SHAPE: Coalition-Aware Expert Pruning for Sparse Mixture-of-Experts LLMs

1st Yuhao Zhang
School of Computer Science and Engineering
Beihang University
Beijing, China
yuhao_zh@buaa.edu.cn

2nd HongXu Jiang*
School of Computer Science and Engineering
Beihang University
Beijing, China
jianghx@buaa.edu.cn

3rd YiXiang Zhang
School of Computer Science and Engineering
Beihang University
Beijing, China
by2242227@buaa.edu.cn

4th Zheng Zhang
School of Computer Science and Engineering
Beihang University
Beijing, China
zz2643037152@buaa.edu.cn

***Abstract*—Sparse Mixture-of-Experts (MoE) large language models achieve high quality with low per-token compute, yet their deployment is often limited by the memory wall: the full expert pool must remain resident to support token-dependent routing. Expert pruning is a direct remedy, but prior criteria often score experts independently and ignore that MoE inference is inherently *coalitional*, where outputs arise from routed top-k expert combinations. We present SHAPE, a task-driven pruning framework that explicitly models *intra-layer* expert cooperation. SHAPE formulates routing traces on a small calibration set as an empirical cooperative game and assigns interaction-aware expert values via a Shapley-style attribution over observed top-k coalitions, enabling the identification of experts that are essential for high-utility collaborations rather than simply frequent. To preserve MoE topology under a global pruning budget, SHAPE further introduces a *quality-coverage* selection rule that retains, in each layer, the minimal expert subset covering an α fraction of Shapley mass, while using bisection to match a target keep rate. Experiments on three modern MoE backbones (Qwen3-30B-A3B, GPT-OSS-20B, and DeepSeek-V2-Lite) across diverse benchmarks show that SHAPE consistently improves robustness over global and layer-wise pruning variants, maintaining accuracy under 20% and 40% expert pruning without additional training and delivering clear reductions in peak GPU memory footprint. The open-source code is available at https://github.com/Alizen-1009/Shapley-Moe.**

***Index Terms*—Mixture-of-Experts, Large Language Models, Pruning, Shapley Value, Expert Cooperation**

I. Introduction

The landscape of Large Language Models (LLMs) has been dramatically reshaped by the widespread adoption of Sparse Mixture-of-Experts (SMoE) architectures, where each MoE layer uses a router to select a subset of experts per token. Foundational works such as GShard [1] and Switch Transformers [2], followed by recent state-of-the-art open models including gpt-oss [3] and DeepSeek-V3.2 [4], have shown that model capacity can be scaled to hundreds of billions or even trillions of parameters with sub-linear growth in training cost. By decoupling total parameter count from per-token active computation, SMoE architectures enable a beneficial trade-off between model capacity and inference latency, leading to strong performance on complex reasoning and coding tasks [5].

However, despite their favorable computational efficiency, the deployment of massive MoE models encounters a critical bottleneck: the "Memory Wall." While conditional computation ensures low floating-point operations (FLOPs) per token, the *entire* expert pool must reside in GPU memory to support dynamic, token-dependent routing decisions. For example, serving a 671B-parameter model such as DeepSeek-V3.2 requires a multi-node GPU cluster with terabytes of VRAM, regardless of its sparse activation pattern [4]. This prohibitive memory requirement fundamentally limits the deployability of high-capacity MoE models in resource-constrained environments, motivating an urgent need for post-training compression techniques that can substantially reduce the served parameter footprint while preserving the functional behavior of routed expert subnetworks.

Among emerging compression strategies, *expert pruning* has emerged as a prominent approach for reducing the parameter footprint of Mixture-of-Experts models by selectively removing experts. Unlike quantization [6], [7], which reduces numerical precision, or expert merging [8]–[10], which fuses expert parameters and may alter underlying functional subspaces [11], pruning aims to reduce model size by selecting a subset of experts to serve at inference time. Early pruning approaches predominantly relied on heuristic, expert-wise metrics, such as activation frequency [12] or gating magnitude [13]. More recent methods incorporate training dynamics or activation statistics to refine expert-wise saliency estimation, including router-weight shift metrics [14] and activation-aware scoring methods [15], [16].

Despite these methodological advancements, a fundamental modeling gap remains: existing pruning criteria predominantly

*Corresponding author: HongXu Jiang (jianghx@buaa.edu.cn).

