REXMOE: REUSING EXPERTS WITH MINIMAL OVERHEAD IN MIXTURE-OF-EXPERTS

Zheyue $Tan^{1,*}$ Zhiyuan Li^2 Tao Yuan 2 Dong Zhou 2 Weilin Liu^2 Yueqing Zhuang 2 Yadong Li^2 Guowei Niu 2 Cheng Qin 3 Zhuyu Yao 2 Congyi Liu^2 Haiyang Xu^2 Boxun Li^2 Guohao Dai 4,5,† Bo Zhao 1,† Yu Wang 6,†

ABSTRACT

Mixture-of-Experts (MoE) architectures have emerged as a promising approach to scale Large Language Models (LLMs). MoE boosts the efficiency by activating a subset of experts per token. Recent works show that fine-grained experts substantially enriches the combinatorial flexibility of active experts and enhances model expressiveness. However, such a design is fundamentally limited by the layer-local routing mechanism: each layer is restricted to its own expert pool. This requires a careful trade-off between expert dimensionality and routing diversity given fixed parameter budgets. We describe REXMOE, a novel MoE architecture that improves routing beyond the existing layer-local approaches by allowing routers to reuse experts across adjacent layers. REXMOE decouples expert dimensionality from per-layer budgets, enabling richer expert combinations without sacrificing individual expert capacity or inflating overall parameters. To this end, we propose a new progressive scaling routing (PSR) strategy to gradually increase the candidate expert pool during training. As a result, REXMOE improves both language modeling and downstream task performance. Extensive experiments on models ranging from 0.5B to 7B parameters across different architectures demonstrate that REXMOE consistently improves performance under fixed architectural dimensions, confirming REXMOE as new design paradigm for parameter-efficient and scalable MoE-based LLMs.

1 INTRODUCTION

Large Language Models (LLMs) have rapidly advanced in scale and capability, reaching hundreds of billions of parameters and demonstrating remarkable progress toward Artificial General Intelligence (AGI). Recent foundation models (Achiam et al., 2023; OpenAI, 2025; Meta AI, 2025; Guo et al., 2025; AI, 2025; Yang et al., 2025) have exhibited strong performance across complex tasks in multiple domains. This progress has been driven by massive investments in data and compute, but such growth also intensifies the tension between model capacity and development practicality. Given the substantial costs involved, Mixture-of-Experts (MoE) architectures have become an increasingly attractive alternative. By dynamically activating only a subset of specialized experts per input, MoEs can match or even exceed the performance of dense counterparts while significantly reducing inference-time computational demands (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022; Jiang et al., 2024; Liu et al., 2024a; b; Yang et al., 2025).

Comparing to the dense counterparts, a key characteristic of MoE architectures is the additional degrees of freedom when replacing the feed-forward networks with MoE blocks: the number of experts, the dimensionality of each expert, and the routing strategy. Recent studies on MoE scaling laws (Clark et al., 2022; Krajewski et al., 2024) highlight that model performance is constrained by trade-offs among these dimensions under a fixed parameter budget. In particular, the size of each expert and the number of experts form a critical axis: increasing the number of smaller experts

¹ Aalto University ² Infinigence-AI ³ Yale University ⁴ Shanghai Jiao Tong University

⁵ Shanghai Innovation Institute ⁶ Tsinghua University

^{*}Work conducted during internship at Infinigence-AI.

[†]Corresponding authors.

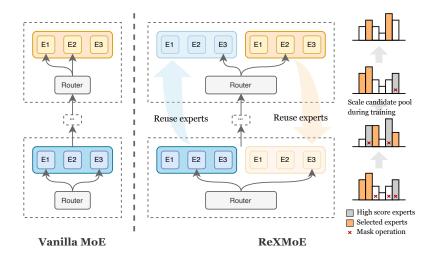


Figure 1: **Overview of REXMOE.** Compared to vanilla MoE, REXMOE enables more flexible expert combinations by reusing experts from adjacent layers. The only additional overhead comes from the router, which learns to route tokens to the expanded candidate pool. During training, REXMOE progressively scale the candidate pool by gradually reducing the number of masked experts until all experts are available.

enriches the space of expert combinations, whereas larger experts preserve stronger representational capacity but limit routing diversity. Such a trade-off is the core of the MoE architectural design.

In practice, recent works show trends toward adopting *finer-grained experts* in MoE design. For example, early Mixtral-of-Experts models (Jiang et al., 2024) employed 8 candidate experts per layer, whereas more recent models such as Qwen3 (Yang et al., 2025) series expand this to 128 experts, DeepSeek-V3 (Liu et al., 2024b) scales the design to 256 experts. From a combinatorial perspective, fragmenting experts into smaller units substantially increases the number and diversity of possible routing combinations, thereby enhancing the expressiveness of MoE models and improving their ability to capture more specialized knowledge (Dai et al., 2024).

A key challenge in existing MoE designs lies in the *layer-local* routing mechanism, where each layer's router is restricted to its own expert pool. This constraint ties architectural choices to perlayer budgets and prevents more flexible balancing between the capacity of individual experts and the combinatorial flexibility of the expert pool. As a result, finer granularity comes at the cost of reduced representational capacity for each expert, since smaller experts correspond to reduced hidden dimensionality in their feed-forward networks. On the other hand, preserving expert dimensionality while simply increasing the number of experts inflates the overall parameter count. This fundamental challenge motivates us to explore new architectural directions that enrich expert combinations without reducing expert capacity or inflating model size.

