STATIC PRESERVATION IS NOT ENOUGH: ADAPTATION-AWARE STRUCTURAL SEARCH FOR MOE PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) language models activate only a few experts per token but retain the full expert bank in memory during inference. Expert pruning reduces this deployment overhead by removing experts, but existing methods commonly evaluate retention decisions using layer-local proxies or a fixed model state. Such estimates can be unreliable because retention decisions interact across layers, while structures that preserve the current model well may not remain effective after subsequent adaptation. We therefore introduce **MASS**, a Model-Wide Adaptation-Aware Structural Search method, in which retention decisions remain revisable as the model evolves. Specifically, we evaluate candidate structures jointly across all MoE layers using full-model next-token prediction feedback. Under the target expert budget, a routing-aware objective guides expert allocation according to token-level native routing demand without favoring capacity expansion. We couple structural search with model adaptation throughout optimization, where current retention preferences define the structure used for model updates and feedback from the adapted model guides their subsequent refinement. Experiments on two distinct MoE architectures under two pruning severities show that MASS achieves the best overall recovered performance. Further analyses show that MASS exhibits favorable representation recovery while preserving native routing demand and limiting post-pruning representation distortion.

1 INTRODUCTION

Mixture-of-Experts (MoE) language models increase parameter capacity while activating only a small subset of experts for each token, thereby decoupling total capacity from per-token computation (Shazeer et al., 2017; Fedus et al., 2022).However, sparse activation does not reduce the expert parameters that must be stored and made accessible during inference (Chen et al., 2022; Lu et al., 2024). Expert pruning mitigates this deployment cost by permanently removing experts from a pretrained model. The core challenge is how to identify a compact expert structure that preserves model capability under a prescribed retention budget.

Existing methods determine which experts to retain using different forms of expert evaluation. Importance-based approaches rank individual experts using routing frequency, gate values, task specialization, or activation statistics (Chen et al., 2022; Lasby et al., 2025). Subset-based methods evaluate expert combinations through layer-reconstruction objectives or gradient-free search, while differentiable approaches optimize continuous expert-importance variables (Lu et al., 2024; Liu et al., 2024; Bai et al., 2025). Despite their different search mechanisms, these methods commonly evaluate candidates using signals local to individual layers or derived from a fixed model state. Such signals describe how experts are currently used, but do not directly characterize the model-wide quality of the retained structure or its behavior after adaptation.

Existing evaluation strategies face two key limitations. First, layer-local evaluation does not capture the model-wide effect of retention decisions. Removing experts from an earlier layer changes the hidden representations and routing distributions received by downstream layers, so locally favorable decisions may behave differently when composed across the network. Second, fixed-state evaluation does not account for how the relative quality of candidate structures may change during subsequent model adaptation. Consequently, a structure favored under static evaluation may not remain optimal after subsequent model adaptation.

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.

To address these limitations, we introduce **MASS** (**M**odel-Wide **A**daptation-Aware **S**tructural **S**earch), which jointly optimizes expert retention across the model while incorporating model adaptation into structural evaluation. We evaluate the model-wide impact of retention decisions across all MoE layers using the model's next-token prediction objective. To keep the continuous search faithful to the final discrete pruning setting, we complement this with two structural objectives. A differentiable budget objective aligns the soft retained capacity with the target pruning budget, while Projected Native Top-K Coverage guides that capacity toward experts demanded by the model's native routing behavior without rewarding capacity expansion. Throughout the search process, structural evaluation and model adaptation are interdependent. The current retention preferences determine the structure under which the model is updated. Meanwhile, the model's response informs subsequent refinement of those preferences.

Experiments spanning two distinct MoE architectures and two pruning severities show that our method achieves the best recovered performance overall among the evaluated pruning approaches. More importantly, the relative performance ranking of retained structures can change during subsequent model updates, showing that strong immediate preservation does not necessarily translate into high post-adaptation utility. Nevertheless, the structures found by MASS exhibit favorable representation recovery under model adaptation. Complementary analyses further show that they also preserve native routing demand and exhibit lower post-pruning representation distortion.

2 Related Work

Expert Pruning. Existing MoE pruning methods identify redundant experts using routing statistics (Chen et al., 2025), reconstruction objectives (Lu et al., 2024), or expert similarity and redundancy (Lee et al., 2025; Zhang et al., 2026). Despite differences in their selection criteria, most evaluate retention decisions at a fixed pretrained state, even though the preferred structure may change as the model adapts. Search-based approaches further optimize over discrete expert subsets, using exhaustive enumeration or genetic search (Liu et al., 2024; Yang et al., 2024). In contrast, we jointly optimize continuous retention variables and model parameters, allowing retention preferences to evolve as the model adapts during structural search.

Weight Pruning. A complementary line of work performs fine-grained pruning at the weight level. SparseGPT (Frantar & Alistarh, 2023) and Wanda (Sun et al., 2024) prune individual weights using reconstruction- or activation-aware criteria, while MoE-Pruner Xie et al., 2024 further incorporates router information into expert-parameter pruning. Accordingly, the two approaches address complementary sources of redundancy, with weight pruning reducing the cost of individual experts and expert pruning reducing the number of experts retained by the model.

3 Method

We formulate expert pruning as a model-wide structural search in which candidate expert structures are evaluated through their effect on next-token prediction and remain revisable as the model parameters adapt. The details of the method are as follows.

3.1 Preliminaries and Model-Wide Structural Search

Consider an MoE model with parameters W and L layers, each containing E experts. Given a token representation $\mathbf{h}_i^{(l)}$, the router in layer l produces logits $\mathbf{r}_i^{(l)} = R^{(l)}(\mathbf{h}_i^{(l)}; W)$ and selects the native Top-K expert set $\mathcal{A}_i^{(l)} = \mathrm{TopK}(\mathbf{r}_i^{(l)}, K)$. Let $\pi_{i,e}^{(l)}$ denote the routing weight normalized over $\mathcal{A}_i^{(l)}$. The layer output is

$$\mathbf{h}_i^{(l+1)} = \sum_{e \in \mathcal{A}_i^{(l)}} \pi_{i,e}^{(l)} E_e^{(l)}\left(\mathbf{h}_i^{(l)}\right). \tag{1}$$

For expert pruning with a target budget of B experts per layer, let $\mathcal{S}^{(l)} \subseteq \{1, \ldots, E\}$ denote the retained expert set in layer l. The collection $\mathcal{S} = \{\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(L)}\}$ defines a model-wide expert

structure, then the standard discrete pruning problem can be formulated as:

$$\begin{aligned} \mathcal{S}^\star &= \arg\min_{\mathcal{S}} \mathcal{L}_{\text{task}}(f_{\mathcal{S}}; \mathcal{D}) \,, \\ \text{s.t.} \quad \mathcal{S}^{(l)} &\subseteq \{1, \ldots, E\}, \qquad |\mathcal{S}^{(l)}| = B, \quad \forall l, \end{aligned} \tag{2}$$

where $f_{\mathcal{S}}$ is the model induced by $\mathcal{S}$. The difficulty extends beyond the $\binom{E}{B}^L$ search space. Since upstream pruning can affect downstream representations and routing, retention decisions are coupled across layers, and their utility changes as W adapts. Therefore, we optimize all layers jointly and maintain a revisable structure during adaptation.

To parameterize the evolving expert structure for gradient-based optimization, we associate expert e in layer l with a learnable structural parameter $\theta_e^{(l)} \in \mathbb{R}$ and define its continuous retention gate as $\alpha_e^{(l)} = \sigma(\theta_e^{(l)}) \in (0, 1)$ The trainable $\boldsymbol{\theta} = \{\theta_e^{(l)}\}_{l,e}$ parameterizes the current candidate structure and represents retention decisions under the evolving model state rather than fixed estimates of intrinsic expert importance.

To evaluate a candidate structure through its actual predictive effect, we embed this retention state directly into each sparse MoE layer. Before Top-K selection, the structural variables shift the native router logits:

$$\widetilde{r}_{i,e}^{(l)} = r_{i,e}^{(l)} + \theta_e^{(l)}, \qquad \widetilde{\mathcal{A}}_i^{(l)} = \text{TopK}\Big(\widetilde{\mathbf{r}}_i^{(l)}, K\Big) \,. \tag{3}$$

The resulting Top-K expert set reflects both the token-dependent preferences of the current router and the model-wide structural state. The same retention state remains attached to the selected experts when their outputs are aggregated:

$$\mathbf{h}_i^{(l+1)} = \sum_{e \in \widetilde{\mathcal{A}}_i^{(l)}} \widetilde{\pi}_{i,e}^{(l)} \alpha_e^{(l)} E_e^{(l)}\Big(\mathbf{h}_i^{(l)}\Big) \,, \tag{4}$$

where $\widetilde{\pi}_{i,e}^{(l)}$ is the corresponding routing weight normalized over $\widetilde{\mathcal{A}}_i^{(l)}$. Since the structural variables participate in every MoE layer, their joint effect can be evaluated through the full-model next-token prediction objective. For a sequence $\mathbf{x} = (x_1, \ldots, x_N)$ drawn from $\mathcal{D}$, we define

$$\mathcal{L}_{\text{NTP}}(W, \boldsymbol{\theta}; \mathcal{D}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[-\frac{1}{N-1} \sum_{i=1}^{N-1} \log p_{W,\boldsymbol{\theta}}(x_{i+1} \mid x_{\le i}) \right] . \tag{5}$$

Unlike layer-wise reconstruction or routing statistics, this objective evaluates retention decisions through their joint effect on model prediction. Gradients update the selected gates and routing weights, while recomputing hard Top-K sets after each step keeps the candidate structure responsive to the current router.

3.2 Capacity-Aware Structural Optimization

The $\mathcal{L}_{\text{NTP}}$ evaluates the predictive utility of a candidate structure at the model level, but does not guarantee capacity matching with the final pruned model. Therefore, we introduce complementary objectives to align the retained capacity with the target budget and guide its allocation across experts.

Capacity Alignment. In order to guide the continuous structural search towards the target budget, we construct a differentiable estimate of the retained capacity. We first approximate each binary retention decision with a soft indicator:

$$\kappa_e^{(l)} = \sigma\left(\frac{\alpha_e^{(l)} - \tau}{T} \right), \qquad C^{(l)} = \sum_{e=1}^{E} \kappa_e^{(l)}, \tag{6}$$

where τ is the retention threshold, T controls the smoothness around that threshold, and $C^{(l)}$ denotes the soft retained-expert count in layer l. We set $\tau = 0.5$ as the neutral midpoint of the gate range and $T = 0.1$ to obtain a sharp yet differentiable approximation to the binary retention indicator. Consequently, gates far from the threshold contribute nearly discretely to the count while preserving useful gradients near the decision boundary. We penalize deviations of the soft retained-expert count from the target budget B:

$$\mathcal{L}_{\text{budget}} = \frac{1}{L} \sum_{l=1}^{L} \left(\frac{C^{(l)} - B}{E} \right)^2 . \tag{7}$$

This term does not enforce the final pruning budget itself, which is satisfied exactly by Top-B discretization. Instead, it reduces the mismatch between the capacity under which retention preferences are learned and the capacity of the resulting discrete structure.

Capacity Allocation. To determine which experts should be retained within the target capacity, we propose Projected Native Top-K Coverage, which derives an allocation signal from the router's current token-level decisions. We first represent the native routing demand using the detached sparse weights:

$$\overline{\pi}_{i,e}^{(l)} = \mathrm{sg}\Big[\pi_{i,e}^{(l)} \mathbb{I}\Big(e \in \mathcal{A}_i^{(l)}\Big)\Big], \tag{8}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. This reference is recomputed from the current router at every structural step and remains fixed within that step. Thus, the coverage signal tracks the routing demand of the evolving model rather than a fixed routing pattern inherited from the pretrained state.

Based on this reference, we quantify routing preservation at the token level and aggregate it into a layer-wise coverage loss as

$$M_i^{(l)} = \sum_{e=1}^{E} \overline{\pi}_{i,e}^{(l)} \kappa_e^{(l)}, \qquad \mathcal{L}_{\mathrm{cov}}^{(l)} = -\frac{1}{N_l} \sum_{i=1}^{N_l} \log\Big(M_i^{(l)} + \epsilon\Big). \tag{9}$$

Here, $M_i^{(l)}$ measures the native Top-K routing mass of token i supported by the current soft structure, with a larger value indicating better preservation of its routing demand.

