Automated Fine-Grained Mixture-of-Experts Quantization

Zhanhao Xie[1], **Yuexiao Ma**[1], **Xiawu Zheng**[1], **Fei Chao**[1], **Wanchen Sui**[2], **Yong Li**[2], **Shen Li**[2], **Rongrong Ji**[1,*]
[1]Key Laboratory of Multimedia Trusted Perception and Effcient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]Alibaba Cloud Computing

Abstract

The Mixture of Experts (MoE) architecture enables efficient model scaling through conditional computation, where only subset of parameters are activated per input. However, this distributed architecture poses unprecedented challenges for model compression, as conventional quantization methods optimized for dense networks prove inadequate. This paper introduces a specialized quantization framework for MoE architectures, motivated by our discovery that weight matrices across expert networks exhibit distinctive channel-wise outlier distributions, necessitating a more nuanced compression approach. Through theoretical analysis incorporating Fisher Information matrices and condition number characteristics, we establish a fundamental relationship between layer functionality and quantization sensitivity, demonstrating that down-projection layers inherently demand higher precision compared to up-projection layers. Leveraging these insights, we develop an automated channel-wise quantization framework that dynamically determines optimal bit-width allocations while maintaining minimal computational overhead through efficient statistical approximations. When evaluated on the Mixtral-8x7b-v0.1 architecture, our methodology demonstrates a 3.96% improvement over existing state-of-the-art approaches across natural language understanding benchmarks, while achieving superior compression ratios. Code is available at: https://github.com/cei1ing4/Fine-Grained-MoE-Quantization

1 Introduction

In recent years, Large Language Models (LLMs) (Touvron et al., 2023a,b; Reid et al., 2024; Zhang et al., 2022; Yang et al., 2024) have demonstrated unprecedented advancements in natural language processing. However, this remarkable progress has been accompanied by an exponential increase in computational resource requirements. The Mixture of Experts (MoE) architecture (Jacobs et al., 1991; Fedus et al., 2022; Lepikhin et al., 2020) has emerged as a transformative paradigm in large language models, introducing efficient parameter scaling through selective activation of context-relevant parameter subsets. Through strategic partitioning of the parameter space into specialized expert networks coupled with dynamic routing mechanisms, MoE models achieve substantial performance gains while maintaining modest training costs and introducing only marginal computational overhead during inference (Jiang et al., 2024; Dai et al., 2024; Liu et al., 2024b). This innovative approach enables the processing of increasingly sophisticated tasks while optimizing resource utilization, marking a significant milestone in scaling language model capabilities (Rajbhandari et al., 2022; Chen et al., 2022).

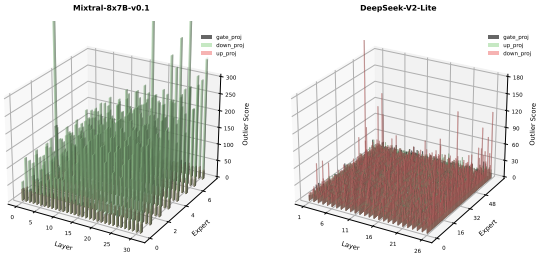


Figure 1: Comparison of different outlier metrics (W_{up}, W_{gate}, and W_{down}) across layers and experts in Mixtral-8x7B and DeepSeek-V2-Lite, revealing consistently elevated values in down-projection metrics across MoE architectures.

The exponential growth of language models has heightened the critical importance of model compression. Various compression techniques, including quantization (Frantar et al., 2023; Yuan et al., 2023; Shao et al., 2023; Ma et al., 2024; Liu et al., 2024c), sparsification (Frantar and Alistarh, 2023;

*Corresponding author: rrji@xmu.edu.cn

Findings of the Association for Computational Linguistics: ACL 2025, pages 27024–27037

Sun et al., 2023), and knowledge distillation (Hsieh et al., 2023; Gu et al., 2024), have demonstrated remarkable efficacy in reducing model footprint while preserving performance. Among these approaches, quantization has emerged as particularly promising, with recent advances successfully addressing the challenging outlier problem in large dense models through sophisticated weight compensation and affine transformation techniques.

However, these quantization methodologies, originally designed for dense architectures, encounter unique challenges when applied to MoE-LLM architectures. The fundamental distinction lies in the dynamic routing mechanism (Shazeer et al., 2017), where input tokens are intelligently directed to specific experts based on semantic content. This sophisticated routing mechanism, combined with sparse computational patterns, renders conventional quantization techniques inadequate without compromising model performance (Li et al., 2023). Moreover, the inherent heterogeneity of expert structures indicates that individual experts within the MoE framework possess distinct parameters and specialized functionalities. Unlike traditional homogeneous neural networks, this architectural diversity suggests that each expert requires a tailored compression strategy, introducing additional complexity to the quantization process. The distinctive characteristics of MoE architectures present unprecedented quantization challenges, particularly in managing outliers effectively across heterogeneous expert networks. Recent research has begun addressing these challenges in MoE compression. (He et al., 2024) introduced a unified compression framework integrating expert slimming with quantization, marking an initial step toward comprehensive MoE compression. Additionally, (Li et al., 2024) investigated variable quantization bit-width allocation based on importance metrics across different structural granularities. However, these approaches predominantly focus on static structural metrics, without fully addressing the unique characteristics of MoE architectures. The field currently lacks a specialized quantization solution engineered specifically for the distinctive nature of MoE models, particularly in effectively managing heterogeneous expert structures.

In response to the unprecedented challenges inherent in quantizing MoE models ((Jiang et al., 2024; Li et al., 2024)), we present a novel automated mixed-precision quantization framework specifically designed for MoE-based Large Language Models. Our approach is fundamentally motivated by a critical empirical observation: weight matrices within MoE architectures exhibit distinctive channel-wise outlier patterns, with substantial heterogeneity across different channels within individual experts. As illustrated in Figure 1 and 2, this phenomenon is consistently observed across diverse MoE architectures, with particularly pronounced patterns in down-projection layers. These findings fundamentally challenge conventional layer-wise quantization paradigms and underscore the necessity for more granular quantization strategies. Through comprehensive theoretical analysis, we establish that various components within MoE architectures demonstrate differential susceptibility to quantization effects. Specifically, we prove that down-projection layers exhibit inherently greater sensitivity to quantization perturbations compared to their up-projection counterparts, as evidenced by their elevated condition numbers and Fisher Information metrics. This theoretical foundation provides principled guidance for optimal bit allocation across the model's architectural components. To reconcile these theoretical insights with practical deployment constraints, we introduce computationally efficient statistical approximation techniques that facilitate automated bit allocation without introducing additional training complexity. Our methodology maintains rigorous theoretical guarantees while achieving the computational efficiency necessary for large-scale model deployment. Experimental results demonstrate significant improvements over existing approaches, achieving a 3.96% performance gain on the Mixtral-8x7B model under extreme low-bit configurations.