To address such a challenge, we propose REXMOE, a novel approach to MoE architecture design that extends routing beyond the conventional *layer-local* boundary. By allowing routers to reuse experts across grouped adjacent layers, REXMOE decouples expert dimensionality from per-layer parameter budgets and introduces a new dimension in MoE design: models can realize richer expert combinations without sacrificing individual expert capacity or inflating the total parameter count. Furthermore, we present a Progressive Scaling Routing (PSR) strategy for training, which enhances the performance of models with reused expert pools. As illustrated in Figure 1, REXMOE reuses experts across adjacent layers with only negligible additional router parameters, while PSR gradually expands the candidate expert pool during training. Extensive experiments on models ranging from 0.5B to 7B parameters across different architectures show that REXMOE consistently improves performance under fixed dimensionality configurations. In addition, ablation studies highlight key design factors and confirm the practicality of our approach. Qualitative analysis further suggests that REXMOE enhances task-specific specialization. Together, these results establish REXMOE as an effective and scalable paradigm for advancing MoE-based LLMs.

In this paper, we make following contributions:

- We design REXMOE, a method that breaks the limitation of *layer-local* routing in MoE architectures. By reusing experts across adjacent layers while adding only negligible router parameters, REXMOE significantly increases the flexibility of expert combinations.
- We propose a *Progressive Scaling Routing* strategy in REXMOE, which gradually enlarges
 the candidate expert pool during training, thereby reducing language modeling loss and
 improving downstream task accuracy.
- We have conducted extensive experiments to demonstrate that REXMOE consistently improves both language modeling ability and downstream task performance across different model sizes and architectures, establishing REXMOE as a practical design paradigm for parameter-efficient and scalable MoE-based LLMs.

2 RELATED WORKS

Mixture-of-Experts. The strength of large models lies in their vast parameter counts, but this also brings the challenge of high computational cost. The Mixture-of-Experts (MoE) framework was introduced to decouple parameter size from per-token computation in large language models (LLMs) during both training and inference. In MoE-based transformer architectures (Vaswani et al., 2017), sparse MoE blocks (Shazeer et al., 2017; Fedus et al., 2022; Lepikhin et al., 2020) replace the Feed-Forward Networks (FFNs), improving efficiency while preserving strong performance. A notable example is Mixtral-of-Experts (Mixtral MoE) (Jiang et al., 2024), an open-source architecture that activates 2 experts from a pool of 8. Compared with dense models of similar computational cost, Mixtral delivers stronger performance across multiple downstream tasks. More recently, DeepSeek-MoE (Dai et al., 2024) adopts a fine-grained MoE design by dividing each FFN into smaller experts, enabling more flexible expert combinations without inflating the total parameter count. The opensource community has continued to advance this trend toward finer-grained experts. For instance, the Qwen3 series (Yang et al., 2025) employs 128 experts, while Kimi-K2 (AI, 2025) scales to 384 experts. Both demonstrate strong performance across diverse domains, reinforcing the idea that richer expert combinations often translate into better results. Overall, the success of these models highlights *finer-grained* design as a reliable and promising direction for future MoE architectures.

Parameter Reusing in Transformers. The standard Transformer constructs token representations using a stack of L distinct layers, each consisting of a self-attention mechanism and a feed-forward network (FFN). Recent studies (Dehghani et al., 2018; Csordás et al., 2024; Bae et al., 2024) have explored reusing a shared set of weights across multiple layers, showing promising gains in parameter efficiency. Universal Transformers (Dehghani et al., 2018) replace the standard stack of unique layers with a single parameter-shared block that is applied recurrently, refining token representations in parallel. This combines the inductive bias of RNNs with the parallelization benefits of Transformers. MoEUT (Csordás et al., 2024) extends this idea by integrating the Mixture-of-Experts (MoE) paradigm into the recurrent Universal Transformer architecture, addressing its parameter-compute scaling bottleneck. This method increases model capacity while maintaining computational efficiency, making parameter sharing feasible for large-scale language modeling. Relaxed Recursive Transformers (Bae et al., 2024) further relax the strict layer-tying constraint by introducing depthwise low-rank adaptation (LoRA) modules, improving performance while keeping the overall model compact. A related approach to parameter reuse in MoE blocks is WideNet (Xue et al., 2022), which derives its strategy from the perspective of reducing total parameters by recurrently reusing the weights of FFNs and self-attention blocks across all Transformer layers. Experiments on small-scale models for both CV and NLP tasks demonstrate its effectiveness, highlighting parameter sharing as a practical way to improve parameter efficiency.

3 Method

In this section, we first revisit the widely used TopK routing strategy for Mixture-of-Experts (MoE) models. We then introduce REXMOE, which enlarges the candidate expert pool by reusing experts across adjacent layers. To further improve performance when increasing the number of routed experts, we propose a Progressive Scaling Routing (PSR) strategy for training. An overview of REXMOE is shown in Figure 1.