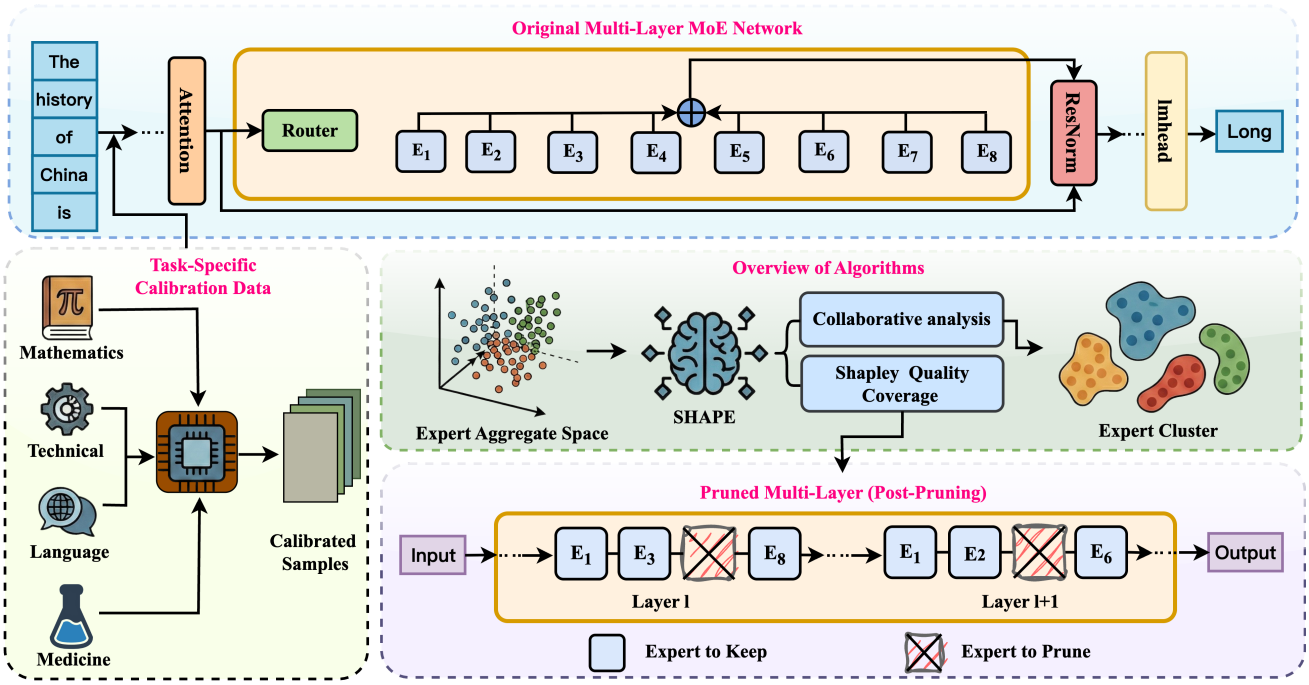

Fig. 1. Overview of SHAPE. We run an offline calibration pass on a small task-specific dataset to extract routing/co-activation traces and estimate cooperative expert contributions. Based on these estimates, SHAPE performs Shapley quality-coverage expert selection to prune redundant experts, producing a compact MoE model for efficient serving.

adopt an *expert-wise* perspective [17]–[19], implicitly assuming that experts contribute independently to the model output. This assumption conflicts with the intrinsic nature of MoE inference, which is inherently *coalitional*. For each token, the model output is produced by a non-linear composition of a specific coalition of top-k experts selected by the router [20]. Consequently, the functional utility of an expert is fundamentally conditioned on the set of co-activated experts within the routed coalition; an expert may exhibit limited standalone utility yet provide indispensable complementary functionality when paired with semantically dominant experts [21]. Conversely, an expert with high routing frequency may be functionally redundant if its contribution is subsumed by other members of the coalition. By ignoring such intra-layer cooperative interactions, expert-wise pruning risks dismantling high-utility coalitions, resulting in disproportionate performance degradation on complex downstream tasks.

To address this modeling limitation, we introduce **SHAPE** (SHapley-Aware Pruning of Experts), a Shapley-guided pruning framework that models expert selection cooperatively for practical deployment. Rather than relying on static, expert-wise heuristics, SHAPE formulates MoE inference on a calibration set [22], [23] as a cooperative game, where the routed top-k experts constitute a coalition of players. We employ Shapley values [24], the unique valuation scheme satisfying standard fairness axioms, to attribute each expert's marginal contribution to the utility of the routed coalition. This interaction-aware valuation enables SHAPE to distinguish experts that are merely frequently selected from those that are cooperatively essential to high-utility coalitions. Furthermore, to avoid structural collapse caused by overly aggressive pruning, we introduce a *Shapley Quality Coverage* mechanism that enforces layer-wise retention based on the distribution of cooperative value.

Figure 1 illustrates the SHAPE workflow. Our contributions are summarized as follows:

1) **Coalitional Expert Valuation.** We introduce a cooperative formulation of MoE pruning based on Shapley values, providing a principled mechanism to attribute expert contributions conditioned on routed coalitions. This work reframes pruning from expert-wise importance estimation to coalition-aware contribution modeling.
2) **Structure-Preserving Pruning Strategy.** We propose a topology-aware selection mechanism that preserves layer-wise cooperative capacity by retaining experts according to their cumulative Shapley value, rather than enforcing a fixed expert count.
3) **Empirical Validation of Cooperative Pruning.** Extensive experiments on modern MoE architectures, including Qwen3-MoE, GPT-OSS, and DeepSeek-V2-Lite, demonstrate that preserving expert cooperation is critical for effective compression. SHAPE achieves up to 40% expert pruning with negligible performance degradation on reasoning-intensive benchmarks.

II. Related Work

A. MoE Compression and Expert Pruning

MoE LLMs achieve *conditional computation* by routing tokens to a sparse subset of experts, but practical serving still

requires storing and accessing a large expert pool, resulting in substantial memory footprint and bandwidth pressure [25], [26]. To reduce deployment cost, prior work explores MoE compression via structured pruning [27], [28], expert merging or reparameterization [8]–[10], quantization [6], [7], and knowledge distillation [29], [30]. Among them, *expert pruning* is particularly attractive for serving because it directly shrinks the number of experts that must be loaded, cached, and dispatched, yielding immediate reductions in memory and inference overhead.