The primary contributions of this work are:

- We elucidate and systematically characterize the channel-wise heterogeneity in MoE weight distributions, revealing distinctive patterns that necessitate the development of fine-grained quantization methodologies.
- We establish a comprehensive theoretical framework that quantifies and explicates layer-wise sensitivity variations in MoE architectures, providing mathematical evidence for the increased quantization precision requirements in down-projection layers.
- We propose an integrated quantization framework that operationalizes these theoretical insights, automatically determining optimal bit-

width allocations while maintaining minimal computational overhead through sophisticated statistical approximations.

- We empirically validate our approach through extensive experimentation, demonstrating superior compression ratios while maintaining model performance across multiple MoE architectures.

2 Related Work

2.1 Mixture-of-Experts Models

Mixture-of-Experts (MoE) architectures have emerged as a significant innovation in Large Language Models (LLMs) (Jiang et al., 2024; Dai et al., 2024; Liu et al., 2024a), offering a balance between model capacity and computational efficiency. Originally proposed by (Jacobs et al., 1991), MoE has evolved substantially through key developments like (Shazeer et al., 2017)'s application to transformer models and (Fedus et al., 2022)'s introduction of sparse gating mechanisms for efficient routing. The theoretical foundations have been strengthened by works such as (Chen et al., 2022; Chowdhery et al., 2022), while recent advances like the Mixtral model (Jiang et al., 2024) have demonstrated that MoE can match full-parameter LLM performance while using significantly fewer active parameters. These developments have sparked growing interest in MoE optimization and compression techniques (Li et al., 2023; Rajbhandari et al., 2022), highlighting the architecture's potential for efficient, large-scale language modeling.

2.2 Post-Training Quantization

Post-training quantization (PTQ) (Wei et al., 2022b; Yao et al., 2022; Ashkboos et al., 2024; Liu et al., 2024c; Sun et al., 2024) has emerged as an efficient technique for model compression, particularly beneficial for LLMs. Unlike quantization-aware training or fine-tuning (Tailor et al., 2020; Ding et al., 2022), PTQ operates on pre-trained models without extensive retraining (Liu et al., 2021; Fang et al., 2020). In computer vision, AdaRound (Nagel et al., 2020) optimizes weight rounding strategies, BRECQ (Li et al., 2021) introduces block-wise reconstruction, and QDROP (Wei et al., 2022a) enhances performance through activation substitution. For LLMs, GPTQ(Frantar et al., 2023) uses approximate second-order information for layerwise quantization, SmoothQuant (Xiao et al., 2023) tackles activation outliers, and AWQ (Lin et al., 2023) preserves critical weights' precision. OmniQuant (Shao et al., 2023) integrates multiple strategies, combining mixed-precision quantization, outlier handling, and adaptive rounding. AffineQuant (Ma et al., 2024) introduces an affine transformation to adjust weight distribution, effectively reducing quantization errors. These advancements have significantly improved LLM deployment efficiency on resource-constrained devices (Kim et al., 2023; Chen et al., 2024).

3 Preliminary

Mixture of Experts. The Mixture-of-Experts (MoE) architecture represents a significant advancement in Transformer models (Fedus et al., 2022; Jiang et al., 2024), superseding the conventional Feed-Forward Network (FFN) sublayer with a more sophisticated dynamic structure. An MoE layer incorporates N parallel FFN modules, designated as experts $E_1, E_2, \ldots, E_N$, alongside a router network (Shazeer et al., 2017) that facilitates dynamic input allocation through a specialized gating mechanism.

For an input vector $x \in \mathbb{R}^{d_{\text{in}}}$, the MoE layer generates output $y \in \mathbb{R}^{d_{\text{out}}}$ through the following computation:

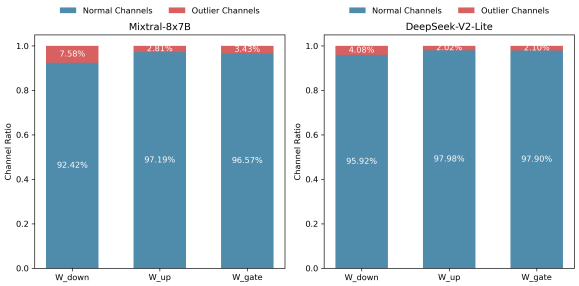
$$y = \sum_{i=1}^{N} r_i(x) E_i(x), \tag{1}$$

where $r_i = [router(x; G)]_i$ denotes the gating weight assigned to expert i, G represents the gating function that determines routing probabilities based on input characteristics, and $E_i(x)$ corresponds to the output of the i-th expert. Following architectural principles established in (Touvron et al., 2023a; Jiang et al., 2024), each expert implements an enhanced FFN structure defined as:

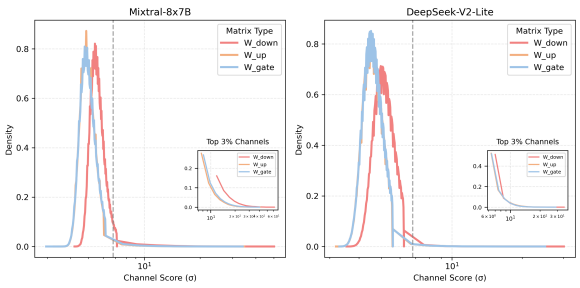
$$E_i(x) = W_d^{(i)}(\text{Act}(W_g^{(i)}x) \odot W_u^{(i)}x), \tag{2}$$

where $\text{Act}(\cdot)$ denotes the activation function, and $\odot$ represents element-wise multiplication. The expert parameters comprise up-projection matrices $W_u^{(i)}, W^{(i)}g \in \mathbb{R}^{d\text{mid}\times d\text{h}}$ and a down-projection matrix $W_d^{(i)} \in \mathbb{R}^{d_{\text{h}}\times d_{\text{mid}}}$, where d_{mid} and d_{h} represent the intermediate and hidden state dimensions, respectively.