3.1 REVIEW OF TOPK ROUTING MIXTURE-OF-EXPERTS

In a standard L-layer transformer-based Mixture-of-Experts (MoE) architecture, the Feed-Forward Network (FFN) blocks are replaced with MoE blocks, each comprising N experts and a router. The candidate expert pool of layer-l, $\mathcal{E}^l = \{ E_1, E_2, \dots, E_N \}$, consists of N experts, each instantiated as an independent FFN. The router is responsible for assigning each input token to a subset of experts. Specifically, the router utilizes the gating network, which is parameterized by trainable weights, computes the probability distribution for the given input, then selects the corresponding experts according to its routing strategy. In TopK routing MoE, the output of the MoE block in l-th layer is computed as follows:

$$\mathbf{h}' = \sum_{i=1}^{N} (\mathbf{g}_i \, \mathbf{E}_i(\mathbf{h})) \tag{1}$$

$$g_i = \begin{cases} s_i, & s_i \in \text{TopK}\left(\left\{s_j \mid 1 \le j \le N\right\}\right), \\ 0, & \text{otherwise}, \end{cases}$$
 (2)

$$\mathbf{s} = \text{Softmax}(\mathbf{W} \cdot \mathbf{h}) \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{N \times d}$ is the weight of the gating network in the router, and g_i is the gating score for expert-i. For brevity, we omit the self-attention and layer normalization in the above formulations.

3.2 EXPANDING CANDIDATE EXPERT POOL

To overcome the limitation of the *layer-local* routing mechanism, we expand the candidate expert pool by allowing the router to select from experts in *grouped* adjacent layers. Consider an L-layer MoE. Let r denote the expert reuse frequency across layers, and let \mathcal{E}^i represent the candidate expert pool of the i-th layer. In REXMOE, the grouped candidate expert pool for layer l is defined as:

$$\mathcal{U} := \bigcup_{i \in G} \mathcal{E}^i, \quad G = \{ \lfloor l/r \rfloor + k \mid 1 \le k \le r \}$$
 (4)

Here, group G is formed by r consecutive layers starting from the $\lfloor l/r \rfloor$ -th layer. In this way, the candidate expert pool of each layer becomes r times larger than in the vanilla setting. The computation of each layer's MoE block is then formulated as:

$$\mathbf{h}' = \sum_{i=1}^{rN} (\mathbf{g}_i \, \mathbf{U}_i(\mathbf{h})) \tag{5}$$

where $U_i \in \mathcal{U}$ is the *i*-th expert in the expanded pool. By increasing r, the enlarged the candidate pool enables more diverse expert combinations and yielding performance gains.

Discussion. A specific router configuration can restrict routing to only local experts, in which case the model reduces to the vanilla MoE. This guarantees that our method always matches the baseline performance under the most constrained setting. When expert reuse is enabled, however, each MoE block includes a larger set of experts, which allows for more diverse expert combinations but can also lead to imbalanced routing patterns. Consequently, load imbalance becomes a critical bottleneck that not only limits generalization but also introduces challenges during training.

3.3 Progressive Scaling Routing Strategy

Another key component of REXMOE is the Progressive Scaling Routing (PSR) strategy, which gradually increases the number of candidate experts during training. When reusing experts from r layers in a TopK MoE with N experts per layer, each router can access up to rN candidates. Instead of training the router to select from all rN candidates from the start, we adopt a progressive scheme: the number of available candidates begins at N and is linearly expanded over the course of training. At iteration t, the candidate expert pool size N_t is defined as:

Table 1: Base MoE architectures used in experiments. "MoE-0.5BA0.07B" denotes a MoE model with 0.5B total parameters and 0.07B active parameters per token. "SE" means "Shared Experts". This naming convention applies to all models.

Model	Hidden Size	Intermediate Size	#Layers	Heads (Q / KV)	#Experts (Shared + Routed / Total)
MoE-0.5BA0.07B	768	384	16	16/2	4/32
MoE-0.5BA0.07B-SE	768	384	16	16/2	1 + 3 / 32
MoE-2.3BA0.3B	512	744	32	16/2	8 / 64
MoE-2.3BA0.3B-SE	512	744	32	16/2	2 + 6 / 64
MoE-7BA3B-SE	2048	1408	32	16 / 4	2+6/64

$$N_{t} = \begin{cases} N, & t \leq t_{s}, \\ \lfloor (1 + \frac{(r-1)(t-t_{s})}{t_{e}-t_{s}}) N \rfloor, & t_{s} < t \leq t_{e}, \\ rN, & t > t_{e}, \end{cases}$$
 (6)

where t_s and t_e specify the start and end iterations of the scaling schedule, respectively. At each iteration, we randomly mask $(rN-N_t)$ experts by setting their gating scores to zero before applying the TopK selection for each token. This design follows the principle of curriculum learning, allowing the model to gradually learn richer and more diverse expert representations.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training environment. All models are trained with Megatron-LM (Shoeybi et al., 2019), an open-source framework for large-scale language model training. We modified the MoE Block and Topk Router implementations to support cross-layer expert reuse and the Progressive Scaling Routing strategy during training. All models are pre-trained from scratch without instruction tuning, using the same hyperparameters across all runs. The sequence length is 4,096 and the total batch size is 512, resulting in a global batch size of 2M tokens. For optimization, we use AdamW (Loshchilov & Hutter, 2017) with $\beta_1=0.9,\,\beta_2=0.95$, weight decay 0.1, and a gradient clipping ratio of 1.0. The learning rate is scheduled to start at 3×10^{-4} and decay to 3×10^{-5} following a cosine schedule. Further details are provided in Appendix B.2. All training jobs are conducted on 4 nodes, each equipped with $32\times NVIDIA$ Hopper GPUs.

