
S'MoRE: Structural Mixture of Residual Experts for Parameter-Efficient LLM Fine-tuning

Hanqing Zeng
Meta AI
zengh@meta.com

Yinglong Xia
Meta AI
yxia@meta.com

Zhuokai Zhao
Meta AI
zhuokai@meta.com

Chuan Jiang
Meta AI
gjiang@meta.com

Qiang Zhang
Meta AI
qiangzhang@meta.com

Jiayi Liu
Meta AI
liujiayi@meta.com

Qunshu Zhang
Meta AI
qunshuzhang@meta.com

Lizhu Zhang
Meta AI
lizhu@meta.com

Xiangjun Fan
Meta AI
maxfan@meta.com

Benyu Zhang
Meta AI
byzhang@meta.com

Abstract

Fine-tuning pre-trained large language models (LLMs) presents a dual challenge of balancing parameter efficiency and model capacity. Existing methods like low-rank adaptations (LoRA) are efficient but lack flexibility, while Mixture-of-Experts (MoE) enhance model capacity at the cost of more & under-utilized parameters. To address these limitations, we propose **Structural Mixture of Residual Experts (S'MoRE)**, a novel framework that seamlessly integrates the efficiency of LoRA with the flexibility of MoE. Conceptually, S'MoRE employs *hierarchical* low-rank decomposition of expert weights, yielding residuals of varying orders *interconnected* in a multi-layer structure. By routing input tokens through sub-trees of residuals, S'MoRE emulates the capacity of numerous experts by instantiating and assembling just a few low-rank matrices. We craft the inter-layer propagation of S'MoRE's residuals as a special type of Graph Neural Network (GNN), and prove that under similar parameter budget, S'MoRE improves *structural flexibility* of traditional MoE (or Mixture-of-LoRA) by exponential order. Comprehensive theoretical analysis and empirical results demonstrate that S'MoRE achieves superior fine-tuning performance, offering a transformative approach for efficient LLM adaptation. Our implementation is available at: <https://github.com/ZimpleX/SMoRE-LLM>.

1 Introduction

Large language models (LLMs) have achieved remarkable success across a wide range of tasks by leveraging extensive pretraining on vast datasets, which equips them with general-purpose knowledge [Achiam et al., 2023, Dubey et al., 2024, Team et al., 2024a, Liu et al., 2024a, Anthropic, 2024]. However, the versatility of pre-trained LLMs often falls short when applied to specialized tasks [Hadi et al., 2023, Ling et al., 2023, Chen et al., 2024]. Fine-tuning addresses this limitation by refining LLM's capabilities to focus on the nuances of different domains [Zhang et al., 2023, Wang et al., 2024a, 2025]. However, it introduces a fundamental tension between balancing parameter efficiency and the need for expanded model capacity to capture task-specific complexity [Wang et al., 2024b].

Low-Rank Adaptations (LoRA) [Hu et al., 2021, Mao et al., 2025] are parameter efficient but lack the capacity required by complex tasks. On the other hand, Mixture-of-Experts (MoE) [Lepikhin et al., 2020, Fedus et al., 2022b, Cai et al., 2024, Dai et al., 2024] architectures improve model

capacity by enabling conditional computation where different tokens activate different experts. However, traditional MoEs are often less parameter- or data-efficient since the multiple experts need to separately learn their own parameters. Moreover, increasing the number of experts poses the challenge of balancing their utilization with low routing overhead.

Thus, to improve model capacity while maintaining parameter efficiency, we look into other scaling dimensions, such as *routing flexibility* unveiled by recent literature. Dai et al. [2024], Ludziejewski et al. [2024], He [2024] empirically show that under the same parameter budget, a large number of small (i.e., fine-grained) experts is more powerful than a small number of large experts. Ludziejewski et al. [2024] quantitatively derives the MoE scaling law with respect to experts’ granularity. DeepSeek-MoE [Dai et al., 2024] elaborates the intuition: Consider an MoE system with 16 rank-128 experts. Under top-2 routing, each token has $\binom{16}{2} = 120$ ways to select its own expert combination. If we break down each rank-128 expert as 4 rank-32 experts (and correspondingly use top-8 routing to keep the same activated parameters), each token now has $\binom{4 \times 16}{8} = 4.4\text{B}$ routing choices.

Yet, “breaking down experts into finer granularity” may not be the most effective way of increasing routing flexibility. There are two potential limitations: 1) in the parameter-efficient fine-tuning (PEFT) scenario, each expert is already of a low rank, making it questionable if finer granularity is desirable; and 2) with more number of experts comes higher requirement on the router’s capability & the learning algorithm – it may be challenging to maintain good expert utilization when the number of experts grows. These lead to our goal: *without* increasing the number of experts, we aim at emulating more routing choices by exploiting the *power of structure*. That is, instead of merely addressing the problem of “which experts to activate” (like most existing literature), we further ask: “how should the activated experts be connected”? To answer it, we first arrange experts into multiple layers. Then the router iterates through the layers and constructs a tree of activated experts, through which the input token propagates. The key observation is that, the same set of experts can interconnect in different ways to form exponentially many *non-isomorphic* tree structures, each yielding a distinct output. We thus formalize such intuition by extending the aforementioned routing flexibility into a new metric, **structural flexibility**, and theoretically quantify its exponential growth enabled by S’MoRE’s design.