Quantization. For a given weight matrix W, the b-bit quantization operation Q_b (Nagel et al., 2021)



(a) Outlier Proportion Across Projection Matrices



(b) L2 Norm Distribution of Outlier Channels

Figure 2: Channel-wise Distribution Analysis across MoE Architectures: (a) Outlier Proportion Across Projection Matrices reveals that outlier channels constitute only a small fraction of total channels, with down-projection matrices (W_{down}) consistently showing higher outlier proportions; (b) L2 Norm Distribution of Outlier Channels demonstrates a pronounced long-tail distribution, where outlier channels exhibit L2 norms orders of magnitude larger than typical channels

is formulated as:

$$Q_b(W) = s \cdot \text{clamp}\Big(\frac{W}{s}\Big), \quad \text{where} \quad s = \frac{\max(|W|) - \min(|W|}{2^b - 1}. \tag{3}$$

Here, s denotes the quantization step size, and clamp$(\cdot)$ represents a function that constrains values within a specified range. The quantization error, denoted as $\epsilon_b(W) = W - Q_b(W)$, exhibits specific statistical properties.

Fisher Information Matrix. In the context of neural networks with parameters W and loss function L, the Fisher Information Matrix is characterized by:

$$F_W = \mathbb{E}\Big[\Big(\frac{\partial L}{\partial W}\Big)\Big(\frac{\partial L}{\partial W}\Big)^T\Big], \tag{4}$$

where the expectation is computed over the underlying data distribution. This matrix provides a principled measure of the local curvature in the loss landscape and quantifies the relative significance of individual parameters within the network architecture.

4 Method

4.1 Channel-wise Mixed-precision Quantization

Extensive empirical analysis of weight matrices in MoE architectures reveals a distinctive structural characteristic: the distribution of outliers exhibits pronounced channel-wise heterogeneity. Our investigation of W_{up}, W_{gate}, and W_{down} matrices across diverse expert networks demonstrates that significant outliers are consistently concentrated within a small subset of channels, while the majority maintain notably uniform distributions. As visualized in Figure 2, these channel-wise outlier patterns exhibit substantial variation, suggesting that conventional layer-wise or tensor-wise quantization approaches, which impose uniform quantization across all channels, are inherently suboptimal for MoE architectures.

Motivated by these empirical observations, we introduce a channel-wise mixed-precision quantization framework. For a weight matrix $W \in \mathbb{R}^{m \times n}$ within expert E_i, our approach independently quantizes each row vector w_j using bit-widths selected from a predefined set $\mathcal{B} = b_1, b_2, ..., b_k$. The quantization operation for channel j with assigned bit-width $b_j \in \mathcal{B}$ is formulated as:

$$Q_{b_j}(w_j) = s_j \cdot \text{round}(\frac{w_j}{s_j}), \tag{5}$$

where the channel-specific quantization step size s_j is computed as:

$$s_j = \frac{\max(|w_j|) - \min(|w_j|)}{2^{b_j} - 1}. \tag{6}$$

This channel-wise quantization paradigm offers significant theoretical and practical advantages. By allocating higher bit-widths to channels containing substantial outliers, the framework preserves the model's critical representational capacity in numerically sensitive regions. Simultaneously, channels exhibiting more uniform distributions can be efficiently quantized with lower bit-widths while maintaining numerical stability, leveraging their inherent robustness to quantization effects. This adaptive precision allocation enables optimal balance between model fidelity and memory efficiency

4.2 Theoretical Analysis of Layer-wise Quantization Sensitivity

A fundamental challenge in quantizing MoE architectures lies in understanding the differential sensitivity of various layer types to precision reduction. Our empirical investigations reveal a striking pattern: down-projection weights W_d consistently exhibit substantially higher magnitude and frequency of outliers compared to up-projection (W_u) and gating (W_g) weights across different experts. This systematic variation in weight distributions suggests an inherent asymmetry in the quantization requirements across different components of MoE architectures.

To rigorously characterize this empirical observation, we establish a theoretical framework that elucidates component-wise quantization sensitivity in MoE architectures. Our analysis synthesizes three fundamental aspects: the relationship between quantization perturbations and loss function dynamics through Fisher Information, the propagation patterns of quantization errors in MLP structures, and the derivation of optimal bit allocation strategies.

Proposition 4.1 (Loss Function Taylor Expansion)**.** *For sufficiently small perturbations ϵ in the parameter space W, the loss function L admits the following second-order approximation:*

$$\begin{aligned} L(W+\epsilon) =& L(W) + tr\left(\frac{\partial L}{\partial W}\epsilon^T\right) + \frac{1}{2}tr\left(\epsilon H_W \epsilon^T\right) \\ &+ O(|\epsilon|^3), \end{aligned} \tag{7}$$

where H_W denotes the Hessian matrix of L with respect to W.

Lemma 4.2 (Fisher Information and Quantization)**.** *In practical neural network deployments where the loss landscape exhibits local convexity and quantization perturbations remain within the regime of Taylor approximation validity, the expected loss increase due to quantization necessarily satisfies:*

$$\mathbb{E}[\Delta L] \geq tr\left(F_W \epsilon_b(W)\epsilon_b(W)^T\right). \tag{8}$$

Proof. Leveraging the Taylor expansion from the previous proposition:

$$\begin{aligned} \Delta L &= L\big(W+\epsilon_b(W)\big) - L(W) \\ &= \tfrac{1}{2}\mathrm{tr}\big(\epsilon_b(W), H_W, \epsilon_b(W)^T\big) \\ &\quad + \mathrm{tr}\big(\tfrac{\partial L}{\partial W}, \epsilon_b(W)^T\big) + O\big(|\epsilon_b(W)|^3\big). \end{aligned} \tag{9}$$

The result follows naturally by taking expectation and applying the definition of the Fisher Information Matrix:

$$\begin{aligned} \mathbb{E}[\Delta L] &\geq \mathbb{E}\Big[\mathrm{tr}\big(\tfrac{\partial L}{\partial W}, \epsilon_b(W)^T\big)\Big] \\ &= \mathrm{tr}\Big(\mathbb{E}\big[\tfrac{\partial L}{\partial W}, \tfrac{\partial L}{\partial W}^T\big], \epsilon_b(W), \epsilon_b(W)^T\Big) \\ &= \mathrm{tr}\big(F_W, \epsilon_b(W), \epsilon_b(W)^T\big). \end{aligned} \tag{10}$$

□

To analyze the quantization sensitivity in FFN structures, we examine the error propagation through the projection layers. Taking the W_{down} and W_{up} matrices as representative examples, we establish the following characterization of quantization-induced errors.