Model architecture. We adopt the widely used Mixture-of-Experts (MoE) transformer architecture with consistent dimensionality settings across all ablation studies. The only differences lie in the router parameters under different reuse configurations. The architectural configurations are summarized in Table 1, where each model name specifies the number of activated and total parameters. The suffix "-SE" indicates that the architecture employs shared experts (Dai et al., 2024; Rajbhandari et al., 2022), and REX models follow the same naming convention. In addition, "-R $\{r\}$ " denotes that experts are reused across r layers.

Training data. We use the sample-100BT partition¹ from fineweb-edu dataset (Lozhkov et al., 2024; Penedo et al., 2024). The tokenizer is from LLaMA-2 (Touvron et al., 2023), with a vocabulary size of 32,000. Since the vocabulary is relatively small, the LLaMA-2 tokenizer does not achieve a high compression ratio. As a result, the processed 100B tokens cover around 87% of the original text. To ensure fair comparison, we fixed the data-parallel size and the shuffle seed, so that all experiments were trained on the same tokens in the same order, making the results directly comparable.

Evaluation metrics. We use lm-evaluation-harness (Gao et al., 2024) to evaluate performance on downstream tasks. Specifically, we report zero-shot accuracy on ARC-Easy (ARC-E) & ARC-Challenge (ARC-C) (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al.,

¹https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu/viewer/sample-100BT

Table 2: Comparison between REXMOE and vanilla MoE models. All models are trained on 100B tokens. Task abbreviations: **Hella.** = HellaSwag, **LAMB.** = LAMBADA, **Lg.QA** = LogiQA, **Op.QA** = OpenBookQA, **Wino.** = WinoGrande. The best accuracy is highlighted in bold.

Model	ARC-E	Hella.	LAMB.	Lg.QA	Op.QA	PIQA	SciQ	SIQA	Wino.	Avg.↑
MoE-0.5BA0.07B	50.67	38.38	32.37	28.42	31.00	65.29	71.20	38.84	53.04	45.47
REX-0.5BA0.07B-R2	52.31	39.06	33.75	25.65	32.80	65.78	71.10	38.33	51.22	45.56
REX-0.5BA0.07B-R4	53.91	39.46	32.76	25.35	32.80	66.81	71.00	38.38	52.17	45.85
MoE-0.5BA0.07B-SE	51.85	38.90	33.26	24.88	32.00	66.05	70.60	39.05	51.54	45.35
REX-0.5BA0.07B-SE-R2	52.06	39.28	32.43	26.57	35.00	66.54	71.80	37.41	51.93	45.89
REX-0.5BA0.07B-SE-R4	53.11	39.39	34.00	28.88	33.40	67.46	71.90	38.69	50.36	46.35
MoE-2.3BA0.3B	58.42	47.14	37.55	27.19	34.80	69.21	75.80	38.69	53.51	49.15
REX-2.3BA0.3B-R2	61.32	46.84	37.20	28.57	35.00	69.48	76.50	39.61	52.33	49.65
REX-2.3BA0.3B-R4	60.94	47.96	38.75	28.42	37.00	70.18	76.30	39.36	53.12	50.23
MoE-2.3BA0.3B-SE	58.42	48.79	38.13	25.35	37.00	69.53	75.00	40.28	52.17	49.41
REX-2.3BA0.3B-SE-R2	59.09	47.99	38.54	27.34	37.60	69.48	74.20	39.56	52.72	49.61
REX-2.3BA0.3B-SE-R4	58.71	48.59	39.01	28.26	39.00	70.67	76.10	39.66	52.80	50.31

2019), LAMBADA (Paperno et al., 2016), LogiQA (Liu et al., 2021), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SciQ (Welbl et al., 2017), SIQA (Sap et al., 2019) and WinoGrande (Sakaguchi et al., 2021). For evaluation of the impact on inference speed after reusing experts from adjacent layers, we adapted REXMOE to vLLM (Kwon et al., 2023) and report the throughput (tokens per second) for prefill and decoding stages. Sampling is disabled in generation.

4.2 Main Results

4.2.1 EVALUATION ON DOWNSTREAM TASKS

Comparisons to vanilla MoEs. We report the accuracy on downstream benchmarks in Table 2. The results show that the proposed REXMOE models consistently outperform vanilla MoE baselines across different model scales and benchmark tasks. Overall, REXMOE achieves stable improvements in both R2 and R4 configurations, with R4 often delivering the highest average accuracy. For example, compared to the base MoE-2.3BA0.3B, the R4 model attains the best results on tasks such as HellaSwag, LAMBADA, OpenBookQA, PIQA, and SIQA, raising the average score to 50.23%, which clearly surpasses the baseline's 49.15%. Similarly, under the "SE" setting, REXMOE-R4 outperforms the corresponding base MoE-2.3BA0.3B-SE. For smaller models in the MoE-0.5BA0.07B series, the advantage of REXMOE also remains consistent, where both R2 and R4 configurations yield notable gains in average accuracy over the baseline. More detailed task-wise accuracy trends during training can be found in Figure 6 and Figure 7 in the appendix. In summary, these results demonstrate that REXMOE consistently improves performance across different model scales and architectures, particularly on reasoning and knowledge-intensive tasks, highlighting its robustness, scalability, and general effectiveness.

