SETPRUNE: BEYOND EXPERT RANKING FOR MIXTURE-OF-EXPERTS PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Expert pruning reduces the memory footprint of Mixture-of-Experts (MoE) models by removing redundant experts, thereby enabling more efficient deployment. Most existing methods assign scalar importance scores to individual experts and prune them according to the resulting ranking. However, pruning is inherently a set-selection problem: the utility of the retained expert set may depend not only on individual experts but also on their collective contributions, while direct subset search becomes combinatorially expensive as the expert pool grows. We propose **SetPrune**, a framework that moves beyond expert ranking by jointly modeling residual individual and collective coupling contributions and directly optimizing the subset of experts to retain. Building on the Möbius decomposition of retained-set utility, SetPrune distinguishes singleton contributions from collective effects associated with expert coalitions. Because exact recovery of the full contribution structure requires evaluating exponentially many coalitions, SetPrune samples coalition-conditioned marginal contributions through Monte Carlo expert permutations and summarizes their collective dependencies using centered marginal covariance. These observations are used to construct residual individual terms and collective couplings within a structured set-level objective, which directly optimizes the retained expert set under a prescribed pruning budget rather than collapsing coalition-dependent information into independent expert scores. We further introduce efficient truncated sampling strategies to reduce the calibration cost of contribution estimation. Experiments demonstrate that SetPrune effectively preserves both downstream and language modeling performance, particularly under aggressive expert pruning, paving the way toward practical set-level expert selection for MoE compression.

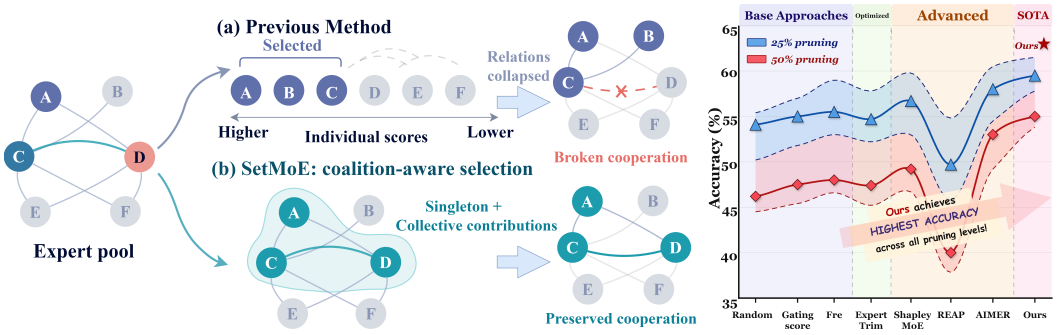
1 INTRODUCTION

Mixture-of-Experts (MoE) models activate only a small subset of experts for each input through conditional router, substantially reducing per-token FLOPs. Extensive work has demonstrated the scalability and computational advantages of this paradigm (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Lewis et al., 2021; Riquelme et al., 2021; Du et al., 2022; Jiang et al., 2024; Dai et al., 2024; DeepSeek-AI, 2024). However, this sparsity applies to computation rather than parameter reduction because the full expert pool still be stored in memory. Expert pruning (Chen et al., 2022) provides a direct and hardware-friendly solution by permanently removing a subset of experts without modifying the remaining model structure.

Most existing expert-pruning methods follow a score-then-select paradigm: each expert is summarized by a scalar importance score derived from routing/activation frequency Muzio et al. (2024), mean gate scores He et al. (2025), router changes Chowdhury et al. (2024), or router-weighted activation norms and reconstruction-based criteria Lasby et al. (2026). Recent cooperative-game approaches incorporate coalition-aware attribution into expert scoring Huang et al. (2025); Zhang (2026), yet the resulting pruning decisions are still largely driven by expert-wise scalar values. Consequently, despite increasingly sophisticated importance estimation, the final decision remains predominantly expert-wise. This creates a mismatch between the decision rule and the pruning objective: ranking favors experts with high scores, whereas pruning ultimately retains an expert set whose utility may depend on how its members contribute together. Direct subset-search methods address this set-level nature more explicitly, but their calibration and search costs can grows combinatorially

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.

with the size of the expert pool. This gap motivates moving beyond expert ranking toward direct modeling and optimization of the utility of the retained expert set. Fig1a illustrates this mismatch, showing why expert-wise ranking can fail to preserve the most valuable expert set.



(a) (b)

Figure 1: **(a)** Motivation of SetPrune: expert-wise ranking evaluates experts independently and may overlook collective contributions in the retained set. **(b)** Performance comparison with representative expert-pruning baselines under the same budget.

To address this challenge, we propose **SetPrune**, a framework for expert set optimization that jointly models residual individual and collective coupling contributions for MoE pruning. SetPrune builds on the Möbius decomposition (Rota, 1964) of utility of the retained expert set, which separates contributions attributable to individual experts from collective contributions arising from expert coalitions. Exact recovery of the full Möbius decomposition would require evaluating combinatorially many coalitions. However, its aggregate contribution structure is observable through coalition-conditioned marginal contributions, where each marginal reflects both the expert's individual contribution and the collective contributions associated with its surrounding coalition. Building on this connection, SetPrune **(i)** samples Monte Carlo expert permutations to obtain marginal observations across diverse coalition contexts; **(ii)** uses the marginal mean to characterize average expert contribution and the covariance of centered marginals to summarize collective dependencies, from which it constructs effective residual individual terms and collective couplings; and **(iii)** integrates these complementary components into a structured quadratic objective for cardinality-constrained expert-set selection. To reduce calibration cost, SetPrune further leverages router-aware affinity to skip expert evaluations with limited utility for marginal-contribution estimation and uses a value-based criterion to terminate severely degraded sampling trajectories early.

Experiments across five MoE architectures, seven zero-shot benchmarks, and WikiText-2 demonstrate the effectiveness of SetPrune, particularly under aggressive pruning: with 50% of experts removed, it retains 92.0% of Qwen3-30B-A3B's unpruned average zero-shot accuracy and reduces WikiText-2 perplexity by 32.7% relative to REAP. Fig 1b provides comparisons with baselines in Qwen3-30B-A3B, illustrating the effectiveness of SetPrune under the same pruning budget. These results highlight SetPrune's ability to preserve model capability under severe expert reduction, suggesting that retaining set-level contribution structure becomes increasingly valuable as pruning becomes more aggressive. Our contributions are:

1. We formulate MoE expert pruning as a budget-constrained set-selection problem and use the Möbius representation to capture residual individual and collective coupling contributions beyond independent expert scores.
2. We develop SetPrune to estimate these contributions from Monte Carlo permutation marginals and directly select experts under a pruning budget, with efficient sampling to reduce calibration cost.
3. Experiments across multiple MoE LLMs and downstream benchmarks demonstrate the practical applicability of SetPrune for efficient MoE compression.

2 RELATED WORK

MoE pruning. Static MoE compression mainly reduces model size through expert merging (Li et al., 2024; Chen et al., 2025) or expert removal. Most pruning methods rank experts using calibration-based importance signals, including activation frequency, gate scores, router changes, marginal contributions, and router-weighted activation statistics (Chen et al., 2022; Koishekenov et al., 2023; Chowdhury et al., 2024; Muzio et al., 2024; He et al., 2025; Huang et al., 2025; Lasby et al., 2026). In contrast, subset-search methods directly optimize retained expert combinations through enumeration, greedy search, evolutionary optimization, or layer-wise genetic search (Lu et al., 2024; Cao et al., 2025; Liu et al., 2024; Yang et al., 2024). While ranking-based methods are efficient, they may overlook dependencies among retained experts, whereas subset-search methods capture such set-level effects at substantially higher search cost. SetPrune aims to bridge this gap by modeling set-level contribution structure while retaining tractable selection.

Möbius Representation of Set Value Functions. Any finite set function admits a unique Möbius representation, which decomposes its value into contributions associated with subsets, with the corresponding coefficients also known as Harsanyi dividends (Rota, 1964; Harsanyi, 1963). Prior work has used this representation to characterize higher-order interactions and developed scalable approximations for arbitrary-order or sparse interaction structures (Fumagalli et al., 2023; Kang et al., 2024). We instead use the Möbius representation to motivate set-level expert selection for MoE pruning, rather than reducing the pruning decision to independent expert scores.

3 PRELIMINARY

In this section, we introduce the set value formulation underlying our expert pruning framework.

3.1 SET VALUE FUNCTION AND MARGINAL CONTRIBUTIONS

Let $\mathcal{N} = \{1, \dots, n\}$ denote the candidate expert set, and let $v : 2^{\mathcal{N}} \to \mathbb{R}$ measure the utility of retaining $S \subseteq \mathcal{N}$. For expert $i \notin S$, its marginal contribution is

$$\Delta_i(S) := v(S \cup \{i\}) - v(S). \tag{1}$$

The marginal contribution $\Delta_i(S)$ depends on the expert subset S to which i is removed. As a result, the same expert may exhibit different marginal contributions across subsets. Given a target number K of retained experts, pruning can be formulated as:

$$S^* = \arg \max_{\substack{S \subseteq \mathcal{N} \\ |S|=K}} v(S). \tag{2}$$

If v is additive, the contribution of each expert is independent of the other retained experts, Eq.2 simplifies to ranking experts by their individual contributions. For a general non-additive v, however, the value of a retained subset also depends on contributions arising from coalitions of experts, making independent expert ranking insufficient for set selection.