Recent pruning criteria are often training-free or lightweight and typically rely on *expert-wise* signals. SEAP [15] leverages sparse activation statistics to prune experts under fixed budgets, and domain-oriented methods use few-shot demonstrations to select experts that best support target evaluations [28]. While effective, these approaches largely score experts in isolation and may overlook interaction effects in top-k routing, where multiple experts are simultaneously activated and aggregated within each MoE layer.

B. Expert Cooperation and Routing-Structured Pruning

A distinctive characteristic of MoE inference is that the output is formed by aggregating a routed top-k *expert coalition*. This coalition structure implies that an expert's utility can be highly context-dependent: removing one expert may alter the behavior of remaining co-activated experts, and the impact may not be predictable from marginal activation frequency alone. Motivated by this, recent studies analyze and exploit structured cooperation patterns in MoE. MoE Pathfinder [31] leverages routing trajectories to guide expert selection, and HSDL [32] reveals hidden collaboration patterns via expert co-activation structures that encode functional roles. These works suggest that *intra-layer* co-activation is a meaningful signal for pruning and that routing distributions contain rich information beyond global expert usage.

Nevertheless, existing routing-aware approaches typically rely on trajectory or co-activation patterns as heuristics, rather than a principled measure of *marginal contribution* under routed coalitions. In particular, the within-layer coalition effect—how much an expert helps *in combination* with other simultaneously activated experts under task-driven routing—remains underexplored in pruning objectives. Our work targets this gap by valuing experts through their cooperative contributions to routed coalitions, leading to pruning decisions that better preserve task-relevant collaboration.

C. Shapley-Based Cooperative Attribution and Scalable Approximation

Shapley value [24] is a canonical cooperative game-theoretic attribution that quantifies each participant's average marginal contribution across all coalitions. It has been applied to neural networks for identifying responsible components and guiding structured pruning, demonstrating advantages in capturing redundancy and interaction effects compared with purely magnitude-based criteria [33], [34]. However, exact Shapley computation is NP-hard, making scalable approximation essential for large models.

Practical Shapley estimation commonly relies on Monte Carlo sampling and related estimators [35], [36]. Despite these advances, most Shapley-based methods focus on neurons/channels or data valuation, and do not directly address conditional computation architectures where coalition formation is governed by routing. MoE models are a natural candidate for cooperative attribution because inference is inherently coalition-based, yet Shapley value has not been systematically developed as an expert-level pruning criterion under task-driven routing distributions. Our work bridges this gap by introducing an efficient Shapley-style approximation tailored to routed expert coalitions for MoE expert selection.

III. Methodology

This section introduces **SHAPE**, a task-conditioned expert pruning framework for Sparse Mixture-of-Experts (SMoE) models. Given a pretrained MoE language model and a target downstream task, SHAPE prunes a large fraction of experts *without retraining* while maintaining task performance.

SHAPE is motivated by the coalitional nature of SMoE inference: at each MoE layer, routing activates a top-k set of experts whose outputs are jointly aggregated, so an expert's value depends on its *cooperative contribution* within routed coalitions rather than in isolation. To capture such interaction effects, SHAPE performs (i) *collaborative analysis* on a small task-specific calibration set to estimate per-layer cooperative contributions via an efficient Shapley-style approximation, and (ii) *Shapley Quality Coverage Pruning* to select a compact expert subset that satisfies a global keep-rate budget while preserving layer-wise contribution mass.

A. Task-Driven Calibration and Notation

We consider a pretrained Sparse Mixture-of-Experts (SMoE) language model with L MoE layers. Each MoE layer $\ell \in \{1, \dots, L\}$ contains a fixed set of experts denoted by $N_\ell = \{E_1, \dots, E_n\}$. Given a token hidden representation h, the router at layer ℓ computes gating scores over experts and activates a fixed number k of experts according to a top-k routing policy. The output of the MoE layer is obtained by aggregating the outputs of the selected experts. In this work, we do not modify the routing policy or expert parameters, and instead focus on the routing behavior induced by the pretrained model.

Let $\mathcal{T}$ denote the set of downstream tasks. For each task $t \in \mathcal{T}$, we construct a small calibration dataset $\mathcal{D}_t^{\text{cal}}$ using the same prompt format as the evaluation set. The purpose of calibration is not to adapt the model, but to expose the routing policy to inputs drawn from the target task distribution. We perform standard inference with the unpruned model on $\mathcal{D}_t^{\text{cal}}$ and record the routing decisions made at all MoE layers for every processed token.

During inference, each token passing through an MoE layer ℓ activates exactly k experts. Collecting these routing traces

yields a set of expert combinations that occur under the task-specific input distribution. For a fixed task t and layer ℓ, we denote by

$$\mathcal{S}_\ell^t = \{A_1, A_2, \ldots, A_J\}, \quad A_j \subset N_\ell,\ |A_j| = k, \tag{1}$$

the set of all distinct expert combinations that appear at least once during inference on $\mathcal{D}_t^{\text{cal}}$. Although the number of possible combinations is $\binom{n}{k}$, the number of observed combinations J is typically much smaller, reflecting the structured routing behavior of the model on a specific task.

For each observed expert combination $A_j \in \mathcal{S}_\ell^t$, we record its activation frequency, defined as the total number of times this exact set of experts is selected by the router across all tokens in $\mathcal{D}_t^{\text{cal}}$. These activation frequencies summarize how expert capacity is allocated by the pretrained model when processing the target task, and serve as empirical statistics characterizing task-conditioned expert cooperation.