Proposed work. We propose **Structural Mixture of Residual Experts** ( S’MoRE), a novel PEFT architecture that improves MoE’s model flexibility & capacity by exploiting experts’ structural relationship, while being as parameter-efficient as LoRA. We start from hierarchical residual decomposition of expert weights, where low-rank parameters of different orders form a multi-layer inter-connection network. We craft the model architecture so that when residuals aggregate and propagate across layers, they 1. remain low-rank to maintain overall efficiency, 2. can generate distinct embeddings for *all* non-isomorphic router-selected sub-trees, which theoretically guarantees high structural flexibility, and 3. are able to express the standard single-layer MoE model variants, which makes S’MoRE a strictly more powerful upgrade. To customize the expert structure for each token, we design a hierarchical router that efficiently and iteratively selects the children residuals when traversing down the selected ancestors. S’MoRE can be conceptually seen as a novel Graph Neural Network (GNN), where the “graph” emerges dynamically from the router’s selection, and the S’MoRE layers can simulate the graph isomorphism test [Huang and Villar, 2021, Xu et al., 2019] to ensure high structural flexibility. Overall, S’MoRE achieves the benefits of both LoRA and fine-grained MoE, and addresses their limitations by exploiting experts’ structure – S’MoRE emulates the capacity of *exponentially more* experts than physically instantiated, while keeping each residual low-rank. We extensively evaluate S’MoRE on 3 base models (LLaMA 3.2 1B, LLaMA 3 8B and Gemma 2 9B), 7 fine-tuning benchmarks, 3 types of router gates, and across different scales. S’MoRE consistently and significantly outperform state-of-the-art models and the 1-layer baselines in terms of both accuracy and parameter efficiency, validating the direction towards better PEFT adapters via structural mixture.

2 Background and Related Work

Parameter efficient fine-tuning (PEFT). Given a pre-trained model, PEFT trains a light-weight adapter whose number of trainable parameters is just a small fraction of the pre-trained weights. LoRA [Hu et al., 2021] and its variants [Liu et al., 2024b, Kopiczko et al., 2024] achieve good empirical performance by learning only a low-rank matrix as the adapter, where rank controls the efficiency-accuracy tradeoff. Despite the high parameter efficiency, their model capacities are limited.

Mixture-of-Experts (MoE). Mixture-of-Experts designs have been shown to boost LLM’s model capacity [Dai et al., 2024, Fedus et al., 2022b, Puigcerver et al., 2024] due to their flexibility in

conditionally activating different sets of parameters for different input tokens. Recent works have tailored MoE for PEFT. MixLoRA [Li et al., 2024a] constructs the adapter as a set of LoRA experts where each token activates its own top- k experts. Similarly, MoLE [Wu et al., 2024b] considers a flat layer of LoRA experts and implements a flexible branch selection scheme. To enhance the mixing flexibility, SMoRA [Zhao et al., 2025] decomposes LoRA into single-rank fine-grained experts, and MoSLoRA [Wu et al., 2024a] integrates a subspace fusing matrix in the low-rank space. MoV [Zadouri et al., 2024] and MoLORA [Zadouri et al., 2024] proposes MoE variants that mix (IA)³ vectors [Liu et al., 2022] or LoRA weights for adapting attention modules. HydraLoRA [Tian et al., 2024] splits LoRA’s up-projection matrix into multiple heads, and then performs weighted sum of each head’s output by the gating weights. In the existing PEFT-MoE models, the expert-selection gates often follow classic designs, such as the noisy top- k [Shazeer et al., 2017] or the switch-transformer [Fedus et al., 2022b] gates. With more experts, it is more challenging to ensure all experts are well utilized [Fedus et al., 2022a]. Such limitation to model scale-up can be addressed by experts’ structural composition in S’MoRE, as we dramatically improve the structural flexibility (see definition in §3.5) without adding more experts. See Appendix A for other MoE variants.

2.1 Preliminaries

We consider one transformer layer and omit the layer index. Let $\mathbf{x} \in \mathbb{R}^d$ be the d -dimensional token embedding input to the adapter. The adapter’s output $\mathbf{x}'$ is added back to the output generated by the pre-trained weights. All adapters here can be applied to weights of both FFN and attention modules.

LoRA formulation. The adapter consists of a down-projection matrix $\mathbf{A} \in \mathbb{R}^{d \times r}$ and an up-projection matrix $\mathbf{B} \in \mathbb{R}^{r \times d}$, with rank $r \ll d$ to achieve parameter efficiency. LoRA maps the input token embedding $\mathbf{x}$ to adapter’s output $\mathbf{x}'$ as follows: $\mathbf{x}' = \mathbf{B} \cdot \mathbf{A} \cdot \mathbf{x}$.

MoE formulation. Let s be the number of experts. The router maps each input token to different experts, where $\text{ROUTE}(\mathbf{x})^i \in \mathbb{R}$ gives expert i ’s score for token $\mathbf{x}$. If the router performs top- k sparse gating, then only top- k values of $\text{ROUTE}(\mathbf{x})^i$ are kept and the rest are cast to 0. Suppose each expert i performs linear transformation via matrix $\mathbf{W}^i$. The MoE layer performs:

$$\mathbf{x}' = \sum_{i=1}^s \text{ROUTE}(\mathbf{x})^i \cdot \mathbf{W}^i \cdot \mathbf{x} \quad (1)$$

3 🍷 S’MoRE

3.1 Low-Rank MoE Variants

Mixture of low-rank experts (MoLRE). To improve parameter efficiency of Eq. 1, we can approximate its $\mathbf{W}^i$ by some low rank $\mathbf{B}^i \cdot \mathbf{A}^i$ as defined in §2.1 (e.g., we can perform SVD on $\mathbf{W}^i$ and derive $\mathbf{B}^i \cdot \mathbf{A}^i$ corresponding to the largest singular values). We term such a model family as *mixture of low-rank experts (MoLRE)* [Wu et al., 2024b, Dou et al., 2024, Li et al., 2024a]. MoLRE’s operation is derived by updating Eq. 1 as follows: $\mathbf{x}' = \sum_{i=1}^s \text{ROUTE}(\mathbf{x})^i \cdot \mathbf{B}^i \cdot \mathbf{A}^i \cdot \mathbf{x}$

