., 2026) jointly optimizes expert bit-widths across the entire model using quantization error and task-loss increments, and adapts the quantized routing behavior through router fine-tuning and progressive quantization. In recent years, mixed-precision quantization has emerged as an effective and increasingly studied approach for compressing MoE models (Xu et al., 2026; Lee et al., 2026; Chen et al., 2026; Yang et al., 2026; Chitty-Venkata et al., 2025; Qi et al., 2026).

3 Method

3.1 Problem Definition

For each routed expert, the bit width is selected from a candidate set $\mathcal{B}$, while the quantizer itself remains unchanged throughout our method. In an MoE model, the feed-forward network is replaced by multiple routed experts. Because different experts process only the tokens routed to them, their contributions to the final output differ. Let the l-th MoE layer contain N_l routed experts. For an input token x, the router selects the expert set $\mathcal{T}_l(x)$ and assigns routing weight $g_{l,i}(x)$ to each selected expert. The layer output is

$$y_l(x) = \sum_{i \in \mathcal{T}_l(x)} g_{l,i}(x) E_{l,i}(x) + E_l^{\mathrm{sh}}(x), \tag{1}$$

where $E_{l,i}(x)$ is the output of routed expert i. When the model contains a shared expert, $E_l^{\mathrm{sh}}(x)$ is its output; otherwise, this term is zero.

Equation (1) shows that quantizing an expert changes its output for the tokens routed to it and consequently affects the weighted MoE layer output. MoE quantization must therefore decide not

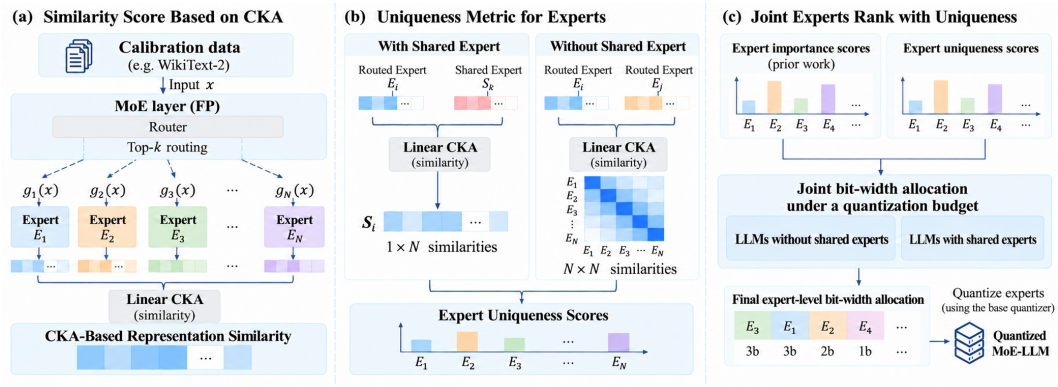


Figure 2: **Overview of the proposed Rep-Q framework.** (a) The routed expert outputs via Top-k routing in a floating-point MoE layer are extracted to compute CKA-based expert similarity score. (b) The uniqueness metric for all the expert is computed according to its structure. (c) The expert importance and uniqueness are jointly modeled to rank the experts and refine the base mixed-precision configuration.

only which low-bit representation to use, but also which bit-width to allocate to each expert. We assign each routed expert a bit-width $b_{l,i} \in \mathcal{B}$, where $\mathcal{B}$ is the set of candidate bit-widths. Let $P_{l,i}$ be the number of parameters in expert i at layer l. The average bit-width over routed experts is

$$\bar{b}(\mathbf{b}) = \frac{\sum_{l=1}^{L} \sum_{i=1}^{N_l} P_{l,i} b_{l,i}}{\sum_{l=1}^{L} \sum_{i=1}^{N_l} P_{l,i}} = B, \tag{2}$$

where B is the target average bit-width. Shared experts, attention modules, gates, and other non-expert parameters retain their base configuration.