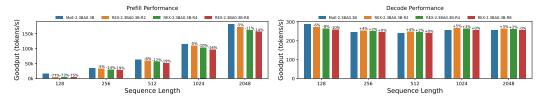
Table 3: Comparisons between REXMOE and open-source models. We report results for models with equivalent total or activated parameters on selected language understanding benchmarks. Our method achieves competitive or superior performance across tasks.

Model	#Act. Params	Data	ARC-E	Hella.	LAMB.	Lg.QA	PIQA	SciQ	Wino.
Llama2-7B (Touvron et al., 2023) MPT-7B-Base (Team, 2023)	7B/7B 7B/7B	2T 1T	76.4 67.3	78.6 76.1	73.9 70.3	30.7	78.1 79.9	93.7	69.3 68.3
DeepSeekMoE-16B (Dai et al., 2024) LLaMA-MoE-8B (Zhu et al., 2024) OpenMoE-8B (Xue et al., 2024)	3B/16B 3B/8B 2.1B/8B	2T - 1T	68.1 60.2 64.1	77.1 70.8 45.5	- 66.6 -	30.6	80.2 77.5 74.2	84.2	70.2 63.6 60.3
REX-7BA3B-SE-R3	3B/7B	1T	75.7	69.0	63.9	33.2	75.0	94.2	65.9

Comparisons to LLMs with equivalent effective parameters We compare REXMOE with representative open-source dense and MoE models of similar total or activated parameter scales in Ta-

ble 3. For a fair comparison, we scale the training data of REX-7BA3B-SE-R3 to 1T tokens sampled from fineweb-edu. The model exhibits well-balanced performance, achieving highest results on LogiQA and SciQ, even when compared to Llama2-7B (Touvron et al., 2023), which uses more activated parameters and is trained on a larger corpus. Meanwhile, REXMOE remains highly competitive across the other benchmarks. These results demonstrate the effectiveness of REXMOE as model size and training data increase, highlighting its scalability for high performance.

4.2.2 IMPACT ON INFERENCE SPEED



(a) Prefill goodput under different sequence length. (b) Decode goodput under different sequence length.

Figure 2: Comparison of prefill and decode goodput between base MoE and REX models. Numbers above the bars indicate the relative speedup over the base MoE.

We adapt REXMOE to the vLLM inference engine (Kwon et al., 2023) to evaluate the impact of expert reuse on practical applications. We fix the output length at 128 tokens and vary the input length to assess both prefill and decoding performance across different sequence lengths. The detailed results are shown in Figure 2. Although the computational overhead compared to vanilla MoE is negligible, REXMOE introduces a larger number of experts into each MoE block, which increases I/O operations during the prefill stage. As a result, the inference speed of the reuse scheme experiences a noticeable decline. As shown in Figure 2(a), a larger candidate expert pool leads to slower prefill speed, with the performance degradation being more pronounced when the input length is relatively short. Since the prefill stage usually accounts for only a small portion of the total time, the decoding stage is of greater practical importance. As illustrated in Figure 2(b), REXMOE achieves comparable performance across different sequence lengths in decoding stage.

4.3 ABLATION STUDIES

4.3.1 EFFECT OF EACH COMPONENT IN REXMOE

To demonstrate the effectiveness of each component in REXMOE, we provide comparative evaluations in Table 4. We utilize the MoE-2.3BA0.3B model as our baseline and set the expert reuse frequency across layers as 4. In addition to using the same average accuracy over the benchmarks reported in Table 2, we evaluate the validation perplexity (PPL) on WikiText (Merity et al., 2016). We find that the simple expansion of the experts' pool results in only a

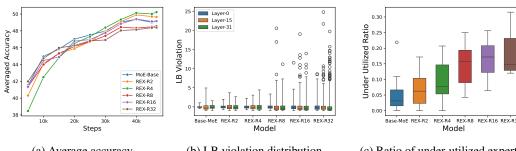
Table 4: Average accuracy on benchmarks and PPL on WikiText.

Model	Avg. Acc.↑	PPL↓
MoE-2.3BA0.3B	49.15	21.19
+ Expert Reuse (4)	49.28	21.12
+ PSR	50.23	20.73

marginal improvement. Specifically, a 0.13% increase in averaged accuracy on downstream tasks and 0.07 in PPL. Furthermore, the incorporation of the PSR strategy yields a significant improvement in model performance by 1.05% in the average accuracy and a drop in PPL of 0.46. Comprehensive benchmarks results of each task can be found in Table 6. These results demonstrate the effectiveness of the expert reuse and PSR strategy.