Lemma 4.3 (Error Propagation in FFN Layers)**.** *For the standard FFN architecture incorporating up-projection matrix $W_u \in \mathbb{R}^{d_h \times d}$ and down-projection matrix $W_d \in \mathbb{R}^{d \times d_h}$ with $d_h > d$, the quantization error propagation manifests as:*

$$\begin{aligned} |\Delta y|_2 \leq& \kappa(W_d)|\epsilon_b(W_d)|_2|x|_2 \\ &+ |W_d|_2|\epsilon_b(W_u)|_2|x|_2, \end{aligned} \tag{11}$$

where $\kappa(W_d) = |W_d|_2|W_d^+|_2$ represents the condition number of W_d.

Proof. The quantization-induced output perturbation naturally decomposes as:

$$\begin{aligned} \Delta y &= (W_d + \epsilon_b(W_d))(W_u + \epsilon_b(W_u))x \\ &\quad - W_d W_u x \\ &= W_d\epsilon_b(W_u)x + \epsilon_b(W_d)W_u x \\ &\quad + \epsilon_b(W_d)\epsilon_b(W_u)x. \end{aligned} \tag{12}$$

Application of the triangle inequality yields:

$$\begin{aligned} |\Delta y|_2 &\leq |W_d\epsilon_b(W_u)x|_2 + |\epsilon_b(W_d)W_u x|_2 \\ &\quad + |\epsilon_b(W_d)\epsilon_b(W_u)x|_2 \\ &\leq |\epsilon_b(W_d)|_2|W_u|_2|x|_2 \\ &\quad + |W_d|_2|\epsilon_b(W_u)|_2|x|_2 + O(|\epsilon_b(W)|^2). \end{aligned} \tag{13}$$

The stated inequality follows from the architectural constraint $|W_u|_2 \leq \kappa(W_d)$. □

Theorem 4.4 (Optimal Bit Allocation)**.** *In typical neural network deployments where input distributions exhibit bounded moments and weight matrices maintain Lipschitz continuity, the optimal bit allocation ratio r^* between $Wdown$ and W_{up} necessarily satisfies:*

$$r^* = \frac{b_d}{b_u} \geq \sqrt{\frac{\kappa(W_d)}{\kappa(W_u)}} \cdot \frac{|FW_d|F}{|FW_u|_F}. \tag{14}$$

The detailed proof can be found in Appendix A.1.

Corollary 4.5. *Under contemporary neural network architectures employing standard initialization schemes, the following relationships naturally emerge:*

$$\kappa(W_d) > \kappa(W_u) \quad \textit{and} \quad |F_{W_d}|F > |FW_u|F, \tag{15}$$

These inequalities, in conjunction with Theorem 4.4, establish that the optimal bit ratio r^* between $Wdown$ and W_{up} necessarily exceeds unity, indicating the imperative for higher bit allocation to the down-projection layer. This result derives from dimensional analysis and established initialization properties, providing theoretical guidance for bit allocation in mixed-precision quantization. The detailed proof is provided in Appendix A.2.

These theoretical findings provide rigorous justification for the empirically observed sensitivity of the down-projection layer. The analysis reveals that this sensitivity emerges from two fundamental factors: (1) the inherent dimensional compression in the down-projection operation, which engenders heightened sensitivity to quantization errors, and (2) its position in the computational graph, where perturbations can be amplified through the network structure. These insights naturally motivate our proposed bit allocation strategy that assigns higher precision to the down-projection layer to preserve model performance..

4.3 Theoretically-Guided Mixed-Precision Quantization Framework

Building on our layer-wise sensitivity analysis and channel heterogeneity characterization, we propose a mixed-precision quantization framework that bridges theoretical insights with deployment efficiency. Our method establishes dual-granularity bit allocation through theoretically-grounded sensitivity metrics.

Channel-wise Theoretical Sensitivity Metric

From Theorem 4.4 connecting quantization sensitivity with Fisher Information F and condition numbers κ, we derive a channel-level sensitivity metric:

$$S_c = \alpha \cdot \underbrace{\text{tr}(F_c)\kappa_c}_{T_c} + (1-\alpha) \cdot O_c \tag{16}$$

where $\alpha \in [0,1]$ regulates the theoretical-empirical trade-off. The theoretical component T_c combines the cumulative gradient variance captured by the Fisher Information trace $\text{tr}(F_c)$ with the numerical stability quantified through the condition number κ_c, which governs parameter perturbation sensitivity. The empirical component O_c implements our channel-wise distribution analysis in Section 4.1 by statistically quantifying quantization-critical outliers. This data-driven mechanism through an integrated analysis of the magnitude and frequency characteristics of outliers, thereby providing essential empirical compensation to theoretical sensitivity estimates.