Under this constraint, it is necessary to determine the mapping between experts and bit-widths within each layer. A base quantization method provides an initial expert-level bit-width configuration according to its own allocation criterion. Let $\mathbf{e}_l^0 = (e_{l,0}^0, \ldots, e_{l,N_l-1}^0)$ be the base expert order at layer l, where an earlier position has higher priority for receiving a larger bit-width, and let $\mathbf{w}_l^0 = (w_{l,0}^0, \ldots, w_{l,N_l-1}^0)$ with $w_{l,j}^0 = b_{l,e_{l,j}^0}^0$ be the bit-width sequence indexed by position. Thus, $\mathbf{e}_l^0$ specifies the expert order and $\mathbf{w}_l^0$ specifies the bit-width associated with each position; together they determine the base configuration $\mathbf{b}_l^0$ under the target average bit-width.

3.2 Similarity Score based on CKA

To measure differences among expert output representations and support subsequent within-layer ranking adjustments, we run the unquantized model on calibration data and record the routed experts and their outputs for each token. To evaluate the functional overlap among expert fastly, linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019) is adopted as a computable proxy for shared representational information and recorded expert outputs is used to measure expert correlation. This procedure only collects outputs and routing statistics; it does not update model parameters.

For routed expert i, the routed output contribution is retained including its actual routing weight, $H_{l,i}^r(x) = g_{l,i}(x) E_{l,i}(x)$, where $E_{l,i}(x)$ is the original output of routed expert i at layer l and $g_{l,i}(x)$ is the routing weight for token x. For a model with a shared expert its output is $H_l^s(x) = E_l^{\mathrm{sh}}(x)$, not multiplied by a routed-expert weight. For a model without shared experts, the outputs of two routed experts in the same layer that process the same tokens are compared.

Stack the outputs of two experts on their common tokens as $X, Y \in \mathbb{R}^{n \times d}$, where n is the number of common tokens and d is the output dimension. Let X_c and Y_c be the matrices centered along the token dimension. We adopt centered linear CKA to measure the correlation between the two output representations:

$$\mathrm{CKA}(X, Y) = \frac{\left\| X_c^\top Y_c \right\|_F^2}{\left\| X_c^\top X_c \right\|_F \left\| Y_c^\top Y_c \right\|_F}. \tag{3}$$

A larger CKA value indicates that the two experts produce more similar output representations on the same tokens.

The number of tokens available for comparison can vary across calibration batches. To avoid giving batches with different numbers of valid tokens equal influence, CKA is computed independently for each batch and aggregate the values by the number of common tokens. For effective batch c, let $n^{(c)}$ be the number of common tokens and let $S^{(c)}$ be the resulting CKA value. The similarity score is defined as

$$S_l(i,j) = \frac{\sum_{c\in\mathcal{C}} n^{(c)} S^{(c)}}{\sum_{c\in\mathcal{C}} n^{(c)}}, \tag{4}$$

where $\mathcal{C}$ is the set of effective batches for the expert pair. Batches with too few common tokens to estimate CKA reliably are excluded. Equation (4) gives batches with more jointly processed tokens greater influence instead of weighting every batch equally.

3.3 Uniqueness Metric for Experts

The designed similarity score measures pairwise similarity between expert output representations, but bit-width adjustment requires a relative expert priority. We map low correlation with reference outputs to a higher expert-priority score. Depending on whether an MoE layer contains a shared expert, the reference set for expert i is defined as

$$\mathcal{R}_l(i) = \begin{cases} \{\text{shared}\}, & \text{if layer } l \text{ contains a shared expert,} \\ \{j \mid j \neq i\}, & \text{otherwise.} \end{cases} \tag{5}$$

When layer l contains a shared expert, $\mathcal{R}_l(i)$ contains only that expert and $S_l(i, \text{shared})$ measures the correlation between its output and that of routed expert i on common tokens. Without a shared expert, $\mathcal{R}_l(i)$ contains the other routed experts in the layer and $S_l(i,j)$ measures the correlation between experts i and j on the tokens routed to both.