4.3.2 EFFECT OF SCALING EXPERT REUSE GROUP SIZE

We investigate the impact of scaling the expert reuse group size in the 2.3B variant of REXMOE, where the reuse frequency ranges from 2 to 32 layers. As presented in Figure 3(a), we illustrate the performance trends on downstream tasks during training for different configurations. In the early training phase, both REX-R2 and REX-R4 underperform the baseline MoE model; however, they eventually surpass it as training progresses, with larger reuse group sizes generally leading



- (a) Average accuracy.
- (b) LB violation distribution.
- (c) Ratio of under-utilized experts.

Figure 3: Average accuracy during training, distribution of load balance violations (degree of expert load imbalance), and distribution of under-utilized experts ratios (larger values indicate more inactive experts) under different cross-layer expert reuse sizes.

to better performance. In contrast, REX-R16 and REX-R32 initially match or even exceed the baseline but later fall behind in the later stages of training. More detailed evaluation results can be found in Table 7 in the appendix. These results suggest that maintaining an appropriate balance in the number of reused layers is critical for sustaining high performance throughout training.

To investigate the cause of the performance decline as the number of reused layers increases, we reserve a validation set from the C4 corpus² to evaluate load balance. Following the MaxVio metric in (Wang et al., 2024), we adopt the Load Balance Violation (LBV) metric to quantify the degree of load imbalance in the MoE block. Specifically, the LBV of expert i is computed as:

$$LBV_i = \frac{Load_i - \overline{Load_i}}{\overline{Load_i}}$$
 (7)

where $Load_i$ denotes the number of tokens assigned to expert i, and $\overline{Load_i}$ is the average expert load. Under perfect balance, LBV_i equals 0. As shown in Figure 3(b), larger r values lead to more significant deviations among outliers in the distribution of LBV_i , indicating that the model tends to activate only a few experts and suffers from a collapse phenomenon during training. In addition, we present the distribution of under-activated experts in Figure 3(c). We observe that as the candidate pool further expands, more experts remain barely activated throughout training. This imbalance in expert utilization explains why the performance of REX-R16 and REX-R32 is lower than baselines.

EFFECT OF PROGRESSIVE SCALING ROUTING 4.3.3

We investigate an alternative Progressive Scaling Routing (PSR) strategy to validate the optimal configuration adopted in our main experiments. Specifically, we introduce PSR-Stepwise, which keeps the number of candidate experts fixed over certain training intervals. In contrast, the strategy described in subsection 3.3 is referred to as PSR-Linear, as it provides a smoother and continuous expansion of the candidate expert pool. We use MoE-2.3BA0.3B as the baseline model and apply different PSR strategies to REX-R4. The corresponding training curves are shown in Figure 3. For both PSR-Stepwise and PSR-Linear, train-

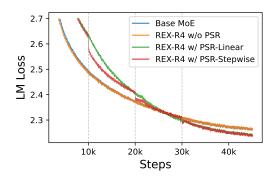


Figure 4: Loss curves for different strategies.

ing starts with 64 candidate experts, and scaling begins at step 10k. In PSR-Linear, the candidate pool is gradually increased to 256 by step 30k. In PSR-Stepwise, the candidate pool is set to 128, 192, and 256 at steps 10k, 20k, and 30k, respectively.

We present the training loss curves of these models in Figure 4. As shown in the figure, the loss curves of Base MoE and REX-R4, where PSR is not enabled, remain almost identical. When

²https://huggingface.co/datasets/allenai/c4

different PSR strategies are applied, model convergence is initially slowed. However, once the candidate pool begins to expand, the models achieve lower loss than those trained without progressive scaling. Notably, PSR-Stepwise accelerates loss reduction during the mid-training phase.

As summarized in Table 5, both PSR strategies deliver clear performance improvements at final convergence, with detailed results provided in Table 6 in the appendix. Nevertheless, the final loss is comparable between the two strategies, while PSR-Linear achieves stronger overall performance on downstream tasks. Therefore, we adopt PSR-Linear to train all models.

Table 5: Average accuracy on benchmarks and PPL on WikiText.

Model	Avg. Acc.↑	PPL↓
REX w/o PSR	49.28	21.12
REX w/ PSR-Stepwise	49.59	20.76
REX w/ PSR-Linear	50.23	20.73

4.4 QUALITATIVE ANALYSIS

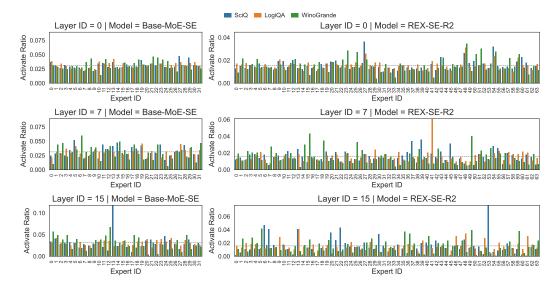


Figure 5: Activate ratio of MoE-SE and REX-SE-R4 across layers in different tasks. The gray dashed lines indicate uniform distribution. REXMOE shows stronger ability in task specialization.

In Figure 5, we present the expert activation ra-

tios of Base-MoE-SE and REX-SE-R2 across layers 0, 7, and 15 on the SciQ, LogiQA, and Wino-Grande tasks. For Base-MoE-SE, the distribution of activated experts remains relatively uniform, with only minor variation across tasks. In contrast, REX-SE-R2 exhibits clear task-specific specialization. For instance, certain experts (e.g., Experts 25 and 49) are activated far more frequently for WinoGrande than for the other tasks, especially in Layers 7 and 15. Similar trends are observed on other tasks, as shown in Figure 8 and Figure 9 in the appendix. These results suggest that the expanded expert pool of REX-SE-R2 allows for more effective task-specific allocation, encouraging the emergence of specialized experts and producing an ensemble-like effect in multi-task scenarios.

