OWEN-VALUE-GUIDED MIXED-PRECISION QUANTIZATION FOR MIXTURE-OF-EXPERTS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-experts (MoE) models activate only a few experts per token, yet low-latency inference can require the entire expert pool in memory. At ultra-low average precision, mixed-precision quantization must decide which experts retain scarce higher-bit slots. Individual sensitivity scores can miss the joint effects of co-routed experts. We propose **Owen-MoE**, which values experts in an availability game while accounting for an output-similarity partition. Two-level Monte Carlo sampling estimates Owen values in $O(TK)$ evaluations per layer for T trials over K experts. A scale-invariant Influence Ratio converts these values into a static $\{2, 4\}$-bit map that can protect locally influential experts within lower-value coalitions. We compare rankings under identical per-layer 4-bit counts and AutoRound settings. Across eight zero-shot benchmarks on Qwen3-30B-A3B, three-seed means at 2.75–2.125 average expert bits, Owen-MoE exceeds Router-norm by 3.57–7.02 points and PMQ by 4.86–10.56 points on average. It also exceeds measured fractional Uniform by 2.78–13.31 points at matched expert-weight payload. Equal-budget controls show an additional gain over flat Shapley of 1.78 points at 2.50 bits and 2.58 points at 2.25 bits. In 96 selected high-coactivation expert pairs where the rankings disagree, Owen's choice reduces round-to-nearest quantization MSE in 78.1% of cases, with a median reduction of 17.6%. Mixtral and Qwen1.5-MoE extend the evaluation to different expert counts and routing configurations. We will release the complete code upon acceptance.

1 INTRODUCTION

Mixture-of-Experts (MoE) architectures (Lepikhin et al., 2021; Fedus et al., 2022; Dai et al., 2024; DeepSeek-AI, 2024a; Yang et al., 2025; Muennighoff et al., 2024) scale parameter count while keeping per-token FLOPs bounded, but they do not decouple memory: low-latency inference places pressure to keep potentially routed experts in VRAM. Qwen3-30B-A3B (Yang et al., 2025) holds 30.5B parameters, approximately 61 GB in BF16, far beyond a single edge accelerator, even though each token activates only 8 of the 128 experts in a layer. This motivates ultra-low-bit mixed-precision quantization for MoE deployment. For the integer weight formats studied here, intermediate average precision requires mixing bit-widths across blocks. The allocation must identify which experts benefit most from higher precision.

Weight-only post-training quantization (PTQ) methods such as GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2023) compress dense Transformers well at 3–4 bits. Subsequent methods address outliers, dense-and-sparse weights, learned transforms, channel reassembly, vector quantization, and salient weights (Xiao et al., 2023; Kim et al., 2023a; Shao et al., 2024; Ma et al., 2024; Liu et al., 2024a; Dettmers et al., 2024; Liu et al., 2024b; Yuan et al., 2024). Uniform precision can degrade sharply below 3 bits on MoE models. Experts can differ in sensitivity, so a common bit-width cannot reflect their relative importance. Expert-level mixed-precision methods address this by allocating a per-expert bit-width from a scalar sensitivity metric such as the router ℓ_2-norm change, activation frequency, or Hessian trace; Hessian-aware assignment was first studied for dense models (Dong et al., 2020). MoE work covers low-bit robustness and benchmarks (Kim et al., 2023b; Li et al., 2024b), expert compression (Huang et al., 2025a; 2026; Chen et al., 2025a), and mixed-precision or expert-aware quantization (Duanmu et al., 2025; Chitty-Venkata et al., 2025; Fu et al., 2025; Chen et al., 2025b; Chu et al., 2025; Chowdhury et al., 2026; Deng et al., 2026; Yang et al., 2026; Zhao

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.

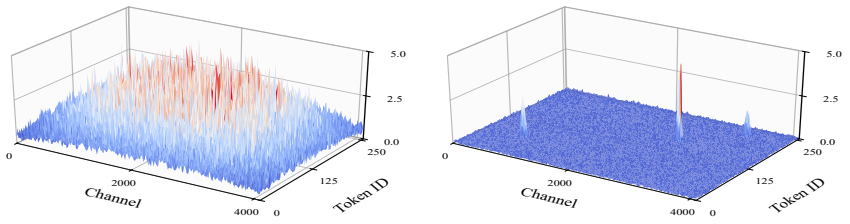

Figure 1: The isolated-evaluation trap on Mixtral. (a) A high-frequency expert activates diffusely. (b) A long-tail local core has large responses on a few tokens. The contrast motivates testing which expert should retain higher precision; response magnitude alone does not determine quantization sensitivity.

et al., 2026; Park et al., 2026). Other approaches merge experts or alter routing (Li et al., 2024a; Zhao et al., 2025; Wang et al., 2025).

These scores can encounter an **isolated-evaluation trap**: expert-wise statistics do not explicitly attribute joint marginal effects. Top-k routing activates experts in groups, and co-routed experts can induce non-zero off-diagonal blocks in the layer's Gauss–Newton matrix. Their joint quantization error can therefore differ from the sum of individual errors. Figure 1 illustrates the distinction between a frequently active expert and a rare **local core** with large responses on a few tokens. Response amplitude motivates our equal-memory quantization interventions, which test which expert should retain higher precision across layers. Importance scores also guide pruning and skipping (Sun et al., 2024; Huang et al., 2025b; Lu et al., 2024; Zhang et al., 2026), with related work on Shapley-style data valuation (Ghorbani & Zou, 2019), calibration sensitivity, and task-based evaluation (Ji et al., 2025; Jaiswal et al., 2024). Here the ranking determines which experts retain higher precision under a fixed per-layer budget.

We address this by treating expert sensitivity as **coalitional** value attribution. The experts of a layer are players in a cooperative game whose characteristic function measures the reduction in calibration language-modeling loss when an expert subset is available. Masked forward passes evaluate this function, while an **a-priori partition** groups experts with similar outputs. Conditional on that partition, the flat Shapley value (Shapley, 1953) and the Owen value (Owen, 1977) are different estimands: Shapley averages over all expert orderings, Owen over partition-consistent ones in which other coalitions enter as whole blocks. The two values attribute interactions in the same game under different ordering distributions; Owen changes how interactions spanning coalitions are allocated (Proposition 2).

The contributions of this paper are summarized as follows. **(1) Structured expert attribution.** We formulate expert importance as an availability game and use an output-similarity partition to define Owen values. A dividend decomposition separates the joint effects that motivate attribution from the higher-order effects through which Owen can differ from Shapley. **(2) Budget-matched quantization.** We estimate Owen values by two-level Monte Carlo shuffling and convert them to a static $\{2, 4\}$-bit map with a scale-invariant Influence Ratio (IR) policy. Baselines receive the same per-layer 4-bit counts, so the comparisons isolate which experts each score protects. **(3) Mechanism diagnostics.** Qwen3 controls isolate the gain over flat Shapley at equal sampling and memory budgets. Cross-layer expert-pair interventions measure the resulting quantization error, while prefix and frequency diagnostics characterize protection choices for rarely routed experts.

Across six Qwen3 budgets, Owen's average advantage over Router-norm and PMQ widens as precision decreases. A measured fractional Uniform comparison establishes the benefit over an index-based allocation, and Mixtral and Qwen1.5-MoE extend the comparison across architectures with different expert counts.

2 Preliminaries and Problem Formulation

MoE layer and the allocation objective. For experts $N = \{1, \dots, K\}$, let $h(x) = \sum_i g_i(x) f_i(x; W_i)$, where x is a layer input, f_i is expert i with weights W_i, and g_i is a sparse

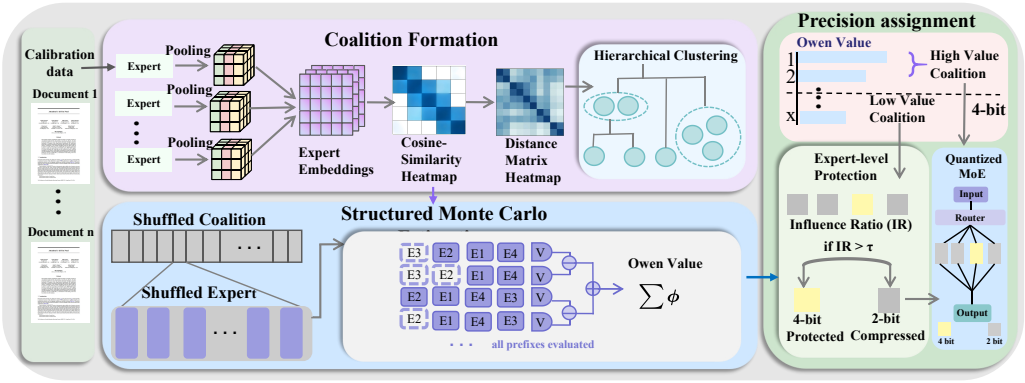

Figure 2: Owen-MoE: cluster calibration outputs, estimate Owen values by two-level shuffling, and protect local cores within low-capacity coalitions using IR. Baselines fill the resulting per-layer 4-bit quotas with their own rankings.

top-k gate. Quantization at bits b_i perturbs weights by $\Delta W_i = \widehat{W}_i - W_i$. The allocation minimizes output distortion under a byte budget:

$$\mathcal{L}(\delta w) = \tfrac{1}{2}\mathbb{E}_x\big\|h(x) - \widehat{h}(x)\big\|_2^2, \qquad \text{s.t.} \quad M(b) = \textstyle\sum_i (b_i/8)\,|W_i| \le \bar{B}, \tag{1}$$

where $\mathbb{E}_x$ averages over layer inputs, $\bar{B}$ is the byte budget, $|W_i|$ the parameter count, $\widehat{h} = \sum_i g_i f_i(x; W_i + \Delta W_i)$, $w_i = \mathrm{vec}(W_i)$, and δw stacks $\delta w_i = \mathrm{vec}(\Delta W_i)$.

Interactions omitted by isolated scores. Conditioning on the full-precision layer input and holding the router logits and Top-k support fixed, a local expansion of (1) at $\delta w = 0$ gives

$$\mathcal{L}(\delta w) = \tfrac{1}{2}\sum_{i,j\in N} \delta w_i^\top G_{ij}\,\delta w_j + o(\|\delta w\|_2^2), \qquad G_{ij} = \mathbb{E}_x\left[g_i(x) g_j(x) J_i(x)^\top J_j(x)\right], \tag{2}$$

where $J_i = \partial f_i / \partial w_i$. At zero perturbation, G is both the positive-semidefinite Hessian and Gauss–Newton matrix. Diagonal scores discard mixed terms, which co-routing can make nonzero and of either sign; orthogonal Jacobians or cancellation can also make them vanish. Proposition 1 establishes a strict allocation failure. We use a forward-pass proxy to avoid re-quantizing every subset.

Expert-importance game. Let $\ell(S)$ be calibration language-modeling loss with only S routable. Unavailable logits receive bias -6×10^4 and their outputs are zeroed. We select $\min(k, |S|)$ available experts and renormalize gates; the empty set gives the residual-only pass. The game $v(S) = \ell(\varnothing) - \ell(S)$, with $v(\varnothing) = 0$, measures availability, including rerouting and substitution, rather than quantization. Flat Shapley averages all expert orderings: $\phi_i = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(K-|S|-1)!}{K!}[v(S\cup\{i\}) - v(S)]$. Owen restricts this average using a partition.

3 Method

Owen-MoE converts calibration outputs into a budgeted precision map through a sequence of connected steps (Figure 2). Output similarity first defines a coalition partition for each layer. That partition restricts the coalition and expert orderings used to compute Owen-value sensitivity, which we estimate by structured Monte Carlo sampling of the availability game. The resulting signed values enter the Influence Ratio rule: low-capacity coalitions are candidates for 2-bit precision, while locally influential experts can remain at 4 bits. Sweeping the coalition quantile produces static maps that are quantized with native-expert AutoRound; matched controls receive the same per-layer 4-bit counts. We then analyze how expert interactions and the partition affect attribution and how proxy error can affect rankings.

3.1 Coalition Partition Construction

For layer l we summarize expert i over the calibration pool $\mathcal{D}$ by $\mu_i^l = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \mathrm{pool}\, f_i(x; W_i)$, where pool averages over token positions, and set the pairwise distance $\rho_{ii'} = 1 - \cos(\mu_i^l, \mu_{i'}^l) \in$

$[0, 2]$ between experts i and i'. Agglomerative clustering with complete linkage on $\{\rho_{ii'}\}$ builds a merge tree that we cut to a fixed $m_l = 5$ coalitions $\mathcal{P}_l = \{C_1, \ldots, C_{m_l}\}$.

Inter-expert distance scales vary across layers, so a common absolute threshold can produce all-singleton partitions, under which Owen reduces to Shapley. A fixed coalition count avoids this degeneracy. We assess the resulting partitions using their singleton fraction and maximum coalition size, with full diagnostics in the supplementary material.

Experts receiving no calibration tokens require separate treatment because their output signatures are undefined. In each affected Qwen3 layer, these experts form one coalition and the remaining experts are clustered into four, preserving $m_l = 5$. Such experts can still receive nonzero marginal contributions when fewer than k experts are available and routing selects every available expert.

3.2 Owen-Value Expert Sensitivity

Let $\mathcal{P} = \{C_1, \ldots, C_m\}$ partition N. For $i \in C_j$ the Owen value is a **two-level** average, over orderings of the coalitions and then of experts within the coalition:

$$\mathrm{Ow}_i(v, \mathcal{P}) = \sum_{R \subseteq \mathcal{P} \setminus \{C_j\}} \sum_{B \subseteq C_j \setminus \{i\}} \big[v(U_R \cup B \cup \{i\}) - v(U_R \cup B)\big]\, p(R, B), \tag{3}$$

where R is a set of other coalitions, B a set of i's own coalition-mates, $U_R = \bigcup_{C \in R} C$, and the two-level weight

$$p(R, B) = \underbrace{\frac{|R|!\,(m - |R| - 1)!}{m!}}_{\text{coalition level}} \cdot \underbrace{\frac{|B|!\,(|C_j| - |B| - 1)!}{|C_j|!}}_{\text{within-coalition level}}. \tag{4}$$

This weight is the probability of the preceding coalitions and members under uniform two-level shuffling. Both extreme partitions recover Shapley.