Mixture of multi-order residues (MoMOR). We can generalize MoLRE’s low-rank approximation into this form $\mathbf{W}^i \approx \sum_{\ell=0}^{L-1} \mathbf{B}_\ell^i \cdot \mathbf{A}_\ell^i$, where each $\mathbf{B}_\ell^i \cdot \mathbf{A}_\ell^i$ has a low rank (so MoLRE corresponds to $L = 0$). We call $\mathbf{B}_\ell^i \cdot \mathbf{A}_\ell^i$ as the $(\ell + 1)$ th-order residual term, and denote its rank as r_ℓ . The sum $\sum_{\ell=0}^{L-1} \mathbf{B}_\ell^i \cdot \mathbf{A}_\ell^i$ can have a rank up to $\sum_{\ell=0}^{L-1} r_\ell$, which is higher than the individual residuals.

We thus introduce *mixture of multi-order residues (MoMOR)*, an extension to MoLRE. Let $\mathcal{R}_\ell = \{\mathbf{B}_\ell^1 \cdot \mathbf{A}_\ell^1, \mathbf{B}_\ell^2 \cdot \mathbf{A}_\ell^2, \dots\}$ be the set of order- $(\ell + 1)$ residues, MoMOR model performs the following:

$$\mathbf{x}' = \sum_{\ell=0}^{L-1} \sum_{i=1}^{s_\ell} \text{ROUTE}_\ell(\mathbf{x})^i \cdot \mathbf{B}_\ell^i \cdot \mathbf{A}_\ell^i \cdot \mathbf{x} \quad (2)$$

where the model dynamically selects and combines different orders of residuals via routing. MoMOR can adaptively distribute computation across different levels of approximation, improving efficiency and expressivity. Notably, when we set $L = 2$ and $\text{ROUTE}_0(\mathbf{x})^i$ as a dense gate, the order-1 experts are activated for all tokens. MoMOR becomes a *shared-expert* MoE. This is a design adopted by DeepSeek-v3 [DeepSeek-AI, 2024] and many others [Rajbhandari et al., 2022, Li et al., 2024a].

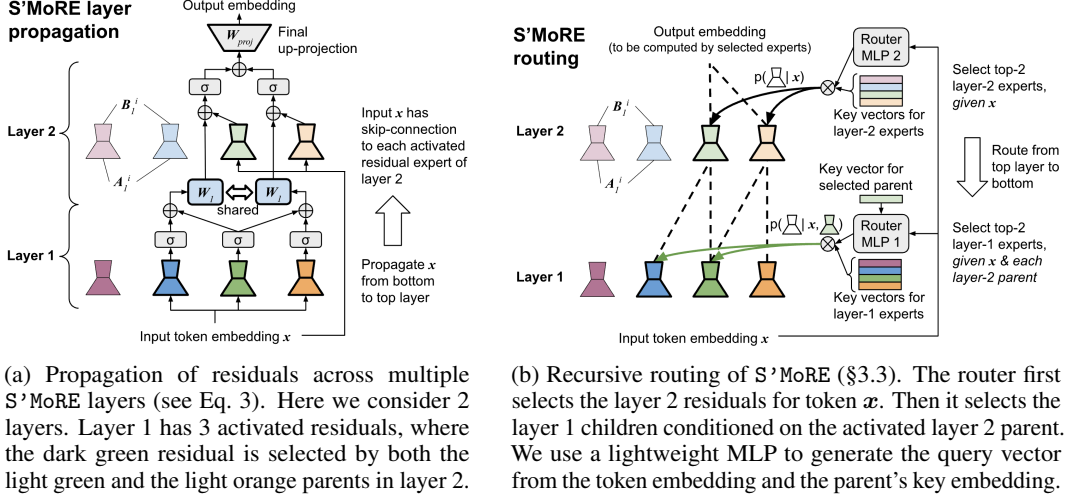


Figure 1: Illustration of the layer propagation and routing process of S'MoRE.

3.2 Structural Mixture

In the following, “layer” refers to a S'MoRE layer with a collection of residual experts, rather than a transformer layer. Based on MoMOR, we arrange all the residues $\mathcal{R}_0, \dots, \mathcal{R}_{L-1}$ into a L -layer structure. For each token x , we activate a sub-structure that interconnects correlated residues in adjacent layers. The token propagates along the sub-structure layer by layer. Each layer implements a lightweight function to aggregate previous-layer residues. Extending the standard MoE to multiple layers improves model capacity by drastically increasing the model’s *structural flexibility* (§3.5).

Parameters. Let $x \in \mathbb{R}^d$ be the d -dimensional token embedding. Layer $\ell + 1$ (for $0 \leq \ell \leq L - 1$) consists of s_ℓ residual experts. Each expert i (with $1 \leq i \leq s_\ell$) consists of a down-projection matrix $A_\ell^i \in \mathbb{R}^{r_\ell \times d}$ and an up-projection matrix $B_\ell^i \in \mathbb{R}^{d_{\ell+1} \times r_\ell}$, where r_ℓ is the experts’ rank and $d_{\ell+1}$ is the output dimension of layer $\ell + 1$. Layer $\ell + 1$ also has a learnable $W_\ell \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ that projects the layer- ℓ output to the $d_{\ell+1}$ -dimensional subspace. Thus, $\mathcal{R}_\ell$ consists of all A_ℓ^i and B_ℓ^i for $1 \leq i \leq s_\ell$.