Based on these routing traces, SHAPE estimates task- and layer-conditioned expert contribution scores, selects retained sets $\{S_{t,\ell}^*\}_{\ell=1}^L$ under a target keep-rate budget, and exports the corresponding pruned model. Throughout the remainder of this section, we restrict analysis to the empirical coalition space induced by top-k routing and do not perform additional forward evaluations beyond standard inference.

B. Coalition-Based Utility and Shapley-Style Expert Valuation

SHAPE evaluates expert importance from an empirical cooperative-game perspective based solely on routing traces observed during task-specific inference. This design targets deployment-oriented pruning scenarios, where repeated forward evaluation or retraining is impractical. Instead of defining coalition utility via loss re-evaluation, we treat the router's activation behavior itself as a proxy for task relevance and construct an empirical game over expert combinations actually used by the model.

For a fixed task t and MoE layer ℓ, let $N_\ell = \{E_1, \ldots, E_n\}$ denote the expert set (players). Running the *unpruned* model on the calibration set $\mathcal{D}_t^{\text{cal}}$, the top-$k$ router induces an *empirical support* of observed coalitions $\mathcal{S}_\ell^t$ (Eq. (1)). We then define a *router-induced empirical game* on this support by assigning each observed coalition an empirical utility proxy $v(A_j)$ equal to its activation frequency (the number of times A_j is selected across all tokens in $\mathcal{D}_t^{\text{cal}}$). This proxy is lightweight and task-conditioned; it reflects where the model allocates conditional capacity under task-matched routing traces.

For an expert $E_i \in N_\ell$, we focus on the subset of coalitions in which it participates,

$$\mathcal{S}_{E_i} = \{A_j \in \mathcal{S}_\ell^t \mid E_i \in A_j\}. \tag{2}$$

While the frequencies $\{v(A_j)\}_{A_j \in \mathcal{S}_{E_i}}$ reflect how often expert E_i appears in routed coalitions, frequency alone does not capture its *contextual indispensability*, as it ignores whether comparable proxy utility can be achieved without this expert. To incorporate coalition context, we consider a (Shapley-inspired) marginal proxy of E_i within an observed coalition A_j,

$$v(A_j) - v(A_j \setminus \{E_i\}), \tag{3}$$

where $A_j \setminus \{E_i\}$ denotes the $(k-1)$-expert sub-coalition obtained by removing E_i.

Such sub-coalitions are not directly observed under top-k routing. We therefore approximate their utility using co-occurrence statistics from $\mathcal{S}_\ell^t$. Specifically, for each $A_j \in \mathcal{S}_{E_i}$, we count how many observed coalitions contain $A_j \setminus \{E_i\}$ as a subset:

$$N_{ij} = \big|\{A \in \mathcal{S}_\ell^t \mid A_j \setminus \{E_i\} \subseteq A\}\big|. \tag{4}$$

Intuitively, N_{ij} captures the conditional co-occurrence strength of the remaining experts in A_j under the task-specific routing policy. A larger value indicates that these experts frequently co-activate in observed coalitions, suggesting a higher degree of substitutability for the removed expert. We emphasize that N_{ij} is *not* the classical utility of the $(k$–$1)$ sub-coalition (which is unobserved under top-k routing); it is a co-activation-based proxy designed to avoid additional forward passes while still reflecting empirical cooperation patterns.

Because N_{ij} aggregates counts over many supersets, it can *overestimate* the sub-coalition proxy, yielding $N_{ij} > v(A_j)$ and thus spurious negative marginals $v(A_j) - N_{ij} < 0$. This issue stems from our proxy construction rather than from the Shapley value definition itself (which does not require monotonicity or superadditivity of v). To obtain a monotonicity-consistent and conservative sub-coalition proxy on the observed support, we introduce an expert-specific scaling factor $\alpha_i \in (0, 1]$ and define

$$\widehat{v}(A_j \setminus \{E_i\}) = \alpha_i \cdot N_{ij}. \tag{5}$$

We select the largest admissible scaling factor that preserves monotonicity across all coalitions involving E_i:

$$\alpha_i = \min_{A_j \in \mathcal{S}_{E_i}} \frac{v(A_j)}{N_{ij}}. \tag{6}$$

This construction guarantees, for all observed coalitions A_j containing E_i, that

$$v(A_j) \geq \widehat{v}(A_j \setminus \{E_i\}) = \alpha_i N_{ij}, \tag{7}$$

thereby preventing the pathological case where the sub-coalition proxy exceeds the coalition proxy. In practice, this makes the resulting attribution *stable and conservative*: it avoids overestimating substitutability from co-occurrence counts, at the cost of potentially underestimating the strength of the $(k-1)$ proxy and thus biasing marginal contributions upward (favoring retention when uncertain).

Finally, we define a Shapley-style approximation of expert E_i's contribution as

$$\phi_{E_i}^{t,\ell} = \frac{1}{|\mathcal{S}_{E_i}|} \sum_{A_j \in \mathcal{S}_{E_i}} \frac{1}{|A_j|} \Big(v(A_j) - \widehat{v}(A_j \setminus \{E_i\}) \Big). \tag{8}$$

The factor $1/|A_j|$ reflects equal sharing of coalition utility among participating experts, analogous to the size-based weighting in the classical Shapley value. The resulting score captures both how frequently expert E_i participates in routed coalitions and how indispensable it is within those coalitions.