3.3 Structured Monte Carlo Approximation

The double sum in (3) has $O(2^m \cdot 2^{\max_j |C_j|})$ terms, and we estimate it by **structured** Monte Carlo using T independent trials. At trial $t \in \{1, \ldots, T\}$ a **two-level shuffle** (a uniform permutation of the m coalitions followed by a uniform permutation of the experts within each coalition) produces a partition-consistent ordering $\pi^{(t)}$. Sweeping $\pi^{(t)}$ we record for each i the marginal

$$\Delta v_i^{(t)} = v_{B_t}\big(P_{\prec i}^{(t)} \cup \{i\}\big) - v_{B_t}\big(P_{\prec i}^{(t)}\big), \tag{5}$$

where B_t is the trial's calibration batch, v_{B_t} its game, and $P_{\prec i}^{(t)}$ the experts preceding i in $\pi^{(t)}$. A partition-consistent ordering induces exactly the pair (R, B) of preceding coalitions and preceding coalition-mates with probability $p(R, B)$. Fixing the calibration pool $\mathcal{D}$ and partition $\mathcal{P}$, each trial independently draws a calibration batch and a two-level shuffle and reuses that batch for every prefix. Let v denote the batch-averaged game and $\sigma_i^2 = \mathrm{Var}(\Delta v_i^{(t)} \mid \mathcal{D}, \mathcal{P})$ include batch and ordering randomness. For $\widehat{\mathrm{Ow}}_i = T^{-1} \sum_{t=1}^{T} \Delta v_i^{(t)}$,

$$\mathbb{E}[\widehat{\mathrm{Ow}}_i \mid \mathcal{D}, \mathcal{P}] = \mathrm{Ow}_i(v, \mathcal{P}), \qquad \mathrm{Var}(\widehat{\mathrm{Ow}}_i \mid \mathcal{D}, \mathcal{P}) = \sigma_i^2 / T. \tag{6}$$

Signed marginals. A marginal in the availability game can be negative. We average signed marginals to estimate the Owen value. The post-hoc scores $\max(0, \Delta v)$ and $|\Delta v|$ are separate ranking ablations. Of 8192 sampled marginals across 256 Mixtral experts, 37.6% are negative, with $32/256$ experts retaining a negative mean (supplementary material). Each trial redraws its calibration batch, re-evaluates $\ell(\varnothing)$, and evaluates all expert prefixes. With K_l experts and T_l trials in layer l, this requires $N_{\mathrm{eval}} = \sum_l T_l(K_l + 1)$ game-value evaluations. This is a one-time offline calibration. The Qwen3 setting requires $48 \cdot 16 \cdot 129 = 99{,}072$ game evaluations. The supplementary timing measurements concern Mixtral; the score-to-bit mapping there costs ~ 4 ms.

3.4 Influence Ratio and Local Core Protection

A specialized expert can have a small absolute value but a large contribution relative to its coalition. The **Influence Ratio** provides a scale-invariant protection rule for this case. Define the coalition

capacity $V(C_j) = \left|\sum_{i \in C_j} \mathrm{Ow}_i\right|$, the magnitude of the coalition's net signed value. Sort capacities in increasing order, breaking ties by coalition index, and call the first $\lfloor \alpha m \rfloor$ coalitions **low-value**. This finite-set quantile convention determines eligibility for compression. Writing $\tau \geq 0$ for a threshold on the ratio below and $Q(i) \in \{2, 4\}$ for the bit-width assigned to expert i (so $b_i = Q(i)$ in (1)), for an expert i in a low-value coalition,

$$\mathrm{IR}_i = \begin{cases} |\mathrm{Ow}_i| \,/\, V(C_j) & V(C_j) > 0, \\ 0 & V(C_j) = 0, \end{cases} \qquad Q(i) = \begin{cases} 4 & C_j \text{ high-value}, \\ 4 & C_j \text{ low-value}, \ |C_j| > 1, \ \mathrm{IR}_i > \tau, \\ 2 & \text{otherwise}. \end{cases} \tag{7}$$

IR_i is invariant to the coalition's overall magnitude, but can exceed one when signed values cancel. A rare expert carrying most of its (small) coalition's net value has $\mathrm{IR}_i \approx 1$ and is recognized as a local core despite its small absolute value. Protection applies only to **non-singleton** low-value coalitions, since a nonzero-value singleton's $\mathrm{IR}_i = 1$ carries no information and its only route to 4-bit is its own value lifting it into the high-value set. That rule is also what makes deep sub-3-bit budgets reachable, because as $\alpha \to 1$ low-value singletons receive 2-bit precision. Thus α controls the number of coalitions eligible for compression, while τ determines which eligible experts retain 4-bit precision. Together they determine the realized budget.

3.5 Memory-Constrained Bit Allocation

Because experts dominate MoE parameter count, the expert-projection footprint uses the same weighted accounting as (1), $M(Q) = \sum_i (Q(i)/8)\,|W_i|$, independent of the input distribution. The quantile α is applied *per layer*, making the $\lfloor \alpha m_l \rfloor$ lowest-capacity coalitions compressible. Raising it can only convert experts from 4- to 2-bit at fixed τ, which makes $M(Q)$ non-increasing in α and the attainable budgets a discrete staircase. We sweep α and report the operating points attained instead of solving for a target $\bar{B}$; displayed bit-widths are measured from the resulting maps. The resulting map has a deterministic expert-weight payload. A target budget is feasible within this policy only if its attainable staircase contains a point at or below that target.

Matched rankings. To compare another score with Owen-MoE at the same memory, we read off the number h_l of 4-bit experts that the Owen-MoE map protects in layer l at a given operating point, let the other score protect its own top h_l experts in that layer, and assign the remaining $K_l - h_l$ to 2 bits. Every arm then has the same per-layer payload, quantizer and calibration protocol and differs only in which experts it protects; for equal-sized experts the layer-average precision is $2 + 2h_l/K_l$. The competing score thereby inherits Owen-MoE's distribution of 4-bit slots across layers in place of its own allocator, and Sections 4.2 and 4.3 apply this matching to isolated heuristics and to flat Shapley and random selection respectively.

3.6 Analysis

All proofs are in the supplementary material.

Proposition 1 (Existential failure of diagonal-only allocation)**.** *There is a three-expert quadratic loss of the form* (2) *with a non-zero cross term for which diagonal-only greedy allocation is strictly sub-optimal at equal memory. With two experts in this equal-sized two-level construction, protecting either expert eliminates the sole cross term; three is the smallest instance of this failure mode.*

The Owen value is the unique attribution satisfying Efficiency, Additivity, Null-player, and symmetry within and across coalitions (Owen, 1977); both extreme partitions recover Shapley. The difference from Shapley is visible in the unanimity basis.

Proposition 2 (Harsanyi-dividend reallocation)**.** *Write $v = \sum_{\varnothing \neq S} d_v(S)\, u_S$ with Harsanyi dividends d_v (Harsanyi, 1963) and unanimity games $u_S(A) = \mathbf{1}\{S \subseteq A\}$ for $A \subseteq N$. For $i \in C_j$, with r_S the number of blocks meeting S,*

$$\mathrm{Ow}_i(v, \mathcal{P}) = \sum_{S \ni i} \frac{d_v(S)}{r_S\,|S \cap C_j|}, \qquad \phi_i(v) = \sum_{S \ni i} \frac{d_v(S)}{|S|}.$$

Dividends inside one block (r_S=1) keep their Shapley allocation; only those spanning $r_S \geq 2$ blocks are reallocated.

First- and second-order dividends receive identical allocations. Differences require higher-order dividends unevenly distributed across blocks: a three-expert unanimity game partitioned as $\{\{1,2\},\{3\}\}$ gives Owen $(1/4, 1/4, 1/2)$ versus Shapley $(1/3, 1/3, 1/3)$. Thus quadratic interactions motivate joint attribution, while matched Owen–Shapley controls test the partition's additional benefit.

Proposition 3 (Proxy stability). $\max_i \left|\mathrm{Ow}_i(v,\mathcal{P}) - \mathrm{Ow}_i(v',\mathcal{P})\right| \le 2\max_S |v(S) - v'(S)|$ *for any two games* v, v' *on* N *and any partition* $\mathcal{P}$. *Consequently, if* $\varepsilon = \max_S |v(S) - v'(S)|$ *and two experts differ by more than* 4ε *under* v, *their order is preserved under* v'.

Availability changes routing and measures language-modeling loss, whereas reconstruction fixes routing. Quantized pair interventions evaluate whether the resulting protection choices reduce error in selected cases.

4 Experiments

4.1 Setup

Qwen3-30B-A3B (Yang et al., 2025) uses top-8 of 128 experts in 48 layers: 6144 modules and 18432 projections, with no shared expert. Transfer models are Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024) (top-2 of eight) and Qwen1.5-MoE-A2.7B (Bai et al., 2024) ($\sim$ 60 experts, with shared experts). We evaluate full test sets with lm-evaluation-harness (Gao et al., 2023) in zero-shot mode: PIQA (Bisk et al., 2020), ARC-e/c (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), MathQA (Amini et al., 2019), and MMLU (Hendrycks et al., 2021).

We use $T = 16/32$ on Qwen3/Mixtral and native-expert AutoRound (Intel, 2024) (200 iterations, group size 128, asymmetric), calibrated on 128 WikiText (Merity et al., 2017) samples of length 512. Routers remain BF16 and attention uses four bits. Scores average three seeds. Coalition construction, estimation, and calibration use only WikiText, with no benchmark labels; each budget's static map is evaluated unchanged across tasks. Reported average bits cover expert weights and exclude group-128 metadata (approximately 0.15 bit per weight).

4.2 Matched-Budget Ranking Comparison

Router-norm+MaxVar and PMQ are evaluated as expert-ranking scores: each fills the 4-bit slots set by Owen-MoE's per-layer map, rather than using its original bit allocator. Under this matching (Section 3.5), Owen-MoE's three-seed mean exceeds Router-norm+MaxVar and PMQ on all 48 task–budget comparisons (Table 1). Its average gains range from 3.57 to 7.02 points over Router-norm and from 4.86 to 10.56 over PMQ across the six budgets. The advantage generally increases as precision decreases: the PMQ gap grows monotonically, while the Router-norm gap is 6.96 points at 2.25 bits and reaches 7.02 points at 2.125 bits. Router-norm also exceeds PMQ at every budget on the eight-task average. Table 5 in the appendix reports the eight-task averages and three-seed standard deviations at 2.75, 2.50, and 2.25 bits.

Task-wise behavior. Tasks differ in their sensitivity to quantization: from BF16 to 2-bit uniform, MMLU drops 44.56 points and MathQA 36.48, while WinoGrande drops 15.31. The largest gain falls on MMLU, the task that degrades most: at 2.125 bits Owen-MoE gains 15.99 points there and 10.44 on BoolQ over Router-norm+MaxVar, against +2.55 on ARC-c. The correspondence is not uniform across tasks, since ARC-e degrades third most yet gains only +4.34. The contrast is sharpest on MMLU across the sweep: Owen-MoE moves from 72.58 to 68.52 between 2.75 and 2.125 bit, a 4.06-point decline, whereas Router-norm falls from 65.35 to 52.53, losing 12.82. Owen leads both ranking baselines on every task at every matched budget; Figure 4 gives the margin over Router-norm for each of the 48 task–budget pairs. At 2.75 bits, the near-equal averages for Owen and Uniform-3 mask a task trade-off: MathQA scores are 43.84 and 50.35, respectively.

Same-memory reference. Measured fractional Uniform assigns $\{2, 3\}$ bits by index, averaging $65.39/63.72/61.85/59.06/55.78/51.46$ over the same descending budgets. Owen gains 2.78–13.31 points at matched payload (Figure 3). Different action spaces make this a system-level comparison. Router-norm trails Uniform by $0.79/0.32$ at the two loosest budgets, then leads by up to 6.29; PMQ trails at five budgets, by at most 2.08.

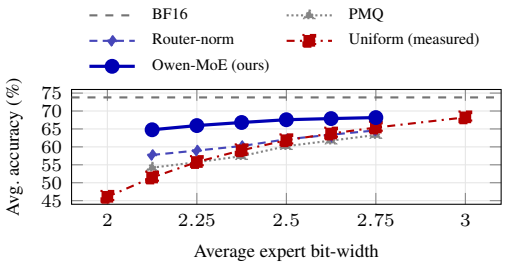

Figure 3: Qwen3-30B-A3B accuracy versus average expert bits. Owen-MoE and the ranking baselines use $\{2, 4\}$ bits; Uniform uses measured $\{2, 3\}$ bits.

Across the sweep, Owen loses 3.40 points versus $6.85/9.10$ for Router-norm/PMQ. At 2.75 bits, Owen scores 68.17, only 0.06 below Uniform-3's 68.23 with 0.25 fewer bits.

Table 1: Zero-shot Qwen3-30B-A3B results (%, `acc_norm` where the task defines it, `acc` otherwise); gray subscripts give the drop from full precision. All ranking methods share the native AutoRound backend, the $\{2, 4\}$ action space, the calibration protocol, and matched per-layer expert budgets; routers stay BF16 and attention is 4-bit in every quantized row. "Bits" averages expert projections. Bold Avg. is best at matched bits.