Propagation. Token x propagates in the L -layer structure in two phases. In the *routing* phase, the router activates the best-matching experts *top-down* (from layer L to 1). At layer L , the router selects experts from $\mathcal{R}_{L-1}$ using standard gates (e.g., Fedus et al. [2022b]). At an intermediate layer $\ell < L$, the router computes the score to activate an expert in $\mathcal{R}_{\ell-1}$, *conditioned* on the already activated ancestors in layers $\ell' > \ell$. This ensures the selected children are connected to their activated parents. Different from the traditional routers, the S'MoRE router customizes a depth- L “residual tree” for each token. See §3.3 for router architecture and §3.5 for structural flexibility of the tree-based routing.

In the *aggregation* phase, the token propagates along the activated residual tree *bottom-up* (from layer 1 to L). Layer $\ell + 1$ aggregates the information from the activated children experts in $\mathcal{R}_\ell$, and generates output embedding for the parent expert in $\mathcal{R}_{\ell+1}$. For each parent expert i , define $\mathcal{N}_\ell^i$ as the set containing the indices of i ’s children experts¹. Layer $\ell + 1$ operates as follows:

$$x_{\ell+1}^i = \sum_{n \in \mathcal{N}_\ell^i} \alpha_\ell^{i,n} \cdot \sigma(B_\ell^n \cdot A_\ell^n \cdot x + W_\ell \cdot x_\ell^n) \quad (3)$$

where $\sigma(\cdot)$ is a non-linear function which can be just an activation (e.g., ReLU [Agarap, 2018]). The scalar $\alpha_\ell^{i,n}$ is the router-generated score, elaborated in §3.3. Inputs to Eq. 3 consist of two parts: 1) Raw token embedding x , which acts as skip connection to residuals $B_\ell^n \cdot A_\ell^n$ of various orders; and 2) x_ℓ^n output from the previous layer, which enables *deep* interaction among multi-order residuals given non-linear $\sigma(\cdot)$ (compared to the shallow aggregation in Eq. 2). For $\ell = 0$, input x_0^n does not exist. To simplify notation, we define $d_0 := 0$, making $x_0^i \in \mathbb{R}^0$ and $W_0 \in \mathbb{R}^{d_1 \times 0}$ as an empty vector /

¹We abuse notation here for ease of description. The nuance is that the same expert can be activated multiple times by different parents / ancestors. So i should refer to the index of a node in the activated tree, rather than the index of just an expert. Similarly, superscript n of x_ℓ^n should be updated to $i \rightarrow n$ as a unique identifier (otherwise it creates ambiguity when expert n is a child of multiple parents). See Eq. 32 in Appendix §3.4.

matrix. Then Eq. 3 applies to all layers $0 \leq \ell \leq L-1$. The last layer L has a single output node (i.e., $s_L = 1$) generated by aggregating information from the entire residual tree. Define $\mathbf{x}_L := \mathbf{x}_L^0$.

Dimensionality d_ℓ . We should set the output d_ℓ to 1) avoid information loss in the aggregation process, and 2) keep the overall L -layer propagation efficient. A naive choice following LoRA is $d_{\ell+1} = d \gg r_\ell$ (e.g., $d = 4096$, $r_\ell = 16$), which makes multiplication with $\mathbf{W}_\ell$ prohibitively expensive. To reduce cost, we should find the smallest $d_{\ell+1}$ that preserves the same amount of information as the vanilla setting $d_{\ell+1} = d$. The problem is equivalent to finding the maximum dimension of the subspace that $\mathbf{x}_{\ell+1}^i$ (Eq. 3) can span for any $\mathcal{N}_\ell^i$, $\mathbf{B}_\ell^n$, $\mathbf{A}_\ell^n$ and activated i . To simplify discussion, ignore activation $\sigma(\cdot)$: 1) For any $\mathbf{x}$, output $\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x}$ maximally spans a d' -dimensional subspace of the original $\mathbb{R}^d$, where $d' = \min\{d_{\ell+1}, r_\ell\}$; 2) There are s_ℓ possible n , leading to s_ℓ different d' -dimensional subspaces. When mutually orthogonal, they maximally span $\min\{d_{\ell+1}, s_\ell \cdot r_\ell\}$ dimensions; 3) $\mathbf{W}_\ell \cdot \mathbf{x}_\ell^n$ can span another subspace of dimension $\min\{d_{\ell+1}, d_\ell\}$ defined by $\mathbf{W}_\ell$ (independent of n). So $\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \mathbf{W}_\ell \cdot \mathbf{x}_\ell^n$ maximally spans $d'' = \min\{d_{\ell+1}, d_\ell + s_\ell \cdot r_\ell\}$ dimensions; 4) Since a subspace is closed under linear combinations, $\sum_{n \in \mathcal{N}_\ell^i} \alpha_\ell^{i,n} (\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \mathbf{W}_\ell \cdot \mathbf{x}_\ell^n)$ remains in the d'' -dimensional subspace, regardless of $\mathcal{N}_\ell^i$. For the vanilla case $d_{\ell+1} = d$ with large enough d , we have $d'' = \min\{d_{\ell+1}, d_\ell + s_\ell \cdot r_\ell\} = d_\ell + s_\ell \cdot r_\ell$. Thus, the minimum $d_{\ell+1}$ is d'' :

$$d_{\ell+1} = d_\ell + s_\ell \cdot r_\ell \quad \Rightarrow \quad d_\ell = \sum_{i=0}^{\ell-1} s_i \cdot r_i, \quad \text{where } d_0 := 0 \text{ and } \ell \in [0, L-1] \quad (4)$$

Final projection. After the last layer L , we map the d_L -dimensional output $\mathbf{x}_L$ to the final output dimension d_{out} (i.e., d_{out} is the dimensionality of $\mathbf{x}'$ in Eq. 1 and Eq. 2). We thus have a projection matrix $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_L \times d}$ that simply performs $\mathbf{x}' = \mathbf{W}_{\text{proj}} \cdot \mathbf{x}_L$.