3.2 MÖBIUS DECOMPOSITION OF THE SET VALUE FUNCTION

Any finite set value function v can be uniquely decomposed through Möbius inversion:

$$v(S) = \sum_{T \subseteq S} m(T), \qquad m(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} v(U), \tag{3}$$

where $m(T)$ denotes the Möbius coefficient associated with expert subset T. Each term captures the incremental value associated with T. Specifically, the coefficient $m(\{i\})$ represents the *singleton Möbius contribution* of expert i, whereas coefficient $m(T)$ with $|T| \geq 2$ represents the *coalition Möbius contributions* arising from the joint presence of the experts in T.

Thus, Eq.1 can be written as:

$$\Delta_i(S) = \sum_{T \subseteq S} m(T \cup \{i\}) = m(\{i\}) + \sum_{j \in S} m(\{i, j\}) + \sum_{\substack{T \subseteq S \\ |T| \geq 2}} m(T \cup \{i\}). \tag{4}$$

Eq.4 shows that each marginal contribution combines the singleton Möbius contribution of expert i with coalition Möbius contributions involving i. Consequently, evaluating the same expert under different subsets provides observations that reflect its collective contribution patterns.

4 METHOD

In this section, we introduce SetPrune, an MoE pruning method based on expert set optimization as shown in Fig.2. Rather than ranking experts independently, SetPrune models residual individual terms and collective couplings from coalition conditioned marginal observations and directly optimizes the retained expert set under a prescribed pruning budget.

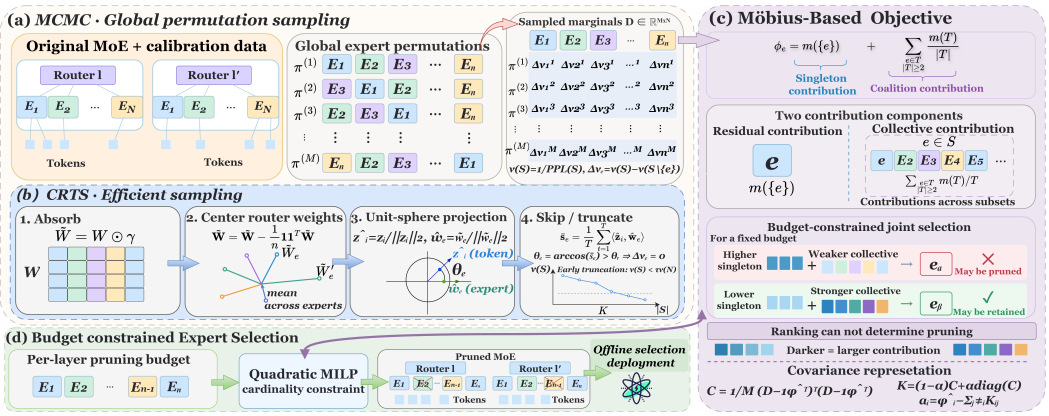


Figure 2: The SetPrune framework, which estimates expert marginal contributions, constructs residual individual terms and collective coupling terms, then performs cardinality-constrained expert-set optimization.

4.1 MONTE CARLO BASED MARGINAL CONTRIBUTION ESTIMATION

Consider an MoE language model with L MoE layers. Layer l contains E_l routed experts and a top-k_l router. Treating each expert as a distinct player, we define the global player set as

$$\mathcal{N} = \{e_i{}^l \mid l \in \{1, \ldots, L\},\ i \in \{1, \ldots, E_l\}\}. \tag{5}$$

We use inverse perplexity as the set value function:

$$v(S) = \frac{1}{\mathrm{PPL}(S)} = \exp\big(-\mathbb{E}_{(x,t)\sim\mathcal{D}}\left[-\log p(x_{t+1} \mid x_{\le t}; S)\right]\big). \tag{6}$$

Maximizing $v(S)$ is therefore strictly equivalent to minimizing $\mathrm{PPL}(S)$, aligning the set value function with the language modeling objective.

To estimate each expert's marginal contribution, we use M Monte Carlo permutations of the global player set. For sample m, let

$$\pi^{(m)} = \big(s_1^{(m)}, s_2^{(m)}, \ldots, s_{|\mathcal{N}|}^{(m)}\big). \tag{7}$$

We traverse the sampled permutation from back to front. At position j, the subset of experts still present is

$$S_j^{(m)} = \{s_1^{(m)}, s_2^{(m)}, \ldots, s_j^{(m)}\}. \tag{8}$$

After removing an expert, we set its routing logit to $-\infty$ so that it is excluded from subsequent routing weight normalization.

For expert $e = s_j^{(m)}$, its marginal contribution in sample m is the change in set value function caused by its removal:

$$\Delta_i^{(m)} = v\big(S_j^{(m)}\big) - v\big(S_j^{(m)} \setminus \{i\}\big) = v\big(S_j^{(m)}\big) - v\big(S_{j-1}^{(m)}\big). \tag{9}$$

Averaging over the M samples derives the marginal contribution estimation:

$$\hat{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} \Delta_i^{(m)}. \tag{10}$$

Because an expert occupies different positions across sampled permutations, expert's marginal contribution reflects its contributions across different subsets.

4.2 Centered Routing and Truncation Strategy

Monte Carlo sampling over global expert permutations incurs substantial computational overhead. We therefore introduce two complementary strategies to accelerate marginal contribution estimation.

Angle-based skipping. Raw router logits depend on both the directions and magnitudes of router weights. Weight of RMSNorm can amplify these magnitudes. As a result, cosine similarity between normalized hidden states and original router weights may provide a biased comparison across experts. To measure affinity in the geometry space for routers, we absorb the RMSNorm weight $\boldsymbol{\gamma}$ into the router weights:

$$\mathrm{RMSNorm}(x)\mathbf{W}^\top = \mathbf{z}(\mathbf{W} \odot \boldsymbol{\gamma})^\top, \tag{11}$$

where $\mathbf{W}$ denotes the router weight matrix and $\mathbf{z}$ denotes the hidden states before application of the RMSNorm weight.

We then remove the direction shared across experts by centering the transformed router weights:

$$\widetilde{\mathbf{W}} = \mathbf{W} \odot \boldsymbol{\gamma}, \qquad \overline{\mathbf{W}} = \widetilde{\mathbf{W}} - \mathrm{mean}_e\big(\widetilde{\mathbf{W}}_{e,:}\big). \tag{12}$$

The centered weights $\overline{\mathbf{W}}$ are used in the subsequent affinity calculation. For token t and expert e, we project the hidden state and router weight onto the unit sphere and compute their cosine similarity:

$$s_{t,e} = \left\langle \frac{\mathbf{z}_t}{\|\mathbf{z}_t\|_2}, \frac{\overline{\mathbf{W}}_{e,:}}{\|\overline{\mathbf{W}}_{e,:}\|_2} \right\rangle. \tag{13}$$

For expert e in layer ℓ, its routing affinity over the T tokens in the current calibration sequence is

$$\bar{s}_{\ell,e} = \frac{1}{T} \sum_{t=1}^{T} s_{t,e}. \tag{14}$$

Since $\bar{s}_{\ell,e} \in [-1, 1]$, we define the corresponding angle as $\theta_{\ell,e} = \arccos(\bar{s}_{\ell,e})$. We skip the forward evaluation for an expert when

$$\theta_{\ell,e} > \theta_\tau. \tag{15}$$

The expert is removed from the current coalition, while its estimated marginal contribution is set to zero. Using the notation of Sec.3.1, we reuse the current coalition value:

$$\Delta_i^{(m)} = 0, \qquad v\big(S_{j-1}^{(m)}\big) \leftarrow v\big(S_j^{(m)}\big). \tag{16}$$

Moreover, under a mild low-exposure condition, the error induced by angle-based skipping admits an explicit bound, as proved in Appendix A.1.

Early truncation. Following the early-truncation idea of Ghorbani & Zou (2020), we stop permutations once the coalition value falls below a fraction of the value of the full expert set:

$$v\big(S_j^{(m)}\big) < \tau\, v(\mathcal{N}), \qquad \tau \in (0, 1). \tag{17}$$

Here, $S_j^{(m)}$ is the expert subset remaining in sample m, $\mathcal{N}$ is the full expert set, and τ is the truncation threshold. This strategy avoids further evaluations after model performance has collapsed, reducing both forward-pass cost and noise in the marginal contribution estimates.

Error introduced by both truncation strategies are controllable and bounded, as proved in Appendix A.2.