Method	Bits	PIQA	ARC-e	ARC-c	BoolQ	HellaS.	Wino.	MathQA	MMLU	Avg.
Full precision	16	80.58	79.04	56.74	88.81	77.77	70.80	58.83	77.77	73.79
Uniform	3	78.94	73.06	48.55	84.74	67.85	68.75	50.35	73.57	68.23$_{-5.56}$
	2	62.89	44.28	31.14	60.55	58.69	55.49	22.35	33.21	46.08$_{-27.71}$
Owen-MoE (ours)	2.75	78.61	73.92	49.47	85.53	73.08	68.29	43.84	72.58	**68.17**$_{-5.62}$
	2.625	78.02	74.26	49.05	86.02	72.39	68.81	43.57	70.96	**67.89**$_{-5.90}$
	2.50	77.48	74.41	48.92	86.16	71.83	69.09	43.51	69.18	**67.57**$_{-6.22}$
	2.375	77.19	72.76	47.87	85.76	71.05	68.81	41.92	69.03	**66.80**$_{-6.99}$
	2.25	76.71	70.72	47.06	85.24	70.51	68.21	40.29	68.76	**65.94**$_{-7.85}$
	2.125	76.29	67.90	45.58	84.71	69.44	67.58	38.13	68.52	**64.77**$_{-9.02}$
Router-norm	2.75	76.16	71.14	47.86	82.24	68.47	66.16	39.44	65.35	64.60$_{-9.19}$
+MaxVar	2.625	75.34	70.16	47.05	80.96	67.53	65.46	37.87	62.85	63.40$_{-10.39}$
	2.50	74.86	68.84	46.26	79.44	66.87	64.63	35.73	60.26	62.11$_{-11.68}$
	2.375	73.65	66.87	44.83	77.36	65.63	63.55	33.17	56.75	60.23$_{-13.56}$
	2.25	72.66	65.04	44.16	75.64	65.07	63.06	31.83	54.36	58.98$_{-14.81}$
	2.125	71.84	63.56	43.03	74.27	64.24	62.34	30.16	52.53	57.75$_{-16.04}$
PMQ	2.75	75.24	70.16	47.16	83.13	70.57	65.73	38.83	55.67	63.31$_{-10.48}$
	2.625	74.56	68.14	45.83	81.68	69.53	65.06	36.17	53.04	61.75$_{-12.04}$
	2.50	73.53	66.47	44.66	80.03	68.76	64.13	33.14	50.87	60.20$_{-13.59}$
	2.375	71.37	62.83	42.24	77.16	66.44	62.97	29.73	46.53	57.41$_{-16.38}$
	2.25	69.73	60.47	40.94	75.16	65.16	61.94	28.06	44.37	55.73$_{-18.06}$
	2.125	68.46	58.14	39.87	73.44	63.96	61.24	26.43	42.16	54.21$_{-19.58}$

4.3 Ablation Study

The Owen map uses the complete IR/LCP allocation policy to set per-layer 4-bit quotas. Owen and flat Shapley then use these quotas with the same calibration data, AutoRound settings, and eight-task evaluation, so the comparison changes only the attribution used to choose protected experts. Both attribution estimators use $T = 16$ on Qwen3 and $T = 32$ on Mixtral. Random selection shares the same downstream configuration and quotas. At 2.50 bits, Owen-MoE improves over flat Shapley by 1.78 points on Qwen3 and 1.10 points on Mixtral. At 2.25 bits, these gains increase to 2.58 and 1.84 points, respectively (Figure 5). Flat Shapley also outperforms random selection on Qwen3 by 3.08 and 4.35 points at the two budgets. These comparisons distinguish the benefit of joint attribution from the additional benefit of the chosen coalition structure.

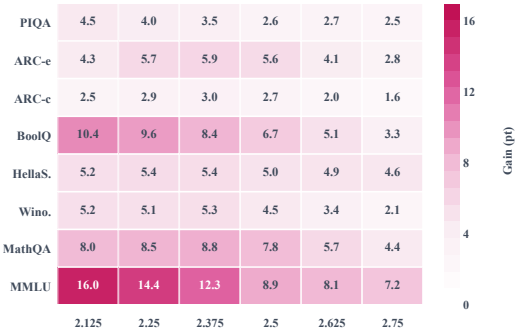

Figure 4: Qwen3 task-wise accuracy gain (points) over matched Router-norm. Darker cells indicate larger gains; all 48 are positive.

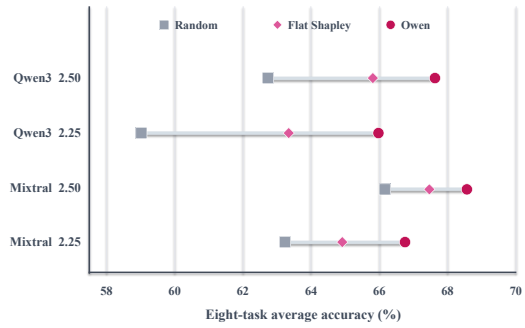

Figure 5: Eight-task means over three seeds at matched per-layer 4-bit counts. Owen and flat Shapley use equal trials; all arms share AutoRound settings.

Estimator checks and sampling stability. Mixtral permits exact finite-game verification. We evaluate all $2^8 = 256$ subsets in each of its 32 MoE layers and apply a subset Möbius transform to obtain Harsanyi dividends and exact Owen and Shapley values. These values satisfy efficiency and recover Shapley under both extreme partitions. The downstream bit maps use $T = 32$ Monte Carlo estimates. On eight Qwen3 layers with three repeats, $T = 16$ versus 48 yields median rank correlation 0.968 (minimum 0.951) and top-16/32/48 overlap $93.0/95.3/96.1\%$. These diagnostics measure ranking and protection-set stability across the tested layers.

4.4 Protection Choices and Quantization Error

Cross-layer interventions. On Qwen3 we select 96 expert pairs across 24 layers, four per layer, from the top 10% by coactivation rate where Owen and Router-norm disagree on which expert to protect. Each pair is evaluated at A4/B2, A2/B4, and A3/B3, holding pairwise average precision fixed. Under round-to-nearest (RTN) quantization, Owen wins on all-token MSE in $75/96$ pairs, ties in four, and loses in 17, using a pre-specified $\pm 1\%$ relative tie band. Table 2 shows a larger improvement on coactivated tokens, linking the ranking choice to the outputs of experts acting together.

Table 2: Qwen3 RTN interventions on selected coactivated, ranking-disagreement pairs. Win-rate denominators include ties; 95% intervals resample layers.

Metric	All	Coactivated
Owen win/tie/loss	75/4/17	80/3/13
Win rate	78.1%	83.3%
95% interval	68.8–86.5%	74.0–90.6%
Median MSE gain	+17.6%	+25.4%
Layers with gain	21/24	22/24
Pairs beating A3/B3	65/96	70/96

A stratified AutoRound recheck selects six RTN wins and four RTN losses across different layers. All six wins persist, and two losses reverse, giving eight wins in ten cases with median all-token/coactivated-token improvements of $13.9/20.7\%$. This recheck connects the intervention to the main quantizer.

Rare experts and a proxy failure mode. Across all 6144 Qwen3 experts, the median within-layer Spearman correlation between Owen value and routing frequency is 0.28 (interquartile range 0.12–0.43). Experts in the lowest frequency quartile account for $18.6/22.4\%$ of protected experts at $2.50/2.25$ bits. Among 22 intervention pairs involving rare protected experts, Owen wins on all-token MSE in 17. In these selected pairs, the ranking identifies rare experts whose protection reduces quantization error.

Prefix diagnostics on eight layers reveal a boundary of the availability game. For intervention cases in the overlapping layers, the median absolute contribution share from prefixes smaller than the routing width ($|P| < 8$) is 14.9% for successful choices and 57.2% for failures. This share has

Spearman correlation -0.46 with relative MSE improvement. Excluding those prefix contributions retains 87.5% of the 2.25-bit protection set. Thus some failures are associated with value assigned when very few experts are available, a setting that differs from retaining all experts at mixed precision. Full diagnostic definitions and statistics are in the supplementary material.

4.5 Transfer Across Expert Granularity

On Mixtral ($K = 8$), Owen exceeds measured index-based $\{2, 3\}$ Uniform by 1.7–3.4 points at 2.125–2.75 bits, matching its accuracy with 0.25–0.375 fewer bits. Its gains over matched Router-norm range from 0.14 to 1.04 points across the six budgets. PMQ and Hessian also score lower at their matched operating points; the supplementary material includes the GPTQ-sourced BSP reference separately. On Qwen1.5-MoE ($K \approx 60$, BF16 average 65.02), Owen-MoE scores 63.39 at 2.75 bits and 60.03 at 2.125 bits. The 2.75-bit result exceeds both Uniform-3 (62.68) and matched Router-norm (62.46).

Across $2.75 \rightarrow 2.125$ bits, Mixtral loses $5.34/5.75/6.06$ points for Owen/Router-norm/PMQ versus Qwen3's $3.40/6.85/9.10$. At 2.75 bits, the Owen–Router gap is $0.39/0.93/3.57$ for $K = 8/{\sim}60/128$.

Edge deployment. On an 8 GB Jetson Orin Nano, the uniform 4-bit DeepSeek-VL2-Small (DeepSeek-AI, 2024b) checkpoint occupies 8.728 GB, whereas Owen-MoE's $\{2, 4\}$-bit map occupies 7.080 GB and scores 62.70 on DocVQA (Mathew et al., 2021). On DeepSeek-VL2-Tiny, its 2.422 GB map runs at approximately 14 tokens/s and scores 380.71 on MME-Reasoning (Fu et al., 2023) versus the BF16 model's 364.64. The supplement reports all nine multimodal benchmarks, including AI2D (Kembhavi et al., 2016) and MMMU (Yue et al., 2024).

4.6 Discussion and Analysis

The matched Shapley controls show that coalition structure contributes beyond joint attribution under the tested per-layer quotas. Its effect depends on the partition: Qwen3 has few singleton coalitions, whereas Mixtral has many. Proposition 2 explains why this distinction matters. Interactions contained within one coalition retain their Shapley allocation, so only interactions spanning coalitions can change attribution. The cross-model comparisons also vary in training, scale, and routing width.

The intervention and prefix diagnostics identify a condition associated with proxy mismatch. In the selected pairs, protection choices fail more often when very small prefixes contribute a larger share of the availability value. Such prefixes leave fewer than k experts available and change routing, whereas mixed-precision quantization retains the full pool. Proposition 3 gives a conditional ranking guarantee when the uniform proxy error is bounded; the interventions characterize selected protection choices under actual quantization.

At the nominal 3.12-bit point, the supplementary IR/LCP diagnostic has mixed task-level effects: the full setting exceeds no-LCP on MathQA and MMLU but trails it on ARC-c, while it exceeds no-IR only on ARC-c. These results characterize the tested operating point.

5 Conclusion

Owen-MoE uses an output-similarity partition to assign Owen values and an Influence Ratio to protect local cores within a static $\{2, 4\}$-bit map. Matched per-layer quotas separate ranking from budget allocation. Across six Qwen3 budgets, Owen leads Router-norm and PMQ on the eight-task average; at 2.50 and 2.25 bits it also exceeds flat Shapley by 1.78 and 2.58 points. Measured fractional-Uniform comparisons show higher accuracy at equal expert-weight payload.

Expert-pair interventions connect these allocations to lower quantization error on coactivated tokens. Mixtral and Qwen1.5-MoE extend the evaluation across expert configurations. Small-prefix diagnostics identify a gap between the proxy and quantization: masking can reroute tokens, while the quantized expert pool remains available. These results support coalition-structured attribution for static mixed-precision allocation under the tested budgets.

Reproducibility Statement

Section 4 and the supplement specify checkpoints, seeds, clustering, trials, AutoRound and harness versions/settings, and proofs. Expert-projection key matching covers $18432/18432$ Qwen3 and $768/768$ Mixtral projections. Calibration uses WikiText only; maps remain fixed across tasks.

Ethics Statement

This work compresses existing open-weight models and introduces no new training data, human subjects, or personally identifying information. Quantization can shift a model's error profile unevenly across domains. The intended use is deployment of existing MoE checkpoints under memory constraints.

AI Use Statement

Large language models assisted manuscript editing and the organization and consistency checking of author-supplied experimental results. The authors take full responsibility for the experimental evidence and all manuscript content.

References

A. Amini, S. Gabriel, P. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *NAACL*, 2019.

J. Bai et al. Qwen1.5-MoE Technical Report. Technical report, Alibaba Group, 2024.

Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*, 2020.

X. Chen et al. MoBE: Mixture-of-Basis-Experts for Compressing MoE-Based LLMs. *arXiv preprint arXiv:2508.05257*, 2025a.

Y. Chen et al. EAC-MoE: Expert-Selection Aware Compressor for Mixture-of-Experts Large Language Models. *arXiv preprint arXiv:2508.01625*, 2025b.

K. T. Chitty-Venkata, J. Ye, and M. Emani. MoPEQ: Mixture of Mixed-Precision Quantized Experts. *arXiv preprint arXiv:2509.02512*, 2025.

M. N. R. Chowdhury, K. El Maghraoui, H. Tsai, N. Wang, G. W. Burr, L. Liu, and M. Wang. Efficient Quantization of Mixture-of-Experts with Theoretical Generalization Guarantees. In *ICLR*, 2026.

S. Chu et al. Dynamic Expert Quantization for Scalable Mixture-of-Experts Inference. *arXiv preprint arXiv:2511.15015*, 2025.

C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*, 2019.

P. Clark et al. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

D. Dai et al. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*, 2024.

DeepSeek-AI. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024a.

DeepSeek-AI. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

J. Deng et al. GEMQ: Global Expert-Level Mixed-Precision Quantization for MoE LLMs. *arXiv preprint arXiv:2605.23078*, 2026.

T. Dettmers, R. Svirschevski, V. Egiazarian, D. Kuznedelev, E. Frantar, S. Ashkboos, A. Borzunov, T. Hoefler, and D. Alistarh. SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1787533e171dcc8549cc2eb5a4840eec-Abstract-Conference.html.

Z. Dong, Z. Yao, D. Arfeen, A. Gholami, M. W. Mahoney, and K. Keutzer. HAWQ-V2: Hessian Aware Trace-Weighted Quantization of Neural Networks. In *NeurIPS*, 2020.

H. Duanmu et al. MxMoE: Mixed-Precision Quantization for MoE with Accuracy and Performance Co-Design. *arXiv preprint arXiv:2505.05799*, 2025.

W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 2022.

E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. *arXiv preprint arXiv:2210.17323*, 2022.

C. Fu et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023.

Z. Fu et al. EAQuant: Enhancing Post-Training Quantization for MoE Models via Expert-Aware Optimization. *arXiv preprint arXiv:2506.13329*, 2025.

L. Gao et al. A Framework for Few-Shot Language Model Evaluation. Zenodo, 2023.

A. Ghorbani and J. Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *ICML*, 2019.

J. C. Harsanyi. A Simplified Bargaining Model for the n-Person Cooperative Game. *International Economic Review*, 1963.

D. Hendrycks et al. Measuring Massive Multitask Language Understanding. In *ICLR*, 2021.

W. Huang et al. Mixture Compressor for Mixture-of-Experts LLMs Gains More. *arXiv preprint arXiv:2410.06270*, 2025a.

Wei Huang, Yue Liao, Yukang Chen, Jianhui Liu, Haoru Tan, Si Liu, Shiming Zhang, Shuicheng Yan, and Xiaojuan Qi. MC#: Mixture Compressor for Mixture-of-Experts Large Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):7607, 2026. doi: 10.1109/TPAMI.2026.3664873.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, Rongrong Ji, and Liujuan Cao. Discovering Important Experts for Mixture-of-Experts Models Pruning Through a Theoretical Perspective. In *NeurIPS*, 2025b.

Intel. AutoRound: Accuracy-First Low-Bit Quantization for LLMs. https://github.com/intel/auto-round, 2024.

A. Jaiswal, Z. Gan, X. Du, B. Zhang, Z. Wang, and Y. Yang. Compressing LLMs: The Truth is Rarely Pure and Never Simple. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9f09f316a3eaf59d9ced5ffaefe97e0f-Abstract-Conference.html.