Directly optimizing the coverage loss, however, conflates expert allocation with capacity expansion. Since $M_i^{(l)}$ is monotonic in the retention indicators, the loss can be reduced by increasing the indicators of routed experts rather than reallocating a fixed amount of capacity. The resulting gradient therefore contains a capacity-expansion component that overlaps with the role of capacity alignment.

To isolate the allocation signal, we remove the component of the coverage gradient that changes the local soft expert count. Let $\mathbf{g}^{(l)} = \nabla_{\boldsymbol{\theta}^{(l)}} \mathcal{L}_{\mathrm{cov}}^{(l)}$ denote the coverage gradient and $\mathbf{u}^{(l)} = \nabla_{\boldsymbol{\theta}^{(l)}} C^{(l)}$ denote the local soft-count direction. We define the projected-coverage loss using

$$\beta^{(l)} = \mathrm{sg}\left[\frac{\langle \mathbf{g}^{(l)}, \mathbf{u}^{(l)} \rangle}{\|\mathbf{u}^{(l)}\|_2^2 + \epsilon}\right], \qquad \mathcal{L}_{\mathrm{pcov}}^{(l)} = \mathcal{L}_{\mathrm{cov}}^{(l)} - \beta^{(l)} C^{(l)}. \tag{10}$$

Since $\beta^{(l)}$ is detached during the current structural update, differentiating $\mathcal{L}_{\mathrm{pcov}}^{(l)}$ subtracts the coverage gradient component aligned with the soft-count direction, up to the numerical stabilization in the denominator. Consequently, the resulting signal promotes the redistribution of structural preference among experts without using capacity expansion as its primary descent direction.

We average the projected-coverage loss across layers as $\mathcal{L}_{\mathrm{pcov}} = L^{-1} \sum_{l=1}^{L} \mathcal{L}_{\mathrm{pcov}}^{(l)}$. The complete structural objective is

$$\mathcal{L}_{\mathrm{struct}} = \mathcal{L}_{\mathrm{NTP}} + \mathcal{L}_{\mathrm{budget}} + \mathcal{L}_{\mathrm{pcov}}. \tag{11}$$

The three terms evaluate model-wide predictive utility, keep the continuous search aligned with the deployment capacity, and allocate structural preference according to the model's current native Top-K routing demand.

3.3 First-Order Alternating Search and Discretization

Since structural utility depends on the model state after adaptation, we capture this dependence through the following bilevel objective:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathcal{L}_{\mathrm{struct}}(W^\star(\boldsymbol{\theta}), \boldsymbol{\theta}; \mathcal{D}_s), \\ \text{s.t.} \quad & W^\star(\boldsymbol{\theta}) = \arg\min_{W} \mathcal{L}_{\mathrm{NTP}}(W, \boldsymbol{\theta}; \mathcal{D}_t), \end{aligned} \tag{12}$$

where $\mathcal{D}_s$ and $\mathcal{D}_t$ are the separate datasets used for structural search and model adaptation, respectively. Exact differentiation of the outer objective requires accounting for the response of $W^\star(\theta)$ to the structural variables, which is impractical at the scale MoE models. We instead adopt a first-order

Table 1: Downstream performance at 50% expert retention. Scores are reported as percentages and P and R denote the immediately pruned and recovered models, respectively. The best and second-best results within each model are shown in **bold** and underline, respectively.

Method	Stage	Downstream Tasks								Avg.
		XStory Cloze	HellaSwag	ARC-C EN	ARC-E EN	BoolQ	NQ Open	TriviaQA	XWinograd EN	
Qwen3-30B-A3B										
Origin	–	79.62±1.04	82.81±0.38	69.20±1.35	89.56±0.63	88.35±0.56	31.19±0.77	62.05±0.36	90.37±0.61	74.14±0.28
Random	*P*	53.34±1.28	33.28±0.47	24.32±1.25	44.53±1.02	51.07±0.87	0.11±0.06	1.10±0.08	58.97±1.02	33.34±0.31
	R	76.70±1.09	75.44±0.43	60.41±1.43	85.14±0.73	84.25±0.64	14.60±0.59	41.32±0.37	88.77±0.65	65.83±0.29
Frequency	*P*	76.90±1.08	78.34±0.41	61.60±1.42	85.44±0.72	84.13±0.64	13.27±0.56	31.26±0.35	89.72±0.63	65.08±0.28
	R	**79.48**±1.04	81.16±0.39	**66.38**±1.38	87.04±0.69	85.47±0.62	21.50±0.68	56.46±0.37	90.49±0.61	71.00±0.28
REAP	*P*	60.56±1.26	58.20±0.49	48.04±1.46	73.57±0.90	79.27±0.71	5.29±0.37	12.46±0.25	78.97±0.85	52.05±0.31
	R	73.00±1.14	73.10±0.44	62.71±1.41	85.23±0.73	82.91±0.66	12.83±0.56	37.95±0.36	86.92±0.70	64.33±0.29
DiEP	*P*	69.23±1.19	67.20±0.47	52.39±1.46	78.54±0.84	81.93±0.67	7.01±0.42	15.23±0.27	85.51±0.73	57.13±0.30
	R	76.17±1.10	76.26±0.42	59.73±1.43	84.85±0.74	84.16±0.64	14.74±0.59	45.05±0.37	88.77±0.65	66.22±0.29
Ours	*P*	73.20±1.14	80.19±0.40	61.77±1.42	84.85±0.74	82.87±0.66	16.68±0.62	43.12±0.37	88.52±0.66	66.40±0.29
	R	79.22±1.04	**81.34**±0.39	65.10±1.39	**87.50**±0.68	**86.33**±0.60	**23.13**±0.70	**57.91**±0.37	**90.75**±0.60	**71.41**±0.28
Qwen3.5-35B-A3B										
Origin	–	78.82±1.05	85.55±0.35	73.55±1.29	91.41±0.57	89.02±0.55	31.11±0.77	67.97±0.35	90.37±0.61	75.98±0.27
Random	*P*	69.69±1.18	68.13±0.47	55.97±1.45	79.25±0.83	86.73±0.59	8.95±0.48	20.85±0.30	86.24±0.71	59.48±0.30
	R	76.64±1.09	78.48±0.41	66.30±1.38	88.30±0.66	86.45±0.60	17.98±0.64	49.51±0.37	89.29±0.64	69.12±0.28
Frequency	*P*	77.56±1.07	83.30±0.37	65.53±1.39	88.09±0.66	85.63±0.61	17.29±0.63	42.63±0.37	90.80±0.60	68.85±0.28
	R	**78.76**±1.05	**84.32**±0.36	68.60±1.36	89.56±0.63	87.74±0.57	21.16±0.68	59.97±0.37	89.72±0.63	72.48±0.27
REAP	*P*	77.04±1.08	81.51±0.39	72.01±1.31	91.62±0.57	86.18±0.60	14.38±0.58	37.88±0.36	90.62±0.60	68.91±0.27
	R	78.16±1.06	83.62±0.37	**71.50**±1.32	**90.95**±0.59	**88.56**±0.56	19.20±0.66	54.46±0.37	**90.41**±0.61	72.11±0.27
DiEP	*P*	67.44±1.21	67.56±0.47	54.69±1.45	80.13±0.82	86.39±0.60	9.25±0.48	28.42±0.34	85.81±0.72	59.96±0.30
	R	75.58±1.11	78.99±0.41	65.87±1.39	88.22±0.66	86.76±0.59	17.73±0.64	54.27±0.37	88.95±0.65	69.55±0.28
Ours	*P*	76.57±1.09	83.80±0.37	64.85±1.40	86.99±0.69	71.16±0.79	16.26±0.61	44.93±0.37	87.53±0.69	66.51±0.29
	R	78.56±1.06	84.23±0.36	70.31±1.34	90.36±0.61	87.34±0.58	**21.88**±0.69	**60.51**±0.36	90.37±0.61	**72.95**±0.27

approximation to the bilevel structural gradient by treating the current model parameters as fixed during each structural update:

$$\frac{d}{d\theta}\mathcal{L}_{\text{struct}}\big(W^{\star}(\theta),\theta;D_s\big) \approx \frac{\partial}{\partial\theta}\mathcal{L}_{\text{struct}}\big(W,\theta;D_s\big)\Big|_{W=W^{(q)}}. \tag{13}$$

This approximation omits the response gradient through model adaptation. At iteration q, we update

$$\boldsymbol{\theta}^{(q+1)} = \boldsymbol{\theta}^{(q)} - \eta_\theta \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{struct}}\Big(W^{(q)},\boldsymbol{\theta}^{(q)};\mathcal{D}_s\Big). \tag{14}$$

We recompute the native routing reference and projection coefficient in Equation 10 from the current model state and detach both during differentiation. Consequently, the update jointly incorporates model-wide predictive utility, target-capacity alignment, and routing-aware allocation under the current model state. We then fix the updated structural state and take a model-parameter update on $\mathcal{D}_t$:

$$W^{(q+1)} = W^{(q)} - \eta_W \nabla_W \mathcal{L}_{\text{NTP}}\Big(W^{(q)},\boldsymbol{\theta}^{(q+1)};\mathcal{D}_t\Big). \tag{15}$$

The updated model state provides the predictive and routing context for the next structural step. Thus, although gradients are not propagated through model adaptation, successive structural updates are evaluated under an evolving model state. During discretization, we rank experts by their gate values and retain the top B experts in each layer.

4 Experiments

4.1 Experimental Setup

Models and Datasets. Our experiments are conducted on two distinct MoE architectures, Qwen3-30B-A3B and Qwen3.5-35B-A3B (Yang et al., 2025; Qwen Team, 2026) at 50% and 75% expert retention. Qwen3-30B-A3B contains 48 MoE layers with 128 routed experts per layer and activates eight experts per token. Qwen3.5-35B-A3B contains 40 MoE layers with 256 routed experts per layer and activates eight routed experts together with one shared expert per token. We prune only the routed experts and leave the shared experts unchanged. For datasets, we sample 2.5B tokens

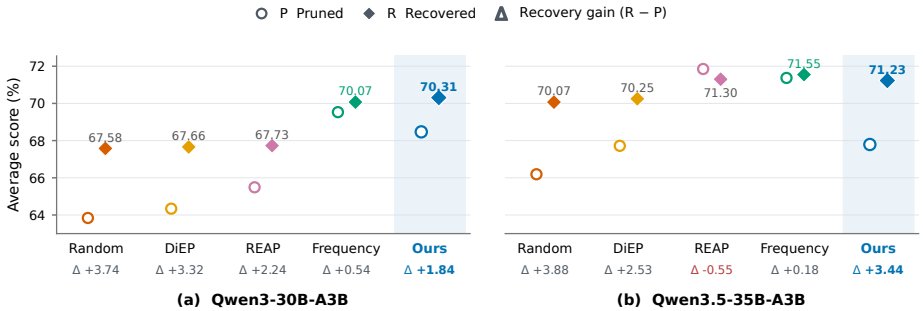

Figure 1: Downstream performance before and after post-pruning recovery at 75% expert retention. P and R denote the immediately pruned and recovered models, respectively, and $\Delta = \mathrm{R} - \mathrm{P}$ is the absolute recovery gain. Each point is the equally weighted average over the all task evaluation suite.

from the English FineWeb-Edu as the structure corpus $\mathcal{D}_s$ and 2.5B tokens from the FineWeb2 as the model-update corpus $\mathcal{D}_t$ (Penedo et al., 2024; 2025). We use $\mathcal{D}_s$ exclusively for expert selection and structural optimization, whereas all model-parameter updates, including matched training and post-pruning recovery, use $\mathcal{D}_t$.

Baselines and Comparison Protocol. We compare against two controls and three expert-pruning methods: **(1) Origin**, the complete unpruned model after matched continued pretraining; **(2) Random**, which uniformly samples the retained experts in each layer; **(3) Frequency**, which ranks experts by routing frequency; **(4) REAP**, which uses router-weighted activation magnitude (Lasby et al., 2025); and **(5) DiEP**, which learns differentiable intra- and inter-layer expert importance (Bai et al., 2025) based on a fixed model state. Since our method updates model weights during structural search, we apply an equal amount of training to each base model before running the competing pruning methods, thereby controlling for the benefit of additional weight updates. The resulting full model serves as both the pruning starting point for all baselines and the unpruned *Origin* reference. The more implementation details are provided in Appendix A.