Algorithm 1 Mixed-precision Quantization Framework for MoE

Require: Model parameters Θ, bit budget B, outlier threshold τ
Ensure: Quantized model parameters Θ_Q

```
for each Transformer block T_b do
  for each Expert E_i in T_b do
    for each MoE linear layer W ∈ W_u, W_g, W_d do
      for each channel c in W do
        K_c ← |w_c|_2 / min_j |w_j|_2
        V_c ← var(wc)/mean(var(w:))
        O_c ← |wij ∈ wc : |wij| > τ|/|w_c|
        S_c ← α · (K_c · V_c) + (1 − α) · O_c
      end for
    end for
    κ_d ← max_i |w_i|_2 / min_i |w_i|_2 for W_d
    κ_u ← max_i |w_i|_2 / min_i |w_i|_2 for W_u
    V_d ← var(W_d)
    V_u ← var(W_u)
    β_i ← sqrt(κ_d/κ_u) · sqrt(V_d/V_u)
    Allocate bits: b_d ← β_i · b_u
  end for
end for
Apply channel-wise quantization using computed sensitivity scores
Return Θ_Q
```

Computationally Efficient Approximation

Direct computation of $\text{tr}(F_c)$ requires $O(N_{params} \cdot N_{samples})$ complexity due to gradient propagation. We develop theoretically-justified approximations:

$$\begin{aligned} \text{tr}(F_c) &\approx V_c = \frac{\text{var}(\mathbf{w}c)}{\text{mean}(\text{var}(\mathbf{w}:))} \\ \kappa_c &\approx K_c = \frac{|\mathbf{w}_c|_2}{\min_j |\mathbf{w}_j|_2} \end{aligned} \tag{17}$$

The variance approximation V_c derives from maximum entropy principles under second-

#BITS	GRANULARITY	METHOD	ACCURACY (%) ↑						
			BOOLQ	PIQA	HELLAS.	WINO.	ARC_E	ARC_C	AVG.
FP16	-	-	85.02	82.59	83.99	76.56	84.13	56.83	78.18
2.86	EXPERT	FREQUENCY	79.36	78.62	74.91	70.88	76.85	45.82	71.07
	METRIC	W-OUTLIER	79.42	79.98	77.65	72.69	78.54	48.63	72.82
	CHANNEL	W-OUTLIER	**82.02**	80.14	78.18	71.11	77.78	49.23	73.10
	CHANNEL	**OURS**	81.53	**80.69**	**78.57**	**71.69**	**79.8**	**50.43**	**73.79**
2.68	METRIC	W-OUTLIER	78.07	78.45	76.25	70.32	76.3	48.12	71.25
	CHANNEL	W-OUTLIER	**81.31**	78.84	77.18	**72.06**	76.18	44.37	71.66
	CHANNEL	**OURS**	80.58	**80.3**	**77.8**	71.11	**78.2**	**48.21**	**72.68**
2.54	EXPERT	FREQUENCY	73.55	78.07	72.51	68.67	72.69	42.15	67.94
	METRIC	W-OUTLIER	74.65	78.07	74.17	71.43	73.15	42.75	69.04
	CHANNEL	W-OUTLIER	76.36	78.24	75.14	**72.06**	75.67	44.8	70.38
	CHANNEL	**OURS**	**80.49**	**79.27**	**76.58**	71.9	**77.57**	**46.76**	**72.10**
2.30	METRIC	W-OUTLIER	71.59	76.71	72.14	68.35	72.81	42.66	67.38
	CHANNEL	W-OUTLIER	76.3	76.93	73.79	69.53	73.19	**41.98**	68.62
	CHANNEL	**OURS**	**79.6**	**78.67**	**75.1**	**70.8**	**73.78**	41.64	**69.93**
2.20	EXPERT	FREQUENCY	69.27	77.31	70.79	69.3	71.3	39.51	66.25
	METRIC	W-OUTLIER	66.27	76.55	70.22	66.69	70.12	39.51	64.89
	CHANNEL	W-OUTLIER	72.45	**77.86**	72.97	70.17	73.32	42.92	68.28
	CHANNEL	**OURS**	**74.31**	77.69	**73.54**	**71.03**	**73.65**	**42.92**	**68.85**

Table 1: Zero-Shot Task Performance of Mixtral-8x7B using Automated Fine-Grained MoE Quantization. **#Bits** denotes bits for weight quantization, **Granularity** represents the level of mixed-precision application, and **Method** indicates the bit allocation strategy. "HellaS." is the short format of "HellaSwag" and "Wino." denotes "Winogrande".

moment constraints, while K_c preserves spectral stability relationships(see Appendix B for detailed derivations). This reduces computational complexity from $O(N^2)$ to $O(N)$. The final implementable metric becomes:

$$S_c = \alpha(K_c V_c) + (1-\alpha)O_c \tag{18}$$

Layer-wise Bit Allocation Strategy

Theorem 4.4 establishes the optimal bit ratio between projection layers:

$$\frac{b_d}{b_u} \geq \sqrt{\frac{\kappa(W_d)}{\kappa(W_u)}} \cdot \frac{|F_{W_d}|F}{|FW_u|_F} \tag{19}$$

We implement this through practical approximations preserving theoretical guarantees:

$$\beta_i = \sqrt{\frac{\kappa_d V_d}{\kappa_u V_u}} \tag{20}$$

where $\kappa_l = \max_j |\mathbf{w}_j|_2 / \min_j |\mathbf{w}_j|_2$ captures numerical stability and V_l approximates Fisher norm $|F_l|_F$. The allocation respects per-expert constraints:

$$b_d^i + b_u^i + b_g^i = B^i \tag{21}$$

where b_d^i, b_u^i, and b_g^i represent the bits allocated to $Wdown$, W_{up}, W_{gate} respectively within expert i, and B^i denotes the total bit budget for that expert.

The algorithm presented in Algorithm 1 integrates theoretical insights with practical efficiency considerations, operating at both the layer and channel granularities while maintaining minimal computational overhead through statistical approximations.