Y. Ji, Y. Xiang, J. Li, Q. Xia, P. Li, X. Duan, Z. Wang, and M. Zhang. Beware of Calibration Data for Pruning Large Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ede933e10afa991a10b6f36b6522129-Abstract-Conference.html.

A. Q. Jiang et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024.

A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A Diagram Is Worth a Dozen Images. In *ECCV*, 2016.

S. Kim, C. Hooper, A. Gholami, et al. SqueezeLLM: Dense-and-Sparse Quantization. *arXiv preprint arXiv:2306.07629*, 2023a.

Y. J. Kim et al. Mixture of Quantized Experts (MoQE): Complementary Effect of Low-Bit Quantization and Robustness. *arXiv preprint arXiv:2310.02410*, 2023b.

D. Lepikhin et al. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *ICLR*, 2021.

P. Li, Z. Zhang, P. Yadav, Y.-L. Sung, Y. Cheng, M. Bansal, and T. Chen. Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3d09a88c3372cdb79401fde592ca4db8-Abstract-Conference.html.

P. Li et al. QuantMoE-Bench: Examining Post-Training Quantization for Mixture-of-Experts. *arXiv preprint arXiv:2406.08155*, 2024b.

J. Lin et al. AWQ: Activation-Aware Weight Quantization for LLM Compression and Acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

J. Liu, R. Gong, X. Wei, Z. Dong, J. Cai, and B. Zhuang. QLLM: Accurate and Efficient Low-Bitwidth Quantization for Large Language Models. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/437bc4ccafd3fc6d4289bd10940be42b-Abstract-Conference.html.

Y. Liu et al. VPTQ: Extreme Low-Bit Vector Post-Training Quantization for Large Language Models. In *EMNLP*, 2024b.

X. Lu, Q. Liu, Y. Xu, A. Zhou, S. Huang, B. Zhang, J. Yan, and H. Li. Not All Experts Are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models. *arXiv preprint arXiv:2402.14800*, 2024.

Y. Ma, H. Li, X. Zheng, F. Ling, X. Xiao, R. Wang, S. Wen, F. Chao, and R. Ji. AffineQuant: Affine Transformation Quantization for Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ddfb58b61e7213dfb9e06c695475b2bd-Abstract-Conference.html.

M. Mathew, D. Karatzas, and C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images. In *WACV*, 2021.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *ICLR*, 2017.

N. Muennighoff et al. OLMoE: Open Mixture-of-Experts Language Models. *arXiv preprint arXiv:2409.02060*, 2024.

G. Owen. Values of Games with a Priori Unions. In *Mathematical Economics and Game Theory*, pp. 76–88. Springer-Verlag, 1977.

H. Park et al. Value-and-Structure Alignment for Routing-Consistent Quantization of Mixture-of-Experts Models. *arXiv preprint arXiv:2606.05688*, 2026.

K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *AAAI*, 2020.

W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, K. Zhang, G. Peng, Y. Qiao, and P. Luo. OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c6483c8a68083af3383f91ee0dc6db95-Abstract-Conference.html.

L. S. Shapley. A Value for n-Person Games. In *Contributions to the Theory of Games*, pp. 307–317. Princeton University Press, 1953.

M. Sun, Z. Liu, A. Bair, and Z. Kolter. A Simple and Effective Pruning Approach for Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/14c856c7a41297804de4c4890e846b25-Abstract-Conference.html.

Z. Wang, J. Zhu, and J. Chen. ReMoE: Fully Differentiable Mixture-of-Experts with ReLU Routing. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94dc604e115237a7f4a758b3146cd976-Abstract-Conference.html.

G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *ICML*, 2023.

A. Yang et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

W. Yang et al. AlphaQ: Calibration-Free Bit Allocation for Mixture-of-Experts Quantization. *arXiv preprint arXiv:2606.04980*, 2026.

Z. Yuan, Y. Shang, and Z. Dong. PB-LLM: Partially Binarized Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3ab228c4703c4459b1a600ebadc5732c-Abstract-Conference.html.

X. Yue et al. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *CVPR*, 2024.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*, 2019.

Zhang et al. SHAPE: Coalition-Aware Expert Pruning for Sparse Mixture-of-Experts LLMs. *arXiv preprint arXiv:2606.09886*, 2026.

J. Zhao et al. BitsMoE: Efficient Spectral Energy-Guided Bit Allocation for MoE LLM Quantization. *arXiv preprint arXiv:2606.00079*, 2026.

Y. Zhao et al. PuzzleMoE: Efficient Compression of Large Mixture-of-Experts Models via Sparse Expert Merging and Bit-Packed Inference. *arXiv preprint arXiv:2511.04805*, 2025.

Supplementary Material

A Theoretical Details

This section makes the estimand, estimator, and precision rule explicit. The analysis conditions on a fixed calibration pool and a fixed coalition partition. Each Monte Carlo trial draws one batch from that pool and uses the same batch for every prefix in the trial, exactly as described in the main paper. The analysis does not assume that the masking game is monotone, and it does not use adaptive truncation or transform signed marginal contributions.

A.1 Availability Game and Partition-Consistent Orderings

Let $N = \{1, \dots, K\}$ be the experts in one MoE layer and let $\mathcal{P} = \{C_1, \dots, C_m\}$ be a disjoint partition of N. For the fixed calibration pool $\mathcal{D}$, let $\mathcal{B}(\mathcal{D})$ be its finite set of calibration batches. For $B \in \mathcal{B}(\mathcal{D})$, denote by $\ell_B(S)$ the empirical language-modeling loss when exactly the experts in $S \subseteq N$ are available to the router. The masked-all state is the residual-only pass. Define the batch game and its calibration-pool average as

$$\begin{aligned} v_B(S) &= \ell_B(\varnothing) - \ell_B(S), \\ v(S) &= \mathbb{E}_B[v_B(S)], \quad B \sim \mathrm{Unif}(\mathcal{B}(\mathcal{D})). \end{aligned} \tag{8}$$

where $v_B(\varnothing) = v(\varnothing) = 0$. All Owen values below refer to the averaged game v conditional on the fixed pool $\mathcal{D}$. Availability can cause rerouting among the remaining experts. Consequently, neither $v(S) \le v(T)$ for $S \subseteq T$ nor nonnegative marginal contributions are assumed.

A partition-consistent ordering is generated by drawing a uniform permutation of the m coalitions, drawing an independent uniform permutation inside every coalition, and concatenating the permuted coalitions. Let $\Sigma(\mathcal{P})$ be the resulting set of orderings. For $\pi \in \Sigma(\mathcal{P})$, define the predecessor set

$$P_i^\pi = \{j \in N : \pi^{-1}(j) < \pi^{-1}(i)\}. \tag{9}$$

The Owen value is the expected signed marginal

$$\mathrm{Ow}_i(v, \mathcal{P}) = \mathbb{E}_{\pi \sim \mathrm{Unif}(\Sigma(\mathcal{P}))} \left[v(P_i^\pi \cup \{i\}) - v(P_i^\pi) \right]. \tag{10}$$

For completeness, suppose $i \in C_j$. Let $R \subseteq \mathcal{P} \setminus \{C_j\}$ be the coalitions preceding C_j, let $S \subseteq C_j \setminus \{i\}$ be the members preceding i inside C_j, and write $U_R = \bigcup_{C \in R} C$. Equation (10) is equivalent to

$$\begin{aligned} \mathrm{Ow}_i(v, \mathcal{P}) = \sum_{R \subseteq \mathcal{P} \setminus \{C_j\}} \sum_{S \subseteq C_j \setminus \{i\}} & p_j(R, S) \\ & \cdot \left[v(U_R \cup S \cup \{i\}) - v(U_R \cup S) \right], \end{aligned} \tag{11}$$

where

$$p_j(R, S) = \frac{|R|!(m - |R| - 1)!}{m!} \frac{|S|!(|C_j| - |S| - 1)!}{|C_j|!}. \tag{12}$$

The two factors are exactly the probabilities of the corresponding block-level and within-block predecessor sets.

Lemma 1 (Extreme partitions)**.** *If $\mathcal{P} = \{\{1\}, \dots, \{K\}\}$ or $\mathcal{P} = \{N\}$, then $\mathrm{Ow}_i(v, \mathcal{P}) = \phi_i(v)$ for every i, where ϕ_i is the Shapley value.*

Proof. For the singleton partition, the coalition permutation is a uniform permutation of the experts. For the grand partition, the single within-coalition permutation is a uniform permutation of the experts. In either case, $\Sigma(\mathcal{P})$ is the set of all $K!$ expert orderings with the uniform distribution, which is the permutation definition of the Shapley value. □

Lemma 2 (Efficiency with signed marginals)**.** *The signed Owen values satisfy*

$$\sum_{i \in N} \mathrm{Ow}_i(v, \mathcal{P}) = v(N) - v(\varnothing) = v(N). \tag{13}$$

Proof. For any fixed $\pi \in \Sigma(\mathcal{P})$, summing the successive marginal contributions telescopes from $v(\varnothing)$ to $v(N)$. Taking the expectation over π preserves the equality. □

Lemma 2 explains why the implementation retains signed marginals. Applying $\max(0, \cdot)$ or $|\cdot|$ before averaging generally breaks the telescoping identity and yields a ranking score, not an Owen value.

A.2 Structured Monte Carlo Estimator

For each trial $t \in \{1, \dots, T\}$, independently draw $B^{(t)} \sim \mathrm{Unif}(\mathcal{B}(\mathcal{D}))$ and $\pi^{(t)} \sim \mathrm{Unif}(\Sigma(\mathcal{P}))$. For expert i, define

$$\begin{aligned} \Delta_i^{(t)} &= v_{B^{(t)}}(P_i^{\pi^{(t)}} \cup \{i\}) - v_{B^{(t)}}(P_i^{\pi^{(t)}}), \\ \widehat{\mathrm{Ow}}_i &= \frac{1}{T} \sum_{t=1}^{T} \Delta_i^{(t)}. \end{aligned} \tag{14}$$

A trial evaluates $v_{B^{(t)}}(\varnothing)$ and then the K successive prefixes of its ordering, always on the same $B^{(t)}$. Thus one layer requires exactly $T(K+1)$ game-value evaluations. The calibration pool is fixed, while the batch is redrawn between trials.

Proposition 4 (Unbiasedness and variance)**.** *Conditional on $\mathcal{D}$ and $\mathcal{P}$,*

$$\mathbb{E}[\widehat{\mathrm{Ow}}_i] = \mathrm{Ow}_i(v, \mathcal{P}), \qquad \mathrm{Var}(\widehat{\mathrm{Ow}}_i) = \frac{\sigma_i^2}{T}, \tag{15}$$

where $\sigma_i^2 = \mathrm{Var}_{B,\pi}(\Delta_i^{(1)})$ includes both batch sampling and partition-consistent ordering variance.

Proof. Conditional on a batch, the two-level shuffle samples each ordering in $\Sigma(\mathcal{P})$ with uniform probability. Averaging first over the ordering and then over the batch gives $\mathbb{E}_{B,\pi}[\Delta_i^{(t)}] = \mathrm{Ow}_i(v, \mathcal{P})$ by linearity of the Owen value in the game. Independence of the T joint draws gives $\mathrm{Var}(T^{-1} \sum_t \Delta_i^{(t)}) = \sigma_i^2 / T$. □

Proposition 5 (Finite-sample concentration)**.** *Write $\Delta_i^{B,\pi} = v_B(P_i^\pi \cup \{i\}) - v_B(P_i^\pi)$. For each expert, let $D_i = \max_{B,\pi} \Delta_i^{B,\pi} - \min_{B,\pi} \Delta_i^{B,\pi}$, where $B \in \mathcal{B}(\mathcal{D})$ and $\pi \in \Sigma(\mathcal{P})$, and suppose $\max_i D_i \le D_\Delta$. For a failure probability $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{i \in N} \left| \widehat{\mathrm{Ow}}_i - \mathrm{Ow}_i(v, \mathcal{P}) \right| \le D_\Delta \sqrt{\frac{\log(2K/\delta)}{2T}}. \tag{16}$$

Proof. Hoeffding's inequality gives $\Pr(|\widehat{\mathrm{Ow}}_i - \mathrm{Ow}_i| \ge \epsilon) \le 2\exp(-2T\epsilon^2 / D_\Delta^2)$ for each i. A union bound over the K experts and solving for ϵ yields Equation (16). □

This concentration statement is conditional: D_Δ is a property of the finite calibration-pool game and can be estimated from sampled marginals, but the bound is not used to tune on downstream evaluation sets.

A.3 How the Partition Changes Attribution

Write the game in the unanimity basis as

$$v = \sum_{\varnothing \ne S \subseteq N} d_v(S) u_S, \qquad u_S(A) = \mathbf{1}\{S \subseteq A\}, \tag{17}$$

where $d_v(S)$ is the Harsanyi dividend of S. For a nonempty S, let

$$r_S = |\{C \in \mathcal{P} : C \cap S \ne \varnothing\}| \tag{18}$$

be the number of coalitions touched by S.

Proposition 6 (Harsanyi-dividend reallocation). *For $i \in C_j$,*

$$\mathrm{Ow}_i(v, \mathcal{P}) = \sum_{S \ni i} \frac{d_v(S)}{r_S |S \cap C_j|}, \qquad \phi_i(v) = \sum_{S \ni i} \frac{d_v(S)}{|S|}. \tag{19}$$

Therefore, a dividend contained in one coalition keeps its Shapley allocation. Only a dividend spanning at least two coalitions can be reallocated. In particular, all first- and second-order dividends retain their Shapley coefficients: a pair within one block has coefficient $1/(1 \cdot 2)$ and a pair across two blocks has coefficient $1/(2 \cdot 1)$. For higher-order dividends, a coefficient differs only when $r_S |S \cap C_j| \neq |S|$. Consequently, cross-block absolute mass includes terms that cannot separate Owen from Shapley.

Proof. Both values are linear in v; therefore, it suffices to evaluate one unanimity game u_S. The marginal of $i \in S$ equals one exactly when i completes S. Under a partition-consistent ordering, C_j is the last of the r_S touched coalitions with probability $1/r_S$. Conditional on this event, i is the last member of $S \cap C_j$ with probability $1/|S \cap C_j|$. Hence $\mathrm{Ow}_i(u_S, \mathcal{P}) = 1/(r_S |S \cap C_j|)$. Under an unrestricted uniform ordering, each member of S is equally likely to complete S, giving $\phi_i(u_S) = 1/|S|$. Summing over the dividends proves the result. □

Proposition 6 is an attribution statement, not a dominance theorem. It identifies exactly when an intermediate partition can differ from flat Shapley, but it does not imply that the resulting ranking must improve downstream accuracy.