3.3 Hierarchical Routing

Fig. 1b illustrates the top-down routing. We start from layer L . The router computes $p(i_{L-1} | \mathbf{x})$, the probability to activate an expert i_{L-1} in $\mathcal{R}_{L-1}$ given token $\mathbf{x}$. The top- f_{L-1} experts with the highest $p(i_{L-1} | \mathbf{x})$ are selected. Next, for each selected expert i_{L-1} , we compute $p(i_{L-2} | i_{L-1}, \mathbf{x})$, which is the conditional probability to activate i_{L-2} in $\mathcal{R}_{L-2}$ given its activated parent i_{L-1} and $\mathbf{x}$. Each activated i_{L-1} further activates f_{L-2} children with the highest $p(i_{L-2} | i_{L-1}, \mathbf{x})$. Generally, the router computes the conditional probability $p(i_{L-\ell} | i_{L-1}, \dots, i_\ell, \mathbf{x})$, with $i_{L-1}, \dots, i_\ell$ being all the activated ancestors of the candidate $i_{L-\ell}$. All activated experts form a depth- L tree. Each depth- ℓ node fans out to $f_{L-\ell-1}$ children experts (the activated layer- L experts are the depth-1 tree nodes).

Let f_ℓ be the fanout factor of each parent expert, F_ℓ be the total number of experts selected from $\mathcal{R}_\ell$ (i.e., F_ℓ is the total number of depth- $(L-\ell)$ experts in the activated tree). The same expert can be selected multiple times by ancestors on different paths – It is possible that $F_\ell > s_\ell$. We derive F_ℓ as:

$$F_\ell = \prod_{i=\ell}^{L-1} f_i \quad (5)$$

Router architecture. For each expert i in $\mathcal{R}_\ell$, we instantiate a learnable m -dimensional key vector $\mathbf{k}_\ell^i \in \mathbb{R}^m$. For the whole candidate pool $\mathcal{R}_\ell$, we instantiate a neural network, $\text{MLP}_\ell(\cdot)$, to generate an m -dimensional query vector based on $\mathbf{x}$ and the ancestors. The routing probability over $\mathcal{R}_\ell$ is computed by the normalized key-query dot product. For a path of activated ancestors, “expert i' in $\mathcal{R}_{\ell+1}$, ..., expert $i'^{\dots'}$ in $\mathcal{R}_{L-1}$ ”, the router generates the query vector $\mathbf{q}$ and the router score α_ℓ^i as follows, where $\text{concat}(\cdot)$ performs vector concatenation and $\text{softmax}(\cdot)$ normalizes over $\mathcal{R}_\ell$.

$$\mathbf{q} = \text{MLP}_\ell \left(\text{concat} \left(\mathbf{x}, \mathbf{k}_{\ell+1}^{i'}, \dots, \mathbf{k}_{L-1}^{i'^{\dots'}} \right) \right) \quad (6)$$

$$\alpha_\ell^i = \text{softmax}(\langle \mathbf{k}_\ell^i, \mathbf{q} \rangle) \quad (7)$$

Computation optimization. Eq. 6 can be computationally expensive when all $\text{MLP}_\ell(\cdot)$ need to process the high-dimensional $\mathbf{x}$. To reduce computation, we first project the d -dimensional $\mathbf{x}$ to a d_{down} -dimensional $\mathbf{x}_{\text{down}}$ (e.g., $d = 4096$, $d_{\text{down}} = 24$), and then replace $\mathbf{x}$ with $\mathbf{x}_{\text{down}}$ in Eq. 6. The dimension of the input to $\text{MLP}_\ell(\cdot)$ then becomes $d_{\text{down}} + (L - \ell - 1) \cdot m$.

Gating types. Our router and layer designs are compatible with various types of gates. In our experiments (§4), we have evaluated: 1. *Dense* gate [Tian et al., 2024], which activates all children

experts ($f_\ell = s_\ell$); 2. *Sparse* noisy top- k gate [Shazeer et al., 2017]; 3. *Sparse* switch gate [Fedus et al., 2022b]. The two sparse gates only activate a subset of the children experts ($f_\ell < s_\ell$) by the top routing scores α . To avoid expert under-utilization and ensure all experts see sufficient amount of tokens during training, we implement an auxiliary load-balance loss according to the original papers [Shazeer et al., 2017, Fedus et al., 2022b]. See Appendix B.1 for more algorithmic details.

3.4 Parameter & Computation Efficiency

Although S’MoRE introduces structural learning modules, our design ensures **similar efficiency to the vanilla LoRA** (w.r.t. both computation and trainable parameters) under the same total rank.

Parameter efficiency. Each S’MoRE layer $\ell + 1$ consists of the following trainable parameters: B_ℓ^n , A_ℓ^n and W_ℓ . The total trainable parameters equals:

$$P_{\ell+1} = s_\ell \cdot (d \cdot r_\ell + r_\ell \cdot d_{\ell+1}) + d_\ell \cdot d_{\ell+1} = s_\ell \cdot d \cdot r_\ell + d_{\ell+1} \cdot (s_\ell \cdot r_\ell + d_\ell) \stackrel{(a)}{=} s_\ell \cdot d \cdot r_\ell + d_{\ell+1}^2 \quad (8)$$

where the last step “(a)” is according to Eq. 4. The final projection matrix (end of §3.2) requires $P_{\text{proj}} = d \cdot d_L$ parameters. So the total number of parameters for all S’MoRE layers equals:

$$P_{\text{proj}} + \sum_{\ell=1}^L P_\ell = d \cdot d_L + d \cdot \left(\sum_{\ell=0}^{L-1} s_\ell \cdot r_\ell \right) + \Delta \stackrel{(b)}{=} 2 \cdot d \cdot d_L + \Delta \stackrel{(c)}{\approx} 2 \cdot d \cdot d_L \quad (9)$$

where $\Delta = \sum_{\ell=1}^L d_\ell^2$. Step “(b)” is by Eq. 4; Δ is the overhead due to multi-layer propagation. Since $d_1 < \dots < d_L \ll d$ (e.g., $d_L = 64$, $d = 4096$), we have $\Delta \ll 2 \cdot d \cdot d_L$. This justifies step “(c)”. In Table 1, we empirically validated the small overhead Δ . With $f_\ell = 2$, $s_\ell = 4$, and $r_\ell = 8$ or 16 for all layers ℓ (consistent with the §4 experiments), Δ is no more than **2%** for 2-layer S’MoRE.

The router’s trainable parameters come from: 1) down-projection for x_{down} , which requires $d \cdot d_{\text{down}}$ parameters, 2) per-layer “query” MLP. By §3.3, the MLP’s input dimension is $d_{\text{down}} + (L - \ell) \cdot m$, where $m \ll d$ is the dimension of the “key” vectors. In practice, we set the MLP hidden dimension as m . Since m and d_{down} are both very small, the router’s parameter count is practically negligible.

Table 1: Overhead Δ compared with the main computation cost $2 \cdot d \cdot d_L$

r_ℓ	L	d_L	$2 \cdot d \cdot d_L$	Δ	Overhead ratio
8	2	64	0.5M	0.005M	1.0%
	3	96	0.8M	0.014M	1.8%
	4	128	1.0M	0.031M	2.9%
16	2	128	1.0M	0.020M	2.0%
	3	192	1.6M	0.057M	3.6%
	4	256	2.1M	0.123M	5.9%

In total, S’MoRE approximately has $2 \cdot d \cdot d_L$ parameters – *the same as the parameter count for a vanilla LoRA* with rank d_L (the 2 factor is due to LoRA’s down- and up-project matrices A and B).

Computation cost. Following similar steps, we can derive the overhead in computation. The computation cost of the baseline LoRA is $2 \cdot d \cdot d_L$. The overhead introduced by S’MoRE is $\Delta' \leq \sum_{\ell=0}^{L-1} F_\ell \cdot d_{\ell+1} \cdot (d_\ell + r_\ell)$, which is again *negligible* in practice. See Appendix C.1 for details.

3.5 Model Capacity

We theoretically show S’MoRE enhances model capacity compared with baselines (see Appendix C.2 for proofs). First, we show that the two low-rank MoE variants in §3.1 are special cases of S’MoRE.

Proposition 3.1. *S’MoRE can express MoLRE, when $L = 1$ and $\sigma(\cdot)$ is the identity mapping.*

Proposition 3.2. *S’MoRE can express MoMOR, when setting $\sigma(\cdot)$ as the identity mapping.*

For any MoLRE (or MoMOR) model, we can find a corresponding S’MoRE that generates identical output as MoLRE (or MoMOR) for any input x . Without σ , we can collapse a multi-layer S’MoRE into a single layer equivalent, where the dimensionality set by Eq. 4 ensures the same rank as MoMOR. Can S’MoRE be theoretically better than MoLRE and MoMOR, if we go beyond the constraints of Propositions 3.1 and 3.2 by setting $L > 1$ and σ as non-linear mapping? To answer it, we analyze an MoE model’s expressive power by quantifying the **structural flexibility**.

Structural flexibility. Let Θ be the collection of all experts’ parameters (B_ℓ^i , A_ℓ^i and W_ℓ for $0 \leq \ell \leq L - 1$ and all i). Given Θ , when a token x comes, different routers may activate different residual experts, and thus generate different output embedding x_L . Therefore, we define $\text{dist}(x; \Theta)$ as the number of *distinct* x_L . The larger $\text{dist}(x; \Theta)$ can be, the more “structurally flexible” the model architecture is. Our focus here is on the multi-layer structure formed by the residual experts, rather than the router network (thus, we assume an ideal router for the following Theorems).

Next we prove S’MoRE’s higher model capacity by quantifying structural flexibility. In the following, we treat α_ℓ^n as binary mask (1 for selected experts, and 0 otherwise) when generating x_L .

Theorem 3.3. *The structural flexibility of MoMOR is upper-bounded by $\Gamma_{\text{MoMOR}} = \max_{\mathbf{x}, \Theta} \text{dist}(\mathbf{x}; \Theta) \leq \binom{s_{L-1}}{f_{L-1}} \cdot \prod_{\ell=0}^{L-2} \left(\sum_{i=f_\ell}^{\min\{F_\ell, s_\ell\}} \binom{s_\ell}{i} \right)$.*

Theorem 3.4. *Setting $\sigma(\cdot)$ as an MLP, there exists some Θ' such that the structural flexibility of S'MoRE is: $\Gamma_{\text{S'MoRE}} = \min_{\mathbf{x}} \text{dist}(\mathbf{x}; \Theta') = \prod_{\ell=0}^{L-1} \binom{s_\ell}{f_\ell}^{F_{\ell+1}}$, where we define $F_L := 1$.*

Above, $\binom{s}{k} = \frac{s!}{k!(s-k)!}$ is the binomial coefficient that quantifies the number of ways to choose k out of s items, ignoring order. F_ℓ is defined in Eq. 5. When increasing the number of layers, $\Gamma_{\text{S'MoRE}}$ exceeds the upper bound Γ_{MoMOR} by orders of magnitude. The reason is that for MoMOR, the $\binom{s_\ell}{i}$ terms are summed over F_ℓ , while for S'MoRE, F_ℓ becomes the **exponent** of $\binom{s_\ell}{f_\ell}$. In Fig. 2, we calculate the theoretical Γ_{MoMOR} and $\Gamma_{\text{S'MoRE}}$ under depth L . Consistent with our experimental setup (§4.1), we set $s_\ell = 4$ and $f_\ell = 2$ for all ℓ . Clearly, $\Gamma_{\text{S'MoRE}}$ is substantially higher than Γ_{MoMOR} even for shallow models ($L = 2$), and $\Gamma_{\text{S'MoRE}}$ grows exponentially faster than Γ_{MoMOR} when increasing L .