C. Expert Selection via Shapley Quality Coverage

Given task and layer specific expert contribution scores $\{\phi_i^{t,\ell}\}$, SHAPE performs expert pruning through a quality-coverage-based selection strategy. This stage is agnostic to how the contribution scores are obtained and focuses solely on using them to derive a structured, budget-controlled pruning decision. The key objective is to retain a subset of experts that preserves most of the cooperative value within each MoE layer, while satisfying a global keep-rate constraint.

For each task t and MoE layer ℓ, we define the total contribution mass of the layer as

$$T_{t,\ell} = \sum_{i \in N_\ell} \phi_i^{t,\ell}. \tag{9}$$

Negative contributions may arise due to estimation noise or destructive interactions and are excluded to avoid destabilizing the selection process.

Algorithm 1 Shapley Quality Coverage Expert Selection

Require: Contribution scores $\{\phi_i^{t,\ell}\}$; target keep rate r; tolerance ϵ
For each layer ℓ, set $\phi_i^{t,\ell} \leftarrow \max(\phi_i^{t,\ell}, 0)$ and $T_{t,\ell} \leftarrow \sum_{i \in N_\ell} \phi_i^{t,\ell}$
Set $\alpha_{\min} \leftarrow 0$, $\alpha_{\max} \leftarrow 1$
while $\alpha_{\max} - \alpha_{\min} > \epsilon$ **do**
  $\alpha \leftarrow (\alpha_{\min} + \alpha_{\max})/2$
  for $\ell = 1$ to L **do**
    Sort experts in N_ℓ by $\phi_i^{t,\ell}$ in descending order
    Select the smallest prefix set $S_{t,\ell}^*(\alpha)$ such that $\sum_{i \in S_{t,\ell}^*(\alpha)} \phi_i^{t,\ell} \geq \alpha \cdot T_{t,\ell}$
  end for
  $\text{keep\_rate}(\alpha) \leftarrow \frac{\sum_{\ell=1}^{L} |S_{t,\ell}^*(\alpha)|}{\sum_{\ell=1}^{L} |N_\ell|}$
  if $\text{keep\_rate}(\alpha) > r$ **then**
    $\alpha_{\max} \leftarrow \alpha$ {keep too many experts; decrease α}
  else
    $\alpha_{\min} \leftarrow \alpha$ {keep too few experts; increase α}
  end if
end while
$\alpha^* \leftarrow \alpha_{\min}$ {largest α with keep rate not exceeding r}
return $\{S_{t,\ell}^*(\alpha^*)\}_{\ell=1}^{L}$

Rather than retaining a fixed number of experts per layer, SHAPE enforces a *quality coverage* criterion. Specifically, for a coverage threshold $\alpha \in (0, 1]$, we select the smallest subset of experts $S_{t,\ell}^*(\alpha) \subseteq N_\ell$ such that

$$\sum_{i \in S_{t,\ell}^*(\alpha)} \phi_i^{t,\ell} \;\geq\; \alpha \cdot T_{t,\ell}. \tag{10}$$

This constraint ensures that an α fraction of the estimated cooperative value in each layer is preserved after pruning. Importantly, coverage is defined in terms of contribution mass rather than expert count, allowing layers with highly concentrated value to retain fewer experts while preventing layers with more diffuse contributions from being overly pruned.

The threshold α controls the trade-off between value preservation and pruning aggressiveness: larger α generally yields larger retained sets. SHAPE selects α by bisection to satisfy a target global keep rate r. For each candidate α, we form $\{S_{t,\ell}^*(\alpha)\}_{\ell=1}^{L}$ by taking, in every layer, the minimal prefix of experts (ranked by $\phi_i^{t,\ell}$) whose cumulative mass reaches α-coverage, and then measure the resulting fraction of retained experts across layers. We decrease α if this fraction is above r and increase it otherwise. Algorithm 1 summarizes the procedure.

Quality coverage serves as a structure-preserving constraint that mitigates layer collapse, a failure mode in which pruning concentrates retained experts in a small subset of layers. By enforcing value preservation independently at each layer, SHAPE maintains routing stability and functional diversity across depth, which is particularly important for deep MoE Transformers where different layers encode distinct transformations.

IV. Experiments

A. Experimental Setup

We evaluate SHAPE on three Mixtral-style sparse MoE LLM backbones (Qwen3-30B-A3B [5], GPT-OSS-20B [3], DeepSeek-V2-Lite [37]) across seven benchmarks covering math reasoning (GSM8K [22], MATH-500 [38]), code generation (HumanEval [23]), knowledge QA (GPQA-Diamond [39]), truthfulness (TruthfulQA [40]), sequence labeling (OntoNotes5 [41]), and medical QA (MedMCQA [42]).

For each task t, we collect routing traces from 25 calibration examples by running the unpruned model and recording per-layer top-k routing decisions and gating weights. We estimate expert importance using our routing-trace-based Shapley-style attribution and rank experts by the contribution mass $\phi_i^{t,\ell}$. SHAPE then applies the quality-coverage rule and uses bisection over a global threshold α to meet the target keep rate, performing pruning *without additional training*.

B. Pruning Results Across Backbones

We evaluate SHAPE across three sparse MoE backbones with different model scales and routing characteristics. Table I summarizes task performance under two pruning ratios (20% and 40%) together with the unpruned baselines.

Across all backbones, SHAPE preserves strong overall performance under moderate pruning. At a 20% pruning ratio, the average performance remains comparable to, and in several cases slightly exceeds, that of the unpruned models. This behavior indicates that a non-trivial fraction of experts contributes limited marginal utility under the evaluated tasks, and can be safely removed without harming task accuracy. As pruning becomes more aggressive, performance degradation increases but remains bounded across models, demonstrating that SHAPE can substantially reduce the expert pool while maintaining stable task competence. The impact varies by task family: reasoning- and coding-heavy benchmarks are generally more sensitive to expert removal than structured prediction or domain QA tasks.