4.3 MÖBIUS-BASED CONTRIBUTION OBJECTIVE

MCMC and CRTS provide marginal contribution estimations for all experts, we then use these observations to approximate the residual individual and collective contributions in the set value function and formulate expert pruning as a tractable set optimization problem.

Let $q_i \in \{0, 1\}$ indicate whether expert i is retained, where $q_i = 1$ denotes retention and $q_i = 0$ denotes pruning. For an expert subset $T \subseteq \mathcal{N}$, we further define

$$q_T := \prod_{i \in T} q_i. \tag{18}$$

According to the Möbius decomposition introduced in Sec. 3.2, the loss in set value induced by pruning can be expressed as

$$\Delta V(q) = \sum_{\substack{T \subseteq \mathcal{N} \\ T \neq \emptyset}} m(T)(1 - q_T), \tag{19}$$

where $m(T)$ denotes the Möbius coefficient associated with expert subset T. Thus, Eq. 19 accounts for the loss of both singleton Möbius contribution and coalition Möbius contribution caused by pruning.

However, computing all Möbius coefficients exactly requires evaluating an exponential number of expert subsets. We therefore use the marginal contributions sampled in Sec.4.1 to construct a tractable objective that captures the average marginal contributions of individual experts and the dependencies induced by different coalition contexts, rather than an exact Möbius estimate. For the l-th MoE layer, let $\mathbf{\Delta}_l^{(m)} = [\Delta_{l,1}^{(m)}, \ldots, \Delta_{l,E_l}^{(m)}]^\top$ denote the marginal contribution vector obtained from the m-th sample, and let $\bar{\mathbf{\Delta}}_l = [\bar{\Delta}_{l,1}, \ldots, \bar{\Delta}_{l,E_l}]^\top$ denote the corresponding mean marginal contribution vector.

According to the Möbius decomposition, since singleton Möbius contribution is invariant across permutations, centering the marginal contributions exactly removes the permutation invariant singleton component together with the mean coalition allocation, so that the remaining variation is induced entirely by collective contributions, leaving only coalition induced fluctuations. We characterize the dependence among these fluctuations using the empirical covariance matrix

$$\widehat{\Sigma}^{(l)} = \frac{1}{M} \sum_{m=1}^{M} \left(\mathbf{\Delta}_l^{(m)} - \bar{\mathbf{\Delta}}_l \right) \left(\mathbf{\Delta}_l^{(m)} - \bar{\mathbf{\Delta}}_l \right)^\top. \tag{20}$$

The off-diagonal elements $\widehat{\Sigma}_{ij}^{(l)}$ measures the statistical dependence coalition-induced fluctuations of experts i and j across different expert subsets. As the permutation changes, higher order coalition Möbius contribution is allocated differently among the experts involved, inducing structured fluctuations in their marginal contributions. The covariance matrix summarizes these dependencies, allowing the exponentially large coalition Möbius contribution structure estimated directly from the existing samples. We use the covariance as a proxy for the higher-order coalition Möbius contribution structure among expert coalitions and incorporate it into the subsequent set-aware optimization objective. A detailed derivation of this covariance coupling is provided in Appendix A.3.

Because a finite number of samples can introduce substantial variance into the covariance estimates, we further apply diagonal shrinkage:

$$\widetilde{\Sigma}^{(l)} = (1 - \alpha)\widehat{\Sigma}^{(l)} + \alpha \, \mathrm{diag}\left(\widehat{\Sigma}^{(l)} \right), \qquad \alpha \in [0, 1], \tag{21}$$

where α controls the shrinkage of the off-diagonal covariance terms.

Given the approximated collective contributions, we define the residual individual contribution of expert i as

$$\eta_{l,i} = \bar{\Delta}_{l,i} - \sum_{\substack{j=1 \\ j \neq i}}^{E_l} \widetilde{\Sigma}_{ij}^{(l)}, \tag{22}$$

Here, $\eta_{l,i}$ doesn't represent an exact estimate of the singleton Möbius contribution coefficient $m(\{i\})$. Instead, it represents the residual unary contribution after the estimated collective couplings have been accounted for.

4.4 CARDINALITY-CONSTRAINED EXPERT-SET OPTIMIZATION

All in all, the pruning induced loss of set value in layer l can then be approximated by the following quadratic objective:

$$\widehat{\Delta V}_l(q) = \sum_{i=1}^{E_l} \eta_{l,i}(1 - q_i) + \sum_{1 \le i < j \le E_l} \widetilde{\Sigma}_{ij}^{(l)}(1 - q_i q_j), \tag{23}$$

The first term represents the loss of residual individual contribution, whereas the second term accounts for the collective coupling contribution. Unlike ranking methods, this objective evaluates a retained expert set by jointly considering both types of contributions.

Given a pruning ratio $r \in [0, 1]$, the number of experts removed from layer l is

$$P_l = \lfloor r E_l \rfloor, \tag{24}$$

and the number of retained experts is $R_l = E_l - P_l$. We therefore formulate expert pruning as the following cardinality-constrained binary quadratic optimization problem:

$$\min_{q \in \{0,1\}^{E_l}} \widehat{\Delta V}_l(q) \qquad \text{s.t.} \sum_{i=1}^{E_l} q_i = R_l. \tag{25}$$

To solve Eq. 25 efficiently with mixed-integer linear programming (Bolusani et al., 2024), we introduce an auxiliary variable $y_{ij} = q_i q_j$ for each expert pair. The product term can be exactly linearized by

$$y_{ij} \ge q_i + q_j - 1, \qquad y_{ij} \le q_i, \qquad y_{ij} \le q_j, \qquad y_{ij} \in \{0, 1\}. \tag{26}$$

These constraints ensure that $y_{ij} = 1$ if and only if both experts i and j are retained. Optimizing the resulting objective therefore selects an expert subset that jointly preserves residual individual and collective contributions under the prescribed pruning budget.

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP

Settings: We randomly sample 128 sequences from C4 Raffel et al. (2020), each sequence truncated to 512 for calibration. We use $M = 20$ Monte Carlo permutations, with $\theta_\tau = 92^\circ$ and $\tau = 0.5$ by default. The resulting cardinality constrained quadratic programming problem is solved using a mixed integer linear programming (MILP) formulation.

Models and Baselines: We evaluate SetPrune on Qwen1.5-MoE-A2.7B Qwen Team (2024), Qwen3-30B-A3B Qwen Team (2025), Qwen3.5-35B-A3B Qwen Team (2026), DeepSeekMoE-16B Dai et al. (2024), and GPT-OSS-20B OpenAI (2025). We compare against Random Lu et al. (2024), Gating Score Dong et al. (2025), Frequency Muzio et al. (2024), Expert Trim He et al. (2025), Shapley-MoE Huang et al. (2025), REAP Lasby et al. (2026), and AIMER Liu et al. (2026) followed the same evaluation setting in Shapely-MoE. Detailed model configurations are provided in Appendix A.5.

Benchmarks: We evaluate zero-shot performance on seven commonsense reasoning benchmarks, including HellaSwag Zellers et al. (2019), WinoGrande Sakaguchi et al. (2021), BoolQ Clark et al. (2019), OpenbookQA Mihaylov et al. (2018), PIQA Bisk et al. (2020), ARC-Easy, and ARC-Challenge Clark et al. (2018). We report the average accuracy across all seven tasks. All evaluations are conducted using the EleutherAI LM Evaluation Harness Gao et al. (2024), together with WikiText-2 perplexity for language modeling evaluation Merity et al. (2016).

5.2 ZERO-SHOT TASKS

Table 1 reports the average zero-shot accuracy across seven downstream tasks. SetPrune remains competitive under 25% pruning and shows a larger advantage under 50% pruning, achieving 57.46% and 61.17% on Qwen1.5-MoE-A2.7B and Qwen3-30B-A3B, respectively, outperforming the strongest baselines by 3.95 and 8.47 points. These results suggest that set-level expert selection becomes increasingly beneficial as the expert budget becomes more constrained.

Table 1: Average zero-shot accuracy (%) across seven tasks. For each model we report 25% / 50% pruning. Bold and underlined numbers denote the optimal and sub-optimal in each column.

	Qwen1.5-A2.7B		Qwen3-30B		Qwen3.5-35B		DeepSeek-16B		GPT-OSS-20B	
Method	25%	50%	25%	50%	25%	50%	25%	50%	25%	50%
None	61.62		66.46		67.78		61.92		58.29	
Random	54.79	45.25	57.69	44.96	66.03	59.80	53.37	41.50	54.81	44.57
Gating Score	55.42	46.30	58.27	46.22	67.29	62.26	54.23	42.51	<u>56.80</u>	<u>53.33</u>
Frequency	55.63	46.84	58.68	46.54	67.07	62.22	54.58	42.84	56.78	53.05
Expert Trim	55.16	46.35	58.27	46.27	67.47	<u>64.05</u>	53.75	42.67	54.43	51.50
Shapley-MoE	57.10	49.01	60.02	48.28	–	–	56.31	44.89	–	–
REAP	46.64	39.66	52.58	36.26	**67.76**	63.56	54.06	43.97	51.34	39.02
AIMER	<u>58.78</u>	<u>53.51</u>	**64.50**	<u>52.70</u>	67.44	62.83	<u>57.70</u>	<u>50.97</u>	50.87	42.31
SetPrune (ours)	**59.54**	**57.46**	<u>64.08</u>	**61.17**	<u>67.55</u>	**64.28**	**59.60**	**54.06**	**58.04**	**54.11**

Table 2: WikiText-2 perplexity of pruned MoE models. For each model we report 25% / 50% pruning. Bold and underlined numbers denote the optimal and sub-optimal in each column.