The average output correlation of expert i is

$$\overline{S}_l(i) = \frac{1}{|\mathcal{R}_l(i)|} \sum_{r\in\mathcal{R}_l(i)} S_l(i,r) = \begin{cases} S_l(i,\text{shared}), & \text{with a shared expert,} \\ \frac{1}{N_l-1}\sum_{j\neq i} S_l(i,j), & \text{otherwise.} \end{cases} \tag{6}$$

The expert uniqueness is defined as the complement of average correlation, $u_{l,i} = 1 - \overline{S}_l(i)$. A higher CKA indicates that an expert is more similar to its reference outputs, whereas a larger $u_{l,i}$ indicates that the expert output is more unique.

3.4 Joint Experts Rank with Uniqueness

There are two typical architectures for expert organization in MoE-LLMs: pure routing expert architecture (each layer only contains routing experts, without shared experts) and shared expert architecture (each layer includes fixed shared experts involved in computation in addition to routing experts). The two strategies share the same core principle: while maintaining the same number of experts in each accuracy level of each layer, adjust the allocation of bit width based on uniqueness sorting, so that experts with higher uniqueness can first obtain higher quantization accuracy.

Pure Routing Expert Architecture. For architectures with only a small number of routing experts in each layer, reordering all experts directly based on uniqueness metric is susceptible to statistical fluctuations, and full layer reordering may disrupt the importance information.

Specifically, first compare the importance ranking with the uniqueness ranking to identify intra layer sorting mismatches. Only perform corrections on layers with significant differences in uniqueness or mismatches exceeding a threshold, while ensuring that the corrections cover different network depths. Within the selected layer, only highly unique experts with obvious mismatches are moved to higher precision positions, while the relative order of the remaining experts remains unchanged. The correction range is adaptively determined based on the distribution of expert numbers in low precision levels (1-bit and 2-bit). The more low precision experts there are, the greater the quantization pressure in that layer, and the correction range is correspondingly extended to more layers.

Shared Expert Architecture. For architectures with multiple routing experts in each layer and shared experts, the larger the expert base, the more stable the statistical estimation of uniqueness metric, thus allowing for more fine-grained adjustment of bit width allocation. Shared experts, as they always participate in the calculation of all tokens, naturally have a high degree of uniqueness in their functions, with a fixed allocation of the highest accuracy and no involvement in reordering. The allocation of bit width for routing experts is optimized based on uniqueness ranking, with the core goal of enabling experts with high uniqueness to achieve higher accuracy as much as possible, while controlling the incremental performance loss.

The correction process follows three principles: (1) the bit width can only be transferred from low uniqueness experts to high uniqueness experts, that is, any adjustment should not allocate high precision to experts with lower uniqueness; (2) The additional quantization error introduced by each adjustment step does not exceed the preset threshold to constrain the deviation of the correction from the original importance allocation; (3) The strength of the correction is positively correlated with the number of experts at each level and the degree of mismatch. The more experts there are and the greater the difference between the uniqueness ranking and the baseline ranking, the more comprehensive the adjustments can be made.

Based on the above principles, the rank is carried out in two steps. Firstly, local swapping is performed within the layer: sorting by uniqueness, high uniqueness experts are moved to higher precision positions, while low uniqueness experts at similar precision levels are used as providers of bit width capacity. Each swapping step must satisfy the requirement of bounded error increment and constant number of experts at each precision level within the layer. If the overall sorting consistency is still below the set threshold after local exchange, further global reallocation is performed: under the precision budget constraint of each layer, the bit width capacity is systematically transferred from the tail of the uniqueness sorting to the head, and the scheme with the highest sorting consistency and the lowest error cost is selected from a set of feasible configurations as the final allocation.