5 Experiments

5.1 Settings

Implementation Details. We implement our method following standard LLM quantization practices (Frantar et al., 2023; Lin et al., 2023; Shao et al., 2023; Ma et al., 2024), using a WikiText2 (Merity et al., 2016) calibration set (128 samples × 2048 tokens) and asymmetric group quantization (group size 128). Attention layers are quantized to 4-bit while maintaining full-precision routers. For FFN layer mixed-precision quantization, we set sensitivity balance coefficient α to 0.5 and employ hardware-friendly 2-bit and 4-bit precision levels. To ensure fair comparison across granularity levels (Li et al., 2024), we maintain consistent ratios of 2-bit and 4-bit parameters while controlling the

#Bits	Granularity	Method	Accuracy (%) ↑					
			PIQA	HELLAS.	WINO.	ARC_E	ARC_C	Avg.
FP16	-	-	80.14	77.28	69.22	73.19	41.64	68.294
2.54	LINEAR	W-OUTLIER	74.32	67.88	63.8	69.2	37.57	63.06
	CHANNEL	W-OUTLIER	75.3	68.13	64.33	69.4	**38.74**	63.71
	CHANNEL	OURS	**75.46**	**68.68**	**64.64**	**69.49**	38.71	**63.94**
2.30	LINEAR	W-OUTLIER	73.99	67.42	62.98	67.61	36.52	62.2
	CHANNEL	W-OUTLIER	75.57	67.72	**63.38**	68.22	36.95	62.76
	CHANNEL	OURS	**75.21**	**68.89**	63.22	**70.22**	**38.95**	**63.79**
2.20	LINEAR	W-OUTLIER	73.88	66.32	61.88	64.9	35.41	60.66
	CHANNEL	W-OUTLIER	73.94	67.2	62.67	65.7	36.69	61.75
	CHANNEL	OURS	**74.16**	**67.54**	**62.98**	**69.07**	**38.31**	**62.88**

Table 2: Zero-Shot Task Performance of DeepSeek-V2-Lite using Automated Fine-Grained MoE Quantization. **#Bits** denotes bits for weight quantization, **Granularity** represents the level of mixed-precision application, and **Method** indicates the bit allocation strategy. "HellaS." is the short format of "HellaSwag" and "Wino." denotes "Winogrande".

overall parameter budget.

Baseline. All experiments were conducted on the Mixtral-8x7B-v0.1 (Jiang et al., 2024), DeepSeek-V2-Lite(Liu et al., 2024a) and Qwen-1.5-MoE(Team, 2024). Our method successfully quantizes this multi-billion parameter model, with all GPTQ (Frantar et al., 2023) experiments completed on a single NVIDIA RTX 3090 GPU with 24GB memory .

Evaluation. To assess the efficacy of our automated fine-grained quantization approach for MoE models, we conducted evaluations across five zero-shot tasks: PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2019), BoolQ (Clark et al., 2019), and ARC (Clark et al., 2018). We utilized the lm-eval-harness (Gao et al., 2021) framework to obtain individual task accuracies and the overall average accuracy. This comprehensive evaluation strategy enables us to gauge the impact of our quantization method on various aspects of model performance, providing insights into its generalization capabilities across diverse natural language understanding tasks.

5.2 Main Results

We conduct comprehensive zero-shot evaluations across various quantization configurations, including different bit-widths, granularity levels, and priority allocation strategies. Our primary baseline is derived from the extensive mixed-precision quantization experiments for MoE models presented in (Li et al., 2024). Their work explores two granularity levels: expert-level allocation, which distributes bits among experts within the same block, and metric-level allocation, which operates across all matrices (W_{down}, W_{up}, W_{gate}) throughout the model. For allocation strategies, they investigate frequency-based (**frequency**) prioritization according to router activation patterns and weight-outlier (**w-outlier**) based ordering. The specific bit-width assignment is determined by the mixed-precision ratio - for instance, a 2.54-bit configuration indicates that 25% of the FFN parameters are quantized to 4-bit precision, with the remaining parameters at 2-bit precision.

The experimental results in Table 1, Table 2 and Table 3 demonstrate the consistent superiority of our approach across various bit-width configurations. Our method achieves the highest average zero-shot accuracy under all bit-width settings, with particularly remarkable performance in lower bit. Notably, at 2.20 bits, our approach outperforms the best metric-level method by 3.96% (68.85% vs. 64.89%) and surpasses the frequency-based expert-level approach by 2.6% (68.85% vs. 66.25%) in accuracy, highlighting its competitive advantage in aggressive quantization scenarios. Furthermore, the effectiveness of channel-level granularity is evident even when using the same weight-outlier parameter selection strategy, consistently outperforming metric-level quantization across all configurations. This improvement suggests that finer-grained mixed-precision allocation enables more precise capture of parameter sensitivity and better preservation of model capabilities under strict bit-width constraints.

5.3 Ablation Study

Sensitivity Analysis of FFN Projections To validate Corollary 4.5 empirically, we conduct an ablation study by quantizing either the down-projection or up-projection layer while maintaining full precision for all other layers. As shown in Figure 3 We measure the mean squared error (MSE) between the original and quantized outputs across different transformer blocks. The results consistently show that down-projection layers exhibit higher quantization sensitivity across all transformer blocks, necessitating more bits in mixed-precision quantization. This experimental observation aligns with our theoretical analysis in Corollary 4.5, providing empirical support for our sensitivity-aware bit allocation strategy.

Analysis of Balance Parameter α We conduct an ablation study on hyperparameter α at 2.54-bit quantization level, varying it from 0.2 to 0.8. As shown in Figure 4, model performance peaks at $\alpha \approx 0.5$, indicating that balancing weight magnitude and gradient information is crucial for effective bit allocation in mixed-precision quantization. Lower α values overemphasize weight magnitudes while higher values excessively prioritize gradient information, both leading to suboptimal quantization results.

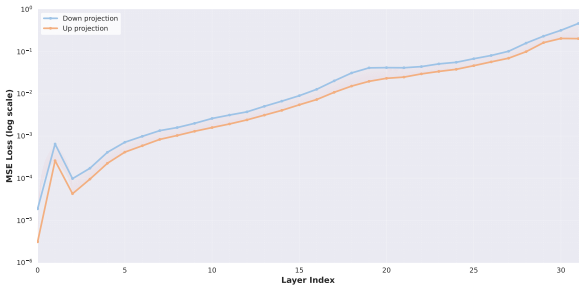


Figure 3: Layer-wise Mean Squared Error (MSE) comparison between down-projection and up-projection quantization, demonstrating consistently higher sensitivity in down-projection layers across transformer blocks.

Analysis of Channel Quantization Direction We investigate the impact of channel direction choice in quantization by comparing row-wise and column-wise quantization strategies, while maintaining the quantization level at 2.54-bit and using weightoutlier as the allocation criterion. Our experiments in Table 4 demonstrate that row-wise quantization consistently achieves superior performance. This finding aligns with the inherent function of FFN layers, as each row vector corresponds to an output dimension that represents a distinct feature transformation. By quantizing along the row direction, we preserve the integrity of individual feature transformations, allowing each output dimension to maintain its unique contribution to the model's representation capacity. This result suggests that preserving the precision of output feature computations is more critical than maintaining the precision of input feature combinations.