	Qwen1.5-A2.7B		Qwen3-30B		Qwen3.5-35B		DeepSeek-16B		GPT-OSS-20B	
Method	25%	50%	25%	50%	25%	50%	25%	50%	25%	50%
Dense	7.26		8.23		7.04		7.08		252.12	
Random	11.01	21.87	13.72	35.74	8.60	14.23	10.61	37.18	<u>160.18</u>	94.03
Gating Score	10.39	20.10	12.39	32.03	**7.29**	**8.47**	9.57	35.12	273.79	369.92
Frequency	10.08	19.78	11.97	31.23	<u>7.31</u>	<u>8.61</u>	9.46	34.27	273.66	301.68
Expert Trim	10.28	19.98	12.29	32.26	7.72	9.99	9.67	35.20	406.24	<u>83.07</u>
Shapley-MoE	<u>9.69</u>	17.57	11.02	28.10	–	–	<u>9.12</u>	31.10	–	–
REAP	15.49	55.88	26.99	172.12	8.24	12.54	13.52	38.62	697.86	1498.83
AIMER	10.05	<u>15.55</u>	<u>10.96</u>	<u>19.98</u>	8.32	13.66	9.91	<u>16.88</u>	705.38	438.45
SetPrune (ours)	**8.60**	**11.43**	**9.39**	**10.92**	7.67	9.85	**7.80**	**10.95**	**13.17**	**18.41**

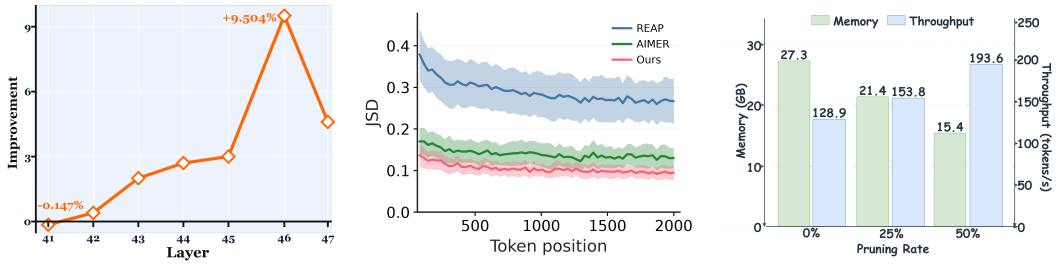
5.3 Overall language ability

Table 2 reports WikiText-2 perplexity under different pruning ratios. SetPrune consistently preserves language modeling performance across the evaluated MoE architectures, with its advantage shows a substantially stronger advantage under 50% pruning ratio. This trend further supports the effectiveness of set-level expert selection under aggressive compression.

5.4 Ablation study

Effect of the angle threshold and truncation threshold. Effect of the angle and truncation thresholds. We examine how the two thresholds impact pruning performance and sampling cost as shown in Table3. For angle-based skipping, decreasing θ_τ reduces sampling cost but overly aggressive skipping degrades performance. Conversely, $\theta_\tau = 94^\circ$ yields almost no skipping but not the highest accuracy under 50% pruning, suggesting that moderate skipping may improve estimation quality by filtering noisy low-affinity evaluations. Overall, $\theta_\tau = 92^\circ$ provides the best trade-off. For early truncation, increasing τ reduces Monte Carlo evaluations, whereas an excessively large τ may discard informative marginal observations. $\tau = 0.5$ achieves the highest zero-shot accuracy at both pruning ratios while substantially reducing sampling cost. We therefore use $\theta_\tau = 92^\circ$ and $\tau = 0.5$ by default.

Effect of individual and collective contributions. We compare SetPrune with an individual-only variant using hidden states MSE between pruned and unpruned Qwen3-30B-A3B WikiText-2 in last layers. Positive value indicates lower representation distortion for SetPrune. As shown in Fig. 3a,



(a) (b) (c)

Figure 3: **(a)** Improvement over individual-only pruning. **(b)** Token-wise JSD to uncompressed Qwen1.5-MoE-A2.7B at 50% pruning. **(c)** Memory and throughput versus pruning rate.

SetPrune achieves lower reconstruction error performs better, with larger gains in deeper layers and a peak improvement of 9.504% at Layer 46, supporting the benefit of SetPrune.

Table 3: Ablations on angle skip and cutoff threshold. Bold and underline represent optimal and sub-optimal results. Time is the saving versus untruncated sample process for cutoff, and saving for angle skip.

Setting	PPL 25% ↓	PPL 50% ↓	Acc 25% ↑	Acc 50% ↑	Time
Angle skip					
90	8.936	27.850	57.31	48.67	−47.43%
91	9.621	<u>13.243</u>	57.32	52.73	−17.41%
92	8.601	**11.427**	**59.54**	**57.46**	−5.09%
93	<u>8.540</u>	12.133	59.17	<u>53.12</u>	−1.39%
94	**8.472**	<u>11.988</u>	<u>59.52</u>	52.05	−0.41%
Cutoff threshold					
0.3	8.545	**11.376**	<u>58.98</u>	<u>54.49</u>	−54.75%
0.4	**8.401**	11.648	58.23	53.77	−57.53%
0.5	<u>8.601</u>	<u>11.427</u>	**59.54**	**57.46**	−61.40%
0.6	8.585	11.853	57.75	54.24	−66.94%
0.7	8.739	11.937	58.18	52.74	−73.05%

5.5 Predictive Fidelity and Inference Efficiency

Token-Level Predictive Fidelity: On WikiText-2, we compared next-token distributions of uncompressed Qwen1.5-MoE-A2.7B with REAP and AIMER. We omit the first 64 tokens of each window, which lack causal context. Fig.3b shows that our method stays closest to the uncompressed model throughout the window, with AIMER in between and REAP substantially farther away. The curves remain comparatively flat, indicating a stable gap in predictive fidelity rather than accumulation over token position.

Memory-Bandwidth Characteristics:We measure the memory reduction and inference acceleration of the pruned Qwen1.5-MoE-A2.7B model, with results shown in Fig . The results are obtained with grouped GEMM over the resident expert tensor on an NVIDIA 6000D GPU. Compared with the original model, higher pruning rates reduce GPU memory usage and improve inference speed.

6 Conclusion

In this work, we introduce SetPrune, which formulates MoE expert pruning as set selection and models residual individual terms together with collective couplings from permutation-based marginal observations. Across diverse MoE models, SetPrune preserves model capability under aggressive pruning while avoiding exhaustive coalition evaluation, highlighting the value of set-level expert interactions for efficient MoE compression.

AI USE STATEMENT

The authors declare the use of large language model (LLM) tools during the preparation of this manuscript. These tools were used only for language editing, including sentence-level revision, word choice, and readability of the narrative text. All parts of the work that bear on its originality and scientific validity were carried out by the authors. This includes the research idea and overall framework; the design, implementation, debugging, and validation of the algorithms and code; and the experimental protocol, data processing, execution of experiments, analysis of results, and the conclusions drawn from them. LLM tools were not used to generate the research idea, the method, the experimental procedure, or the conclusions. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 5.1 specifies the calibration data, sampling configuration, pruning hyperparameters, evaluated models, baselines, benchmarks, and evaluation framework used in our experiments. The appendix provides detailed model configurations (Appendix A.5), complete pseudocode for SetPrune (Appendix A.4), proofs of the theoretical claims, and full per-task and ablation results. All experiments use publicly available models and datasets. We will release the complete implementation, pruning and evaluation scripts, configuration files, and instructions for reproducing the reported results in a public repository following the review process.

ACKNOWLEDGMENTS

Jun Li was partially supported by the National Natural Science Foundation of China under Grant Nos. U24A20330, 62361166670, and the Fundamental Research Funds for the Central Universities, No 30926WZCT202601009.

REFERENCES

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020.

Suresh Bolusani, Mathieu Besançon, Ksenia Bestuzheva, Antonia Chmiela, João Dionísio, Tim Donkiewicz, Jasper van Doornmalen, Leon Eifler, Mohammed Ghannam, Ambros Gleixner, Christoph Graczyk, Katrin Halbig, Ivo Hedtke, Alexander Hoen, Christopher Hojny, Rolf van der Hulst, Dominik Kamp, Thorsten Koch, Kevin Kofler, Jurgen Lentz, Julian Manns, Gioni Mexi, Erik Mühmer, Marc E. Pfetsch, Franziska Schlösser, Felipe Serrano, Yuji Shinano, Mark Turner, Stefan Vigerske, Dieter Weninger, and Liding Xu. The scip optimization suite 9.0. *arXiv preprint arXiv:2402.17702*, 2024. URL https://arxiv.org/abs/2402.17702.