4 Experiments

4.1 Experimental Setup

Models. The proposed method is evaluated on two representative MoE-LLMs: (1) Mixtral 8×7B (Jiang et al., 2024), which consists of 32 MoE layers with 8 routed experts per layer; and (2) DeepSeek-V2-Lite (DeepSeek-AI, 2024), which contains 27 Transformer layers, where the first layer is a dense layer and the remaining 26 layers are MoE layers, each consisting of 64 routed experts and 2 shared experts.

Quantization and Implementation Settings. All experiments use GPTQ (Frantar et al., 2023) as the underlying quantization method. For Mixtral 8×7B, we follow the quantization settings and base expert ordering of Chowdhury et al., with expert bit-width candidates of $\{1, 2, 3\}$ bits, and evaluate multiple average bit-width budgets ranging from 1.5 to 2.5 bits. For DeepSeek-V2-Lite, we follow the quantization settings and base precision allocation of GEMQ and evaluate average expert bit-widths of 1.5, 2.0, and 2.5 bits. All quantization is run on NVIDIA RTX 5000 Ada Generation GPUs and all evaluations on NVIDIA RTX 5880 Ada Generation GPUs; the software environment follows that of MC-MoE (Huang et al., 2025) and GEMQ (Deng et al., 2026).

4.2 Experimental Results

Overall performance comparison. The accuracy of Mixtral 8×7B on eight zero-shot tasks is listed in Table 1. Bold denotes the highest value, and bold with underline denotes multiple methods achieving the same highest value. At the same average bit width, the proposed Rep-Q achieves average accuracy comparable to or better than the optimal baseline in all test configurations. Specifically, within the entire bit width range of 1.5-bit to 2.5-bit, the average accuracy of Rep-Q is consistently higher than that of MC-MoE and Chowdhury. It is worth noting that as the average bit width gradually decreases from 2.5-bit to 1.5-bit, the performance gap between the three methods shows a trend of increase. When the bit width is low (1.5 to 1.875-bit), Rep-Q has the most obvious relative advantage, indicating that considering expert uniqueness can effectively alleviate accuracy loss in scenarios with high quantization pressure.

Table 1: Mixtral 8×7B on eight zero-shot tasks. Bold denotes the highest value, and bold with underline denotes multiple methods achieving the same highest value.