We further observe clear differences across backbones. Larger and more expressive models, such as Qwen3-30B-A3B and GPT-OSS-20B, demonstrate stronger resilience to pruning, maintaining high accuracy even when 40% of experts

TABLE I
TASK ACCURACY OF THREE MOE LLM BACKBONES UNDER BASELINE AND PRUNING. "AVG" IS THE ARITHMETIC MEAN OVER *available* TASK SCORES IN THE ROW.

Model	Setting	GSM8K	HumanEval	GPQA-D	MATH-500	TruthfulQA	OntoNotes5	MedMCQA	Avg
Qwen3-30B-A3B	Baseline	96.21	95.73	58.59	96.80	77.60	87.15	68.35	82.92
	20% pruned	96.74	95.12	60.10	96.80	75.64	84.42	68.18	82.43
	40% pruned	95.60	89.63	67.68	94.40	73.16	82.37	66.32	81.31
GPT-OSS-20B	Baseline	94.24	95.73	61.62	92.40	69.28	93.17	68.40	82.12
	20% pruned	95.53	93.29	62.63	95.00	69.03	93.16	68.47	82.44
	40% pruned	95.07	85.41	56.57	92.20	66.10	91.25	66.54	79.02
DeepSeek-V2-Lite	Baseline	85.52	84.15	28.28	64.20	44.43	86.74	41.26	62.08
	20% pruned	83.17	82.56	31.52	62.55	46.58	87.21	43.50	62.44
	40% pruned	81.54	75.80	26.53	59.53	42.56	86.22	39.52	58.81

are removed. In contrast, DeepSeek-V2-Lite shows earlier performance degradation, reflecting a smaller redundancy margin and a more compact expert allocation. Nevertheless, even in this setting, SHAPE consistently preserves competitive performance, indicating that coalition-aware expert selection remains effective when model capacity is limited.

Overall, these results highlight that effective MoE pruning depends not only on identifying individually strong experts, but also on preserving cooperative structures that emerge during task execution. By accounting for expert contributions across coalition contexts, SHAPE provides a stable pruning mechanism that generalizes across backbones and task families.

C. Comparison with Baseline Pruning Criteria

We compare SHAPE with representative baseline pruning criteria across all evaluated backbones and report detailed task-level results on Qwen3-30B-A3B as a representative case; GPT-OSS-20B and DeepSeek-V2-Lite show similar trends.

Overall, simple pruning heuristics based on random selection, activation frequency, or router gating scores exhibit clear limitations. Although frequency- and gating-based methods outperform random pruning, they still incur noticeable performance degradation even at moderate pruning ratios. In particular, reasoning-intensive tasks such as GSM8K, HumanEval, and GPQA-Diamond are highly sensitive to these heuristics, suggesting that routing statistics alone are insufficient to capture task-relevant expert utility. Similar relative gaps between naive heuristics and task-aware methods are consistently observed across backbones.

Stronger baselines, including RAEP [16] and EASY-EP [28], substantially improve pruning stability by incorporating output-aware or task-aligned signals, but degrade more noticeably at higher pruning ratios.

Across all pruning ratios, SHAPE consistently achieves the strongest or near-strongest performance. Its advantage becomes most evident at 40% pruning, where maintaining a small number of critical experts is decisive. On Qwen3-30B-A3B, SHAPE retains substantially higher accuracy on GPQA-Diamond and GSM8K compared to all baselines, and the same qualitative behavior is observed on the other backbones. This consistency across models suggests that SHAPE captures backbone-agnostic properties of expert cooperation rather than exploiting architecture-specific artifacts.

Taken together, these results indicate that the primary limitation of existing pruning criteria lies in treating experts as largely independent units. In sparse MoE inference, however, effective computation emerges from structured cooperation among a small subset of experts. By explicitly valuing experts through their marginal contributions across coalition contexts, SHAPE more reliably preserves these cooperative structures, leading to improved robustness under aggressive pruning.

D. Ablation on Pruning Strategies

We analyze the effect of different pruning strategies under the same expert budget to isolate the role of the quality-coverage constraint in SHAPE. Table III reports average task performance at 20% and 40% pruning across three backbones, comparing SHAPE with two simplified variants: global pruning based on a single Shapley ranking and layer-wise pruning that selects top experts independently within each layer.

Across backbones and pruning ratios, SHAPE consistently achieves the most stable performance. In contrast, global pruning exhibits increasing instability as pruning becomes more aggressive. By selecting experts solely according to a global ranking, this strategy can over-prune layers that contain long-tail specialists while over-allocating capacity to others, leading to structural imbalance in the MoE topology. This effect becomes more pronounced at 40% pruning, where removing a small number of layer-specific experts can disproportionately degrade end-to-end performance.

Layer-wise pruning mitigates extreme layer collapse by enforcing a fixed budget per layer, but still underperforms SHAPE in most settings. While selecting experts independently within each layer preserves locally strong specialists, it fails to account for how capacity should be distributed across layers according to task demands. As a result, layers that require higher cooperative capacity for a given task may remain under-provisioned, limiting the model's ability to form effective expert coalitions.