6 Conclusion

This paper presents an automated fine-grained quantization framework for MoE models. Our key theoretical contribution establishes the relationship between Fisher Information, condition numbers, and quantization sensitivity, providing principled guidance for bit allocation across different model components. Through extensive empirical analysis, we demonstrate that MoE weight matrices exhibit distinct channel-wise outlier patterns, necessitating fine-grained quantization approaches. Our proposed framework successfully bridges theoretical insights with practical efficiency considerations through statistical approximations, enabling automated bit allocation without additional training overhead.

Limitations

While our work demonstrates promising results in MoE quantization, several limitations warrant discussion. First, our experimental validation, although thorough on Mixtral-8x7B, DeepSeek-V2 and Qwen-1.5-MoE, is constrained to a limited set of MoE architectures. Despite the block-wise quantization approach being computationally feasible on a single GPU, resource constraints prevented us from extending our evaluation to emerging models such as DeepSeek-V3, potentially limiting our understanding of the framework's generalizability across different MoE architectures. Additionally, while our work focuses specifically on quantization-based compression, it does not explore potential synergies with other compression techniques such as pruning, knowledge distillation, or structured sparsification. Future work could investigate the integration of our fine-grained quantization framework with these complementary compression methods, potentially yielding more comprehensive and efficient MoE compression solutions.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *arXiv preprint arXiv:2404.00456*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Hong Chen, Chengtao Lv, Liang Ding, Haotong Qin, Xiabin Zhou, Yifu Ding, Xuebo Liu, Min Zhang, Jinyang Guo, Xianglong Liu, and 1 others. 2024. Db-llm: Accurate dual-binarization for efficient llms. *arXiv preprint arXiv:2402.11960*.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. 2022. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, and 1 others. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Shaojin Ding, Phoenix Meadowlark, Yanzhang He, Lukasz Lew, Shivani Agrawal, and Oleg Rybakov. 2022. 4-bit conformer with native quantization aware training for speech recognition. *arXiv preprint arXiv:2203.15952*.

Jun Fang, Ali Shafiee, Hamzah Abdel-Aziz, David Thorsley, Georgios Georgiadis, and Joseph H Hassoun. 2020. Post-training piecewise linear quantization for deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 69–86. Springer.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. *International Conference on Learning Representations*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training compression for generative pretrained transformers. In *International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, and 1 others. 2021. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Shwai He, Daize Dong, Liang Ding, and Ang Li. 2024. Demystifying the compression of mixture-of-experts through a unified framework. *arXiv preprint arXiv:2406.02500*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

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, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. 2023. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Pingzhi Li, Xiaolong Jin, Yu Cheng, and Tianlong Chen. 2024. Examining post-training quantization for mixture-of-experts: A benchmark. *arXiv preprint arXiv:2406.08155*.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. 2023. Merge, then compress: Demystify efficient smoe with hints from its routing policy. *arXiv preprint arXiv:2310.01334*.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. 2021. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*.

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

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024b. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2024c. Spinquant–llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. 2021. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103.

Yuexiao Ma, Huixia Li, Xiawu Zheng, Feng Ling, Xuefeng Xiao, Rui Wang, Shilei Wen, Fei Chao, and Rongrong Ji. 2024. Affinequant: Affine transformation quantization for large language models. *arXiv preprint arXiv:2403.12544*.

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

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. 2020. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pages 7197–7206. PMLR.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. 2021. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. 2022. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pages 18332–18346. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2023. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, and 1 others. 2024. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*.

Shyam A Tailor, Javier Fernandez-Marques, and Nicholas D Lane. 2020. Degree-quant: Quantization-aware training for graph neural networks. *arXiv preprint arXiv:2008.05000*.

Qwen Team. 2024. Qwen1.5-moe: Matching 7b model performance with 1/3 activated parameters".

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. 2022a. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*.

Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. 2022b. Outlier suppression: Pushing the limit of low-bit transformer language models. In *Advances in Neural Information Processing Systems*.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*.

Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. 2023. Rptq: Reorder-based post-training quantization for large language models. *arXiv preprint arXiv:2304.01089*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *CoRR*, abs/1905.07830.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendix A: Detailed Proofs for Optimal Bit Allocation

A.1 Proof of Theorem 4.4

Proof. The proof consists of three main steps:
Step 1: Error Propagation. As we have previously derived in Equation 12, from matrix perturbation analysis, we have:

$$\begin{aligned}\Delta y &= (W_d + \epsilon_b(W_d))(W_u + \epsilon_b(W_u))x \\ &\quad - W_d W_u x \\ &= W_d \epsilon_b(W_u)x + \epsilon_b(W_d)W_u x \\ &\quad + \epsilon_b(W_d)\epsilon_b(W_u)x.\end{aligned} \tag{22}$$

Step 2: Bounding the Expected Loss Increase. Using Lemma 4.2 and 4.3, we can bound the expected loss increase. First, note that for uniformly distributed quantization error, we have:

$$\mathbb{E}[\epsilon_b(W)\epsilon_b(W)^T] = \frac{1}{2^{2b}}\text{diag}(\text{var}(W)) \tag{23}$$

For W_d:

$$\begin{aligned}\mathbb{E}[\Delta L_d] &\geq \text{tr}(F_{W_d}\epsilon_b(W_d)\epsilon_b(W_d)^T) \\ &= \text{tr}(F_{W_d}\mathbb{E}[\epsilon_b(W_d)\epsilon_b(W_d)^T]) \\ &= \frac{1}{2^{2b_d}}\text{tr}(F_{W_d}\text{diag}(\text{var}(W_d))) \\ &= \frac{|F_{W_d}|_F}{2^{2b_d}}\end{aligned} \tag{24}$$

Similarly for W_u:

$$\mathbb{E}[\Delta L_u] \geq \text{tr}(F_{W_u}\epsilon_b(W_u)\epsilon_b(W_u)^T) = \frac{|F_{W_u}|_F}{2^{2b_u}} \tag{25}$$