Method	Avg. bits/exp.	PIQA	Wino.	MathQA	ARC-e	ARC-c	BoolQ	MMLU	HellaS.	Avg.
FP16	16	83.79	76.48	41.84	83.46	59.90	85.11	67.87	84.02	72.81
MC-MoE	1.5	72.80	**67.88**	**26.57**	**64.10**	**39.85**	**71.99**	31.16	64.70	54.88
	1.625	**76.44**	**70.32**	**26.93**	**67.47**	41.72	66.45	37.56	67.93	56.85
	1.75	75.19	68.51	**29.98**	68.86	42.66	**76.02**	41.16	70.27	59.08
	1.875	**77.04**	**70.64**	**30.39**	**70.33**	**45.65**	74.25	45.55	73.45	**60.91**
	2.0	**78.51**	**73.01**	**32.83**	73.19	47.35	78.75	47.06	74.38	63.13
	2.125	**78.67**	**73.24**	**32.46**	**74.87**	**50.34**	75.35	51.12	**76.27**	64.04
	2.25	80.03	**74.43**	**32.70**	74.79	48.72	79.91	53.37	**77.73**	65.21
	2.375	79.92	74.11	**35.51**	76.26	51.71	75.26	53.75	78.18	65.59
	2.5	**81.66**	73.56	**37.15**	76.56	52.56	82.23	56.37	79.68	67.47
Chowdhury	1.5	72.74	63.77	24.19	60.27	36.43	66.70	36.43	64.48	53.13
	1.625	73.50	63.61	25.66	64.02	41.21	72.35	39.23	67.58	55.90
	1.75	74.54	**70.09**	27.81	68.81	42.41	70.43	**45.10**	69.76	58.62
	1.875	76.17	70.40	26.53	68.60	44.62	75.11	**47.34**	73.60	60.30
	2.0	77.31	69.77	29.28	69.78	46.93	**79.79**	48.61	74.15	61.95
	2.125	78.56	72.14	30.02	72.43	48.98	**83.06**	52.16	75.92	64.16
	2.25	78.78	73.40	31.56	73.27	49.23	82.84	**55.60**	77.56	65.28
	2.375	**81.72**	**75.14**	32.96	76.77	**53.84**	**85.54**	57.47	78.24	67.71
	2.5	81.18	**74.59**	36.62	**78.75**	**54.61**	85.84	59.61	79.59	**68.85**
Ours	1.5	**73.72**	65.19	26.20	62.04	39.68	71.22	**38.31**	**65.90**	**55.28**
	1.625	73.50	66.38	26.10	66.50	**42.06**	**73.21**	**39.52**	**68.07**	**56.92**
	1.75	**76.01**	69.38	27.87	**69.23**	**42.75**	75.26	43.31	**70.48**	**59.29**
	1.875	76.33	68.51	28.98	69.49	44.71	**75.90**	46.85	**73.85**	60.58
	2.0	77.86	71.98	31.49	**73.70**	**48.72**	76.15	**51.94**	**74.84**	**63.33**
	2.125	78.56	71.98	30.75	73.40	49.66	81.16	**52.44**	76.13	**64.26**
	2.25	**80.74**	73.32	31.12	**76.35**	**51.79**	**84.28**	53.11	77.20	**65.99**
	2.375	80.25	74.51	34.67	**77.99**	53.41	84.59	**57.83**	**79.05**	**67.79**
	2.5	80.79	**74.59**	36.78	77.74	54.52	**86.30**	**60.21**	**79.86**	**68.85**

Table 2: Mixtral 8×7B on seven zero-shot tasks.

Method	Avg. bits/exp.	PIQA	ARC-e	ARC-c	HellaS.	Wino.	MathQA	MMLU	Avg.
GEMQ	1.5	73.50	59.81	35.75	62.97	**65.67**	25.23	35.82	51.25
	2.0	77.42	68.73	43.60	72.54	69.30	29.15	48.64	58.48
	2.5	**81.83**	**79.04**	54.27	79.54	74.11	35.31	57.79	65.98
Ours	1.5	**73.72**	**62.04**	**39.68**	**65.90**	65.19	**26.20**	**38.31**	**53.01**
	2.0	**77.86**	**73.70**	**48.72**	**74.84**	**71.98**	**31.49**	**51.94**	**61.50**
	2.5	80.79	77.74	**54.52**	**79.86**	**74.59**	**36.78**	**60.21**	**66.36**

Comparison with GEMQ. Table 2 compares Rep-Q and GEMQ on seven zero-shot tasks in Mixtral 8×7B. GEMQ is a quantization method based on model level error estimation, and its precision allocation depends on quantization sensitivity estimation. The experimental results show that under the 1.5-bit, 2.0-bit, and 2.5-bit configurations, the average accuracy of Rep-Q reaches 53.01%, 61.50%, and 66.36%, respectively, all higher than the corresponding results of GEMQ (51.25%, 58.48%, and 65.98%). This result indicates that even when compared with methods that rely on fine error estimation, Rep-Q can achieve better accuracy allocation through uniqueness ranking correction alone, further demonstrating the effectiveness of expert uniqueness as a basis for allocation.