A.4 Exact Owen and Shapley Computation on Mixtral

Mixtral has $K = 8$ experts per MoE layer, so `a1_analyze.py` enumerates all $2^8 = 256$ subsets in each of the 32 layers. For every $S \subseteq N$, the project evaluator performs an actual masked forward pass: unavailable router logits receive a finite -60000 bias and an expert-blocking hook enforces the active set. The stored finite-game value is

$$v(S) = \ell(\varnothing) - \ell(S). \tag{20}$$

Thus, the analysis uses 8192 measured layer–subset configurations rather than sampled permutations. "Exact" here means exact for this fully enumerated, calibration-pool game; it does not turn the finite calibration loss into a population expectation.

A fast subset Möbius transform computes all Harsanyi dividends in $O(K2^K)$ time,

$$d_v(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} v(T). \tag{21}$$

Exact Shapley and Owen values then follow directly from Equation (19), with no permutation sampling. The implementation hard-asserts efficiency, $\sum_i \phi_i = \sum_i \mathrm{Ow}_i = v(N)$, and also asserts $\mathrm{Ow}_i = \phi_i$ under both the all-singleton and single-coalition partitions on the measured Mixtral games. These exact values serve verification and the diagnostics below; the bit maps and downstream results use the $T = 32$ Monte Carlo estimates.

The same dividends support three diagnostics without additional model forwards. We measure cross-block interaction mass as

$$\rho_{\mathrm{cross}} = \frac{\sum_{S : r_S \geq 2} |d_v(S)|}{\sum_{\varnothing \neq S \subseteq N} |d_v(S)|}, \tag{22}$$

decompose absolute mass by interaction order $|S|$, and measure pairwise synergy by $v(\{i, j\}) - v(\{i\}) - v(\{j\})$. Because each dividend is an alternating sum whose noise can grow rapidly with interaction order, raw high-order absolute mass is not treated as evidence by itself. We interpret this mass alongside the exact attribution identities, rather than as a ranking result.

A.5 Stability of the Masking Proxy

Let v be the availability game and let v' be any second game on the same expert set, for example a hypothetical game defined by subset-specific quantization loss. Set

$$\|v - v'\|_\infty = \max_{S \subseteq N} |v(S) - v'(S)|. \tag{23}$$

Proposition 7 (Proxy stability)**.** *For any fixed partition $\mathcal{P}$,*

$$\max_{i \in N} |\mathrm{Ow}_i(v, \mathcal{P}) - \mathrm{Ow}_i(v', \mathcal{P})| \leq 2\|v - v'\|_\infty. \tag{24}$$

Proof. Let $\varepsilon = \|v - v'\|_\infty$. For any predecessor set A and expert $i \notin A$,

$$\begin{aligned} &|[v(A \cup \{i\}) - v(A)] - [v'(A \cup \{i\}) - v'(A)]| \\ &\quad \leq |v(A \cup \{i\}) - v'(A \cup \{i\})| \\ &\qquad + |v(A) - v'(A)| \leq 2\varepsilon. \end{aligned} \tag{25}$$

Averaging this inequality over partition-consistent orderings proves Equation (24). □

Corollary 3 (Ranking preservation)**.** *If* $\mathrm{Ow}_i(v, \mathcal{P}) - \mathrm{Ow}_{i'}(v, \mathcal{P}) > 4\varepsilon$ *with* $\varepsilon = \|v - v'\|_\infty$*, then* $\mathrm{Ow}_i(v', \mathcal{P}) > \mathrm{Ow}_{i'}(v', \mathcal{P})$.

Proof. Proposition 7 can move each of the two values by at most 2ε. Their difference can therefore decrease by at most 4ε. □

The premise of Corollary 3 is not asserted automatically. It states a verifiable sufficient condition: transfer from the masking game to a precision game requires uniform proxy accuracy and a sufficiently large attribution margin.

A.6 Local Core Protection and Memory

Let $a_i = \mathrm{Ow}_i(v, \mathcal{P})$ and define the signed-net capacity of coalition C_k as

$$V_k = \left| \sum_{i \in C_k} a_i \right|. \tag{26}$$

Sort the coalitions in nondecreasing order of (V_k, k), where the index provides a deterministic tie-break. For $\alpha \in [0, 1]$, let $r_\alpha = \lfloor \alpha m \rfloor$ and let $\mathcal{L}_\alpha$ contain the first r_α coalitions in this order. These are the low-value coalitions eligible for compression. This deterministic order-statistic convention implements the main paper's finite-set capacity quantile and resolves ties by coalition index. For $i \in C_k$, define

$$\mathrm{IR}_i = \begin{cases} |a_i|/V_k, & V_k > 0, \\ 0, & V_k = 0. \end{cases} \tag{27}$$

Given $\tau \geq 0$, the two-level precision rule is

$$Q_{\alpha,\tau}(i) = \begin{cases} 4, & C_k \notin \mathcal{L}_\alpha, \\ 4, & C_k \in \mathcal{L}_\alpha,\ |C_k| > 1, \text{ and } \mathrm{IR}_i > \tau, \\ 2, & \text{otherwise.} \end{cases} \tag{28}$$

As stated in the main paper, IR_i can exceed one under sign cancellation. A singleton may have a defined ratio, but the precision rule explicitly excludes it from LCP because it has no within-coalition teammates. A zero-capacity coalition has no local core under this rule. This convention can be discontinuous under cancellation: for values $(1, -1 + \eta)$, ratios grow as $1/|\eta|$ when $\eta \neq 0$, but are set to zero at $\eta = 0$. Scale invariance therefore does not imply stability to estimation noise near zero capacity.

Proposition 8 (Positive scale invariance)**.** *Replacing v by $v' = \lambda v$ for any $\lambda > 0$ leaves $Q_{\alpha,\tau}$ unchanged.*

Proof. Linearity of the Owen value gives $a'_i = \lambda a_i$, hence $V'_k = \lambda V_k$. The capacity ordering is unchanged, and every nonzero ratio satisfies $|a'_i|/V'_k = |a_i|/V_k$. Therefore both $\mathcal{L}_\alpha$ and the protection decisions remain unchanged. □

Proposition 9 (Monotone budget staircase). *Fix v, $\mathcal{P}$, and τ. If $0 \le \alpha_1 \le \alpha_2 \le 1$, then*

$$Q_{\alpha_2,\tau}(i) \le Q_{\alpha_1,\tau}(i) \quad \textit{for all } i \in N. \tag{29}$$

Consequently, for

$$M(Q) = \sum_{i \in N} \frac{Q(i)}{8} |W_i|, \tag{30}$$

where $|W_i|$ is the number of scalar weight parameters in expert i, $M(Q_{\alpha,\tau})$ is non-increasing in α.

Proof. The deterministic capacity order implies $\mathcal{L}_{\alpha_1} \subseteq \mathcal{L}_{\alpha_2}$. Experts in coalitions already in $\mathcal{L}_{\alpha_1}$ keep the same IR decision. An expert in a newly admitted low-value coalition either remains at 4-bit because it satisfies the protection rule or changes from 4-bit to 2-bit. No expert can change from 2-bit to 4-bit. Since every $|W_i|$ is nonnegative, the memory claim follows. □

Define the corresponding no-LCP rule Q^0_α by assigning every member of $\mathcal{L}_\alpha$ to 2-bit and every other member to 4-bit. If

$$H_{\alpha,\tau} = \{i \in C_k : C_k \in \mathcal{L}_\alpha,\ |C_k| > 1,\ \mathrm{IR}_i > \tau\} \tag{31}$$

is the protected set, then the exact memory overhead of LCP is

$$M(Q_{\alpha,\tau}) - M(Q^0_\alpha) = \frac{1}{4} \sum_{i \in H_{\alpha,\tau}} |W_i|. \tag{32}$$

Equation (32) is important for controlled ablations: Full and no-LCP runs must be matched by realized memory, or by the number of 4-bit experts per layer when experts are equal-sized, before an accuracy difference can be attributed to protection.

A.7 Why Diagonal-Only Allocation Can Fail

The following construction proves the first proposition in the main paper. This construction demonstrates an interaction-driven allocation failure.

Proposition 10 (Failure of diagonal-only allocation). *There exists a positive-definite three-expert quadratic loss with a nonzero cross term for which diagonal-only greedy allocation is strictly suboptimal under an equal-memory two-level budget.*

Proof. Let $e_i(b) = 0$ for $b_i = 4$ and $e_i(b) = \delta$ for $b_i = 2$, with $\delta > 0$. Define $e(b) = (e_1(b), e_2(b), e_3(b))^\top$ and

$$\mathcal{L}(b) = \frac{1}{2} e(b)^\top G e(b), \qquad G = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 2.9 \end{bmatrix}. \tag{33}$$

The matrix is positive definite: the first coordinate is positive and the lower 2×2 block has determinant $1.9 > 0$ and positive trace.

Suppose the budget assigns 4-bit to exactly one equal-sized expert. A diagonal-only rule protects expert 1 because its diagonal coefficient, 3, is the largest. This gives $b_{\mathrm{iso}} = (4, 2, 2)$ and

$$\mathcal{L}(b_{\mathrm{iso}}) = \frac{1}{2}\delta^2(1 + 2.9 + 2) = \frac{1}{2}\delta^2(5.9). \tag{34}$$

The equally feasible allocation $b^\star = (2, 2, 4)$ removes the interaction between experts 2 and 3 and gives

$$\mathcal{L}(b^\star) = \frac{1}{2}\delta^2(3 + 1) = \frac{1}{2}\delta^2(4). \tag{35}$$

Therefore, $\mathcal{L}(b_{\mathrm{iso}}) - \mathcal{L}(b^\star) = 0.95\delta^2 > 0$. □

With two experts, equal low-precision perturbations, and a budget that protects one expert, either feasible choice removes the only cross term. A diagonal-only rule then minimizes the remaining diagonal contribution. The three-expert construction is therefore the smallest example of this particular failure mode.

A.8 Scope of the Guarantees

The results above establish four precise facts: the two-level shuffle is an unbiased estimator of the Owen value of the fixed calibration-pool-averaged game; its variance decreases as $1/T$; an intermediate partition reallocates only cross-coalition interaction dividends; and the precision rule is scale-invariant with a non-increasing memory staircase in α. The matched empirical controls separately evaluate the chosen partition and its downstream ranking.

B Precision Accounting and Inference

Budget semantics. Let $\mathcal{E}$ denote all expert modules included in the reported budget. For an expert map Q, the reported average expert precision is

$$\begin{aligned} \bar{b}(Q) &= \frac{\sum_{i\in\mathcal{E}} |W_i| Q(i)}{\sum_{i\in\mathcal{E}} |W_i|}, \\ M_{\text{expert}}(Q) &= \frac{\bar{b}(Q)}{8} \sum_{i\in\mathcal{E}} |W_i|. \end{aligned} \tag{36}$$

Thus, matching $\bar{b}(Q)$ matches the expert-weight payload; for equal-sized Mixtral experts, matching the number of 4-bit experts per layer is equivalent. Whole-model residency can be written as $M_{\text{total}} = M_{\text{fixed}} + M_{\text{expert}} + M_{\text{backend}}$. Non-MoE modules are contained in the common fixed term, while packing metadata, temporary buffers, and kernel workspaces belong to the backend term. The theoretical budget concerns expert weights. Peak VRAM and throughput also depend on the backend; deployment results report measurements where available.

LCP edge cases and inference cost. The Influence Ratio is a scale-free decision statistic, not a probability; it need not sum to one and can exceed one under signed cancellation. Exact zero capacity is assigned ratio zero. A small positive denominator can yield a large ratio, which makes LCP conservatively retain a member of a low-value non-singleton coalition. Singletons are excluded from protection, and high-value coalitions remain 4-bit by construction. Equation (32) gives the additional expert-weight payload. After the map Q is frozen, coalition construction, Monte Carlo estimation, IR computation, and LCP are removed from inference. The router and Top-k execution graph are unchanged; the method introduces no dynamic precision selection or coalition lookup at runtime.

C Algorithm

Algorithm 1 summarizes the complete per-layer procedure from coalition construction to the frozen AutoRound precision map.

Algorithm 1 Owen-MoE bit allocation (per layer)

Require: experts N (K=$|N|$), calibration pool $\mathcal{D}$, masking loss $\ell(\cdot)$, compression quantile α, protection threshold τ, and trial budget T

$\{\rho_{ii'}\} \leftarrow$ cosine distance of mean-pooled native expert outputs
$\mathcal{P} \leftarrow$ agglomerative-cluster($\{\rho_{ii'}\}$, complete linkage, m=5)
{Structured MC: estimate Owen values}
$\mathrm{Ow}_i \leftarrow 0,\ n_i \leftarrow 0$ for all i
for $t = 1$ to T **do**
  $\pi \leftarrow$ two-level shuffle of N consistent with $\mathcal{P}$
  $B_t \leftarrow$ uniform batch from $\mathcal{D}$
  $P \leftarrow \emptyset;\ \ell_{\text{prev}} \leftarrow \ell_{B_t}(\emptyset)$ {masked-all = residual-only pass}
  for expert i in order π **do**
    $P \leftarrow P \cup \{i\};\ \ell_{\text{cur}} \leftarrow \ell_{B_t}(P)$
    $\Delta \leftarrow \ell_{\text{prev}} - \ell_{\text{cur}};\ \ell_{\text{prev}} \leftarrow \ell_{\text{cur}}$
    $\mathrm{Ow}_i \leftarrow \frac{n_i\,\mathrm{Ow}_i + \Delta}{n_i + 1};\ n_i \leftarrow n_i + 1$
  end for
end for
{IR-based Local Core Protection → two-level bit map}
$V(C) \leftarrow \left|\sum_{j \in C} \mathrm{Ow}_j\right|$ for all $C \in \mathcal{P}$; flag the $\lfloor \alpha m \rfloor$ lowest-capacity coalitions as low-value
for all $i \in N$; let $c(i)$ satisfy $i \in C_{c(i)}$ **do**
  $Q(i) \leftarrow 4$ **if** $C_{c(i)}$ high-value **or** $\big(|C_{c(i)}| > 1$ **and** $V(C_{c(i)}) > 0$ **and** $|\mathrm{Ow}_i|/V(C_{c(i)}) > \tau\big)$, **else** 2 {low-value singletons → 2-bit}
end for
Expand each $Q(i)$ to its w_1, w_2, w_3 projections in `layer_config`; AutoRound-quantize the model
return expert bit map Q and quantized model

D Full Experimental Protocol

Table 3 records the Mixtral configuration used for the comparisons below. Table 4 gives the measured Mixtral bit–accuracy frontier. For the fractional Uniform curve, an average expert precision $b \in (2, 3)$ assigns a fraction $3 - b$ of experts to 2-bit and the remainder to 3-bit, using a fixed expert-index order and no value-aware score. This is the measured $\{2, 3\}$ system baseline defined in the main paper. Owen-MoE uses the $\{2, 4\}$ action space. At $3.0/2.5$ bits the ranking controls inherit the Owen-MoE map's per-layer counts, which average four/two of eight experts at 4 bits, and vary only which experts fill them.