By the independence of quantization operations between layers and the additivity of expectation:

$$\begin{aligned}\mathbb{E}[\Delta L] &= \mathbb{E}[\Delta L_d] + \mathbb{E}[\Delta L_u] \\ &\geq \frac{|F_{W_d}|F}{2^{2b_d}} + \frac{|FW_u|_F}{2^{2b_u}}\end{aligned} \tag{26}$$

□

A.2 Proof of Corollary 4.5

Proof. **Step 1: Dimensional Analysis.** Recall $W_d \in \mathbb{R}^{d\times d_h}$ and $W_u \in \mathbb{R}^{d_h\times d}$, where $d_h > d$.
Step 2: Xavier Initialization. Under Xavier initialization, weights are typically sampled as:

$$W_{ij} \sim \mathcal{N}\left(0, \frac{2}{n_{\text{in}} + n_{\text{out}}}\right). \tag{27}$$

Step 3: Condition Number Comparison. For W_d, this leads to:

$$\kappa(W_d) \approx \frac{\sqrt{d_h}}{d}\frac{\max\sigma_i}{\min\sigma_i}, \tag{28}$$

while for W_u:

$$\kappa(W_u) \approx \frac{\sqrt{d}}{d_h}\frac{\max\sigma_i}{\min\sigma_i}, \tag{29}$$

so $\kappa(W_d) > \kappa(W_u)$ since $d_h > d$.
Step 4: Fisher Information Comparison. Due to the bottleneck from d_h to d, W_d compresses information more aggressively, leading to:

$$|F_{W_d}|F > |FW_u|_F. \tag{30}$$

Hence, both terms in the inequality of Theorem 4.4 exceed 1, implying $r^* = \frac{b_d}{b_u} > 1$. □

B Appendix B: Approximation Derivation

The approximations are derived from fundamental theorems in optimization and matrix theory:

B.1 Fisher Information Trace Approximation

1. Starting with the exact definition:
$$\text{tr}(F_c) = \mathbb{E}[\|\partial L/\partial w_c\|^2]$$
2. Applying optimization theory and Taylor expansion near optimality:
$$\partial L/\partial w_c = \delta_c \otimes x \approx H_c(w_c - w_c^*)$$
where H_c is the Hessian and w_c^* are optimal parameters.
3. This leads to the key approximation:
$$\begin{aligned}\text{tr}(F_c) &= \mathbb{E}[\|H_c(w_c - w_c^*)\|^2] \\ &= \mathbb{E}[\delta_c^2]\mathbb{E}[\|x\|^2] \propto \text{var}(w_c)\end{aligned} \tag{31}$$
4. The normalization term emerges naturally: $\text{mean}(\text{var}(w_:))$ ensures scale invariance across layers

B.2 Condition Number Approximation

1. From matrix perturbation theory:
$$\kappa_c = \sigma_{\max}(w_c)/\sigma_{\min}(w_c)$$
2. For channel vectors, this simplifies to:
$$\kappa_c = \|w_c\|_2 / \min_j \|w_j\|_2$$
preserving the relative scaling relationships between channels.

C Other expertment results

#Bits	Granularity	Method	Accuracy (%) ↑					
			PIQA	HELLAS.	WINO.	ARC_E	ARC_C	Avg.
FP16	-	-	80.14	77.28	69.22	73.19	41.64	**68.294**
2.69	LINEAR	W-OUTLIER	76.71	72.60	65.43	59.60	32.34	61.336
	CHANNEL	W-OUTLIER	77.48	73.67	62.59	63.49	35.41	62.528
	CHANNEL	OURS	**77.55**	**74.13**	**65.51**	**64.95**	**35.84**	**63.596**
2.58	LINEAR	W-OUTLIER	77.53	72.77	62.12	61.07	31.23	60.944
	CHANNEL	W-OUTLIER	76.33	72.78	63.33	61.62	34.39	61.690
	CHANNEL	OURS	**77.84**	**73.24**	**63.38**	**63.56**	**33.36**	**62.276**
2.47	LINEAR	W-OUTLIER	75.90	72.60	62.12	61.871	31.28	60.7542
	CHANNEL	W-OUTLIER	75.97	72.46	61.56	62.09	32.79	60.974
	CHANNEL	OURS	**75.97**	**73.21**	**62.19**	**62.18**	**33.59**	**61.428**

Table 3: Zero-Shot Task Performance of Qwen-1.5-MoE using Automated Fine-Grained MoE Quantization. **#Bits** denotes bits for weight quantization, **Granularity** represents the level of mixed-precision application, and **Method** indicates the bit allocation strategy. "HellaS." is the short format of "HellaSwag" and "Wino." denotes "Winogrande".

Direction	Accuracy (%) ↑					
	PIQA	HELLAS.	WINO.	ARC_E	ARC_C	Avg.
Down Row	76.55	73.52	69.61	75.34	43.86	67.78
Down Column	**78.84**	75.25	68.98	74.75	44.2	68.40
All Column	78.13	**75.48**	71.51	75.38	44.28	68.95
All Row	78.24	75.14	**72.06**	**75.67**	**44.8**	**69.19**

Table 4: Performance comparison of different channel quantization directions in FFN layers of Mixtral-8x7B-v0.1. All Row represents using row-wise quantization for all channel quantization matrices, Down Row represents row-wise quantization for W_{down} only, column-wise for W_{up} and W_{gate}. "HellaS." is the short format of "HellaSwag" and "Wino." denotes "Winogrande".

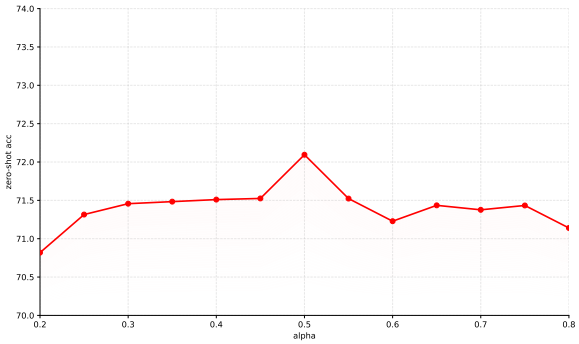


Figure 4: Impact of balance parameter α on zero-shot accuracy under 2.54-bit quantization, demonstrating optimal performance at α nearby 0.5.