Evaluation. We evaluate the pruned models both before and after recovery. The former assesses how well the pruned structure preserves model performance, while the latter measures its performance after adaptation to the retained experts. The primary evaluation covers eight benchmarks based on `lm-evaluation-harness` (Gao et al., 2024): XWinograd (Tikhonov & Ryabinin, 2021), TriviaQA (Joshi et al., 2017), Natural Questions Open (Lee et al., 2019), XStoryCloze (Lin et al., 2022), HellaSwag (Zellers et al., 2019), BoolQ (Clark et al., 2019), ARC-Challenge, and ARC-Easy (Clark et al., 2018). Results on broader tasks are provided in Appendices F and G

4.2 Main Results

Table 1 reports immediately pruned and recovered performance at 50% expert retention, while Figure 1 demonstrates the corresponding results at 75% retention. Since our central question is whether a retained structure remains effective after model adaptation, we treat recovered performance as the primary outcome and use immediately pruned performance to characterize static preservation. Although the structural search is performed on a general-purpose corpus, our method remains effective across held-out downstream tasks, indicating that its learned retention decisions transfer beyond the search corpus.

Across the four model–retention settings, MASS achieves the strongest recovered performance overall. At 50% retention, it achieves the highest recovered average on both Qwen3-30B-A3B and Qwen3.5-35B-A3B, reaching 71.41 and 72.95, respectively, and outperforming the strongest competing method by 0.41 and 0.47 points. At 75% retention, MASS again ranks first on Qwen3-30B-A3B with an average of 70.31 and reaches 71.23 on Qwen3.5-35B-A3B, within 0.32 points of the best result. Such consistent performance across settings indicates the robust post-adaptation capability of MASS under different MoE configurations and pruning severities.

Immediate preservation does not reliably predict post-adaptation performance. This is particularly clear when comparing MASS with Frequency. At 75% retention on Qwen3-30B-A3B, MASS

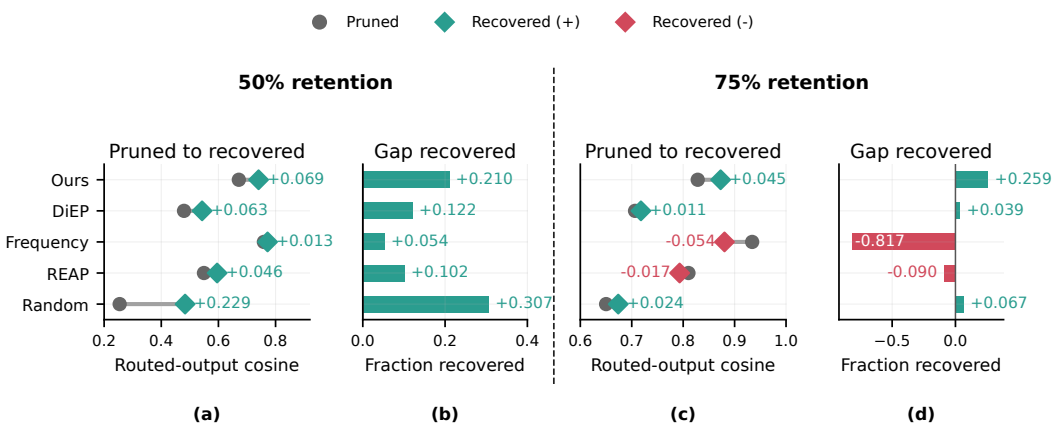

Figure 2: Representation recovery under the common post-pruning update stage. Paired points show routed-output cosine before and after recovery; bars show the recovered fraction of the pruning-induced cosine gap, with negative values indicating further deviation from the full model.

trails Frequency by 1.06 points immediately after pruning, but its larger recovery gain (1.84 versus 0.54) reverses the ranking and yields a 0.24-point recovered advantage. A similar reversal occurs at 50% retention on Qwen3.5-35B-A3B, where MASS moves from 2.34 points below Frequency before recovery to 0.47 points above it afterward. These changes show that a structure that better preserves the current model immediately after pruning is not necessarily the one that remains most effective after subsequent adaptation.

The comparison with DiEP more directly isolates the benefit of evaluating structural utility beyond a fixed model state. Both methods optimize differentiable structural variables, but DiEP estimates expert importance under a fixed model state, whereas our search jointly optimizes expert retention and model adaptation, allowing the retained structure to be selected according to its utility for the adapted model. As our search includes model-parameter updates, we control for this additional training through the matched comparison protocol in Appendix A.1 . Under this controlled setting, MASS achieves higher recovered performance than DiEP in all four model–retention settings, with average improvements ranging from 0.98 to 5.19 points. The difference is particularly clear at 75% retention on Qwen3.5-35B-A3B, shown in Figure 1: the two methods achieve nearly identical immediately pruned averages, yet MASS gains 3.44 points during recovery compared with 2.53 points for DiEP. These results indicate that structural preferences inferred from a fixed model state need not remain optimal after adaptation, and support jointly searching for expert retention under the model adaptation process.

5 Analysis Studies

5.1 Structural Compatibility with Post-Pruning Adaptation

The main results show that strong immediately pruned performance does not necessarily translate into strong recovered performance. We further examine whether retained structures differ in how their internal representations evolve during post-pruning adaptation. Unless otherwise specified, all analysis experiments are conducted on Qwen3-30B-A3B.

Figure 2(a,c) reports the routed-output cosine similarity between each pruned or recovered model and the corresponding unpruned model, averaged over valid tokens and MoE layers. Higher values indicate better preservation of the full unpruned model's intermediate representations. Figure 2(b,d) reports the recovered fraction of the pruning-induced cosine gap, $(c_{\text{recovered}} - c_{\text{pruned}}) \,/\, (1 - c_{\text{pruned}})$, where negative values indicate that recovery moves the model farther from the unpruned reference.

The representation dynamics reflect the performance changes observed in the main results. Frequency achieves the highest routed-output similarity immediately after pruning, but this advantage does not persist during recovery. It recovers only 5.4% of the pruning-induced cosine gap at 50% retention and exhibits a −81.7% recovered fraction at 75% retention, indicating limited adaptation headroom and poor robustness to subsequent model updates. In contrast, MASS combines strong immediately

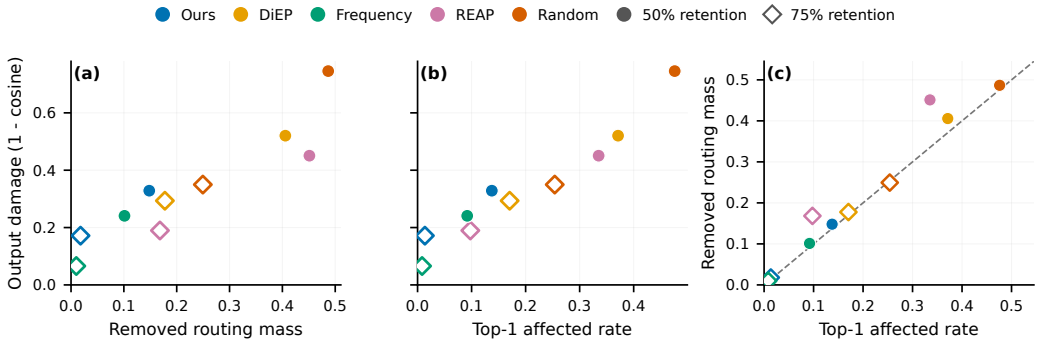

Figure 3: Static routing damage and post-pruning output deviation. (a–b) Lower removed routing mass (RRM) and Top-1 affected rate coincide with lower post-pruning output deviation across both retention ratios. (c) RRM versus Top-1 affected rate. Points above the diagonal certify routing-mass loss from secondary native Top-K assignments.

pruned representation preservation with substantial subsequent recovery, recovering 21.0% and 25.9% of the gap at the two retention ratios, respectively. Random shows a larger recovered fraction at 50% retention, but starts from substantially greater pruning-induced representation damage.

The contrast with DiEP further supports the benefit of adaptation-aware structural search. MASS exhibits higher routed-output similarity than DiEP both before and after recovery at both retention ratios, and the difference becomes more pronounced during adaptation. More importantly, at 75% retention, recovery increases the routed-output similarity of MASS by 0.045, compared with only 0.011 for DiEP. This result is consistent with the benefit of allowing retention decisions to evolve with the model state rather than selecting them from a fixed model state.

Together, these results distinguish immediate representation preservation from compatibility with subsequent adaptation. A structure can remain close to the unpruned model immediately after pruning yet move away from it during subsequent adaptation. Our method instead shows positive aggregate representation recovery at both retention ratios. Appendix D further shows that this behavior is broadly distributed across model depth and persists across intermediate search checkpoints, rather than being driven by a small subset of layers or a particular search checkpoint.

5.2 Native Routing Preservation and Post-Pruning Structural Distortion

We next examine the routing and representation properties of the immediately pruned structures. Since Projected Native Top-K Coverage uses the router's token-level demand to guide expert allocation, we measure how much native routing support is removed by each pruning method and how this relates to the resulting representation distortion.

We quantify routing preservation using *removed routing mass* (RRM), the expected native routing weight assigned to removed experts, $\mathrm{RRM} = \mathbb{E}_{l,\, i \in \mathcal{I}_l}[\sum_{e \in \mathcal{A}_i^{(l)} \setminus \mathcal{S}^{(l)}} \pi_{i,e}^{(l)}]$, and *Top-1 affected rate*, the fraction of tokens whose native Top-1 expert is removed. Here, $\mathcal{I}_l$ denotes the valid-token set in layer l. Lower values indicate stronger preservation of native routing demand. We additionally measure routed-output cosine similarity between the pruned and unpruned models to characterize the resulting intermediate representation distortion.

Across methods, greater routing disruption generally coincides with larger post-pruning output deviation. As shown in Figure 3(a–b), MASS maintains low RRM and Top-1 affected rates at both retention ratios, particularly under the more aggressive 50% setting, where its RRM and Top-1 affected rate are 0.148 and 0.137, respectively, compared with 0.406 and 0.371 for DiEP. Frequency also preserves native routing well immediately after pruning, providing a useful contrast with its weak representation recovery observed in Section 5.1.

Figure 3(c) further distinguishes preservation of the native Top-K routing mass from preservation of only the dominant route. REAP exhibits substantially higher RRM than its Top-1 affected rate would suggest, indicating additional loss from secondary native expert assignments. In contrast, MASS

Table 2: Ablation study of the training objectives. Pruned and recovered results are averaged over all evaluation tasks. Mean $|\Delta B|$ denotes the mean absolute budget gap across MoE layers. RRM, Top-1 affected rate, and output cosine measure the perturbation induced by structural pruning.

Variant	Performance		Budget	Pruning Disruption		
	Pruned (%) ↑	**Recovered** (%) ↑	**Mean** $\lvert\Delta B\rvert$ ↓	**RRM** (%) ↓	**Top-1 aff.** (%) ↓	**Output cos.** ↑
Full	**71.73**	73.84	16.47	**1.89**	**1.43**	**0.835**
w/o $\mathcal{L}_{\text{pcov}}$	58.21	69.14	**1.39**	38.69	43.86	0.427
w/o $\mathcal{L}_{\text{budget}}$	70.41	**73.98**	33.23	3.98	4.00	0.803
w/o both	67.60	72.46	9.81	28.65	33.28	0.558

remains low on both metrics, consistent with Projected Native Top-K Coverage preserving token-conditional routing support beyond the dominant expert. These observations are consistent with the intended role of Projected Native Top-K Coverage: guiding expert allocation toward structures that retain more of the router's token-level demand. However, Frequency also achieves strong immediate routing preservation while exhibiting weak or even negative representation recovery. Native routing preservation therefore characterizes the quality of the immediately pruned structure, but does not by itself determine post-adaptation utility.

5.3 Ablation Studies

We ablate $\mathcal{L}_{\text{pcov}}$ and $\mathcal{L}_{\text{budget}}$ to examine their respective roles in structural search. Removing $\mathcal{L}_{\text{pcov}}$ causes the largest performance degradation, reducing the pruned and recovered averages by 13.52 and 4.71 points, respectively. Additionally, removing $\mathcal{L}_{\text{budget}}$ results in the overall retention mass drifts during search, weakening the correspondence between the structure optimized during search and the final discrete structure.