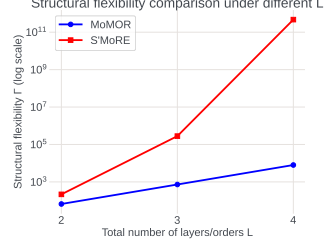


Figure 2: $\Gamma_{\text{S'MoRE}}$ and Γ_{MoMOR} w.r.t. L (with $s_\ell = 4$, $f_\ell = 2$).

We explain the intuition of the proof, and defer the details to Appendix C.2. *First*, $\Gamma_{\text{S'MoRE}}$ quantifies the number of non-isomorphic depth- L trees that can be formed by any router. Each node at tree-level ℓ (i.e., an expert in $\mathcal{R}_{L-\ell}$; the same expert may appear multiple times at tree-level ℓ under different ancestor paths) has $\omega = \binom{s_{L-\ell-1}}{f_{L-\ell-1}}$ ways of selecting its children set. All nodes at tree-level ℓ jointly contribute to a $\omega^{F_{L-\ell}}$ factor. *Secondly*, S'MoRE can generate distinct outputs for all non-isomorphic sub-trees. We borrow conclusions from the Graph Neural Network literature. We view Eq. 3 as defining a variant of Graph Isomorphism Network (GIN) [Xu et al., 2019]. S'MoRE's L -layer propagation simulates the L -iteration Weisfeiler-Lehman (WL) test [Huang and Villar, 2021], where including non-linearly activated σ is the key to ensure an injective ‘‘color refinement’’ process in WL. It then follows that the L layer S'MoRE can distinguish non-isomorphic trees of depth L . *Third*, without activation σ , S'MoRE degrades to MoMOR, and is unable to distinguish many non-isomorphic depth- L trees. Fig. 3 shows 3 examples with $L = 2$. Node 2 is the final output node (tree root). When we activate the same set of experts (‘‘0,1’’, ‘‘0,2’’, ‘‘0,3’’, ‘‘0,4’’, ‘‘1,1’’, ‘‘1,2’’) but connect them differently (non-isomorphic), MoMOR always generates the same output while S'MoRE can produce different ones. This shows $\Gamma_{\text{MoMOR}} < \Gamma_{\text{S'MoRE}}$ and S'MoRE's higher expressivity.

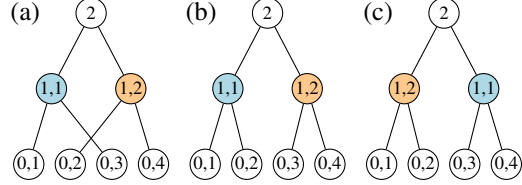


Figure 3: Examples where the same set of activated experts interconnect differently. MoMOR always generates the same output for (a), (b) and (c), while S'MoRE can distinguish all the three cases. A variant of S'MoRE that performs activation σ differently (§3.6) can differentiate (a) from (b) or (c), but cannot differentiate (b) from (c). Note that (b) and (c) differ by swapped ‘‘1,1’’ and ‘‘1,2’’.

3.6 Model Variants

How activation σ affects structural learning. Theorem 3.4 concretely shows the benefit of including activation σ in Eq. 3. What if we tweak Eq. 3 to let σ operate on $\mathbf{x}_\ell^n$ rather than $\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \mathbf{W}_\ell \cdot \mathbf{x}_\ell^n$?

$$\mathbf{x}_{\ell+1}^i = \sum_{n \in \mathcal{N}_\ell^i} \alpha_\ell^{i,n} \cdot (\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \mathbf{W}_\ell \cdot \sigma(\mathbf{x}_\ell^n)) \quad (10)$$

We can then decompose Eq. 38 as $\sum_{n \in \mathcal{N}_\ell^i} \mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \sum_{n \in \mathcal{N}_\ell^i} \mathbf{W}_\ell \cdot \sigma(\mathbf{x}_\ell^n)$ (ignoring $\alpha_\ell^{i,n}$ for simplicity), and use Fig. 3 as an example to understand its expressive power. Trees (a) and (b) have the same layer-2 experts, ‘‘1,1’’ and ‘‘1,2’’, making their **gray** terms equivalent. Yet, their different layer-1 children combinations (tree (a) has ‘‘0,1’’ + ‘‘0,3’’ and ‘‘0,2’’ + ‘‘0,4’’, while tree (b) has ‘‘0,1’’ + ‘‘0,2’’ and ‘‘0,3’’ + ‘‘0,4’’) make their **green** terms different. This enables Eq. 38 to differentiate (a) from (b). Following this reasoning, for (b) and (c), their gray and green terms are both equal. Thus, Eq. 38 yields identical outputs for the two trees, even though they are non-isomorphic.

Corollary 3.5. Let $\Gamma_{S'MoRE*}^\ell$ be the structural flexibility of ℓ -layer S'MoRE variant under Eq. 38. It satisfies the following recursion: $\Gamma_{S'MoRE*}^\ell = \binom{s_{\ell-1}}{f_{\ell-1}} \cdot \left(\Gamma_{S'MoRE*}^{\ell-1} + f_{\ell-1} - 1 \right)$, where $\Gamma_{S'MoRE*}^0 := 1$.