Table 3: Mixtral protocol. Qwen3 uses the same pipeline with $T = 16$ and the model-specific settings of Section E; competing scores inherit the per-layer 4-bit counts of the Owen-MoE map produced by the allocation row.

Component	Setting
Quantizer	AutoRound, iters 200, group 128, asym
Expert impl.	native-expert path; individual w_1, w_2, w_3 projections
Non-MoE layers	fixed 4-bit
Owen expert bits	$\{2, 4\}$ (baseline sets as reported)
Coalitions	agglomerative, complete linkage, $m = 5$
Owen est.	Signed full-sweep MC, $T = 32$; one batch redrawn per trial
Allocation	$r_\alpha = \lfloor \alpha m \rfloor$ per layer; τ swept
Calibration	128 samples, seq. len 512, WikiText
Eval	lm-eval-harness, full sets, zero-shot

Reporting conventions. The three-task controlled tables report `acc_norm` for ARC-c and MathQA and `acc` for MMLU, following the fixed extraction script; all values are percentages. The eight-task Avg. is the unweighted arithmetic mean of the eight displayed task scores, rounded to two decimal places. "Bits" averages the assigned precision over expert projections. In Mixtral, all 768 expert projections have the same parameter count; therefore, this arithmetic mean is also the parameter-weighted expert average. Non-MoE modules remain 4-bit and are excluded from the reported expert bit-width.

Evaluation separation. Coalition construction, Owen estimation, and AutoRound calibration use only WikiText calibration examples and model outputs. Downstream benchmark labels are not used to form coalitions, estimate values, fit the quantizer, or construct task-specific precision maps. A bit map is generated once for each budget and evaluated unchanged on every downstream task. This separation prevents task-specific fitting from entering the reported zero-shot comparison.

Table 4: Owen-MoE vs. the measured same-memory **Uniform** $2/3$-bit mix (no importance-based ranking). Avg. zero-shot accuracy (%); ‡ = integer uniform endpoints (Owen has no point there).

Avg. bits	Uniform	Owen-MoE	Δ
3.00‡	70.85	—	—
2.75	68.69	70.40	+1.71
2.625	67.52	69.42	+1.90
2.50	66.25	68.52	+2.27
2.375	64.89	67.76	+2.87
2.25	63.41	66.75	+3.34
2.125	61.65	65.06	+3.41
2.00‡	58.73	—	—

E QWEN3-30B-A3B FULL RESULTS

Qwen3-30B-A3B routes each token to $k = 8$ of $K = 128$ experts in each of its 48 MoE layers. It has no shared expert, and every expert has its own gate, up and down projection, for 6144 expert modules and 18432 expert projections. All configurations keep the routers in BF16 and quantize attention to 4 bits (group 128, asymmetric), so rows differ only in how expert projections are split between 2 and 4 bits. Per-task Qwen3 results are in the main text's matched-budget table. This section adds the estimation protocol, item-level error calculation, measured fractional-Uniform comparison (Table 6), and MathQA analysis. Sections F and G give the attribution controls and quantization interventions.

Protocol. Owen values use $T = 16$ partition-consistent trials per layer, over coalitions built by the output-similarity procedure of the main text. With 8 of 128 experts active per token, 62 of the 6144 experts receive no calibration token at all; their output vectors are zero and their similarity to other experts is undefined, so each layer places them in one separate coalition.

Matched per-layer budget. The Owen-MoE maps on Qwen3 come from the same pipeline as on Mixtral: partition, Owen estimation, and the coalition-capacity and Influence Ratio policy of the main text's Section 3. Every competing score is then held to that map's budget: in each layer it protects as many experts at 4 bits as the Owen-MoE map does, chosen by its own score. The per-layer counts average $8/16/24/32/40/48$ of 128, so the average expert bit-width is exactly $2.125/2.25/2.375/2.50/2.625/2.75$ for every method and the only free variable is *which* experts a score protects. The resulting differences measure ranking performance under fixed per-layer budgets. Each baseline inherits Owen-MoE's distribution of 4-bit slots across layers.

The Router-norm+MaxVar and PMQ arms run through the same pipeline with experts ranked by their own scores. Quantization is AutoRound v0.15.0 (iters 200, group 128, asymmetric) on 128 WikiText sequences of length 512, and all 18432 expert-projection keys of each Owen-MoE map matched a model module. Evaluation is zero-shot on full test sets, and scores are means over three seeds.

Table 5: Matched Qwen3-30B-A3B rankings at three representative budgets. Avg. is the eight-task average from the displayed task scores, followed by the standard deviation across three seeds (percentage points). Drop is the average-score difference from the full-precision reference of 73.79. Methods share the same per-layer 4-bit counts at each budget.

Method	Bits	Avg. ↑	Drop ↓
Owen-MoE	2.75	$\mathbf{68.17 \pm 0.26}$	5.62
Router-norm+MaxVar	2.75	64.60 ± 0.28	9.19
PMQ	2.75	63.31 ± 0.39	10.48
Owen-MoE	2.50	$\mathbf{67.57 \pm 0.31}$	6.22
Router-norm+MaxVar	2.50	62.11 ± 0.22	11.68
PMQ	2.50	60.20 ± 0.41	13.59
Owen-MoE	2.25	$\mathbf{65.94 \pm 0.29}$	7.85
Router-norm+MaxVar	2.25	58.98 ± 0.34	14.81
PMQ	2.25	55.73 ± 0.61	18.06

Test-set uncertainty. Owen-value ranking scores above Router-norm on all 48 task–budget pairs (six budgets $\times$ eight tasks; zero reversals), and above PMQ on all 48. Using the harness's binomial standard errors on the full test sets and treating the two arms as independent, 37 of the 48 Owen–Router differences exceed two standard errors. The other eleven still favor Owen but by less, and they are not spread evenly: ARC-c falls below the threshold at all six budgets (0.8–1.5 standard errors; it is the smallest test set at 1172 items), PIQA at the three loosest ($1.8/1.9/1.9$) and WinoGrande at the two loosest ($1.1/1.8$). At 2.125 bits, six of eight task differences exceed two standard errors, with MMLU gaining $+15.99$ points. These standard errors quantify uncertainty from test items; Table 5 reports the corresponding three-seed standard deviations at three representative budgets.

Budget efficiency. Owen-MoE at 2.25 bit (65.94) exceeds Router-norm at 2.75 bit (64.60) on the eight-task average while spending 0.5 fewer bits per expert, and leads on six of the eight tasks, trailing on ARC-e (-0.42) and ARC-c (-0.80). Extending the comparison to 2.125 bit leaves only a 0.17-point average lead over Router-norm at 2.75 and reverses three tasks, so we place the equal-accuracy crossing near 0.5 fewer bits per expert. At 2.75 bits Owen-MoE is level with full Uniform-3 (68.17 vs. 68.23) at 0.25 fewer bits per expert.

Table 6: **Gain over the measured fractional-Uniform baseline**, eight-benchmark average (points). Fractional Uniform draws from $\{2, 3\}$ bits and picks which experts to demote by expert index, using no importance score; at each budget it is run through the same quantizer, calibration set and evaluation harness as the ranked arms, and its realized expert-weight payload is checked against theirs. Its own averages are $68.69/67.52/66.25/64.89/63.41/61.65$ on Mixtral and $65.39/63.72/61.85/59.06/55.78/51.46$ on Qwen3, top to bottom. Positive means a score beats allocating by index at the same average expert precision.

	Mixtral-8x7B ($K = 8$)			Qwen3-30B-A3B ($K = 128$)		
Bits	Router-norm	PMQ	Owen-MoE	Router-norm	PMQ	Owen-MoE
2.75	$+1.32$	$+1.16$	$\mathbf{+1.71}$	-0.79	-2.08	$\mathbf{+2.78}$
2.625	$+1.52$	$+1.15$	$\mathbf{+1.90}$	-0.32	-1.97	$\mathbf{+4.17}$
2.50	$+2.13$	$+1.28$	$\mathbf{+2.27}$	$+0.26$	-1.65	$\mathbf{+5.72}$
2.375	$+1.83$	$+1.04$	$\mathbf{+2.87}$	$+1.17$	-1.65	$\mathbf{+7.74}$
2.25	$+2.38$	$+1.75$	$\mathbf{+3.34}$	$+3.20$	-0.05	$\mathbf{+10.16}$
2.125	$+2.61$	$+2.14$	$\mathbf{+3.41}$	$+6.29$	$+2.75$	$\mathbf{+13.31}$

Reading the fractional-Uniform comparison. Table 6 compares allocation systems at matched average expert precision. The two models show different patterns. On Mixtral every ranked arm clears the index-order baseline at every budget, Router-norm by 1.3–2.6 points, PMQ by 1.0–2.1, and Owen-MoE by 1.7–3.4; the ordering is Owen-MoE $>$ Router-norm $>$ PMQ $>$ index order throughout, and the gains are roughly flat across the sweep. On Qwen3 the arms separate instead. Owen-MoE's gain grows monotonically from $+2.78$ to $+13.31$, while the isolated scores stay near

the baseline and cross below it at the loosest budgets: Router-norm is 0.79 and 0.32 points below at 2.75 and 2.625 bit before rising to $+6.29$ at 2.125, and PMQ is below at five of six budgets by at most 2.08.

The small crossings occur at the loosest budgets. The action spaces differ between the systems; the fixed-$\{2, 4\}$ controls isolate expert ranking under matched per-layer quotas. On Qwen3, Owen-MoE's advantage over index order grows as the budget tightens; the corresponding gains of isolated scores are smaller. Training data, scale, and routing also differ across the models, so expert count alone does not explain their different patterns.

MathQA. At 2.75 bits, Owen-MoE scores 43.84 on MathQA against Uniform-3's 50.35. At 2.25 bits, Owen-MoE reaches 40.29 against 31.83 for Router-norm and 28.06 for PMQ, leads of 8.46 and 12.23 points under matched per-layer quotas.

F Qwen3 Attribution Controls and Sampling Stability

F.1 Owen, Flat Shapley, and Random Selection

The Owen map uses the complete IR/LCP allocation policy to set the per-layer protection quotas. Owen and flat Shapley both use the full calibration, quantization, and evaluation configuration with the same $T = 16$ trial count; their attribution scores determine which experts occupy the matched slots. All three arms share the downstream protocol and protect, in each layer, as many experts at 4 bits as the Owen-MoE map, 32 per layer on average at 2.50 bits and 16 at 2.25 bits. Table 7 reports the corresponding eight-task means. Random denotes random expert selection under this quota, not a random coalition partition. The corresponding Owen configuration is the one used in the main results. Before this comparison, both extreme-partition identities were checked on the same finite Mixtral game.

Table 7: Qwen3 eight-task means over three seeds at matched per-layer protection quotas.

Bits	Owen	Flat Shapley	Random	Owen–Shapley
2.50	**67.57**	65.79	62.71	+1.78
2.25	**65.94**	63.36	59.01	+2.58

Flat Shapley improves over random selection by $3.08/4.35$ points. The additional Owen gains isolate the effect of replacing the flat attribution with the chosen structured attribution at equal sampling and memory budgets. The gap grows as the protection quota shrinks.

F.2 Monte Carlo Budget Stability

Eight layers spanning model depth are evaluated at $T = 16$ and $T = 48$, with three independent random repeats at each trial budget. Comparisons use within-layer ranks and the actual top-16, top-32 and top-48 protection sets (Table 8).

Table 8: Qwen3 sampling-budget diagnostic over eight layers and three independent repeats. Overlaps compare protection sets from $T = 16$ and $T = 48$.

Metric	Value
Median within-layer Spearman correlation	0.968
Minimum correlation across tested layers and repeats	0.951
Mean top-16 overlap	93.0%
Mean top-32 overlap	95.3%
Mean top-48 overlap	96.1%

These measurements support using $T = 16$ for the matched attribution comparison by showing ordering and selection stability in the tested layers.

G QWEN3 QUANTIZATION INTERVENTIONS AND PROXY DIAGNOSTICS

G.1 CROSS-LAYER EXPERT-PAIR INTERVENTIONS

Within each layer, candidate pairs come from the top 10% of coactivation rates and must have opposing Owen and Router-norm protection choices. The resulting set contains 96 pairs from 24 layers, four pairs per layer. Each pair is evaluated under A4/B2, A2/B4 and A3/B3, with equal average precision across the pair. The initial quantizer is round-to-nearest (RTN). All-token MSE is the primary outcome, and coactivated-token MSE measures behavior where both experts contribute. Records include expert identities, routing frequencies, Owen values, Router-norm scores and coactivation rates.

Let the subscripts indicate the allocation selected by each ranking. Relative improvement is

$$I = 100\% \, \frac{\mathrm{MSE}_{\mathrm{Router}} - \mathrm{MSE}_{\mathrm{Owen}}}{\mathrm{MSE}_{\mathrm{Router}}}. \tag{37}$$

Differences within the pre-specified $\pm 1\%$ relative band count as ties. Confidence intervals resample layers to account for the grouping of four pairs within each layer. Table 9 summarizes the outcomes.

Table 9: RTN intervention outcomes in the selected Qwen3 expert-pair set. Win-rate denominators include ties.

Metric	All tokens	Coactivated tokens
Owen win / tie / loss	75 / 4 / 17	80 / 3 / 13
Owen win rate	78.1%	83.3%
Layer-resampled 95% interval	68.8–86.5%	74.0–90.6%
Median relative MSE improvement	+17.6%	+25.4%
Layers with positive median improvement	21/24	22/24
Pairs where Owen beats A3/B3	65/96	70/96

Improvements occur across layers and are larger on coactivated tokens. The reported rates describe high-coactivation pairs with ranking disagreements.

G.2 AUTOROUND RECHECK

A pre-specified stratified selection takes six RTN wins and four RTN losses from different layers. Replacing RTN with AutoRound while keeping the other conditions fixed gives the outcomes in Table 10.

Table 10: AutoRound recheck on ten pairs stratified by their RTN outcome.

RTN outcome	Pairs	AutoRound wins	AutoRound losses
Owen win	6	6	0
Owen loss	4	2	2
Total	10	8	2

Median MSE improvement is 13.9% on all tokens and 20.7% on coactivated tokens. The six RTN wins persist and two failures reverse, showing both agreement across quantizers and sensitivity of individual cases to quantization details.