Mingyu Cao, Gen Li, Jie Ji, Jiaqi Zhang, Ajay Jaiswal, Li Shen, Xiaolong Ma, Shiwei Liu, and Lu Yin. Condense, don't just prune: Enhancing efficiency and performance in MoE layer pruning. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2412.00069. arXiv:2412.00069.

I-Chun Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, and Chun-Yi Lee. Retraining-free merging of sparse MoE via hierarchical clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 8594–8620, 2025. URL https://github.com/wazenmai/HC-SMoE.

Tianyu Chen, Shaohan Huang, Yuan Xie, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. Task-specific expert pruning for sparse mixture-of-experts. *arXiv preprint arXiv:2206.00277*, 2022. URL https://arxiv.org/abs/2206.00277.

Mohammed Nowaz Rabbani Chowdhury, Meng Wang, Kaoutar El Maghraoui, Naigang Wang, Pin-Yu Chen, and Christopher Carothers. A provably effective method for pruning experts in fine-tuned sparse mixture-of-experts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 8815–8847, 2024. URL https://proceedings.mlr.press/v235/chowdhury24a.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. URL https://arxiv.org/abs/2401.06066.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. URL https://arxiv.org/abs/2405.04434.

Zican Dong, Han Peng, Peiyu Liu, Wayne Xin Zhao, Dong Wu, Feng Xiao, and Zhifeng Wang. Domain-specific pruning of large mixture-of-experts models with few-shot demonstrations, 2025. URL https://arxiv.org/abs/2504.06792.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 5547–5569, 2022. URL https://proceedings.mlr.press/v162/du22c.html.

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

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order shapley interactions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness. https://github.com/EleutherAI/lm-evaluation-harness, 2024.

Amirata Ghorbani and James Y. Zou. Neuron shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5922–5932, 2020.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

Shwai He, Daize Dong, Liang Ding, and Ang Li. Towards efficient mixture of experts: A holistic study of compression techniques. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=HTpMOl6xSI.

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 *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=7kQjbCQwtT.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Justin Singh Kang, Yigit Efe Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the möbius transform. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1467.

Yeskendir Koishekenov, Alexandre Bérard, and Vassilina Nikoulina. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3567–3585, Toronto, Canada, 2023. URL https://aclanthology.org/2023.acl-long.198/.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ukGxWd2aDG.

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

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6265–6274, 2021. URL https://proceedings.mlr.press/v139/lewis21a.html.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient SMoE with hints from its routing policy. In *The Twelfth International Conference on Learning Representations*, 2024.

Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Matthew B. Blaschko, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs. *arXiv preprint arXiv:2407.00945*, 2024. URL https://arxiv.org/abs/2407.00945.

Zongfang Liu, Shengkun Tang, Yifan Shen, Huan Wang, and Xin Yuan. AIMER: Calibration-free task-agnostic MoE pruning, 2026. URL https://arxiv.org/abs/2603.18492.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.334. URL https://aclanthology.org/2024.acl-long.334/.

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Alexandre Muzio, Alex Sun, and Churan He. SEER-MoE: Sparse expert efficiency through regularization for mixture-of-experts. *arXiv preprint arXiv:2404.05089*, 2024. URL https://arxiv.org/abs/2404.05089.

OpenAI. gpt-oss-20b model card. https://openai.com/index/gpt-oss-20b/, 2025.

Qwen Team. Qwen1.5-MoE: A2.7b moe model. https://qwenlm.github.io/blog/qwen-moe/, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.05974.

Gian-Carlo Rota. On the foundations of combinatorial theory i: Theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. doi: 10.1007/BF00531932.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2021. URL https://arxiv.org/abs/1907.10641. Communications of the ACM, 64(9):99–106, 2021.

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

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, and Bo Yuan. MoE-I^2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10456–10466, 2024. URL https://arxiv.org/abs/2411.01016.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Yuhao Zhang. Shape: Coalition-aware expert pruning for sparse mixture-of-experts llms. *arXiv preprint arXiv:2606.09886*, 2026.

A Appendix

A.1 Controlled Error of Angle-Based Skipping

Let S be an expert coalition containing expert i in layer ℓ, and let x be a calibration sequence of length T. The marginal contribution of expert i is

$$\Delta_i(S, x) = v_x(S) - v_x(S \setminus \{i\}). \tag{27}$$

Let $p^S_{t,j}(x)$ denote the actual routing weight of expert j after top-k selection and normalization over the selected experts. Define the average routing exposure of expert i as

$$r_i(S, x) = \frac{1}{T} \sum_{t=1}^{T} p^S_{t,i}(x). \tag{28}$$

Theorem 1 (Controlled error of angle-based skipping). Suppose that removing expert i leaves the input to layer ℓ and the routing logits of all remaining experts unchanged. The router selects the experts with the largest logits and normalizes their exponentiated logits to sum to one. If fewer than k experts remain, all remaining experts are selected. Suppose further that

$$\|f_j(z_t)\| \le G_\ell \tag{29}$$

for every candidate expert j and token t, and that the value function satisfies

$$|v_x(S) - v_x(S \setminus \{i\})| \le \frac{L_v}{T} \sum_{t=1}^{T} \|y^S_t - y^{S\setminus\{i\}}_t\|, \tag{30}$$

where y^S_t and $y^{S\setminus\{i\}}_t$ are the corresponding outputs of layer ℓ.

Assume that the angle-based skip rule is applied only under a small exposure condition, every coalition S and sequence x to which the rule is applied, in Sec5.4 our experiments provide empirical support for its practical validity.

$$\theta_{\ell,i}(S, x) > \theta_e \quad \Longrightarrow \quad r_i(S, x) \le \rho_e. \tag{31}$$

Then the marginal contribution omitted by angle-based skipping satisfies

$$|\Delta_i(S, x)| \le 2L_v G_\ell \rho_e. \tag{32}$$

For each token t, let $q^S_{t,j}(x)$ denote the softmax probability over all candidate experts in layer ℓ, and let $p^S_{t,j}(x)$ denote the actual mixing weight after top-k selection and optional renormalization. Define

$$u^S_{t,i}(x) = \begin{cases} q^S_{t,i}(x), & \text{without post-top-}k\text{ renormalization,} \\ p^S_{t,i}(x), & \text{with post-top-}k\text{ renormalization,} \end{cases} \qquad r_i(S, x) = \frac{1}{T} \sum_{t=1}^{T} u^S_{t,i}(x). \tag{33}$$

Assume that the angle-based skip rule satisfies

$$\theta_{\ell,i}(S, x) > \theta_e \quad \Longrightarrow \quad r_i(S, x) \le \rho_e. \tag{34}$$

Then

$$|\Delta_i(S, x)| \le C L_v G_\ell \rho_e, \qquad C = \begin{cases} 3, & \text{without post-top-}k\text{ renormalization,} \\ 2, & \text{with post-top-}k\text{ renormalization.} \end{cases} \tag{35}$$

Proof. Fix a token t and suppress the token, coalition, and sequence indices. Let A denote the selected expert set before removal, and let p, p' be the actual mixing vectors before and after removing i, indexed over the same candidate set with $p'_i = 0$. Let $w_j = \exp(a_j)$ and $q_j = w_j / \sum_u w_u$.

Without post-top-k renormalization. Here $p_j = q_j \mathbf{1}\{j \in A\}$, and removing i gives $q'_j = q_j/(1 - q_i)$ for every remaining expert. If $i \notin A$, the selected set is unchanged, so

$$\|p - p'\|_1 = \frac{q_i}{1 - q_i} \sum_{j \in A} q_j \le q_i. \tag{36}$$

If $i \in A$, let h be the highest-ranked expert outside A, and write $q_h = 0$ if no replacement is available. Set $B = \sum_{j \in A \setminus \{i\}} q_j$. Since $q_h \le q_i$ and $B + q_h \le 1 - q_i$,

$$\|p - p'\|_1 = q_i + \frac{q_i B + q_h}{1 - q_i} \tag{37}$$

$$\le q_i + \frac{q_i(1 - q_i - q_h) + q_h}{1 - q_i} \tag{38}$$

$$= 2q_i + q_h \le 3q_i. \tag{39}$$

With post-top-k renormalization. Here $p_j = w_j / Z$ for $j \in A$, where $Z = \sum_{j \in A} w_j$, and $p_j = 0$ otherwise. If $i \notin A$, the selected set and its mixing weights are unchanged, giving $\|p - p'\|_1 = 0 = 2p_i$. If $i \in A$, let w_h be the exponentiated logit of its replacement, or zero if no replacement is available. Since $w_h \le w_i$, the new denominator satisfies

$$Z' = Z - w_i + w_h \le Z. \tag{40}$$

Thus $p'_j \ge p_j$ for all $j \ne i$. Both mixing vectors sum to one, and therefore

$$\|p - p'\|_1 = p_i + \sum_{j \ne i} (p'_j - p_j) = p_i + 1 - (1 - p_i) = 2p_i. \tag{41}$$