The structural diagnostics further clarify this distinction. Without $\mathcal{L}_{\text{pcov}}$, all disruption metrics deteriorate sharply: RRM and the Top-1 affected rate increase from 1.89% to 38.69% and from 1.43% to 43.86%, respectively, while output cosine drops from 0.84 to 0.43. Notably, this variant still achieves the smallest mean budget gap of 1.39. This contrast shows that matching the target budget alone does not guarantee a functionally well-preserved pruned structure. The effect of $\mathcal{L}_{\text{budget}}$, by contrast, is most evident in budget adherence. Adding this term reduces the mean budget gap under both settings of $\mathcal{L}_{\text{pcov}}$: from 33.23 to 16.47 with projected coverage and from 9.81 to 1.39 without it. The aggregate budget-gap trajectories in Figure 4 provide complementary optimization-level evidence, showing that $\mathcal{L}_{\text{budget}}$ keeps structural search more closely aligned with the target budget throughout optimization. Overall, the combination of the two objectives yields a stronger balance between pruning quality and budget adherence than either objective alone.

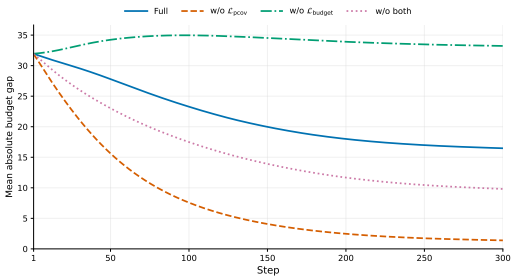

Figure 4: Mean absolute budget gap during structural search. The budget term directly controls soft capacity, while budget proximity alone does not determine routing quality.

6 Conclusion

In this work, we introduced MASS, a model-wide adaptation-aware structural search framework for MoE expert pruning. Rather than relying on layer-local or fixed-state estimates, MASS evaluates retention decisions jointly across the model and allows the retained structure and model parameters to co-adapt throughout search. Experiments across different MoE configurations and pruning severities show strong recovered performance, while further analyses indicate that immediate preservation alone does not reliably predict post-adaptation utility. These results support evaluating expert structures under evolving model states when constructing compact MoE models.

AI USE DISCLOSURE

In this work, we used generative AI tools to assist in improving linguistic clarity and fluency, providing feedback on experimental design and analysis, and interpreting experimental results. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or support qualitative or thematic data analysis. All AI-assisted suggestions and analyses were independently reviewed and validated by the authors, who take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide comprehensive implementation and experimental details in the appendix to facilitate reproduction of our results. These include the implementation of MASS and its structural search and adaptation procedure, the construction of the final pruned model, and the associated optimization settings. The appendix further specifies the datasets and evaluation protocols, model and pruning configurations, training and search hyperparameters, and baseline implementations used in our experiments. Detailed settings for the ablation and analysis experiments are also provided to support verification of our empirical findings.

REFERENCES

Sikai Bai, Haoxi Li, Jie Zhang, Zicong Hong, and Song Guo. DiEP: Adaptive mixture-of-experts compression through differentiable expert pruning. *arXiv preprint arXiv:2509.16105*, 2025.

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.

Yuanteng Chen, Yuantian Shao, Peisong Wang, and Jian Cheng. EAC-MoE: Expert-selection aware compressor for mixture-of-experts large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12942–12963, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.633. URL https://aclanthology.org/2025.acl-long.633/.

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, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

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. URL https://arxiv.org/abs/1803.05457.

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

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/frantar23a.html.

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, 07 2024. URL https://zenodo.org/records/12608602.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbinliu, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. MuBench: Assessment of multilingual capabilities of large language models across 61 languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16163–16192, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.794. URL https://aclanthology.org/2026.findings-acl.794/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12359–12374, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.760. URL https://aclanthology.org/2023.emnlp-main.760/.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5622–5640, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.334. URL https://aclanthology.org/2024.findings-acl.334/.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. *arXiv preprint arXiv:2510.13999*, 2025.

Jaeseong Lee, Seung-won Hwang, Aurick Qiao, Daniel F. Campos, Zhewei Yao, and Yuxiong He. STUN: Structured-then-unstructured pruning for scalable MoE pruning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13660–13676, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.671. URL https://aclanthology.org/2025.acl-long.671/.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612/.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616/.

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.

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, 2024.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.17557.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One pipeline to scale them all—adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025. URL https://arxiv.org/abs/2506.20920.

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

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.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024.

Alexey Tikhonov and Max Ryabinin. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3534–3546, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.310. URL https://aclanthology.org/2021.findings-acl.310/.

Yanyue Xie, Zhi Zhang, Ding Zhou, Cong Xie, Ziang Song, Xin Liu, Yanzhi Wang, Xue Lin, and An Xu. MoE-Pruner: Pruning mixture-of-experts large language model using the hints from its router. *arXiv preprint arXiv:2410.12013*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.612. URL https://aclanthology.org/2024.findings-emnlp.612/.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pp. 9–35, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sicon-1.2. URL https://aclanthology.org/2024.sicon-1.2/.

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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Zeliang Zhang, Nikhil Ghosh, Jiani Liu, Bin Yu, and Xiaodong Liu. Does a global perspective help prune sparse moes elegantly?, 2026. URL https://arxiv.org/abs/2604.06542.

Table 3: Effect of matched continued pretraining on aggregate downstream performance. Scores are reported as decimals.

Model	Stage	Downstream Tasks											Avg.
		XStory Cloze	HellaSwag EN	OpenBook QA	ARC-C EN	ARC-E EN	BoolQ	MMLU EN	NQ Open	TriviaQA	XNLI EN	XWinograd EN	
Qwen3-30B-A3B													
	Before	0.7988	0.8303	0.5260	0.6766	0.8885	0.8817	0.8123	0.3166	0.6582	0.5586	0.9006	0.7135
	After	0.7962	0.8281	0.5360	0.6920	0.8956	0.8835	0.8097	0.3119	0.6205	0.5683	0.9037	0.7132
	Δ	-0.0026	-0.0022	**+0.0100**	**+0.0154**	**+0.0071**	**+0.0018**	-0.0026	-0.0047	-0.0377	**+0.0097**	**+0.0031**	-0.0002
Qwen3.5-35B-A3B													
	Before	0.8054	0.8547	0.5260	0.7389	0.9162	0.8924	0.7502	0.3105	0.6988	0.5876	0.8976	0.7253
	After	0.7882	0.8555	0.5300	0.7355	0.9141	0.8902	0.7829	0.3111	0.6797	0.5739	0.9037	0.7241
	Δ	-0.0172	**+0.0008**	**+0.0040**	-0.0034	-0.0021	-0.0022	**+0.0327**	**+0.0006**	-0.0191	-0.0137	**+0.0061**	-0.0012

A Additional Experimental Protocols

A.1 Controlled Comparison Protocol

MASS performs model-parameter updates alongside structural optimization, whereas competing pruning methods estimate the retained structure without this additional training. Directly pruning the original base models with these methods would therefore conflate structural selection with the effect of the model updates performed during MASS search. To control for this factor, we apply a matched continued-pretraining stage on D_t to each base model using the same model-update configuration as in MASS. The resulting complete model is used as the pruning starting point for Random, Frequency, REAP, and DiEP, and also serves as the unpruned Origin reference. MASS instead starts from the same original base model and incorporates the corresponding model-update steps within its alternating search.

Table 3 verifies that this matched training preserves the overall starting capability. Following matched continued pretraining, the 11-task average changes by only -0.02 and -0.12 percentage points for Qwen3-30B-A3B and Qwen3.5-35B-A3B, respectively. Thus, the matched update stage controls for additional model training without materially changing the aggregate capability of the models used for pruning comparisons.

After expert selection, all methods retain the same proportion of routed experts in every layer and undergo an identical recovery stage on D_t. The resulting protocol therefore matches retained capacity, model-update budget, and post-pruning recovery while preserving each method's original expert-selection procedure.

A.2 Implementation Details

All experiments are implemented in PyTorch and run on NVIDIA H100 GPUs. Structural search and the matched continued-pretraining baseline are each run for 300 steps with a sequence length of 4,096 and a global batch size of 2,048. Model parameters are optimized with AdamW $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 0.1, using a cosine learning-rate schedule with 0.5% warmup from a peak of 1×10^{-5} to 1×10^{-6}. Structural parameters are initialized as $\theta \sim \mathcal{N}(0, 0.05^2)$ and updated once per step with a separate AdamW optimizer $(\beta_1, \beta_2) = (0.9, 0.999)$ under the same weight decay and ϵ, with the learning rate decaying from 3×10^{-3} to 3×10^{-4} under the same schedule. After discretization, all pruned models undergo the same 300-step recovery on $\mathcal{D}_t$ using the model-parameter optimizer configuration, with a peak learning rate of 1×10^{-5}.

A.3 Calibration-Budget Sensitivity

Frequency and REAP estimate expert importance by accumulating routing and activation statistics over calibration tokens. These static procedures require full-model forward computation and layer-wise statistic collection for every additional token, making large calibration corpora increasingly time-consuming. We therefore use 1M FineWeb-Edu tokens as a practical calibration budget and measure the marginal structural change obtained by increasing this budget fivefold.

Table 4: Mean layer-wise overlap (%) between the Top-B expert sets selected using 1M and 5M calibration tokens. The high overlap shows that a fivefold increase in calibration data produces only limited changes to the retained structures.

Model	Experts retained	Frequency ↑	REAP ↑
Qwen3-30B-A3B	75%	98.50	95.62
Qwen3-30B-A3B	50%	98.34	92.64
Qwen3.5-35B-A3B	75%	98.31	97.58
Qwen3.5-35B-A3B	50%	98.05	95.68

For each model and retention ratio, we independently select the layer-wise Top-B expert sets using 1M and 5M tokens sampled from $\mathcal{D}_s$. Let $\mathcal{S}_l^{1\mathrm{M}}$ and $\mathcal{S}_l^{5\mathrm{M}}$ denote the two retained sets in layer l. We compute their normalized intersection as $|\mathcal{S}_l^{1\mathrm{M}} \cap \mathcal{S}_l^{5\mathrm{M}}|/B_l$ and average it across all MoE layers. Table 4 shows that increasing the calibration budget from 1M to 5M tokens preserves 92.64%–98.50% of the selected experts. These results indicate limited sensitivity to the calibration budget over the tested range. We therefore use 1M calibration tokens for Frequency and REAP throughout the experiments.

B Computational Efficiency

We further evaluate the computational cost of MASS during both structural search and subsequent recovery. Since most pruning baselines determine the retained structure without optimizing trainable structural parameters, we compare search-stage efficiency with DIEP, which likewise performs parameterized optimization during pruning. All methods are evaluated under the same hardware and training configuration. As shown in Table 5, MASS achieves slightly higher search throughput than DiEP, while requiring moderately higher peak allocated memory. This suggests that incorporating model adaptation into structural search preserves comparable runtime efficiency, with the main additional cost being a modest increase in search-stage memory usage.

Table 5: Efficiency during structural search and recovery.

Method	Throughput ↑ (k tokens/s)	Peak Memory Allocated (GB) ↓
Search Stage		
DiEP	127.12	**54.38**
Ours	**129.33**	56.64
Recovery Stage		
Random	**126.16**	58.53
DiEP	125.41	57.91
REAP	125.17	58.54
Frequency	123.32	57.71
Ours	123.10	**57.64**

During recovery, the throughput of all pruned models remains within a narrow range of 123.10k–126.16k tokens/s. MASS achieves comparable recovery throughput while using the lowest peak allocated memory among the evaluated methods. These results demonstrate that the structure identified by MASS can be recovered with computational costs comparable to those of conventional pruning methods. Overall, the adaptation-aware search introduced by MASS incurs only limited search-stage overhead and does not compromise the efficiency of subsequent model adaptation.

C Routing Balance During Recovery Training

We further examine how different pruning strategies affect routing-balance quality throughout recovery training. Figure 5 summarizes the layer-wise routing statistics over 300 training steps. Specifically, the entropy balance measures the uniformity of expert utilization, while overload and underload deviations quantify the largest positive and negative log-deviations from the mean expert load, respectively. Our method and Frequency consistently maintain the highest entropy balance, while DiEP, REAP, and Random exhibit lower median entropy and substantially worse behavior among their adverse layers. A similar ordering appears in the maximum overload deviation: our method remains comparable to Frequency and consistently below the other baselines across the full recovery process. These results indicate that the advantage is sustained throughout recovery rather than arising from an isolated training checkpoint.