G.3 PREFIX DEPENDENCE OF MARGINAL CONTRIBUTIONS

Eight layers spanning depth are analyzed by predecessor-set size. For each prefix bucket, the diagnostic records sample count, mean marginal contribution, signed contribution sum and absolute contribution share. For expert i, define

$$r_i^{<8} = \frac{\sum_{t=1}^{T} |\Delta_i^{(t)}| \, \mathbf{1}\{|P_i^{\pi^{(t)}}| < 8\}}{\sum_{t=1}^{T} |\Delta_i^{(t)}|}. \tag{38}$$

Absolute contributions avoid cancellation in this diagnostic. The original Owen estimator continues to average signed marginals. The threshold 8 is Qwen3's routing width, so smaller prefixes describe availability states with fewer experts than normally active for a token. Table 11 compares the contribution shares across the analyzed expert sets.

Table 11: Median small-prefix absolute-contribution shares. Success and failure subsets use intervention cases in the overlapping analyzed layers.

Expert set	Median $r_i^{<8}$
All analyzed experts in eight layers	16.8%
Experts protected at 2.25 bits	21.4%
Successful intervention protection choices	14.9%
Failed intervention protection choices	57.2%

Small-prefix share has Spearman correlation -0.46 with relative MSE improvement. Removing contributions from prefixes of size less than eight retains 87.5% of the 2.25-bit protection set. Most tested protection choices persist, while failures depend more strongly on sparse availability states. The prefix-excluded score is a diagnostic variant; its measured outcome is protection-set agreement at 2.25 bits.

G.4 Routing Frequency and Rare-Expert Protection

All 6144 Qwen3 experts are analyzed using within-layer ranks. An expert is rare if its routing frequency lies in the lowest quartile of its layer. This avoids comparing raw Owen magnitudes across layers. Table 12 summarizes the routing and protection diagnostics.

Table 12: Frequency and protection diagnostics on Qwen3. Rare-expert shares use the protected set as the denominator.

Metric	Value
Median within-layer Owen–frequency Spearman correlation	0.28
Interquartile range across layers	0.12–0.43
Rare-expert share of the 2.50-bit protected set	18.6%
Rare-expert share of the 2.25-bit protected set	22.4%
Intervention pairs involving rare protected experts	22
All-token Owen wins among these pairs	17/22

Owen and routing frequency carry different ordering information. Some selected rare experts reduce actual quantization error, but their protected-set shares remain below the 25% population share. The evidence supports selective identification of useful rare experts, rather than a general preference for low frequency.

H Mixtral-8x7B Full Results

Mixtral-8x7B-Instruct-v0.1 routes top-2 of $K = 8$ experts per layer and is the coarse-grained end of the granularity range. Its $2^8 = 256$ active-expert subsets per layer can be enumerated, and we use the resulting exact values to verify the estimator (Section A.4); every result reported here, including the equal-budget ranking control of the main text, uses the $T = 32$ Monte Carlo estimates. Table 13 gives per-task results and Figure 6 the bit–accuracy frontier. Owen values use $T = 32$ trials, complete-linkage coalitions cut to five groups per layer, and $\alpha = 0.6$, $\tau = 0.8$ unless swept.

The measured fractional-Uniform baseline. Uniform quantization is realizable only at integer bit-widths, so any sub-3-bit target already mixes blocks. The **Uniform** rows use an index-based allocation: for an average b, a fraction $3 - b$ of experts is demoted to 2-bit by expert index, drawing from $\{2, 3\}$ while Owen-MoE draws from $\{2, 4\}$. Every intermediate row was quantized and evaluated using the protocol above. Owen-MoE clears it at every budget in 2.125–2.75 bit by $+1.7$ to $+3.4$ average points, which is the same accuracy at ≈ 0.25–0.375 fewer bits per expert. Because the

two policies use different action spaces this is a system-level memory comparison, not an isolated test of ranking quality; the matched-budget rows below supply that test.

Table 13: Zero-shot Mixtral-8x7B results (acc/acc_norm, %); gray subscripts give the drop from full precision. All ranking methods share the native AutoRound backend, the $\{2, 4\}$ action space, the calibration protocol, and matched per-layer expert budgets. "Bits" averages expert projections; fractional Uniform is a measured $\{2, 3\}$ baseline. Bold Avg. is best at matched bits.

Method	Bits	PIQA	ARC-e	ARC-c	BoolQ	HellaS.	Wino.	MathQA	MMLU	Avg.
Full precision	16	83.68	83.50	59.64	85.05	83.99	76.44	41.61	67.85	72.72
Uniform	3	82.32	80.05	57.42	86.09	81.51	75.14	39.43	64.84	70.85$_{1.87\downarrow}$
	2.75	81.45	77.74	54.87	84.09	80.01	73.52	37.38	60.49	68.69$_{4.03\downarrow}$
	2.625	80.88	76.71	53.23	83.03	79.33	72.53	36.44	58.00	67.52$_{5.20\downarrow}$
	2.50	80.39	75.19	51.85	81.67	78.54	71.63	35.06	55.67	66.25$_{6.47\downarrow}$
	2.375	79.72	73.94	49.99	80.40	77.49	70.61	34.00	52.94	64.89$_{7.83\downarrow}$
	2.25	79.10	72.34	48.51	78.70	76.48	69.32	32.53	50.32	63.41$_{9.31\downarrow}$
	2.125	78.19	70.59	46.41	76.76	75.18	67.90	31.02	47.13	61.65$_{11.07\downarrow}$
	2	76.44	67.68	43.60	72.51	72.93	65.27	28.58	42.79	58.73$_{13.99\downarrow}$
Owen-MoE	2.75	82.91	80.34	59.82	82.14	81.65	74.53	39.05	62.74	**70.40**$_{2.32\downarrow}$
	2.625	82.15	79.48	57.91	82.02	81.18	74.16	37.64	60.83	**69.42**$_{3.30\downarrow}$
	2.50	81.42	78.53	56.18	81.98	80.51	73.68	36.94	58.92	**68.52**$_{4.20\downarrow}$
	2.375	81.14	77.65	55.36	81.76	79.24	73.35	35.42	58.15	**67.76**$_{4.96\downarrow}$
	2.25	80.85	74.83	54.67	81.61	77.88	72.84	33.86	57.49	**66.75**$_{5.97\downarrow}$
	2.125	80.12	74.58	52.88	78.91	77.05	69.84	32.95	54.12	**65.06**$_{7.66\downarrow}$
Router-norm+MaxVar	2.75	81.83	80.47	56.31	85.57	81.05	74.98	38.29	61.55	70.01$_{2.71\downarrow}$
	2.625	81.45	78.62	54.86	85.60	80.57	74.66	36.75	59.78	69.04$_{3.68\downarrow}$
	2.50	80.79	78.41	54.44	85.14	79.36	74.35	36.28	58.23	68.38$_{4.34\downarrow}$
	2.375	80.20	75.38	51.19	84.92	78.69	73.80	33.47	56.07	66.72$_{6.00\downarrow}$
	2.25	80.41	72.90	50.09	84.04	77.46	73.95	31.96	55.53	65.79$_{6.93\downarrow}$
	2.125	78.94	74.45	51.02	80.12	76.56	70.17	31.29	51.56	64.26$_{8.46\downarrow}$

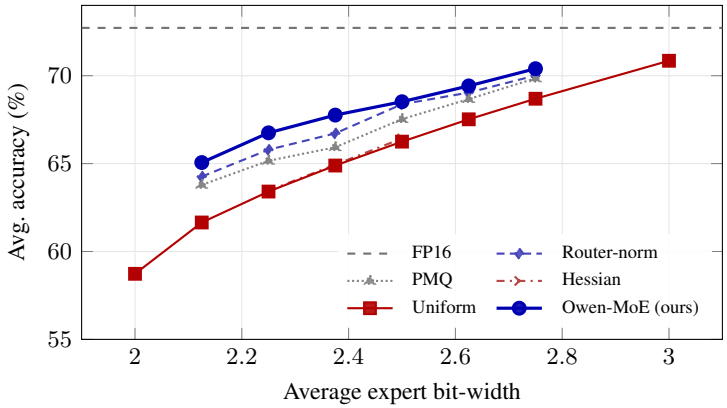

Figure 6: Bit–accuracy on Mixtral. The red Uniform curve is the measured $\{2, 3\}$-bit index allocation. Owen-MoE exceeds it by 1.7–3.4 points and leads the matched ranking baselines across the reported budgets.

Matched-budget rankings. Rerunning Router-norm+MaxVar, activation frequency, flat Shapley, PMQ and Hessian through the same $\{2, 4\}$ bit-map generator, calibration set, AutoRound settings and harness, Owen-MoE leads at every matched point: by $+0.39/+0.38/+0.14/+1.04/+0.96/+0.80$ over Router-norm+MaxVar at $2.75/2.625/2.50/2.375/2.25/2.125$ bit, and by more over PMQ ($69.85/68.67/67.53/65.93/65.16/63.79$ average over the same six budgets) and Hessian ($66.43/63.46$ at $2.50/2.25$). Their per-task columns and the GPTQ-sourced BSP reference point at 2.54 bit are in Table 14. Read task-wise, the margin concentrates on compression-sensitive tasks: at 2.125 bit Owen gains $+0.80$ on average but $+2.56$ on MMLU, while losing 1.21 on BoolQ, a task on which 3-bit uniform already exceeds full precision.

H.1 Additional Mixtral Comparisons

Matched PMQ and Hessian results, a GPTQ-based BSP reference, and three-task controls extend the Mixtral comparison above. The equal-budget three-task results are in Table 15.

Table 14: Additional Mixtral results (acc/acc_norm, %). PMQ and Hessian use matched per-layer budgets and AutoRound. BSP is a GPTQ-based reference at 2.54 expert bits from the MC# source and is excluded from matched-backend comparisons.

Method	Bits	PIQA	ARC-e	ARC-c	BoolQ	HellaS.	Wino.	MathQA	MMLU	Avg.
PMQ	2.75	82.05	78.87	56.48	84.80	81.15	75.30	38.39	61.79	69.85
	2.625	81.56	78.41	52.99	83.67	80.04	74.66	38.26	59.76	68.67
	2.50	80.63	76.94	53.33	83.15	80.02	74.98	37.15	54.05	67.53
	2.375	80.47	73.32	50.00	81.93	78.54	74.66	35.18	53.34	65.93
	2.25	80.14	72.14	49.32	83.15	77.62	74.19	33.70	51.00	65.16
	2.125	79.00	75.51	49.91	72.26	76.76	72.30	34.07	50.53	63.79
Hessian	2.50	80.21	76.38	51.20	81.11	78.05	72.97	35.27	56.21	66.43
	2.25	79.21	72.41	46.70	79.15	76.38	71.25	31.97	50.60	63.46
BSP (GPTQ, not AutoRound)	2.54	68.23	54.97	28.38	68.16	55.61	62.19	24.07	27.74	48.67

Table 15: **Three-task equal-budget ranking control** on Mixtral (complementing the main text's ranking controls). Every row uses AutoRound with the same calibration protocol and $\{2, 4\}$ action space and selects the same number of 4-bit experts per layer, so only the ranking changes. Bold marks the best result within each budget.

Ranking	Avg. bits	ARC-c	MathQA	MMLU	Avg.
Owen	3.00	**62.83**	**42.07**	**65.74**	**56.88**
Flat Shapley	3.00	62.19	41.56	65.13	56.29
Random	3.00	60.04	39.73	62.87	54.21
Owen	2.50	**56.18**	**36.94**	**58.92**	**50.68**
Flat Shapley	2.50	55.63	36.47	58.36	50.15
Random	2.50	53.47	34.86	56.23	48.19

I Qwen1.5-MoE Full Results

Table 16 summarizes the Qwen1.5-MoE accuracy results. The equal-budget perplexity comparison is reported in Table 17.

Table 16: Zero-shot reasoning on Qwen1.5-MoE-A2.7B (Avg. over the 8 benchmarks, %). Owen-MoE at 2.75 bit beats uniform 3-bit at less memory and exceeds the routing-based baseline at matched bits.

Method	Bits	Avg.
Full precision	16	65.02
Uniform	3	62.68
	2	51.53
Owen-MoE	2.75	**63.39**
	2.50	**62.32**
	2.25	**61.22**
	2.125	**60.03**
Router-norm+ MaxVar	2.75	62.46

Table 18 records the frozen Mixtral configuration.

Table 17: Perplexity (PPL; lower is better) at 2.50 average expert bits (complementing the main text's ranking controls). All quantized rows use the same AutoRound backend and per-layer expert budget; FP16 is the unquantized reference.

Model	Owen-MoE	Flat Shapley	Random	FP16
Mixtral-8x7B	**5.0520**	5.0717	5.0736	4.1370
Qwen1.5-MoE-A2.7B	**7.8180**	7.8537	8.0364	7.2200

Table 18: Reproducibility snapshot for the native-expert Mixtral experiments.

Item	Frozen setting
Model	Mixtral-8x7B-Instruct-v0.1; 32 MoE layers, 8 experts/layer, top-2 routing
Calibration	WikiText train; 128 sequences $\times$ 512 tokens; three seeded runs averaged
Coalition/value	Complete linkage, $m = 5$; signed full-sweep MC, $T = 32$
AutoRound	v0.12.3; 200 iterations; group 128; asymmetric quantization
Evaluation	lm-eval v0.4.10; full test sets; zero-shot; one map per budget
Runtime stack	Python 3.12.3; PyTorch 2.11.0+cu128; Transformers 5.12.1; Accelerate 1.14.0
Config audit	768/768 projection keys matched; zero unmatched entries

J Backend Sensitivity and Reproducibility

Our experiments use the native-expert path. AutoRound exposes separate gate, up, and down projections corresponding to w_1, w_3, w_2; all 768 `layer_config` entries match live modules. Fused implementations instead store batched tensors and require reconstruction; incompatible naming can silently leave experts unquantized or coarsen the intended grouping. We freeze and report the `transformers` and AutoRound versions, expert naming convention, match rate (zero unmatched), group size, and asymmetry setting. With a fixed partition, score vector, and bit map, quantizer realization is comparable; score generation remains dependent on the checkpoint, calibration pool, and extraction path.

K Matched Ranking-Control Protocol

The matched controls vary only the protected-expert ranking. Owen, flat Shapley, random, Router-norm+MaxVar, activation frequency, PMQ, and Hessian share the checkpoint, calibration inputs, $\{2, 4\}$ action space, per-layer 4-bit count, AutoRound settings, and evaluation harness. Differences isolate ranking at fixed memory. Fractional Uniform is reported separately as a practical same-memory $\{2, 3\}$ system baseline.