Generalization on shared expert architecture. Table 3 shows the results of DeepSeek-V2-Lite on seven zero-shot tasks as well as WikiText-2 (Merity et al., 2017) and C4 perplexity. This model contains 64 routed experts and 2 shared experts in each MoE layer, and is a representative testing platform for verifying the generalization ability of the method on MoE-LLMs. In task accuracy, Rep-Q achieved average accuracies of 46.87%, 54.70%, and 61.14% at 1.5-bit, 2.0-bit, and 2.5-bit, respectively, which are better or equal to the corresponding results of GEMQ (46.27%, 54.65%, and 61.05%). In perplexity index, the WikiText-2 perplexity of Rep-Q at 2.0-bit and 2.5-bit is 7.25 and 6.64, respectively, which is on par with GEMQ; On C4, it achieved results equivalent to or better than GEMQ. It is worth noting that under the extremely low bit setting of 1.5-bit, Rep-Q is on par with GEMQ on WikiText-2 (9.16), while the difference on C4 (17.52 *vs.* 17.49) is negligible. This indicates that in architectures with shared experts, Rep-Q's fine-grained reallocation strategy can effectively utilize the unique differences between a large number of routing experts, achieving more reasonable bit width allocation without increasing quantization budget.

Table 3: DeepSeek-V2-Lite on seven zero-shot tasks and perplexity.

Method	Avg. bits/exp.	PIQA	ARC-e	ARC-c	HellaS.	Wino.	MathQA	MMLU	Avg.	WT2	C4
FP16	16	80.03	76.43	49.06	77.98	71.11	39.83	54.89	64.19	6.31	9.32
GEMQ	1.5	70.62	**56.48**	31.06	54.05	62.27	**25.19**	24.18	46.27	**9.16**	**17.49**
	2.0	77.31	67.68	**40.79**	68.49	64.88	**28.04**	**35.35**	54.65	**7.24**	11.83
	2.5	78.73	**74.12**	46.08	74.40	68.35	35.78	**49.93**	61.05	**6.64**	10.48
Ours	1.5	**71.06**	**56.48**	**31.83**	**56.00**	**63.06**	24.49	**25.16**	**46.87**	**9.16**	17.52
	2.0	**77.80**	**68.77**	40.19	**68.50**	**66.85**	26.26	34.53	**54.70**	7.25	**11.80**
	2.5	**79.16**	73.65	**46.50**	**74.58**	**68.98**	**35.95**	49.19	**61.14**	**6.64**	**10.47**

Table 4: Ablation on the direction of uniqueness on Mixtral 8×7B.

Bits	Ordering	PIQA	ARC-e	ARC-c	BoolQ	HellaS.	Wino.	MathQA	MMLU	Avg.
2.5	Fwd uniq	80.90	78.28	53.16	86.33	79.48	73.64	37.05	60.05	**68.61**
	Rev uniq	80.52	76.73	53.07	86.36	79.00	75.37	36.78	60.33	68.52
2.0	Fwd uniq	75.73	70.83	46.59	78.96	74.13	69.30	29.51	47.14	**61.52**
	Rev uniq	77.20	69.49	44.20	78.62	75.35	68.90	29.88	46.65	61.29
1.5	Fwd uniq	72.63	60.73	38.05	71.90	65.87	64.80	27.60	38.07	**54.96**
	Rev uniq	51.31	29.17	28.07	57.00	47.15	58.17	21.88	24.18	39.62

4.3 Ablation Study

To verify the core hypothesis that experts with higher uniqueness should receive higher precision protection, we designed a sorting direction ablation experiment in Figure 3: replacing the allocation order of experts' bit widths with sorting in ascending order of uniqueness scores (forward, *i.e.* high uniqueness corresponds to high precision) or descending order (reverse, *i.e.* high uniqueness corresponds to low precision), while keeping the rest of the settings consistent with the main experiment. Table 4 lists the ablation results on the Mixtral 8×7B. Under the 1.5-bit, 2.0-bit, and 2.5-bit configurations, the average accuracy of forward uniqueness sorting is 54.96%, 61.52%, and 68.61%, respectively, all higher than reverse sorting (39.62%, 61.29%, and 68.52%). It is worth noting that under the extremely low bit setting of 1.5-bit, reverse sorting causes the model performance to approach collapse (average accuracy drops sharply to 39.62%), while forward sorting still maintains 54.96% of available performance. As the bit width increases, the performance loss of reverse sorting gradually narrows (the gap at 2.5-bit is reduced to 0.09%), indicating that when the quantization budget is sufficient, the influence of sorting direction weakens, which is consistent with the conclusion in the main experiment.