The distinction is more pronounced for expert underloading. As shown in Figure 5(c), our method and Frequency maintain substantially smaller underload deviations with narrower interquartile

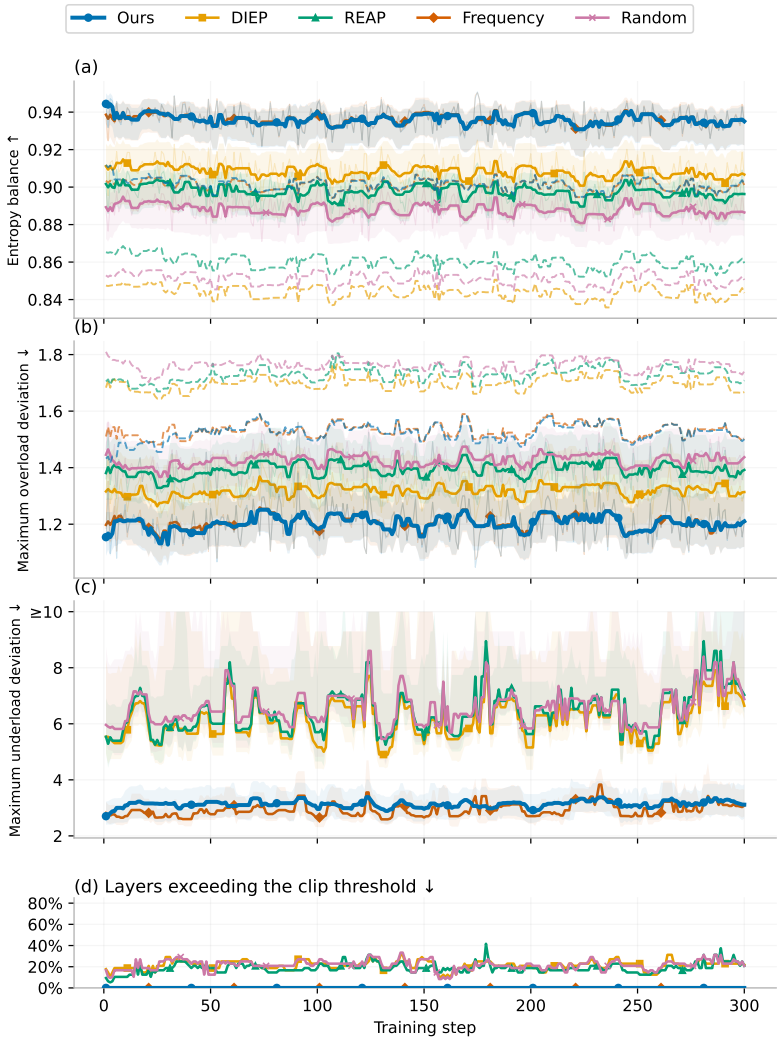

Figure 5: **Routing balance during recovery training.** Solid lines show the 7-step rolling median across MoE layers, shaded regions denote the rolling interquartile range, and dashed lines in (a–b) show the mean over the worst 10% of layers. Panel (c) clips underload deviations at 10 for readability, while panel (d) reports the fraction of layers exceeding this threshold. Higher is better in (a), whereas lower is better in (b–d).

ranges. In contrast, DiEP, REAP, and Random frequently produce extreme values that exceed the visualization threshold. Figure 5(d) confirms that approximately 10–30% of their MoE layers exceed this threshold during much of recovery training, with occasional larger spikes, whereas our method and Frequency largely avoid such clipped layers. Taken together, these observations show that our pruning strategy preserves routing balance at a level comparable to the strongest baseline while substantially outperforming the remaining pruning methods. Importantly, this analysis characterizes the quality and stability of routing during recovery; it does not by itself imply faster convergence.

D Layer-Wise Structural Effects and Search Stability

This analysis tests whether the aggregate recovery gains arise consistently across model depth and search time. The results show that the advantage is broadly distributed across layers and persists across checkpoints, supporting the stability of the learned expert structure rather than a checkpoint-specific effect.

The layer-wise comparison in Figure 6 contextualizes the aggregate metrics. Specifically, our MASS improves the routed output in 47 of 48 layers at 50% retention and 46 of 48 layers at 75% retention,

with median layer-wise cosine gains of 0.074 and 0.050. Frequency preserves a higher pruned cosine at 75% retention, yet recovery degrades all 48 layers; REAP exhibits the same model-wide regression. This contrast shows that strong static preservation does not by itself ensure compatibility with subsequent model updates. MASS instead maintains a favorable balance between limited initial distortion and broadly distributed recovery.

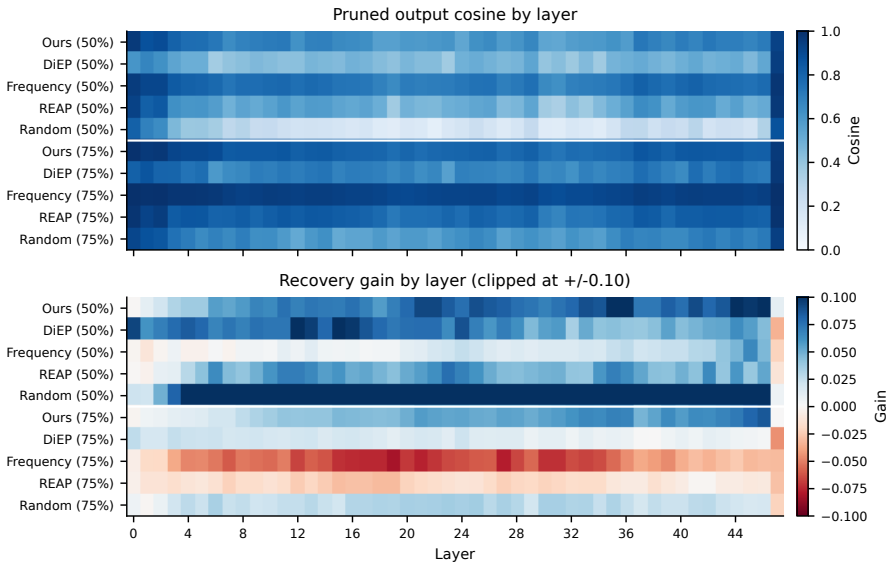

Figure 6: Layer-wise structural distortion and post-pruning recovery. The upper heatmap reports pruned routed-output cosine, and the lower heatmap reports recovery cosine gain; the white separator divides 50% and 75% expert retention. Recovery gains outside $[-0.10, 0.10]$ are saturated to preserve visibility around zero. Ours recovers positively across nearly all layers, whereas Frequency and REAP regress throughout the model at 75% retention.

The same behavior persists as the structural search progresses, as shown in Figure 7. At checkpoints 100, 200, and 300, MASS maintains positive overall recovery gains under both retention ratios. Its recovered output cosine at 75% retention remains within 0.871–0.873, while the number of positively recovered layers increases from 43 to 46 and then 47. In contrast, DiEP develops a negative overall gain at the final 75%-retention checkpoint, where only 3 of 48 layers improve. Consistency across both model depth and search time shows that the recovery behavior arises from a stable expert structure rather than a few favorable layers or a particular discretization checkpoint.

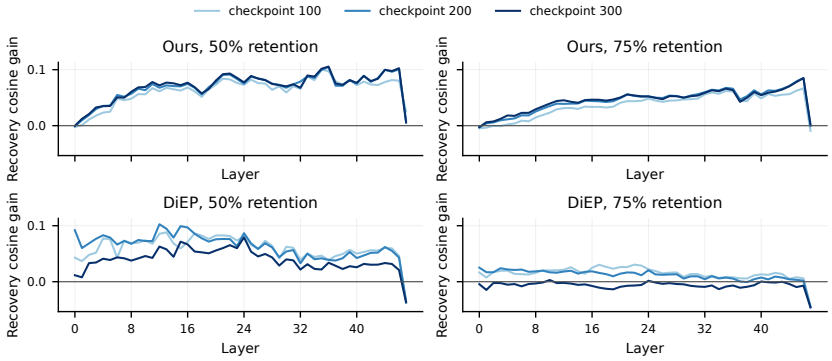

Figure 7: Layer-wise recovery gain across search checkpoints. Ours exhibits broadly positive recovery across model depth and remains stable as search progresses. DiEP becomes less recovery-compatible at the final 75%-retention checkpoint, where most layers regress.

E Evaluation Protocol and Reporting Conventions

We report task-level results over 52 tasks. The English knowledge group contains MMLU and its MMLU-LE task variants (Hendrycks et al., 2021). The multilingual knowledge group contains ArabicMMLU (Koto et al., 2024), IndoMMLU (Koto et al., 2023), and JMMLU (Yin et al., 2024). The multilingual scientific and commonsense groups use the MuBench versions (Han et al., 2026) of ARC-Challenge (Clark et al., 2018) and HellaSwag (Zellers et al., 2019), respectively. All task definitions are executed through `lm-evaluation-harness` (Gao et al., 2024). For readability, the tasks are organized as English Knowledge (25), Multilingual Knowledge (9), Multilingual Scientific Reasoning (6), and Multilingual Commonsense Reasoning (12). These presentation groups do not change aggregation: each task receives equal weight in the Overall-52 macro-average.

All metrics are reported on a 0–100 scale. Each table includes the selected metric for every task; we use normalized accuracy when available and otherwise report accuracy. We evaluate expert retention ratios of 75% and 50%. Each method cell reports the immediately pruned and recovered results on one line as P / R. Origin denotes the corresponding unpruned reference.

E.1 Multiple-Choice Scoring

For log-likelihood-based multiple-choice tasks, let x denote the context and $c_i = (c_{i,1}, \ldots, c_{i,T_i})$ the tokenized continuation for choice i. We concatenate each choice with the context using the task-configured target delimiter and compute the teacher-forced log-likelihood over the continuation tokens only:

$$s_i = \mathrm{LL}(c_i \mid x) = \sum_{t=1}^{T_i} \log p_\theta(c_{i,t} \mid x, c_{i,<t}). \tag{16}$$

The standard prediction selects the choice with the largest total log-likelihood, $\widehat{y} = \arg\max_i s_i$. Accuracy is the fraction of examples for which this prediction equals the gold choice. For length-normalized accuracy, we score each choice by $s_i/|c_i|_{\mathrm{char}}$, where $|c_i|_{\mathrm{char}}$ is the number of characters in the choice text, excluding the context and target delimiter. We select the choice with the largest normalized score and report the fraction of examples for which it equals the gold choice.

F Task-Level Results on Qwen3-30B-A3B

F.1 75% Expert Retention Ratio

Table 6: Task-level performance of Qwen3-30B-A3B at a 75% expert retention ratio. Each method cell reports the immediately pruned and recovered results on one line as P / R. The best and second-best results are marked independently within each stage using bold and underlined text, respectively. Origin is an unpruned reference and is excluded from ranking.