In either case, restoring the indices yields

$$\|p_t^S - p_t^{S \setminus \{i\}}\|_1 \le C u_{t,i}^S(x). \tag{42}$$

Because the layer input is unchanged and every candidate expert satisfies $\|f_j(z_t)\| \le G_\ell$,

$$\|y_t^S - y_t^{S \setminus \{i\}}\| = \left\| \sum_j \left(p_{t,j}^S - p_{t,j}^{S \setminus \{i\}}\right) f_j(z_t) \right\| \tag{43}$$

$$\le G_\ell \|p_t^S - p_t^{S \setminus \{i\}}\|_1 \tag{44}$$

$$\le C G_\ell u_{t,i}^S(x). \tag{45}$$

Applying the assumed Lipschitz bound and averaging over tokens gives

$$|\Delta_i(S, x)| \le \frac{L_v}{T} \sum_{t=1}^{T} \|y_t^S - y_t^{S \setminus \{i\}}\| \tag{46}$$

$$\le C L_v G_\ell r_i(S, x) \tag{47}$$

$$\le C L_v G_\ell \rho_e, \tag{48}$$

where the last inequality follows from equation 34. □

A.2 Error Bound for Truncated Marginal-Mean Estimation

Let x be a randomly sampled calibration sequence, and let π be a uniformly sampled permutation of all experts. For expert i, let P_i^π denote the set of experts preceding i in π. Its exact marginal contribution is

$$\Delta_i(\pi, x) = v_x(P_i^\pi \cup \{i\}) - v_x(P_i^\pi), \tag{49}$$

and its Shapley value is

$$\phi_i = \mathbb{E}_{\pi, x}\left[\Delta_i(\pi, x)\right]. \tag{50}$$

Let $\mathcal{A}_i$ denote the event that expert i is skipped by the angle-based rule, and let $\mathcal{E}_i$ denote the event that the current permutation is terminated before the marginal contribution of expert i is evaluated.

Given M independent samples $\{(\pi_m, x_m)\}_{m=1}^M$, define

$$\Delta_i^{(m)} = \Delta_i(\pi_m, x_m). \tag{51}$$

The double-truncation estimator is

$$\widehat{\phi}_i^{\mathrm{DT}} = \frac{1}{M} \sum_{m=1}^{M} \Delta_i^{(m)} \mathbf{1}\left\{ \left(\mathcal{A}_i^{(m)}\right)^c \cap \left(\mathcal{E}_i^{(m)}\right)^c \right\}. \tag{52}$$

Define the probabilities of the two truncation events as

$$p_{\theta,i} = \Pr(\mathcal{A}_i), \qquad p_{v,i} = \Pr(\mathcal{E}_i). \tag{53}$$

Theorem 2 (Error Bound for Truncated Marginal-Mean Estimation). Suppose that the marginal contribution is uniformly bounded:

$$|\Delta_i(\pi, x)| \le R_i, \tag{54}$$

and that the marginal contributions omitted by early termination satisfy

$$\mathbb{E}\left[|\Delta_i(\pi, x)| \,|\, \mathcal{E}_i\right] \le \varepsilon_{v,i}. \tag{55}$$

If the angle-based skipping rule satisfies the conditions of Theorem 1, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|\phi_i - \widehat{\phi}_i^{\mathrm{DT}}\right| \le 2p_{\theta,i}L_v G_\ell \rho_e + p_{v,i}\varepsilon_{v,i} + R_i\sqrt{\frac{2\log(2/\delta)}{M}}. \tag{56}$$

Proof. Define

$$Y_i = \Delta_i(\pi, x)\mathbf{1}\left\{\mathcal{A}_i^c \cap \mathcal{E}_i^c\right\}. \tag{57}$$

The total estimation error can be decomposed as

$$\left|\phi_i - \widehat{\phi}_i^{\mathrm{DT}}\right| \le \underbrace{|\mathbb{E}[\Delta_i] - \mathbb{E}[Y_i]|}_{\text{Truncation bias}} + \underbrace{\left|\mathbb{E}[Y_i] - \frac{1}{M}\sum_{m=1}^{M} Y_i^{(m)}\right|}_{\text{Monte Carlo estimation error}}. \tag{58}$$

Since

$$\Delta_i - Y_i = \Delta_i \mathbf{1}\left\{\mathcal{A}_i \cup \mathcal{E}_i\right\}, \tag{59}$$

the triangle inequality and the union bound give

$$|\mathbb{E}[\Delta_i] - \mathbb{E}[Y_i]| \le \mathbb{E}\left[|\Delta_i|\mathbf{1}\{\mathcal{A}_i\}\right] + \mathbb{E}\left[|\Delta_i|\mathbf{1}\{\mathcal{E}_i\}\right] \tag{60}$$

$$= p_{\theta,i}\mathbb{E}\left[|\Delta_i| \,|\, \mathcal{A}_i\right] + p_{v,i}\mathbb{E}\left[|\Delta_i| \,|\, \mathcal{E}_i\right]. \tag{61}$$

By Theorem 1,

$$\mathbb{E}\left[|\Delta_i| \,|\, \mathcal{A}_i\right] \le 2L_v G_\ell \rho_e. \tag{62}$$

Combining this result with equation 55 yields

$$|\mathbb{E}[\Delta_i] - \mathbb{E}[Y_i]| \le 2p_{\theta,i}L_v G_\ell \rho_e + p_{v,i}\varepsilon_{v,i}. \tag{63}$$

This argument does not require $\mathcal{A}_i$ and $\mathcal{E}_i$ to be independent.

Moreover, $Y_i \in [-R_i, R_i]$. Applying Hoeffding's inequality to the M independent samples gives, with probability at least $1 - \delta$,

$$\left|\mathbb{E}[Y_i] - \frac{1}{M}\sum_{m=1}^{M} Y_i^{(m)}\right| \le R_i\sqrt{\frac{2\log(2/\delta)}{M}}. \tag{64}$$

Combining the angle-skipping bound established in Theorem A.1, the truncation-bias bound in Eq. equation 63, and the Monte Carlo bound in Eq. equation 64 yields Eq. equation 56. □

A.3 Covariance Coupling of Collective Contributions

Let $N = \{1, \ldots, n\}$ be the expert set of a single MoE layer, and let $v : 2^N \to \mathbb{R}$ be the cooperative game induced by a *fixed* calibration set, with Möbius coefficients

$$m(T) = \sum_{S \subseteq T}(-1)^{|T|-|S|}v(S), \qquad v(S) = \sum_{T \subseteq S} m(T). \tag{65}$$

We prove that the covariance of permutation marginals couples every collective set of Möbius term at least two into a pairwise matrix C.

A.3.1 Order allocation of collective contributions. Let π be a uniform random permutation of N, and let P_i^π be the set of experts preceding i in π. The marginal contribution of expert i is

$$\Delta_i(\pi) = v(P_i^\pi \cup \{i\}) - v(P_i^\pi)\,. \tag{66}$$

For any two distinct experts i and j, a single permutation realizes exactly one of

$$j \in P_i^\pi \;\text{ and }\; i \notin P_j^\pi, \qquad \text{or} \qquad i \in P_j^\pi \;\text{ and }\; j \notin P_i^\pi. \tag{67}$$

Thus a cooperative increment involving both i and j cannot appear as a new marginal contribution of both experts in the same permutation. Changing the permutation reallocates that contribution among the members of T.

Singleton–cooperation decomposition For every expert i and every permutation π,

$$\Delta_i(\pi) = m_i + u_i(\pi), \tag{68}$$

where $m_i := m(\{i\})$ is the permutation-invariant singleton Möbius contribution and

$$u_i(\pi) := \sum_{\substack{i \in T \\ |T| \geq 2}} m(T)\, \mathbf{1}\{T \setminus \{i\} \subseteq P_i^\pi\} \tag{69}$$

is the sum of all coalition Möbius contribution terms allocated to i under π.

A.3.2 Meaning of centering. The Shapley value is the expected marginal,

$$\phi_i := \mathbb{E}_\pi[\Delta_i(\pi)] = m_i + \bar{u}_i, \qquad \bar{u}_i := \mathbb{E}_\pi[u_i(\pi)]. \tag{70}$$

Under a uniform permutation, i is last inside a fixed coalition $i \in T$ with probability $1/|T|$, so $\bar{u}_i = \sum_{i \in T,\, |T| \geq 2} m(T)/|T|$. Centering the marginal therefore yields the identity

$$\Delta_i(\pi) - \phi_i = u_i(\pi) - \bar{u}_i. \tag{71}$$

Centering removes two permutation-invariant singleton m_i and the mean collective allocation $\bar{u}_i$. Which remains is not the absolute importance of expert i, but the deviation of the current collective-allocation from its average.