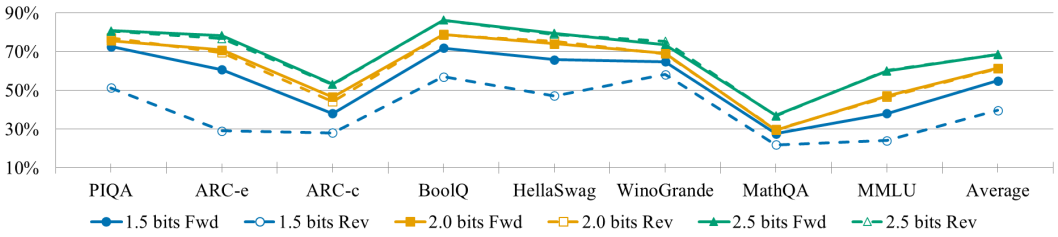


Figure 3: Ablation on uniqueness sorting (Mixtral 8×7B). Fwd (high uniqueness to high precision) surpasses Rev (reverse) across 1.5/2.0/2.5 bits.

5 Conclusion

This paper proposes Rep-Q, a mixed-precision quantization framework for MoE-LLMs that exploits expert uniqueness as a complementary criterion to traditional importance measures. Unlike prior work relying solely on activation frequency or loss sensitivity, Rep-Q constructs a similarity-based metric to quantify inter-expert replaceability, assigning higher bit-width to experts with high importance and low replaceability under a given quantization budget. Experiments on Mixtral 8×7B and DeepSeek-V2-Lite demonstrate that Rep-Q consistently outperforms existing expert-level mixed-precision baselines in both perplexity and downstream accuracy, validating uniqueness as an effective and orthogonal allocation signal.

AI USE STATEMENT

In this work, we used generative AI tools to polish and translate the manuscript to improve its readability.

We have not used generative AI tools for other tasks with required disclosure, and the remaining required disclosure tasks are not applicable to this work.

We have reviewed all AI-assisted work.

We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yuanteng Chen, Yuantian Shao, Peisong Wang, and Jian Cheng. Eac-moe: Expert-selection aware compressor for mixture-of-experts large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12942–12963, 2025a.

Yuanteng Chen, Peisong Wang, Zhilei Liu, Nanxin Zeng, Yuantian Shao, Shiqiang Lang, Tao Liu, Chuangyi Li, Qinghao Hu, Gang Li, et al. Mode: Modality-decomposed expert-level mixed-precision quantization for moe multimodal llms. *arXiv preprint arXiv:2606.17118*, 2026.

Zhixuan Chen, Xing Hu, Dawei Yang, Zukang Xu, Xu Chen, Zhihang Yuan, Sifan Zhou, and Jiangyong Yu. MoEQuant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=0epuNvt5Dj.

Krishna Teja Chitty-Venkata, Jie Ye, and Murali Emani. Mopeq: Mixture of mixed precision quantized experts. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4082–4091. IEEE, 2025.

Mohammed Nowaz Rabbani Chowdhury, Kaoutar El Maghraoui, Hsinyu Tsai, Naigang Wang, Geoffrey W. Burr, Liu Liu, and Meng Wang. Efficient quantization of mixture-of-experts with theoretical generalization guarantees. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yiMlVBAoQi.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. URL https://arxiv.org/abs/2405.04434.

Jianing Deng, Song Wang, Dongwei Wang, Zijie Liu, Tianlong Chen, Huanrui Yang, and Jingtong Hu. GEMQ: Global expert-level mixed-precision quantization for moe LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5Q4ts0pr9T.