TABLE II
BASELINE COMPARISON ON QWEN3-30B-A3B (INCLUDING THE UNPRUNED BASELINE AND 20%/40% PRUNING). BEST RESULTS UNDER THE SAME PRUNING RATE ARE IN BOLD. "AVG" IS THE ARITHMETIC MEAN OVER THE REPORTED TASKS.

Rate	Method	GSM8K	HumanEval	GPQA-D	MATH-500	TruthfulQA	OntoNotes5	MedMCQA	Avg
-	**Unpruned**	96.21	95.73	58.59	96.80	77.60	87.15	68.35	82.92
20%	Random	61.24	58.36	21.42	60.27	42.31	71.16	38.18	50.42
	Frequency	86.12	86.37	40.12	88.14	64.05	82.17	58.12	72.16
	Gating	87.42	87.15	42.18	89.36	65.41	83.24	59.27	73.43
	RAEP	95.47	94.52	57.56	95.42	76.42	86.14	67.51	81.86
	EASY-EP	95.82	95.01	58.12	95.86	**77.14**	**86.54**	67.96	82.35
	SHAPE (Ours)	**96.74**	**95.12**	**60.10**	**96.80**	75.64	84.42	**68.18**	**82.43**
40%	Random	35.43	28.27	8.28	32.15	21.29	60.41	15.46	28.76
	Frequency	75.14	72.11	20.16	72.29	45.18	76.28	40.17	57.33
	Gating	77.19	74.32	22.54	74.28	47.21	77.31	42.15	59.29
	RAEP	93.42	90.47	52.42	92.58	43.12	83.15	64.41	74.22
	EASY-EP	94.12	**91.07**	53.12	93.14	**73.28**	**83.51**	65.14	79.05
	SHAPE (Ours)	**95.60**	89.63	**67.68**	**94.40**	73.16	82.37	**66.32**	**81.31**

By contrast, SHAPE explicitly enforces a minimum retained Shapley mass per layer through the quality-coverage constraint. This mechanism prevents both global imbalance and rigid per-layer allocation, encouraging a balanced expert set that maintains cooperative structures across the network. The resulting pruned models exhibit consistently higher average performance, particularly under aggressive pruning, demonstrating that coverage-aware expert selection is crucial for preserving MoE functionality under constrained expert budgets.

TABLE III
ABLATION OF PRUNING STRATEGIES USING THE AVG METRIC AT 20% AND 40% PRUNING.

Strategy	Rate	Qwen3-Moe	GPT-OSS	DeepSeekV2
SHAPE	20%	82.43	82.44	62.44
	40%	81.31	77.53	58.81
Global	20%	82.21	81.62	61.53
	40%	80.49	76.92	57.64
Layer-wise	20%	81.53	80.85	60.86
	40%	79.58	75.83	56.84

E. Results and Analysis

Figure 2 reports the measured peak VRAM under 0/20/40% expert pruning for both 32K and 8K contexts. Although sparse MoE models activate only a small subset of experts per token, serving the full expert pool still incurs substantial memory overhead due to parameter storage and runtime bookkeeping. Across backbones and context lengths, pruning yields a clear and steady reduction in peak memory usage, with larger savings under more aggressive expert removal. This indicates that memory consumption in practical MoE serving is dominated by the size of the resident expert pool rather than by the number of activated experts during inference.

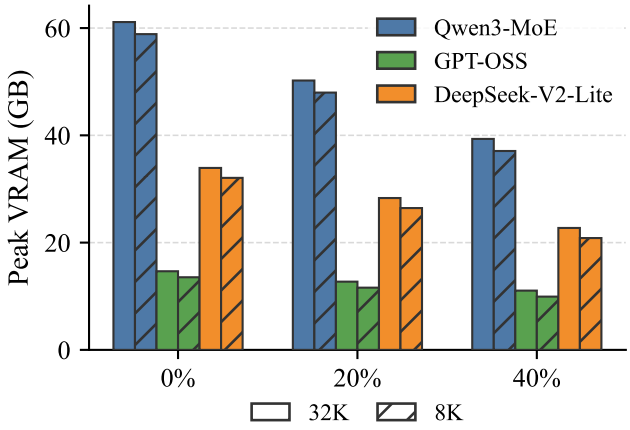

Fig. 2. Peak VRAM (GB) under 0/20/40% expert pruning (BS=1). Solid bars: 32K; hatched bars: 8K.

Importantly, the observed memory reduction is achieved without additional training or architectural modification, highlighting the practicality of pruning as a deployment-oriented optimization. Combined with the accuracy results in the previous sections, these findings suggest that SHAPE enables a favorable trade-off between model capacity and resource footprint, making large sparse MoE models more amenable to deployment under constrained GPU memory budgets.

V. CONCLUSION

We proposed **SHAPE**, a task-driven expert pruning framework for sparse MoE language models that respects the coalitional nature of routed top-k inference by valuing experts through cooperative contributions estimated from task-specific routing traces, and selecting experts with a layer-aware quality-coverage rule that matches a target keep rate via bisection.

Across multiple modern MoE backbones and diverse benchmarks, SHAPE consistently preserves accuracy better than global and layer-wise pruning variants under 20% and 40% pruning, while reducing the served expert footprint and peak GPU memory, making MoE deployment more practical in resource-constrained settings.

REFERENCES

[1] D. Lepikhin, Lee, and et.al, "Gshard: Scaling giant models with conditional computation and automatic sharding," in *International Conference on Learning Representations*, 2021.