5 Conclusion

In this work, we present REXMOE, a novel MoE design paradigm that overcomes the limitation of *layer-local* routing. By allowing routers to reuse experts across grouped adjacent layers, REXMOE decouples expert dimensionality from per-layer budgets and substantially enlarges the candidate expert pool with only negligible router overhead. Combined with the Progressive Scaling Routing strategy, it further enhances training stability and performance. Extensive experiments across diverse architectures and model scales show that REXMOE consistently improves language modeling perplexity, downstream task accuracy, and the ability to learn task-specialized experts. Overall, these results establish REXMOE as a parameter-efficient and practically scalable paradigm for designing MoE-based LLMs.

ACKNOWLEDGEMENT

This work is partially funded by Research Council of Finland (grant number 362729) and Business Finland (grant number 169/31/2024).

REPRODUCIBILITY STATEMENT

We provide sufficient details for reproducing our key experiments. Training configurations are described in subsection 4.1 and subsection B.2, while the data processing pipeline is detailed in subsection B.1.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Moonshot AI. Kimi-k2: Open-source models by moonshot ai. https://github.com/ MoonshotAI/Kimi-K2, 2025.
- Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive transformers: Effective parameter sharing with layer-wise lora. *arXiv preprint arXiv:2410.20672*, 2024.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.
- Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pp. 4057–4086. PMLR, 2022.
- 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 NAACL-HLT*, pp. 2924–2936, 2019.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.
- Róbert Csordás, Kazuki Irie, Jürgen Schmidhuber, Christopher Potts, and Christopher D Manning. Moeut: Mixture-of-experts universal transformers. Advances in Neural Information Processing Systems, 37:28589–28614, 2024.
- Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.
- Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.
- 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.

- 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.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.
- Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.
- Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3622–3628, 2021.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
- Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.
- Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, April 2025.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.
- OpenAI. Introducing openai o3 and o4-mini, May 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/.

- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, 2016.
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.
- Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pp. 18332–18346. PMLR, 2022.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *COMMUNICATIONS OF THE ACM*, 64(9), 2021.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.
- 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.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *W-NUT 2017*, pp. 94, 2017.
- Fuzhao Xue, Ziji Shi, Futao Wei, Yuxuan Lou, Yong Liu, and Yang You. Go wider instead of deeper. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8779–8787, 2022.
- Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv* preprint arXiv:2402.01739, 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, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025.

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

Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. Llama-moe: Building mixture-of-experts from llama with continual pre-training. *arXiv* preprint *arXiv*:2406.16554, 2024. URL https://arxiv.org/abs/2406.16554.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

We acknowledge the use of Large Language Models (LLMs) to assist in writing and polishing this paper. Their role was limited to improving the clarity and readability of the manuscript; they were not involved in the design of the methodology or in the scientific analysis.

B ADDITIONAL EXPERIMENTS DETAILS

B.1 DATA PROCESSING

We use the sample-100BT partition³ of fineweb-edu (Lozhkov et al., 2024) for our main experiments. Each sample in the dataset is tokenized independently and then randomly concatenated into sequences of 4,096 tokens, which are used for training.

B.2 Hyper-parameters and Parallelism Configurations

We use the same hyper-parameters for all model training runs. The training sequence length is set to 4,096, and the global batch size is 512, resulting in a training batch size of 2M tokens. The base frequency for Rotary Positional Embedding (ROPE) (Su et al., 2024) is 10,000. For optimization, we use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1, gradient clip ratio is 1.0. We adopt a warmup–cosine-decay learning rate scheduler, with an initial learning rate of 3×10^{-4} that decays to 3×10^{-5} by the end of training. The number of warmup steps is fixed at 100 for all experiments. When the number of routed experts exceeds 8, we enable Expert Parallelism (EP) with a parallelism size of 8 to accelerate training. No other parallelism strategies, such as Tensor Parallelism (TP) or Pipeline Parallelism (PP), are used in these runs. We globally fix the random seed to 42.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 FULL EVALUATION RESULTS FOR DIFFERENT PSR VARIANTS

Table 6: Comparisons between base MoE and variants of REXMOE.

Model	ARC-E	Hella.	LAMB.	Lg.QA	Op.QA	PIQA	SciQ	SIQA	Wino. Avg.↑
Base MoE	58.42	47.14	37.55	27.19	34.80	69.21	75.80	38.69	53.51 49.15
REX-R4 w/o PSR	58.16	46.94	38.52	25.96	36.40	70.67	74.50	39.46	52.88 49.28
REX-R4 w/ PSR-Stepwise	60.65	48.25	37.67	27.04	34.40	70.84	74.60	39.10	53.75 49.59
REX-R4 w/ PSR-Linear	60.94	47.96	38.75	28.42	37.00	70.18	76.30	39.36	53.12 50.23

Complete evaluation results for different PSR variants are provided in Table 6, with the base model being MoE-2.3B-A0.3B.