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
English Knowledge (25 tasks)							
MMLU (10 tasks)							
Business Ethics	Acc.	82.00	78.00 / 80.00	**81.00 / 83.00**	79.00 / 79.00	**81.00** / 74.00	80.00 / 82.00
Global Facts	Acc.	58.00	33.00 / 46.00	**56.00 / 49.00**	44.00 / 45.00	52.00 / 42.00	46.00 / 48.00
High-School Geography	Acc.	92.93	90.40 / 88.38	**93.94** / 90.91	91.92 / 89.90	92.93 / 92.42	91.41 / **93.43**
Miscellaneous	Acc.	89.66	83.78 / 84.93	86.84 / 86.97	86.72 / 86.59	**87.23** / 85.19	87.10 / **87.23**
Professional Accounting	Acc.	71.28	53.90 / 56.74	**67.02 / 65.96**	59.93 / 58.87	64.54 / 56.74	62.41 / 64.54
Professional Law	Acc.	60.17	52.41 / 53.98	**58.60 / 57.89**	55.41 / 53.78	55.48 / 54.63	56.26 / 57.11
Professional Psychology	Acc.	83.17	79.08 / 80.88	**85.46** / 83.50	80.56 / 79.90	83.17 / 83.01	84.15 / **84.64**
Public Relations	Acc.	72.73	66.36 / 70.91	**75.45 / 73.64**	71.82 / 68.18	71.82 / 67.27	74.55 / **73.64**
Sociology	Acc.	91.54	88.56 / 86.57	**89.55 / 89.06**	86.07 / 86.57	87.56 / 86.07	87.56 / **89.06**
World Religions	Acc.	87.72	75.44 / 83.04	82.46 / 84.21	**84.21** / 86.55	82.46 / 81.29	**84.21 / 90.06**
MMLU-LE (15 tasks)							
Astronomy	Norm. acc.	72.37	65.13 / 68.42	67.76 / 69.08	**70.39 / 70.39**	66.45 / 69.08	65.79 / 69.74
Clinical Knowledge	Norm. acc.	59.62	52.45 / 55.85	55.09 / 55.47	**58.11 / 59.62**	55.85 / 54.72	56.60 / 56.23
High-School Biology	Norm. acc.	72.58	63.87 / 65.48	66.45 / **69.36**	64.52 / 67.42	65.48 / 66.13	**67.74** / 69.03
High-School Geography	Norm. acc.	67.68	58.59 / 63.64	**64.65** / 64.65	61.11 / 66.16	62.63 / 64.14	61.62 / **66.67**
High-School Macroeconomics	Norm. acc.	62.56	**57.44** / 56.15	56.15 / 57.18	52.31 / 56.15	54.10 / 54.87	54.10 / **57.69**
Jurisprudence	Norm. acc.	56.48	50.93 / 50.00	50.93 / 50.93	48.15 / 50.93	**52.78 / 55.56**	51.85 / 53.70
Logical Fallacies	Norm. acc.	56.44	55.22 / 54.60	**56.44** / 55.83	50.31 / 55.83	53.99 / **57.67**	53.37 / **57.67**

Continued on the next page

Table 6 continued

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
Marketing	Norm. acc.	73.50	69.23 / 69.66	**71.80** / 70.51	66.67 / 68.38	70.08 / 68.80	71.37 / **72.65**
Miscellaneous	Norm. acc.	76.37	69.60 / 73.69	**75.35** / 75.86	70.88 / 74.58	70.75 / 75.10	72.80 / **76.37**
Moral Scenarios	Norm. acc.	62.23	55.98 / 49.72	66.26 / **62.91**	**67.60** / 55.98	58.99 / 49.72	63.80 / 62.12
Prehistory	Norm. acc.	61.42	45.99 / 56.17	**59.26** / **60.49**	55.56 / 55.56	54.01 / 56.79	57.72 / **60.49**
Professional Medicine	Norm. acc.	63.23	49.27 / 58.82	**61.77** / 62.13	58.82 / 61.03	52.94 / 57.35	59.56 / **63.23**
Professional Psychology	Norm. acc.	57.68	51.96 / 53.76	**55.56** / **56.54**	52.12 / 55.56	51.96 / 55.56	53.10 / **56.54**
Security Studies	Norm. acc.	33.47	**31.84** / 33.88	31.43 / 34.69	30.20 / 31.84	31.02 / 31.84	**31.84** / **35.51**
Sociology	Norm. acc.	55.22	50.75 / 51.24	**52.24** / **53.23**	**52.24** / 52.24	49.75 / 50.25	49.25 / 52.74
Multilingual Knowledge (9 tasks)							
Arabic MMLU (6 tasks)							
Anatomy	Norm. acc.	40.00	32.59 / 34.81	**38.52** / **39.26**	35.56 / 38.52	37.04 / **39.26**	35.56 / **39.26**
Electrical Engineering	Norm. acc.	43.45	36.55 / 36.55	30.34 / 35.17	**39.31** / 37.93	33.79 / **39.31**	33.10 / **39.31**
High-School Macroeconomics	Norm. acc.	43.85	35.13 / 40.00	**41.28** / 41.28	32.82 / 38.20	39.23 / 41.54	36.92 / **42.05**
Human Sexuality	Norm. acc.	38.93	36.64 / 38.93	**46.56** / 40.46	36.64 / 30.53	32.82 / 35.88	37.41 / **41.98**
Sociology	Norm. acc.	34.33	32.84 / **33.83**	30.85 / 29.35	28.86 / 30.35	**34.83** / 31.84	30.85 / 31.84
Virology	Norm. acc.	33.73	**35.54** / 30.12	31.32 / 33.73	32.53 / 30.72	32.53 / **37.35**	31.93 / 34.34
IndoMMLU (1 task)							
IndoMMLU	Norm. acc.	47.11	40.13 / 39.60	**44.38** / 43.98	39.31 / 41.34	39.02 / 38.66	42.49 / **44.30**
Japanese MMLU (2 tasks)							
Clinical Knowledge	Norm. acc.	47.33	36.67 / 40.00	35.33 / 38.00	34.67 / 37.33	**40.00** / **43.33**	34.00 / 42.00
Philosophy	Norm. acc.	42.67	**36.00** / 34.67	34.67 / 32.67	**36.00** / **38.67**	34.00 / 34.00	32.67 / 37.33
Multilingual Scientific Reasoning (6 tasks)							
ARC-Challenge (6 tasks)							
German	Norm. acc.	60.58	53.24 / 56.83	**57.34** / 58.36	50.17 / 54.86	52.73 / 54.78	55.12 / **58.53**
French	Norm. acc.	61.43	52.39 / 57.77	**58.19** / **59.81**	55.46 / 57.68	55.03 / 56.91	54.44 / 59.64
Indonesian	Norm. acc.	60.67	51.20 / 56.40	**55.29** / 57.00	49.83 / 53.58	49.49 / 50.34	50.51 / **57.59**
Malay	Norm. acc.	56.57	48.21 / 51.54	**50.68** / 51.54	40.44 / 47.35	44.28 / 45.22	46.76 / **52.05**
Tamil	Norm. acc.	47.44	34.64 / 40.02	36.86 / 39.85	36.26 / 41.98	34.81 / 37.71	**40.19** / **44.45**
Turkish	Norm. acc.	54.44	**46.50** / 49.40	44.80 / 49.83	37.29 / 44.54	39.51 / 44.03	41.64 / **50.43**
Multilingual Commonsense Reasoning (12 tasks)							
HellaSwag (12 tasks)							
Arabic	Norm. acc.	59.37	50.25 / 53.48	**54.30** / **56.01**	49.88 / 53.66	51.39 / 53.27	49.70 / 54.05
German	Norm. acc.	69.53	61.26 / 65.03	**68.13** / **68.78**	52.68 / 62.19	61.24 / 63.88	66.56 / 68.37
English	Norm. acc.	82.81	76.92 / 79.97	**82.39** / 82.28	77.50 / 80.21	79.19 / 80.55	82.34 / **82.38**
French	Norm. acc.	73.25	65.56 / 68.75	**72.38** / **72.55**	61.00 / 67.46	64.16 / 68.22	70.63 / 72.40
Indonesian	Norm. acc.	68.12	59.80 / 62.79	**65.96** / 66.45	52.77 / 59.68	56.48 / 59.59	65.25 / **66.76**
Italian	Norm. acc.	71.09	64.33 / 67.51	**70.52** / **70.75**	57.89 / 65.67	63.72 / 66.75	68.88 / 70.60
Malay	Norm. acc.	64.01	55.62 / 58.40	**62.19** / **62.40**	47.41 / 54.85	52.51 / 55.16	60.30 / 62.34
Dutch	Norm. acc.	68.43	57.46 / 61.85	**67.15** / **67.40**	49.41 / 57.65	58.50 / 60.50	65.62 / 67.31
Portuguese	Norm. acc.	73.35	65.20 / 68.93	**69.80** / 70.30	61.43 / 67.54	60.29 / 64.65	66.77 / **70.94**
Russian	Norm. acc.	63.88	55.54 / 60.48	**58.47** / 61.98	55.16 / 61.03	50.43 / 56.06	56.72 / **62.14**
Tamil	Norm. acc.	52.86	43.06 / 47.08	45.17 / 47.26	43.36 / 47.29	41.73 / 45.91	**50.22** / **51.97**
Turkish	Norm. acc.	54.99	46.70 / 50.50	46.66 / 50.34	37.97 / 44.53	40.29 / 45.70	**48.87** / **51.82**

F.2 50% Expert Retention Ratio

Table 7: Task-level performance of Qwen3-30B-A3B at a 50% expert retention ratio. Each method cell reports the immediately pruned and recovered results on one line as P / R. The best and second-best results are marked independently within each stage using bold and underlined text, respectively. Origin is an unpruned reference and is excluded from ranking.

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
English Knowledge (25 tasks)							
MMLU (10 tasks)							
Business Ethics	Acc.	82.00	23.00 / 70.00	**73.00** / **78.00**	61.00 / 66.00	68.00 / 74.00	56.00 / 72.00
Global Facts	Acc.	58.00	18.00 / **45.00**	**44.00** / 40.00	36.00 / 44.00	43.00 / 35.00	35.00 / 36.00
High-School Geography	Acc.	92.93	20.20 / 79.80	**88.38** / **88.89**	77.27 / 85.35	84.85 / **88.89**	80.30 / 87.37
Miscellaneous	Acc.	89.66	26.82 / 78.67	72.41 / **82.12**	70.50 / 80.20	**76.88** / 79.31	68.84 / 80.59
Professional Accounting	Acc.	71.28	26.60 / 48.94	46.45 / 48.23	47.16 / 50.71	**51.77** / 51.06	42.55 / **51.77**
Professional Law	Acc.	60.17	23.21 / 47.07	**50.13** / **52.22**	41.20 / 48.24	49.48 / 49.94	46.94 / 51.96
Professional Psychology	Acc.	83.17	25.00 / 72.06	**77.12** / **78.59**	55.56 / 65.20	68.95 / 75.82	70.26 / 77.61
Public Relations	Acc.	72.73	22.73 / 62.73	**72.73** / 70.00	57.27 / 65.45	62.73 / **70.91**	52.73 / 69.09
Sociology	Acc.	91.54	20.90 / 82.09	82.09 / **86.57**	68.66 / 80.60	**85.08** / **86.57**	74.13 / 84.58
World Religions	Acc.	87.72	22.81 / 75.44	71.34 / 76.02	72.52 / 80.12	**82.46** / **80.70**	69.59 / 78.95
MMLU-LE (15 tasks)							
Astronomy	Norm. acc.	72.37	25.66 / 54.60	51.32 / 63.82	**58.55** / 59.87	57.24 / 63.82	50.66 / **65.79**
Clinical Knowledge	Norm. acc.	59.62	27.55 / 49.81	45.66 / 55.09	43.40 / 50.19	47.55 / 50.57	**49.43** / **56.23**
High-School Biology	Norm. acc.	72.58	30.97 / 56.77	57.42 / 61.29	47.74 / 56.45	51.61 / 60.65	**60.65** / **61.61**