A.3.3 Covariance as higher-order coupling. Define

$$C_{ij} := \mathrm{Cov}_\pi(\Delta_i(\pi), \Delta_j(\pi)) = \mathbb{E}_\pi\big[(\Delta_i - \phi_i)(\Delta_j - \phi_j)\big]. \tag{72}$$

Then, for every $i \neq j$,

$$C_{ij} = \mathbb{E}_\pi\big[(u_i - \bar{u}_i)(u_j - \bar{u}_j)\big] = \mathrm{Cov}_\pi(u_i, u_j). \tag{73}$$

Equation equation 73 is the coupling statement. C_{ij} aggregates all co-fluctuations generated by the exclusive order allocation equation 67, as π varies, a given cooperative increment $m(T)$ is reassigned among the members of T, and the covariance summarizes the resulting dependence. Thus C_{ij} is an *effective coupling* of the full-order game, not an exact second-order Möbius inversion, and not a unique represetation of $\{m(T)\}$.

A.4 Algorithm of SetPrune

A.5 Details of the MoE models architecture

Table 4 summarizes the MoE models used in our experiments, including the total and activated parameter counts as well as the number of routed experts and the number of experts activated per token.

A.6 Detailed results of the zero-shot task

Algorithm 1 SetPrune: Set-Level Expert Pruning for MoE Models

Require: MoE model with global expert set $\mathcal{N} = \{E_l^i\}$, calibration dataset $\mathcal{D}$, Monte Carlo samples M, angle threshold θ_τ, truncation threshold τ, shrinkage coefficient α, pruning ratio r
Ensure: Pruned MoE model

Compute full-model value $V_{\text{full}} \leftarrow v(\mathcal{N}) = 1/\text{PPL}(\mathcal{N})$ on $\mathcal{D}$ ▷ Full-set calibration
for each MoE layer $l = 1$ to L **do**
  $\mathbf{W}^{f(l)} \leftarrow \mathbf{W}^{(l)} \odot \boldsymbol{\gamma}^{(l)}$ ▷ Eq. (11)
  $\bar{\mathbf{W}}^{(l)} \leftarrow \mathbf{W}^{f(l)} - \text{mean}_e(\mathbf{W}^{f(l)}_{e,:})$ ▷ Eq. (12)
end for
Initialize $\Delta_i^{(m)} \leftarrow 0$ for all experts i and samples m
for $m = 1$ to M **do** ▷ Monte Carlo loop
  Sample a permutation $\pi^{(m)}$ of the global expert set $\mathcal{N}$ ▷ Eq. (7)
  $\mathcal{S} \leftarrow \mathcal{N}$, $V \leftarrow V_{\text{full}}$
  for $j = |\mathcal{N}|$ down to 1 **do** ▷ Permutation traversal
    $i \leftarrow \pi_j^{(m)}$
    if $V < \tau V_{\text{full}}$ **then**
      break ▷ Early truncation, Eq. (17)
    end if
    Compute routing angle $\theta_{l,i}$ using current hidden states and $\bar{\mathbf{W}}^{(l)}$ ▷ Eqs. (13)–(15)
    $\mathcal{S}' \leftarrow \mathcal{S} \setminus \{i\}$
    if $\theta_{l,i} > \theta_\tau$ **then**
      $\Delta_i^{(m)} \leftarrow 0$ ▷ Angle-based skipping
      $\mathcal{S} \leftarrow \mathcal{S}'$ ▷ Eq. (16)
    else
      Evaluate $V' \leftarrow v(\mathcal{S}')$ on $\mathcal{D}$
      $\Delta_i^{(m)} \leftarrow V - V'$ ▷ Eq. (9)
      $\mathcal{S} \leftarrow \mathcal{S}'$, $V \leftarrow V'$
    end if
  end for
end for
for each MoE layer $l = 1$ to L **do** ▷ Möbius-based contribution construction
  $\bar{\boldsymbol{\Delta}}_l \leftarrow \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\Delta}_l^{(m)}$
  $\widehat{\boldsymbol{\Sigma}}^{(l)} \leftarrow \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{\Delta}_l^{(m)} - \bar{\boldsymbol{\Delta}}_l)(\boldsymbol{\Delta}_l^{(m)} - \bar{\boldsymbol{\Delta}}_l)^\top$ ▷ Eq. (20)
  $\widetilde{\boldsymbol{\Sigma}}^{(l)} \leftarrow (1-\alpha)\widehat{\boldsymbol{\Sigma}}^{(l)} + \alpha\ \text{diag}(\widehat{\boldsymbol{\Sigma}}^{(l)})$ ▷ Eq. (21)
  for each expert i in layer l **do**
    $\eta_{l,i} \leftarrow \bar{\Delta}_{l,i} - \sum_{j\neq i} \widetilde{\Sigma}_{ij}^{(l)}$ ▷ Eq. (22)
  end for
  $R_l \leftarrow E_l - \lfloor rE_l \rfloor$ ▷ Retained-expert budget
  Solve

$$\mathbf{q}^{*(l)} = \arg\min_{\mathbf{q}\in\{0,1\}^{E_l}} \sum_i \eta_{l,i}(1-q_i) + \sum_{i<j} \widetilde{\Sigma}_{ij}^{(l)}(1-q_iq_j)$$

  s.t. $\sum_i q_i = R_l$ ▷ Eqs. (23)–(25)
  Linearize q_iq_j with auxiliary variables y_{ij} and solve by MILP ▷ Eq. (26)
  Prune all experts with $q_i^{*(l)} = 0$
end for
return Pruned MoE model

Table 4: Architecture of the MoE models used in our study. Experts and activated experts refer to *routed* experts (top-k). Shared experts, if any, are always active and are not pruned.

Model	Total params	Activated params	Routed experts	Top-k	Shared experts
Qwen1.5-MoE-A2.7B Qwen Team (2024)	14.3B	2.7B	60	4	1
Qwen3-30B-A3B Qwen Team (2025)	30.5B	3.3B	128	8	0
Qwen3.5-35B-A3B Qwen Team (2026)	35B	3B	256	8	1
DeepSeekMoE-16B Dai et al. (2024)	16.4B	2.8B	64	6	2
GPT-OSS-20B OpenAI (2025)	20.9B	3.6B	32	4	0

Table 5: Zero-shot 25% pruned performance on seven commonsense benchmarks. Optimal and sub-optimal pruned results are shown in bold and underlined respectively.

Model	Method	HellaSwag	Winogrande	BoolQ	OBQA	PIQA	ARC-e	ARC-c	Mean
Qwen1.5-MoE-A2.7B	None	58.06	68.90	79.69	30.40	79.71	73.27	41.30	61.62
	Random	51.20	62.38	71.35	25.60	72.88	63.98	36.11	54.79
	Gating Score	52.10	63.01	72.41	25.80	73.01	64.35	37.24	55.42
	Frequency	52.21	63.10	71.90	25.90	73.19	65.31	37.78	55.63
	Expert Trim	51.89	63.12	71.90	25.00	73.20	63.90	37.14	55.16
	Shapley-MoE	53.90	65.68	74.89	27.00	74.09	65.01	39.12	57.10
	REAP	44.02	58.41	63.30	19.00	66.97	49.41	25.34	46.64
	AIMER	52.94	**70.17**	**77.65**	25.60	76.61	68.81	39.68	58.78
	SetPrune(Ours)	**55.01**	67.17	69.79	**29.00**	**79.76**	**74.79**	**41.30**	**59.54**
Qwen3-30B-A3B	None	59.69	70.88	88.56	34.60	79.71	79.29	52.47	66.46
	Random	49.52	62.77	81.87	28.60	69.46	69.24	42.43	57.69
	Gating Score	50.23	62.98	82.18	29.20	70.03	70.23	43.01	58.27
	Frequency	50.68	63.20	82.56	29.60	70.27	70.56	43.90	58.68
	Expert Trim	50.64	63.37	82.01	29.10	69.89	69.79	43.09	58.27
	Shapley-MoE	52.35	65.01	84.12	30.10	71.56	71.23	45.80	60.02
	REAP	45.40	59.59	73.76	23.00	70.08	61.57	34.64	52.58
	AIMER	56.39	69.77	**87.13**	31.60	77.53	**77.48**	**51.62**	**64.50**
	SetPrune(Ours)	**57.70**	**71.43**	86.79	**35.00**	**79.49**	74.24	43.94	64.08
Qwen3.5-35B-A3B	None	63.39	74.43	85.32	33.00	82.15	81.73	54.44	67.78
	Random	56.72	73.72	**89.51**	30.60	79.38	79.29	52.99	66.03
	Gating Score	61.10	71.61	87.19	32.30	79.76	81.76	**57.31**	67.29
	Frequency	61.46	72.43	87.45	32.20	80.32	80.62	55.00	67.07
	Expert Trim	**62.95**	**75.14**	88.53	**33.20**	**82.26**	79.80	50.43	67.47
	Shapley-MoE	–	–	–	–	–	–	–	–
	REAP	62.04	73.64	84.92	**33.20**	80.85	**83.12**	56.57	**67.76**
	AIMER	61.31	72.77	87.77	32.60	80.03	81.10	56.48	67.44
	SetPrune(Ours)	62.53	73.98	87.26	32.30	81.11	81.33	54.37	67.55
DeepSeekMoE-16B	None	58.08	69.93	73.73	32.60	78.62	74.96	43.34	61.61
	Random	46.42	64.04	69.46	24.80	70.33	63.60	34.92	53.37
	Gating Score	47.09	65.35	70.56	25.30	71.26	64.32	35.75	54.23
	Frequency	47.67	66.09	71.05	25.40	71.87	64.09	35.90	54.58
	Expert Trim	46.45	64.95	69.89	24.80	70.43	63.89	35.86	53.75
	Shapley-MoE	50.13	67.15	71.45	28.00	72.09	66.45	38.90	56.31
	REAP	47.84	62.19	67.16	23.60	75.08	68.39	34.13	54.06
	AIMER	53.35	**69.06**	**72.48**	26.60	73.94	67.80	**40.70**	57.70
	SetPrune(Ours)	**55.78**	68.19	71.38	**31.40**	**78.62**	**72.31**	39.51	**59.60**
GPT-OSS-20B	None	41.49	65.98	75.41	26.40	75.41	77.69	45.65	58.29
	Random	40.68	64.40	74.25	25.40	72.52	69.61	36.77	54.81
	Gating Score	40.72	65.27	74.56	26.40	74.76	75.34	40.53	56.80
	Frequency	40.38	**65.82**	73.94	**27.40**	74.92	74.83	40.19	56.78
	Expert trim	39.47	62.35	73.70	25.00	73.34	70.54	36.60	54.43
	Shapley-MoE	–	–	–	–	–	–	–	–
	REAP	40.89	64.09	64.95	22.20	71.11	63.55	32.59	51.34
	AIMER	39.05	61.88	65.26	23.00	70.57	63.17	33.19	50.87
	SetPrune(Ours)	**46.51**	65.75	**75.26**	24.80	**75.73**	**76.18**	**42.06**	**58.04**