C.2 Full Evaluation Results for Different Reuse Sizes

Table 7: Comparisons between base MoE and REXMOE with different reuse sizes.

Model	ARC-E	Hella.	LAMB.	Lg.QA	Op.QA	PIQA	SciQ	SIQA	Wino.	Avg.↑
REX-R8	58.75	46.80	37.07	26.27	35.00	69.97	72.50	37.97	52.64	48.55
REX-R16	58.59	46.79	38.48	27.80	35.40	70.02	72.20	39.36	53.83	49.16
REX-R32	58.21	46.28	35.26	27.04	35.60	70.35	72.80	39.15	50.91	48.40

Complete evaluation results for different reuse sizes are provided in Table 7, with the base model being MoE-2.3B-A0.3B.

Table 8: Architecture of Top2 MoE model used in the additional experiments.

Model	Hidden Size	Intermediate Size	#Layers	Heads (Q / KV)	#Experts (Shared + Routed / Total)
MoE-0.5BA0.13B	768	1536	16	16/2	2/8

Table 9: Comparisons between base MoE and REXMOE with different reuse sizes.

Model	ARC-E	Hella.	LAMB.	Lg.QA	Op.QA	PIQA	SciQ	SIQA	Wino. Avg.↑
MoE-0.5BA0.13B									
REX-0.5BA0.13B-R2	52.82	39.26	33.18	27.96	32.00	66.05	70.60	38.08	52.80 45.86
REX-0.5BA0.13B-R2 REX-0.5BA0.13B-R4	51.94	39.34	32.25	27.04	32.80	65.56	70.60	38.69	50.51 45.41

C.3 EVALUATION ON TOP2 MOE

We further apply REX to a Top2 MoE, with its architecture detailed in Table 8. The corresponding evaluation results are reported in Table 9.

C.4 TASK-WISE ACCURACY

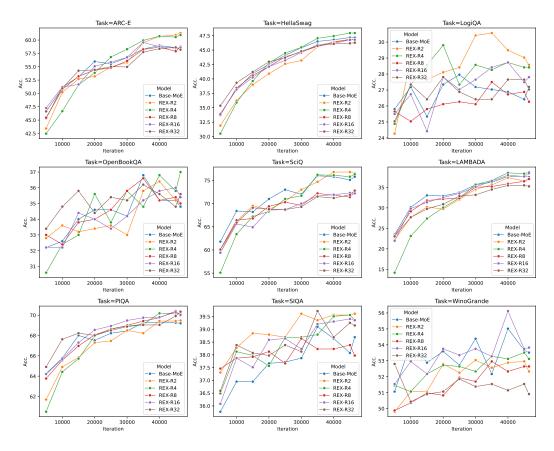


Figure 6: **Task-wise accuracy change as training progresses.** Base-MoE is MoE-2.3BA0.3B.

³https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu/viewer/sample-100BT

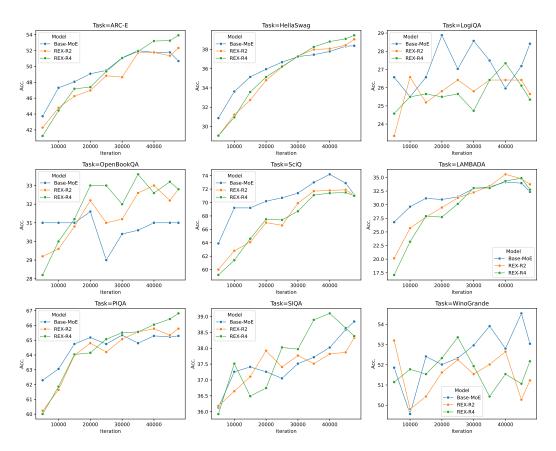


Figure 7: Task-wise accuracy change as training progresses. Base-MoE is MoE-0.5BA0.1B.

C.5 TASK-WISE EXPERTS SELECTION VISUALIZATION

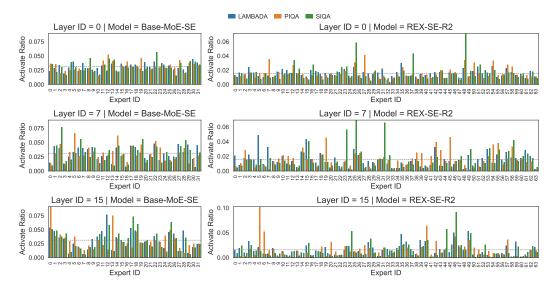


Figure 8: Activate ratio of MoE-SE and REX-SE-R4 across layers in different tasks. The gray dashed lines indicate uniform distribution.

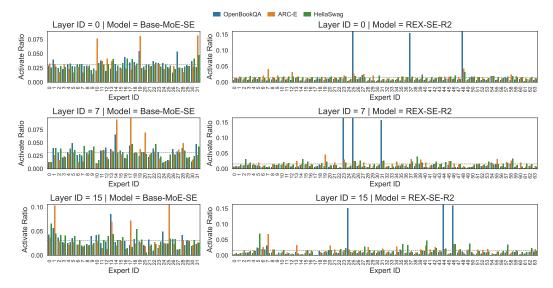


Figure 9: Activate ratio of MoE-SE and REX-SE-R4 across layers in different tasks. The gray dashed lines indicate uniform distribution.