Haojie Duanmu, Xiuhong Li, Zhihang Yuan, Size Zheng, Jiangfei Duan, Xingcheng Zhang, and Dahua Lin. MxMoE: Mixed-precision quantization for MoE with accuracy and performance co-design. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14793–14806. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/duanmu25a.html.

Yi-Zeng Fang and Juinn-Dar Huang. Beyond freezing the router: Rank-aligned post-training quantization for mixture-of-experts models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=bPsPPI65hf.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Elias Frantar and Dan Alistarh. Qmoe: Sub-1-bit compression of trillion parameter models. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 439–451, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/c74b624843218d9b6713fcf299d6d5e4-Paper-Conference.pdf.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Zhongqian Fu, Tianyi Zhao, Ning Ding, Xianzhi Yu, Xiaosong Li, Yehui Tang, and Yunhe Wang. EAQuant: Enhancing post-training quantization for MoE models via expert-aware optimization. *arXiv preprint arXiv:2506.13329*, 2025. URL https://arxiv.org/abs/2506.13329.

Wei Huang, Yue Liao, Jianhui Liu, Ruifei He, Haoru Tan, Shiming Zhang, Hongsheng Li, Si Liu, and Xiaojuan Qi. Mixture compressor for mixture-of-experts llms gains more. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 68628–68648, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/abc1943857a42935ceacff03c524bb44-Paper-Conference.pdf.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Young Jin Kim, Rawn Henry, Raffy Fahim, and Hany Hassan Awadalla. Who says elephants can't run: Bringing large scale moe models into cloud scale production. In *Proceedings of the third workshop on simple and efficient natural language processing (SustaiNLP)*, pp. 36–43, 2022.

Young Jin Kim, Raffy Fahim, and Hany Hassan Awadalla. Mixture of quantized experts (MoQE): Complementary effect of low-bit quantization and robustness. *arXiv preprint arXiv:2310.02410*, 2023. URL https://arxiv.org/abs/2310.02410.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Deokjae Lee, Sihun Chu, and Hyun Oh Song. Q-strata: Hierarchical bit allocation for mixed-precision quantization of mixture-of-experts llms. *arXiv preprint arXiv:2608.30564*, 2026.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Pingzhi Li, Xiaolong Jin, Zhen Tan, Yu Cheng, and Tianlong Chen. Quantmoe-bench: Examining post-training quantization for mixture-of-experts. *arXiv preprint arXiv:2406.08155*, 2025. URL https://arxiv.org/abs/2406.08155.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Huachen Qi, Ruiyu Zhuo, Bowen Shi, Xiang Chang, Fei Chao, Changjing Shang, and Qiang Shen. Profiling-free mixed-precision quantization for MoE LLMs via fuzzy rule interpolation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21484–21499, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.982. URL https://aclanthology.org/2026.acl-long.982/.

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. URL http://jmlr.org/papers/v21/20-074.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Yuzhuang Xu, Xu Han, Yuanchi Zhang, Yixuan Wang, Yijun Liu, Shiyu Ji, Qingfu Zhu, and Wanxiang Che. Camera: Multi-matrix joint compression for moe models via micro-expert redundancy analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27395–27404, 2026.

Wanqi Yang, Yuexiao Ma, Alexander Conzelmann, Xiawu Zheng, Michael W. Mahoney, T. Konstantin Rusch, and Shiwei Liu. Alphaq: Calibration-free bit allocation for mixture-of-experts quantization. In *ICLR 2026 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2026. URL https://openreview.net/forum?id=rbE8Pxs8vx.

Xiangyang Yin, Xingyu Liu, Tianhua Xia, BO BAO, Vithursan Thangarasa, Valavan Manohararajah, Eric Sather, and Sai Qian Zhang. Codequant: Unified clustering and quantization for enhanced outlier smoothing in low-precision mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ATpchFiBQi.