What each baseline score is, and what the projection removes. References for all four methods are in the main text. **Router-norm+MaxVar** ranks experts primarily by the change in the router's ℓ_2 norm, assigning higher precision to experts whose router norm changes least, and then promotes experts with high maximum intra-neuron variance (MaxVar), the largest deviation of a first-layer weight row from its row mean. Both quantities are read from the pretrained checkpoint. **PMQ**, the mixed-precision stage of MC-MoE, combines two calibration statistics: access frequency, the fraction of calibration tokens routed to an expert, and an activation-weighted term averaging the routing weight the expert receives, which are multiplied under two balancing exponents. It then selects bit-widths with an integer program whose objective also uses each MoE layer's post-quantization output reconstruction error. **Hessian** ranks by a diagonal Hessian-trace sensitivity. **BSP**, the reference point quoted for Mixtral, is neither of these: it allocates at layer rather than expert granularity, holding 25% of MoE layers at 4 bits and the rest at 2, and its published numbers come from a GPTQ backend, so we quote it for context only and exclude it from matched comparisons.

PMQ and Router-norm+MaxVar ordinarily choose both the number and identities of protected experts. Here their scores select expert identities under the per-layer 4-bit counts of the Owen-MoE map. The measured fractional-Uniform comparison separately evaluates complete $\{2, 3\}$ and $\{2, 4\}$ allocation systems at matched expert-weight payload.

L RELEASED EVALUATION ARTIFACT

The anonymous artifact verifies quantization and evaluation from the frozen 2.5-bit map. It includes the per-projection assignment, AutoRound/lm-eval runner, MathQA adapter, versions, metadata, and SHA-256 checksums. A dry run matches all 768 expert projections (576 two-bit and 192 four-bit), enabling exact frozen-map evaluation. Algorithm 1 specifies the preceding coalition, Owen-estimation, and mapping stages, which are outside this evaluation-focused release.

Table 19: Multimodal results on DeepSeek-VL2-Tiny (1/3): AI2D, DocVQA, and InfoVQA. Best values among the ~ 2.4 GB methods are bold.

Method	Assignment	Expert Prec.	Size(GB)	AI2D	DocVQA	InfoVQA
Baseline	Uniform	16	6.278	84.326	88.467	63.892
Baseline	Uniform-AutoRound	8	3.854	84.294	88.632	64.003
Baseline	Uniform-AutoRound	4	2.642	83.549	88.162	62.959
Baseline	Activation Freq. (a)	2/3/4	2.427	**84.003**	87.652	61.955
Baseline	Activation Freq. (b)	2/3/4	2.410	83.776	87.666	**62.004**
MoPEQ	Hessian Sens. (a)	2/3/4	2.367	83.970	87.334	61.645
MoPEQ	Hessian Sens. (b)	2/3/4	2.349	83.323	87.633	60.983
MoPEQ	Norm. Freq-Sens (a)	2/3/4	2.311	83.258	87.106	60.858
MoPEQ	Norm. Freq-Sens (b)	2/3/4	2.348	83.355	**87.950**	61.695
Owen-MoE	**Owen value**	**2/4**	2.422	83.358	87.410	60.430

Table 20: Multimodal results on DeepSeek-VL2-Tiny (2/3): MME-Reasoning, MME-Perception, and MMMU. Best values among the ~ 2.4 GB methods are bold.

Method	Assignment	Expert Prec.	Size(GB)	MME-Reas.	MME-Perc.	MMMU
Baseline	Uniform	16	6.278	364.643	1,554.758	40.222
Baseline	Uniform-AutoRound	8	3.854	357.143	1,549.785	40.111
Baseline	Uniform-AutoRound	4	2.642	357.857	1,563.248	39.111
Baseline	Activation Freq. (a)	2/3/4	2.427	358.214	1,561.679	37.111
Baseline	Activation Freq. (b)	2/3/4	2.410	340.000	1,561.810	38.111
MoPEQ	Hessian Sens. (a)	2/3/4	2.367	371.786	1,561.112	**38.556**
MoPEQ	Hessian Sens. (b)	2/3/4	2.349	374.643	1,560.403	32.000
MoPEQ	Norm. Freq-Sens (a)	2/3/4	2.311	370.000	1,532.439	38.111
MoPEQ	Norm. Freq-Sens (b)	2/3/4	2.348	376.786	**1,585.692**	33.333
Owen-MoE	**Owen value**	**2/4**	2.422	**380.710**	1,547.820	36.670

Table 21: Multimodal results on DeepSeek-VL2-Tiny (3/3): RealWorldQA, ScienceQA, and BLINK. Best values among the ~ 2.4 GB methods are bold.

Method	Assignment	Expert Prec.	Size(GB)	RealWorldQA	SciQA	BLINK
Baseline	Uniform	16	6.278	65.490	88.460	40.505
Baseline	Uniform-AutoRound	8	3.854	65.229	88.317	40.768
Baseline	Uniform-AutoRound	4	2.642	65.098	87.458	42.083
Baseline	Activation Freq. (a)	2/3/4	2.427	**66.144**	86.934	41.189
Baseline	Activation Freq. (b)	2/3/4	2.410	66.013	86.552	42.188
MoPEQ	Hessian Sens. (a)	2/3/4	2.367	63.791	87.172	**42.609**
MoPEQ	Hessian Sens. (b)	2/3/4	2.349	64.837	87.458	40.505
MoPEQ	Norm. Freq-Sens (a)	2/3/4	2.311	64.052	86.028	42.188
MoPEQ	Norm. Freq-Sens (b)	2/3/4	2.348	64.575	87.029	40.768
Owen-MoE	**Owen value**	**2/4**	2.422	64.710	**87.600**	41.710

Table 22: Multimodal results on DeepSeek-VL2-Small (1/3): AI2D, DocVQA, and InfoVQA. Best values among the $\sim$7 GB methods are bold.

Method	Assignment	Expert Prec.	Size(GB)	AI2D	DocVQA	InfoVQA
Baseline	Uniform	16	30.078	88.342	92.296	72.380
Baseline	Uniform-AutoRound	8	15.845	73.154	63.372	44.328
Baseline	Uniform-AutoRound	4	8.728	71.503	62.580	42.274
Baseline	Activation Freq. (a)	2/3/4	7.357	51.036	13.193	24.568
Baseline	Activation Freq. (b)	2/3/4	7.305	**70.531**	62.005	**41.174**
MoPEQ	Hessian Sens. (a)	2/3/4	7.215	41.767	62.018	40.589
MoPEQ	Hessian Sens. (b)	2/3/4	7.141	69.560	61.737	40.054
MoPEQ	Norm. Freq-Sens (a)	2/3/4	6.941	55.019	14.194	24.255
MoPEQ	Norm. Freq-Sens (b)	2/3/4	7.113	70.045	62.356	41.169
Owen-MoE	**Owen value**	**2/4**	7.080	61.840	**62.700**	40.860

Table 23: Multimodal results on DeepSeek-VL2-Small (2/3): MME-Reasoning, MME-Perception, and MMMU. Best values among the $\sim$7 GB methods are bold.

Method	Assignment	Expert Prec.	Size(GB)	MME-Reas.	MME-Perc.	MMMU
Baseline	Uniform	16	30.078	513.214	1,649.135	48.000
Baseline	Uniform-AutoRound	8	15.845	494.643	1,643.240	33.222
Baseline	Uniform-AutoRound	4	8.728	485.357	1,640.147	34.000
Baseline	Activation Freq. (a)	2/3/4	7.357	372.857	1,532.495	9.333
Baseline	Activation Freq. (b)	2/3/4	7.305	**486.071**	**1,678.309**	**32.556**
MoPEQ	Hessian Sens. (a)	2/3/4	7.215	481.071	1,675.200	30.000
MoPEQ	Hessian Sens. (b)	2/3/4	7.141	466.786	1,659.337	32.000
MoPEQ	Norm. Freq-Sens (a)	2/3/4	6.941	372.857	1,552.618	18.889
MoPEQ	Norm. Freq-Sens (b)	2/3/4	7.113	485.714	1,666.169	30.667
Owen-MoE	**Owen value**	**2/4**	7.080	472.860	1,668.800	31.220

Table 24: Multimodal results on DeepSeek-VL2-Small (3/3): RealWorldQA, ScienceQA, and BLINK. Best values among the $\sim$7 GB methods are bold.

Method	Assignment	Expert Prec.	Size(GB)	RealWorldQA	SciQA	BLINK
Baseline	Uniform	16	30.078	69.804	94.373	52.236
Baseline	Uniform-AutoRound	8	15.845	58.431	94.564	42.557
Baseline	Uniform-AutoRound	4	8.728	57.124	95.279	41.820
Baseline	Activation Freq. (a)	2/3/4	7.357	43.007	85.312	38.611
Baseline	Activation Freq. (b)	2/3/4	7.305	55.425	**95.136**	41.189
MoPEQ	Hessian Sens. (a)	2/3/4	7.215	**56.601**	93.515	41.767
MoPEQ	Hessian Sens. (b)	2/3/4	7.141	55.425	93.562	**42.767**
MoPEQ	Norm. Freq-Sens (a)	2/3/4	6.941	46.797	87.935	39.663
MoPEQ	Norm. Freq-Sens (b)	2/3/4	7.113	55.163	94.278	42.083
Owen-MoE	**Owen value**	**2/4**	7.080	55.200	94.527	41.645

M Additional Diagnostics and Ablations

M.1 Coalition Formation Diagnostics

Mixtral's per-layer median inter-expert distance varies by approximately $2.3\times$. In the eight-layer threshold scan, the absolute $\theta = 0.20$ cut passes the structural gate in one layer. Fixed-count clustering yields a maximum coalition size of four and approximately 65% singleton coalitions. Qwen3 has a different structure: only 10 of its 240 coalitions are singletons (4.2%), and the largest coalition contains 95 experts. Across its 6144 experts, 62 receive no calibration token and are grouped separately within their layers. These statistics describe the partitions used in the attribution analysis.

Table 25 reports the eight-layer threshold scan separately from the 32-layer fixed-count audit. Absolute thresholds are brittle under layer-dependent distance scales. Fixed-count complete linkage instead yields five coalitions per layer, a non-singleton in every layer, and no group of seven or eight experts. This establishes non-degeneracy, not semantic optimality.

Table 25: Partition-health audit. "Good" denotes the structural gate used by the sweep: a non-degenerate number of coalitions, at least one non-singleton coalition, and no group of seven or more experts. Scans with different layer counts are reported separately rather than pooled.

Cut rule	Linkage	Layers	Avg. m_l	Singleton	Max. $\lvert C\rvert$	Non-singleton layers	Good layers
"Singleton" is the fraction of *coalitions* of size one, not of experts; with $m_l = 5$ over $K = 8$ experts the two differ by $8/m_l$.							
Absolute $\theta = 0.20$	complete	8	6.88	92.7%	6	2/8	1/8
Absolute $\theta = 0.30$	average	8	8.00	100.0%	1	0/8	0/8
Per-layer quantile 0.35	complete	32	4.75	61.2%	5	32/32	29/32
Fixed $m_l = 5$	complete	32	5.00	65.0%	4	32/32	32/32

M.2 Sensitivity to α and τ

Table 26 audits allocation sensitivity on one frozen score artifact. Increasing α enlarges the low-value set and lowers precision; increasing τ protects fewer local leaders and further reduces average bits. The sweep uses the fixed scores without downstream labels.

Table 26: Sensitivity of the realized map. "Leaders" is the number of low-value experts retained at 4-bit by LCP.

α	τ	Avg. bits	Leaders
0.4	0.6	3.5078	21
0.4	0.8	3.4609	15
0.4	0.9	3.4531	14
0.6	0.6	3.2500	27
0.6	0.8	3.1875	19
0.6	0.9	3.1719	17
0.8	0.6	2.9297	41
0.8	0.8	2.8047	25
0.8	0.9	2.7656	20

M.3 Signed Marginal Diagnostics

Table 27 compares score transforms against the signed estimator. Across 256 experts and 32 structured trials per expert, $3{,}080/8{,}192 = 37.6\%$ of sampled marginals are negative; $254/256$ experts have at least one negative trial, and $32/256$ experts retain a negative mean signed Owen value. This supports the theoretical decision to preserve signs in the estimator.

M.4 Component Ablations

Table 28 compares the LCP and IR policies at the nominal 3.12-bit operating point. The displayed bit average is rounded; realized payloads can differ slightly across policies because assignments are

Table 27: How post-hoc marginal transforms change the expert scores. Pearson correlation and top-128 overlap are computed over the same 256 experts.

Comparison to signed	Pearson r	Top-128 overlap
Clamp-zero	0.8208	109/128
Absolute value	0.3393	84/128

discrete. The full setting uses signed Owen values with IR and LCP enabled. Table 29 compares the signed and transformed policies.

Table 28: Local Core Protection diagnostic at the nominal 3.12-bit operating point.

Setting	Avg. bits	ARC-c	MathQA	MMLU
NO-LCP	3.12	55.12	37.86	62.17
NO-IR	3.12	54.78	38.29	62.38
FULL	3.12	54.86	38.09	62.31

Table 29: Negative-marginal policy diagnostic at the nominal operating point. The Keep signed row reproduces the full setting in Table 28; the task scores therefore agree across tables. Clamp-zero and absolute value are post-hoc score transformations.

Aggregation	Avg. bits	ARC-c	MathQA	MMLU	Avg.
Clamp zero	3.12	53.67	36.18	61.55	50.47
Absolute value	3.12	55.03	37.22	62.10	51.45
Keep signed	3.12	54.86	38.09	62.31	51.75

M.5 Calibration Cost

Table 30 reports calibration time and ranking stability. The calibration pool is fixed, and rank correlation is computed against MC$=64$. Calibration size has little timing effect; Monte Carlo count dominates. The rank correlations at MC$=16$ and MC$=32$ are 0.94 and 0.93, respectively.

Table 30: Wall-clock calibration cost (seconds) and rank correlation relative to 64 Monte Carlo trials.

Cal. samples \ MC	8	16	32	64
16	304	610	1216	2394
128	321	627	1227	2463
Corr. vs. MC$=64$	0.60	0.94	0.93	1.00

Multimodal deployment is measured on an 8 GB Jetson Orin Nano. Complete task-wise results appear in Tables 19, 20, and 21 for Tiny and Tables 22, 23, and 24 for Small; on Tiny the released Owen-MoE configuration runs at approximately 14 tokens/s, while the FP16 model exceeds the device memory capacity.