Table 6: Zero-shot 50% pruned performance on seven commonsense benchmarks. Optimal and sub-optimal pruned results are shown in bold and underlined respectively.

Model	Method	HellaSwag	Winogrande	BoolQ	OBQA	PIQA	ARC-e	ARC-c	Mean
Qwen1.5-MoE-A2.7B	None	58.06	68.90	79.69	30.40	79.71	73.27	41.30	61.62
	Random	40.27	57.72	60.25	18.80	62.49	49.88	27.33	45.25
	Gating Score	42.90	58.93	61.65	19.40	62.90	50.28	28.01	46.30
	Frequency	43.88	59.46	61.90	19.80	63.37	50.80	28.69	46.84
	Expert Trim	42.98	58.57	61.89	19.10	62.96	50.57	28.35	46.35
	Shapley-MoE	45.78	60.90	65.09	21.50	64.97	53.96	30.89	49.01
	REAP	32.21	52.57	61.87	17.40	57.45	34.26	21.84	39.66
	AIMER	45.10	64.96	72.81	22.40	71.44	61.45	36.43	53.51
	SetPrune(Ours)	**48.71**	**65.27**	**73.12**	**28.20**	**77.58**	**70.96**	**38.40**	**57.46**
Qwen3-30B-A3B	None	59.69	70.88	88.56	34.60	79.71	79.29	52.47	66.46
	Random	36.36	56.06	66.16	20.00	58.68	48.78	28.69	44.96
	Gating Score	38.89	56.89	68.98	20.70	58.87	49.56	29.65	46.22
	Frequency	39.45	57.90	67.90	20.80	58.98	49.79	30.98	46.54
	Expert Trim	38.87	57.10	69.12	20.90	58.78	49.16	29.98	46.27
	Shapley-MoE	41.95	59.12	71.23	22.30	60.14	51.24	31.98	48.28
	REAP	27.84	49.41	51.16	13.60	56.91	37.16	17.75	36.26
	AIMER	43.05	64.40	75.66	24.20	66.65	61.87	33.11	52.70
	SetPrune(Ours)	**53.70**	**68.75**	**83.24**	**33.00**	**77.91**	**70.12**	**41.47**	**61.17**
Qwen3.5-35B-A3B	None	63.39	74.43	85.32	33.00	82.15	81.73	54.44	67.78
	Random	47.35	68.82	**87.52**	22.40	73.78	74.20	44.54	59.80
	Gating Score	56.41	69.85	83.17	28.40	75.61	74.35	48.01	62.26
	Frequency	56.50	69.77	82.93	27.40	75.50	74.30	49.12	62.22
	Expert Trim	**58.85**	**72.93**	87.16	29.80	**78.67**	74.92	45.99	64.05
	Shapley-MoE	–	–	–	–	–	–	–	–
	REAP	53.00	69.93	83.12	29.60	75.19	**80.72**	53.33	63.56
	AIMER	47.95	68.90	83.88	31.20	73.12	80.30	**54.44**	62.83
	SetPrune(Ours)	57.70	72.37	83.73	**31.30**	77.71	78.67	48.45	**64.28**
DeepSeekMoE-16B	None	58.13	69.93	73.46	32.40	78.40	74.79	43.86	61.57
	Random	33.67	55.35	58.93	16.00	59.55	43.49	23.49	41.50
	Gating Score	34.67	56.36	60.01	17.10	60.12	44.23	25.09	42.51
	Frequency	33.90	57.76	60.13	18.00	59.89	44.21	26.00	42.84
	Expert Trim	34.56	57.01	59.79	18.10	59.81	43.13	26.29	42.67
	Shapley-MoE	36.98	59.01	61.34	19.30	62.02	46.67	28.90	44.89
	REAP	36.81	57.06	56.61	19.00	68.23	47.39	22.70	43.97
	AIMER	43.39	65.67	**71.50**	20.20	68.50	54.97	**32.59**	50.97
	SetPrune(Ours)	**49.92**	**65.75**	64.59	**25.00**	**75.41**	**66.04**	31.74	**54.06**
GPT-OSS-20B	None	41.49	65.98	75.41	26.40	75.41	77.69	45.65	58.29
	Random	32.65	57.46	66.91	17.60	63.06	52.31	22.01	44.57
	Gating Score	37.92	62.43	**75.29**	23.80	72.52	67.85	33.53	53.33
	Frequency	38.34	60.93	74.56	**24.60**	72.63	67.89	32.42	53.05
	Expert trim	38.77	59.19	75.05	21.80	71.93	63.76	30.03	51.50
	Shapley-MoE	–	–	–	–	–	–	–	–
	REAP	29.48	52.72	54.46	16.20	61.43	40.82	18.00	39.02
	AIMER	31.68	56.35	60.46	17.00	62.35	47.14	21.16	42.31
	SetPrune(Ours)	**44.42**	**64.72**	66.09	23.60	**74.37**	**69.49**	**36.09**	**54.11**

Table 7: Zero-shot accuracy (%) of angle-threshold and performance-ratio ablations on Qwen1.5-MoE-A2.7B. Angle 92° is excluded. Best score in each column within a prune ratio is in **bold**. Angle 92° and Ratio 0.5 are our default setting which we reported in main paper.

Prune	Setting	HellaSwag	WinoG.	BoolQ	OBQA	PIQA	ARC-e	ARC-c	Avg
25%	Angle 90°	52.50	65.43	**75.11**	28.20	76.77	68.43	34.73	57.31
	Angle 91°	53.64	64.01	69.05	27.80	78.51	69.91	38.31	57.32
	Angle 93°	55.68	65.82	73.52	29.60	78.73	71.17	39.68	59.17
	Angle 94°	**55.84**	66.93	74.22	29.60	79.27	71.21	39.59	**59.52**
	Ratio 0.3	55.63	66.22	66.91	30.80	79.11	**74.20**	**40.02**	58.98
	Ratio 0.4	55.28	67.56	65.99	30.80	**79.43**	71.59	36.95	58.23
	Ratio 0.6	55.29	66.54	67.19	30.40	79.16	69.11	36.60	57.75
	Ratio 0.7	55.58	**67.88**	67.13	**31.00**	79.16	70.16	36.35	58.18
50%	Angle 90°	41.74	54.70	64.77	21.80	73.07	56.40	28.24	48.67
	Angle 91°	47.69	63.46	63.85	25.20	75.52	61.99	31.40	52.73
	Angle 93°	49.49	62.12	63.43	25.20	76.44	61.99	33.19	53.12
	Angle 94°	**49.68**	63.06	59.11	24.40	76.44	59.43	32.25	52.05
	Ratio 0.3	49.42	**63.85**	**66.57**	25.00	75.63	64.98	**36.01**	**54.49**
	Ratio 0.4	48.98	62.83	65.72	25.40	75.24	64.31	33.87	53.77
	Ratio 0.6	49.64	62.35	64.59	**26.60**	**76.55**	**65.57**	34.39	54.24
	Ratio 0.7	48.78	62.67	62.26	24.60	74.76	63.89	32.25	52.74