Continued on the next page

Table 7 continued

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
High-School Geography	Norm. acc.	67.68	31.31 / 57.58	**59.60 / 65.15**	44.44 / 60.10	52.02 / 56.57	58.08 / 61.11
High-School Macroeconomics	Norm. acc.	62.56	26.67 / 50.51	47.95 / 51.03	38.97 / 48.46	44.10 / 50.51	**49.74 / 53.59**
Jurisprudence	Norm. acc.	56.48	22.22 / 44.44	50.93 / **53.70**	29.63 / 42.59	49.07 / 51.85	**54.63** / 50.00
Logical Fallacies	Norm. acc.	56.44	33.13 / 50.92	**50.31** / 54.60	41.10 / 44.17	47.24 / 52.76	**50.31 / 55.83**
Marketing	Norm. acc.	73.50	34.62 / 66.24	**70.08 / 72.65**	52.99 / 64.10	61.54 / 66.24	66.24 / 69.66
Miscellaneous	Norm. acc.	76.37	28.35 / 67.56	**64.37** / 71.65	50.45 / 63.22	56.83 / 69.09	63.86 / **71.78**
Moral Scenarios	Norm. acc.	62.23	27.60 / 36.31	**51.06** / 41.34	39.22 / 48.83	45.92 / 43.91	35.20 / **49.61**
Prehistory	Norm. acc.	61.42	23.15 / 47.84	**51.85 / 57.10**	38.27 / 48.15	43.21 / 53.09	51.54 / 54.63
Professional Medicine	Norm. acc.	63.23	27.94 / 48.53	52.21 / **60.29**	38.97 / 49.63	34.56 / 52.21	**54.04** / 59.19
Professional Psychology	Norm. acc.	57.68	25.98 / 50.82	**51.47 / 54.09**	33.50 / 43.30	43.63 / 49.35	50.65 / 53.43
Security Studies	Norm. acc.	33.47	21.63 / 32.65	30.61 / **33.47**	24.90 / 29.80	26.12 / 30.61	**31.84** / 32.24
Sociology	Norm. acc.	55.22	27.86 / 44.28	**49.75 / 49.25**	37.31 / 40.80	39.80 / 45.77	46.27 / 47.76
Multilingual Knowledge (9 tasks)							
Arabic MMLU (6 tasks)							
Anatomy	Norm. acc.	40.00	21.48 / 30.37	**28.15** / 24.44	25.93 / **35.56**	25.18 / 28.15	26.67 / 34.81
Electrical Engineering	Norm. acc.	43.45	22.76 / 31.03	25.52 / 26.21	**31.72 / 33.10**	26.90 / 31.72	27.59 / 31.72
High-School Macroeconomics	Norm. acc.	43.85	21.54 / 32.56	23.59 / 30.77	**28.20** / 31.03	26.67 / **33.85**	21.03 / 29.74
Human Sexuality	Norm. acc.	38.93	29.77 / 30.53	29.01 / **37.41**	**32.82** / 31.30	30.53 / 34.35	31.30 / 30.53
Sociology	Norm. acc.	34.33	25.87 / 24.38	25.37 / 21.89	25.87 / 28.86	**28.36 / 29.85**	22.39 / 23.38
Virology	Norm. acc.	33.73	28.92 / 30.72	24.10 / 27.71	27.71 / 30.12	**31.32 / 35.54**	28.31 / 28.92
IndoMMLU (1 task)							
IndoMMLU	Norm. acc.	47.11	25.90 / **35.04**	**29.09** / 31.88	28.65 / 33.17	28.24 / 32.64	27.66 / 32.07
Japanese MMLU (2 tasks)							
Clinical Knowledge	Norm. acc.	47.33	24.67 / 36.67	**29.33** / 37.33	**29.33 / 40.00**	26.00 / 36.67	26.67 / 34.00
Philosophy	Norm. acc.	42.67	26.67 / **32.67**	26.00 / 26.67	26.67 / 28.67	24.67 / 28.67	**27.33** / 28.00
Multilingual Scientific Reasoning (6 tasks)							
ARC-Challenge (6 tasks)							
German	Norm. acc.	60.58	22.18 / **46.42**	32.34 / 43.52	29.61 / 42.92	**36.43** / 46.16	29.52 / 40.10
French	Norm. acc.	61.43	25.09 / 46.76	34.64 / 45.90	33.79 / **46.84**	34.90 / 44.97	**36.09** / 45.65
Indonesian	Norm. acc.	60.67	24.15 / **48.21**	28.58 / 36.77	**30.29** / 41.30	29.61 / 38.82	29.35 / 37.71
Malay	Norm. acc.	56.57	25.00 / **43.77**	24.83 / 32.08	**27.73** / 36.35	25.85 / 34.13	27.47 / 32.17
Tamil	Norm. acc.	47.44	23.29 / 31.23	23.80 / 27.82	**24.83** / 28.93	24.66 / **33.11**	24.23 / 28.41
Turkish	Norm. acc.	54.44	24.23 / **37.54**	25.00 / 30.20	24.74 / 30.97	**25.85** / 32.94	25.26 / 29.44
Multilingual Commonsense Reasoning (12 tasks)							
HellaSwag (12 tasks)							
Arabic	Norm. acc.	59.37	28.39 / **44.63**	32.41 / 40.14	32.41 / 41.47	**37.07** / 43.58	31.56 / 41.39
German	Norm. acc.	69.53	26.84 / 55.80	**46.31 / 56.40**	32.54 / 46.69	45.56 / 55.97	41.34 / 53.15
English	Norm. acc.	82.81	33.28 / 75.44	78.34 / 81.16	58.20 / 73.10	68.46 / 76.76	**79.55 / 81.51**
French	Norm. acc.	73.25	29.86 / 58.44	43.14 / **58.53**	37.57 / 50.50	43.12 / 55.97	**45.59** / 58.19
Indonesian	Norm. acc.	68.12	29.94 / **55.51**	36.12 / 44.09	33.60 / 44.60	35.81 / 45.41	**36.34** / 44.75
Italian	Norm. acc.	71.09	30.82 / **58.18**	41.91 / 54.18	36.39 / 50.00	39.28 / 52.57	**42.84** / 53.24
Malay	Norm. acc.	64.01	28.96 / **50.93**	34.83 / 41.35	33.22 / 41.21	34.56 / 42.10	**35.36** / 41.43
Dutch	Norm. acc.	68.43	27.57 / **52.51**	**44.23** / 52.42	33.18 / 41.62	42.54 / 51.37	39.48 / 47.63
Portuguese	Norm. acc.	73.35	29.90 / **58.52**	43.86 / 55.12	37.61 / 51.74	41.45 / 53.37	**44.53** / 55.16
Russian	Norm. acc.	63.88	27.99 / **50.22**	34.00 / 47.66	31.79 / 45.42	**39.51** / 49.01	34.87 / 48.70
Tamil	Norm. acc.	52.86	29.30 / **39.72**	33.03 / 36.39	31.35 / 34.83	**33.73** / 38.20	32.87 / 36.48
Turkish	Norm. acc.	54.99	29.31 / **40.11**	31.59 / 34.90	30.96 / 34.08	**31.77** / 35.79	31.23 / 34.24

G Task-Level Results on Qwen3.5-35B-A3B

G.1 75% Expert Retention Ratio

Table 8: Task-level performance of Qwen3.5-35B-A3B at a 75% expert retention ratio. Each method cell reports the immediately pruned and recovered results on one line as P / R. The best and second-best results are marked independently within each stage using bold and underlined text, respectively. Origin is an unpruned reference and is excluded from ranking.

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
English Knowledge (25 tasks)							
MMLU (10 tasks)							
Business Ethics	Acc.	89.00	76.00 / 80.00	82.00 / 80.00	80.00 / 82.00	**83.00** / 83.00	81.00 / **86.00**
Global Facts	Acc.	59.00	45.00 / 47.00	48.00 / 45.00	**57.00 / 53.00**	52.00 / **53.00**	46.00 / **53.00**
High-School Geography	Acc.	87.88	85.86 / 92.93	**94.95** / 91.92	94.44 / 92.42	**94.95** / 92.93	87.37 / **93.94**
Miscellaneous	Acc.	91.70	84.16 / 88.38	90.42 / 90.55	90.42 / 90.81	**92.47** / 90.93	89.14 / **91.57**
Professional Accounting	Acc.	60.28	65.60 / 69.86	62.06 / 67.38	69.50 / 68.08	**71.28** / 59.57	60.28 / **70.57**
Professional Law	Acc.	67.60	59.13 / 60.76	62.71 / 62.52	**65.06 / 64.21**	61.08 / 61.73	61.86 / 63.69

Continued on the next page

Table 8 continued

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
Professional Psychology	Acc.	86.60	82.19 / 84.97	86.11 / 84.48	**87.09** / **87.91**	83.82 / 84.31	84.48 / 86.28
Public Relations	Acc.	64.55	65.45 / 72.73	73.64 / 69.09	**77.27** / 73.64	74.55 / **75.45**	74.55 / 74.55
Sociology	Acc.	75.12	78.61 / 89.06	89.55 / 91.54	89.55 / 92.04	**90.55** / 92.54	86.57 / **93.03**
World Religions	Acc.	85.38	78.36 / **90.06**	**88.89** / 88.30	87.14 / 87.72	85.97 / **90.06**	86.55 / 88.89
MMLU-LE (15 tasks)							
Astronomy	Norm. acc.	74.34	63.82 / 67.11	69.08 / 68.42	71.05 / 68.42	**71.71** / 69.74	69.74 / **70.39**
Clinical Knowledge	Norm. acc.	62.64	56.98 / 58.49	58.87 / 59.62	**63.40** / **61.13**	60.76 / **61.13**	60.38 / 60.00
High-School Biology	Norm. acc.	75.81	61.61 / 70.00	**73.55** / 71.29	73.23 / **72.58**	68.06 / 68.39	69.03 / 72.26
High-School Geography	Norm. acc.	69.19	59.09 / 65.15	66.16 / 66.16	**69.19** / 66.67	65.66 / 66.67	64.65 / **68.69**
High-School Macroeconomics	Norm. acc.	62.56	55.64 / 59.74	62.82 / **61.03**	**63.85** / 59.49	59.74 / 60.00	58.20 / **61.03**
Jurisprudence	Norm. acc.	58.33	51.85 / 57.41	57.41 / 58.33	**58.33** / 58.33	52.78 / 53.70	54.63 / **59.26**
Logical Fallacies	Norm. acc.	61.35	59.51 / 58.28	**61.96** / 60.12	60.74 / 60.74	55.22 / 60.12	60.74 / **61.96**
Marketing	Norm. acc.	77.35	74.79 / 74.36	73.93 / 74.79	**78.20** / 76.50	74.79 / 75.21	77.35 / **77.35**
Miscellaneous	Norm. acc.	79.06	69.60 / 76.37	76.88 / 78.29	**77.39** / 78.80	76.88 / 78.54	77.27 / **79.06**
Moral Scenarios	Norm. acc.	74.75	53.18 / 60.45	**72.63** / 70.39	71.84 / 68.72	68.05 / 67.93	66.03 / **71.28**
Prehistory	Norm. acc.	65.43	52.47 / 59.26	61.73 / 61.42	62.04 / 63.58	59.57 / 63.89	**65.12** / **66.36**
Professional Medicine	Norm. acc.	69.12	57.35 / 65.44	**70.59** / 68.75	**70.59** / 69.85	66.18 / 67.65	**70.59** / **70.22**
Professional Psychology	Norm. acc.	62.26	53.92 / 57.35	60.46 / 59.97	**62.91** / 60.95	54.41 / 56.05	59.64 / **61.44**
Security Studies	Norm. acc.	37.96	32.24 / 35.51	35.92 / 33.88	36.73 / 35.92	33.06 / 35.10	**37.55** / **36.33**
Sociology	Norm. acc.	66.67	53.73 / 58.21	58.21 / 57.71	61.69 / **62.19**	53.73 / 56.22	**63.18** / **62.19**
Multilingual Knowledge (9 tasks)							
Arabic MMLU (6 tasks)							
Anatomy	Norm. acc.	39.26	37.78 / 35.56	38.52 / 33.33	**42.96** / 40.74	39.26 / **42.22**	39.26 / 39.26
Electrical Engineering	Norm. acc.	48.28	39.31 / 35.86	37.24 / 39.31	**43.45** / **42.76**	42.76 / 37.93	41.38 / 40.69
High-School Macroeconomics	Norm. acc.	46.15	38.72 / 42.31	44.87 / 41.54	**45.64** / **44.10**	40.51 / 43.59	41.80 / **44.10**
Human Sexuality	Norm. acc.	38.17	34.35 / 34.35	**41.22** / 35.88	35.12 / 37.41	36.64 / 37.41	38.17 / **41.22**
Sociology	Norm. acc.	40.80	29.35 / 29.35	36.82 / 39.80	35.32 / 34.83	35.32 / 35.32	**37.81** / **41.29**
Virology	Norm. acc.	33.13	25.90 / 32.53	**33.73** / 31.93	28.31 / 30.12	31.93 / 34.34	31.93 / **36.75**
IndoMMLU (1 task)							
IndoMMLU	Norm. acc.	49.38	41.25 / 41.89	**47.90** / **46.79**	44.34 / 42.85	45.02 / 44.63	43.58 / 46.19
Japanese MMLU (2 tasks)							
Clinical Knowledge	Norm. acc.	54.00	47.33 / 52.00	55.33 / 54.67	46.67 / 52.67	56.00 / 54.00	**57.33** / **55.33**
Philosophy	Norm. acc.	52.67	37.33 / 42.00	**47.33** / **48.67**	36.67 / 42.67	36.67 / 40.67	46.67 / 45.33
Multilingual Scientific Reasoning (6 tasks)							
ARC-Challenge (6 tasks)							
German	Norm. acc.	66.21	57.00 / 62.54	**65.27** / **65.02**	61.18 / 62.54	61.77 / 61.86	60.15 / 63.91
French	Norm. acc.	65.96	56.66 / 61.77	**65.96** / 65.27	65.44 / 65.27	60.75 / 62.80	61.35 / **65.36**
Indonesian	Norm. acc.	64.08	52.90 / 58.62	**64.25** / **62.97**	56.83 / 59.81	59.47 / 59.98	59.73 / 62.54
Malay	Norm. acc.	63.48	48.38 / 53.58	**61.43** / **61.26**	54.69 / 56.74	55.12 / 56.23	57.25 / 59.98
Tamil	Norm. acc.	52.39	39.42 / 45.39	49.32 / **51.62**	29.69 / 36.18	**49.40** / 48.55	43.34 / 45.99
Turkish	Norm. acc.	61.52	46.08 / 53.33	**61.01** / **59.13**	40.78 / 47.35	50.34 / 51.54	52.39 / 58.11
Multilingual Commonsense Reasoning (12 tasks)							
HellaSwag (12 tasks)							
Arabic	Norm. acc.	63.40	55.58 / 57.80	59.51 / 60.14	**61.48** / 60.53	56.74 / 58.38	60.53 / **61.05**
German	Norm. acc.	73.52	65.69 / 69.24	**71.82** / 71.89	68.80 / 69.78	67.02 / 69.04	71.52 / **72.52**
English	Norm. acc.	85.55	80.43 / 82.47	85.27 / 84.94	85.21 / 85.02	81.84 / 83.45	**85.87** / **85.06**
French	Norm. acc.	77.15	67.69 / 72.10	**75.69** / 75.45	74.85 / 75.08	71.05 / 72.67	75.37 / **75.56**
Indonesian	Norm. acc.	71.60	61.90 / 65.42	**71.22** / **70.55**	62.65 / 66.79	64.96 / 67.11	69.71 / 70.53
Italian	Norm. acc.	74.70	66.80 / 70.22	**72.75** / 72.39	72.22 / **72.83**	66.28 / 69.23	71.49 / 72.40
Malay	Norm. acc.	67.83	58.24 / 61.29	**67.63** / **66.64**	59.67 / 61.67	59.88 / 62.10	65.30 / 66.56
Dutch	Norm. acc.	73.59	65.10 / 68.23	**71.95** / **71.58**	63.08 / 66.38	59.19 / 65.33	68.26 / 69.76
Portuguese	Norm. acc.	76.53	68.05 / 71.81	**76.00** / **75.30**	70.28 / 72.30	65.70 / 70.78	72.46 / 73.40
Russian	Norm. acc.	66.80	57.22 / 63.16	64.44 / 65.75	62.76 / 65.00	62.01 / 65.09	**65.08** / **67.07**
Tamil	Norm. acc.	56.19	46.93 / 50.45	**55.16** / **55.35**	35.57 / 41.54	51.27 / 53.66	51.89 / 53.11
Turkish	Norm. acc.	60.52	50.55 / 54.70	**60.00** / **60.03**	43.89 / 49.77	50.33 / 54.02	58.67 / 59.62