[2] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[3] OpenAI *et al.*, "gpt-oss-120b and gpt-oss-20b model card," 2025.

[4] DeepSeek-AI *et al.*, "Deepseek-v3.2: Pushing the frontier of open large language models," 2025.

[5] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, and e. a. Chengen Huang, "Qwen3 technical report," 2025.

[6] E. Frantar and D. Alistarh, "Qmoe: Practical sub-1-bit compression of trillion-parameter models," in *Conference on Neural Information Processing Systems*, 2023.

[7] X. Hu, Chen, and et.al, "Moequant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[8] W. Li *et al.*, "Moe-svd: Structured mixture-of-experts llms compression via singular value decomposition," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[9] Y. Zhou, Z. Zhao, Cheng, and et.al, "Dropping experts, recombining neurons: Retraining-free pruning for sparse mixture-of-experts llms," in *Findings of the Association for Computational Linguistics: EMNLP 2025*.

[10] W. Chen *et al.*, "Retraining-free merging of sparse mixture-of-experts via hierarchical clustering," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[11] S. Bai, H. Li, J. Zhang, Z. Hong, and S. Guo, "Diep: Adaptive mixture-of-experts compression through differentiable expert pruning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[12] T. Chen, S. Huang, Y. Xie, B. Jiao, D. Jiang, H. Zhou, J. Li, and F. Wei, "Task-specific expert pruning for sparse mixture-of-experts," *CoRR*, vol. abs/2206.00277, 2022.

[13] A. Muzio, A. Sun, and C. He, "Seer-moe: Sparse expert efficiency through regularization for mixture-of-experts," 2024.

[14] M. N. R. Chowdhury and et.al, "A provably effective method for pruning experts in fine-tuned sparse mixture-of-experts," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[15] Liang *et al.*, "Seap: Training-free sparse expert activation pruning unlock the brainpower of large language models," *arXiv preprint arXiv:2503.07605*, 2025.

[16] M. Lasby, I. Lazarevich, N. Sinnadurai, S. Lie, Y. Ioannou, and V. Thangarasa, "Reap the experts: Why pruning prevails for one-shot moe compression," *arXiv preprint arXiv:2510.13999*, 2025.

[17] Y. Xu, X. Han, Y. Zhang, Y. Wang, Y. Liu, S. Ji, Q. Zhu, and W. Che, "Camera: Multi-matrix joint compression for moe models via micro-expert redundancy analysis," 2025.

[18] M. Cao, G. Li, J. Ji, J. Zhang, X. Ma, S. Liu, and L. Yin, "Condense, don't just prune: Enhancing efficiency and performance in moe layer pruning," *Transactions on Machine Learning Research (TMLR)*, 2025.

[19] Z. Zhang, X. Liu, H. Cheng, C. Xu, and J. Gao, "Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts," in *Findings of the Association for Computational Linguistics: ACL 2025*.

[20] N. Shazeer, Mirhoseini, and et.al, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations (ICLR)*, 2017.

[21] Z. Su, Q. Li, H. Zhang, W. Ye, Q. Xue, Y. Qian, Y. Xie, N. Wong, and K. Yuan, "Unveiling super experts in mixture-of-experts large language models," 2025.

[22] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," 2021.

[23] Q. Peng, Y. Chai, and X. Li, "Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics*.

[24] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games*, H. W. Kuhn and A. W. Tucker, Eds. Princeton University Press, 1953.

[25] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus, "St-moe: Designing stable and transferable sparse expert models," *Transactions on Machine Learning Research (TMLR)*, 2022.

[26] Mistral AI, "Mixtral of experts," 2023, model release / technical report.

[27] M. N. R. Chowdhury, M. Wang, K. E. Maghraoui, N. Wang, P.-Y. Chen, and C. Carothers, "A provably effective method for pruning experts in fine-tuned sparse mixture-of-experts," 2024.

[28] Z. Dong, H. Peng, P. Liu, X. Zhao, D. Wu, F. Xiao, and Z. Wang, "Domain-specific pruning of large mixture-of-experts models with few-shot demonstrations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[29] F. Xue, X. He, X. Ren, Y. Lou, and Y. You, "One student knows all experts know: From sparse to dense," in *Tiny Papers @ ICLR 2023*.

[30] L. Yang and J. Song, "Rethinking the knowledge distillation from the perspective of model calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[31] X. Yang, Y. Tian, and Y. Song, "Moe pathfinder: Trajectory-driven expert pruning," 2025.

[32] Y. Tang, Y. Tang, N. Zhang, M. Chen, and Y. Li, "Unveiling hidden collaboration within mixture-of-experts in large language models," 2025.

[33] A. Ghorbani and J. Zou, "Neuron shapley: Discovering the responsible neurons," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[34] M. Ancona, C. Oztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[35] A. Ghorbani, J. Zou, and A. Esteva, "Data shapley valuation for efficient batch active learning," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2022.

[36] K. Fan, Y. Yang, and C. Ma, "Enhancing interpretability for vision models via shapley value optimization," in *Proceedings of the 40th Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[37] DeepSeekAI *et al.*, "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," 2024.

[38] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," 2021.

[39] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q&a benchmark," in *First Conference on Language Modeling*, 2024.

[40] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," in *Proceedings of the 60th annual meeting of the association for computational linguistics*, 2022, pp. 3214–3252.

[41] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "OntoNotes: The 90% solution," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, Jun. 2006.

[42] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on health, inference, and learning*. PMLR, 2022, pp. 248–260.