G.2 50% Expert Retention Ratio

Table 9: Task-level performance of Qwen3.5-35B-A3B at a 50% expert retention ratio. Each method cell reports the immediately pruned and recovered results on one line as P / R. The best and second-best results are marked independently within each stage using bold and underlined text, respectively. Origin is an unpruned reference and is excluded from ranking.

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
English Knowledge (25 tasks)							
MMLU (10 tasks)							
Business Ethics	Acc.	89.00	73.00 / **79.00**	65.00 / 75.00	**76.00** / 76.00	75.00 / 78.00	59.00 / 72.00

Continued on the next page

Table 9 continued

Task	Metric ↑	Origin	Random P / R	Frequency P / R	REAP P / R	DiEP P / R	Ours P / R
Global Facts	Acc.	59.00	38.00 / 43.00	42.00 / 46.00	**44.00** / 46.00	35.00 / **50.00**	30.00 / 48.00
High-School Geography	Acc.	87.88	84.34 / 88.89	80.81 / **90.91**	**89.39** / **90.91**	87.37 / 89.90	61.62 / **90.91**
Miscellaneous	Acc.	91.70	77.14 / 84.42	72.41 / **86.59**	81.35 / 85.82	**81.48** / 84.04	68.84 / 86.08
Professional Accounting	Acc.	60.28	**52.13** / **57.09**	37.59 / 55.32	43.26 / 52.13	42.55 / 43.97	36.52 / 53.19
Professional Law	Acc.	67.60	52.09 / 56.00	49.15 / **58.02**	**55.09** / 57.43	40.48 / 45.89	45.96 / 56.19
Professional Psychology	Acc.	86.60	71.89 / 78.43	65.36 / 80.23	**80.39** / **83.01**	72.88 / 78.27	48.37 / 74.18
Public Relations	Acc.	64.55	69.09 / **72.73**	69.09 / 67.27	**70.91** / 66.36	65.45 / 66.36	40.91 / 66.36
Sociology	Acc.	75.12	79.60 / 85.57	56.22 / 90.55	**85.57** / **91.05**	81.59 / 85.57	46.27 / 89.55
World Religions	Acc.	85.38	71.93 / 85.38	78.36 / 84.80	77.78 / 84.80	**81.87** / 87.14	73.10 / **88.30**
MMLU-LE (15 tasks)							
Astronomy	Norm. acc.	74.34	44.74 / 59.21	57.90 / 62.50	**67.11** / 64.47	57.90 / **65.79**	59.87 / 61.84
Clinical Knowledge	Norm. acc.	62.64	47.17 / 55.47	54.72 / 59.24	**63.40** / **61.13**	53.96 / 54.72	58.87 / 59.24
High-School Biology	Norm. acc.	75.81	50.00 / 60.65	61.29 / 63.55	**69.68** / 68.71	57.74 / 64.52	61.29 / **69.36**
High-School Geography	Norm. acc.	69.19	55.05 / 60.61	**65.15** / **64.14**	64.65 / 62.63	56.06 / 60.10	58.59 / 63.13
High-School Macroeconomics	Norm. acc.	62.56	44.10 / 51.80	57.69 / 56.15	**58.72** / **60.26**	47.69 / 56.15	54.62 / 58.46
Jurisprudence	Norm. acc.	58.33	47.22 / 50.00	54.63 / 55.56	**56.48** / 55.56	37.96 / 45.37	50.00 / **56.48**
Logical Fallacies	Norm. acc.	61.35	47.24 / 53.99	56.44 / **60.12**	**57.67** / 56.44	49.69 / 57.67	55.22 / 57.05
Marketing	Norm. acc.	77.35	67.52 / 73.50	73.08 / 73.50	**73.50** / 73.08	67.52 / 73.50	71.80 / **75.21**
Miscellaneous	Norm. acc.	79.06	56.83 / 70.37	67.69 / 72.80	**68.58** / 73.56	62.84 / **74.07**	67.05 / 73.44
Moral Scenarios	Norm. acc.	74.75	27.71 / 59.55	62.68 / 62.57	**69.94** / **68.49**	55.53 / 55.31	53.74 / 60.89
Prehistory	Norm. acc.	65.43	40.43 / 54.01	58.33 / 59.88	**58.64** / **60.80**	47.53 / 57.41	58.02 / 60.19
Professional Medicine	Norm. acc.	69.12	40.07 / 56.98	63.60 / **67.65**	**68.38** / 66.91	50.37 / 59.19	63.97 / 65.44
Professional Psychology	Norm. acc.	62.26	43.30 / 49.84	**57.68** / 55.72	57.19 / 56.70	45.10 / 50.82	54.09 / **57.35**
Security Studies	Norm. acc.	37.96	27.75 / 33.47	34.69 / 34.69	33.47 / 33.06	25.71 / **35.10**	**35.10** / **35.10**
Sociology	Norm. acc.	66.67	48.76 / 50.25	54.73 / 53.73	**55.22** / 56.22	45.27 / 54.23	53.73 / **57.21**
Multilingual Knowledge (9 tasks)							
Arabic MMLU (6 tasks)							
Anatomy	Norm. acc.	39.26	30.37 / 29.63	**34.81** / 34.07	32.59 / **36.30**	31.11 / 32.59	31.11 / 35.56
Electrical Engineering	Norm. acc.	48.28	26.90 / 32.41	27.59 / 31.03	**40.00** / 33.79	34.48 / **36.55**	28.97 / 28.97
High-School Macroeconomics	Norm. acc.	46.15	30.00 / 35.38	**36.92** / 36.92	33.08 / 36.92	28.46 / **37.69**	32.05 / 35.13
Human Sexuality	Norm. acc.	38.17	**38.17** / 35.12	36.64 / 34.35	33.59 / 35.12	**38.17** / 33.59	34.35 / **36.64**
Sociology	Norm. acc.	40.80	24.88 / 29.35	**33.33** / **30.85**	30.35 / 28.36	23.88 / 29.35	31.84 / 30.35
Virology	Norm. acc.	33.13	28.31 / 25.90	28.92 / **32.53**	27.71 / **32.53**	27.11 / 31.93	**30.12** / 30.12
IndoMMLU (1 task)							
IndoMMLU	Norm. acc.	49.38	35.10 / 37.94	**42.77** / **43.14**	37.45 / 38.10	33.90 / 37.97	37.61 / 41.53
Japanese MMLU (2 tasks)							
Clinical Knowledge	Norm. acc.	54.00	38.67 / 44.00	**45.33** / 42.67	**45.33** / 44.67	42.00 / **50.67**	42.67 / 48.00
Philosophy	Norm. acc.	52.67	30.67 / 37.33	**38.67** / **40.00**	32.00 / 34.67	29.33 / 33.33	31.33 / 38.00
Multilingual Scientific Reasoning (6 tasks)							
ARC-Challenge (6 tasks)							
German	Norm. acc.	66.21	41.21 / 51.62	**53.67** / 59.39	42.49 / 51.45	39.68 / 53.67	48.72 / **59.47**
French	Norm. acc.	65.96	44.03 / 54.86	50.51 / 54.86	**53.75** / **56.14**	43.60 / 55.38	45.99 / 53.16
Indonesian	Norm. acc.	64.08	42.23 / 50.68	**56.83** / 58.45	46.76 / 55.20	40.36 / 50.43	47.44 / **58.62**
Malay	Norm. acc.	63.48	38.31 / 46.33	**51.45** / **55.80**	41.55 / 46.16	34.47 / 43.00	45.22 / 54.61
Tamil	Norm. acc.	52.39	24.91 / 32.08	33.36 / 39.25	24.83 / 29.44	26.11 / 31.23	**33.96** / **40.70**
Turkish	Norm. acc.	61.52	34.04 / 42.58	**39.25** / **46.25**	32.25 / 32.42	30.97 / 39.76	33.28 / 40.87
Multilingual Commonsense Reasoning (12 tasks)							
HellaSwag (12 tasks)							
Arabic	Norm. acc.	63.40	40.42 / 48.96	**48.11** / **53.71**	47.11 / 50.91	40.75 / 49.84	42.66 / 52.24
German	Norm. acc.	73.52	47.60 / 60.22	**63.74** / **68.27**	48.88 / 56.96	46.19 / 58.05	60.62 / 67.81
English	Norm. acc.	85.55	68.13 / 78.48	83.30 / **84.32**	81.51 / 83.62	69.45 / 79.67	**83.72** / 84.13
French	Norm. acc.	77.15	54.55 / 65.94	**64.48** / **69.27**	57.57 / 64.38	51.03 / 63.48	56.02 / 65.50
Indonesian	Norm. acc.	71.60	49.85 / 59.32	**68.04** / **68.30**	51.53 / 57.84	44.74 / 55.39	61.14 / 66.59
Italian	Norm. acc.	74.70	51.90 / **62.52**	**54.77** / 61.03	52.63 / 59.61	46.66 / 58.10	53.48 / 60.35
Malay	Norm. acc.	67.83	46.56 / 54.99	**62.78** / **63.63**	47.40 / 53.05	41.01 / 49.73	55.38 / 62.16
Dutch	Norm. acc.	73.59	47.97 / 58.84	**64.45** / **67.65**	50.93 / 57.42	43.83 / 56.47	57.88 / 66.61
Portuguese	Norm. acc.	76.53	54.05 / 64.72	**65.50** / **69.93**	53.58 / 61.10	48.07 / 59.84	54.94 / 63.64
Russian	Norm. acc.	66.80	45.10 / 56.03	44.36 / 57.32	47.28 / 56.86	41.91 / 55.04	**49.24** / **57.44**
Tamil	Norm. acc.	56.19	32.78 / 38.54	42.06 / 47.00	33.29 / 37.70	33.00 / 38.96	**44.67** / **49.20**
Turkish	Norm. acc.	60.52	39.51 / 46.41	**46.93** / **51.94**	35.38 / 38.00	36.33 / 43.66	40.19 / 45.73