It is easy to see that S'MoRE under Eq. 3 is more expressive than S'MoRE* under Eq. 38. Further, both S'MoRE variants are stronger than the baseline 1-layer MoEs. This is also illustrated by Fig. 3.

S'MoRE with cross-layer parameter sharing. We introduce S'MoRE[#], another useful variant which lets the experts of different layers share the same parameters. i.e., $s := s_\ell$ and $r := r_\ell$ are the same for all layers ℓ . And $A^i := A_\ell^i$ and $B^i := B_\ell^i$ for all ℓ and $1 \leq i \leq s$. This means experts in different layers now operate in the same embedding subspace, and hence the intermediate hidden dimension $d := d_\ell$ is the same for all ℓ – we update Eq. 4 as $d = s \cdot r$. The layers still propagate by Eq. 3.

We summarize the properties of S'MoRE[#]. Following similar derivation² in §3.4 (plugging in d above), we conclude S'MoRE[#] has comparable parameter & computation efficiency as the vanilla LoRA. The structural flexibility below also has a similar form as Theorem 3.4 – S'MoRE and S'MoRE[#] exponentially boost structural flexibility of the 1-layer baselines, MoMOR and MoLRE, respectively.

Corollary 3.6. The structural flexibility of S'MoRE[#] equals $\prod_{\ell=0}^{L-1} \binom{s}{f}^{F_{\ell+1}}$ where $F_L := 1$.

Alternative router design (bottom-up version). In addition to the top-down router in §3.3, we can also perform bottom-up routing, making the routing and layer propagation flow along the same direction. The bottom-up router still aims at customizing different children experts for different parents. Yet, when routing bottom-up, the parent index is unknown when we select the children. So now the key vector k (see Eq. 6) is not directly associated with any specific parent expert. It instead represents a node position in the routing tree. See Appendix B.2 for details and tradeoff discussion.

4 Experiments

4.1 Experimental Setup

Datasets. We fine-tune on a diverse set of benchmarks, including ARC-c/e [Clark et al., 2018], Commonsense QA (CSQA) [Talmor et al., 2018], OpenBook QA (OBQA) [Mihaylov et al., 2018], Winogrande [Sakaguchi et al., 2021], GSM8K [Cobbe et al., 2021], and HumanEval [Chen et al., 2021]. For HumanEval, we follow Tian et al. [2024] to train the base LLM on CodeAlpaca [Chaudhary, 2023], and evaluate “Pass@1” on HumanEval. For all other datasets, we fine-tune on the training split and evaluate “Accuracy” on the test split. See Appendix §D.1 for more details.

Base models & baselines. We use LLaMA 3.2-1B, LLaMA 3-8B [Dubey et al., 2024] and Gemma 2-9B [Team et al., 2024b] as the base models. We insert adapters of different kinds: 1. LoRA [Hu et al., 2021]; 2. mixture of LoRA experts (MixLoRA [Li et al., 2024a]): the state-of-the-art parameter efficient MoE adapter, which is essentially the single-layer version of S'MoRE; 3. HydraLoRA [Tian et al., 2024]: another state-of-the-art PEFT adapter implementing a MoE variant of LoRA by splitting LoRA’s up-projection B into multiple heads, and combining the multi-head outputs via scores from a dense gate; and 4. S'MoRE: the multi-layer extension of the above. To further evaluate the generalizability, we implement 3 variants of MixLoRA and S'MoRE using different gates (see §3.3 and Appendix B.1): 2 sparse gates (noisy top- k [Shazeer et al., 2017] and switch-transformer [Fedus et al., 2022b] gates), and 1 dense gate (same as HydraLoRA [Tian et al., 2024]). See Appendix D.2.

Training & evaluation methodology. For hyperparameter tuning, we train all models using the same number of epochs, learning rate schedule, gradient accumulation steps and machine type. All models are trained under the LLaMA-Factory [Zheng et al., 2024] framework and evaluated with OpenCompass [Contributors, 2023b]. For hyperparameter search, we set an equal budget of trainable parameters, and vary the expert rank, the number of experts, the number of activated experts, etc. See Appendix D.2 for details of the hyperparameter range, and the hardware / software configuration.

4.2 Main Results

Table 2 presents the comprehensive comparison on accuracy and parameter efficiency. For all the base model and the gate type, we consistently observe that S'MoRE achieves **significant accuracy improvement without sacrificing parameter efficiency**. Specifically, 1. Among all the methods, while LoRA’s parameter counts are low, its average accuracy is also the lowest. This implies the

²In S'MoRE[#], the same expert may be activated in multiple layers. To avoid redundancy, we first collect the set of activated experts across all layers, and then compute $B^i \cdot A^i \cdot x$ only once for each activated expert i .

Table 2: Comparison under two base models & three gate types. The hyperparameter search sets the same parameter budget for all models. The “Param.” column denotes the trainable parameters (B) for the highest-accuracy model. In the “Method” column, number in parentheses denote the number of experts / heads (“4-4” denotes a 2-layer S’MoRE, each with 4 experts). Highest accuracy under the same gate is highlighted in **bold**, and highest accuracy across all gates is highlighted in **red**.

Gate	Method	ARC-c		ARC-e		CSQA		OBQA		Winogrande		Avg	Avg	
		Acc.	Param.	Acc.	Param.	Acc.	Param.	Acc.	Param.	Acc.	Param.	Acc.	Param.	
LLaMA 3.2 1B	Base	32.54	0	66.31	0	23.67	0	43.80	0	50.75	0	43.41	0	
	LoRA	36.27	0.004	74.78	0.002	63.80	0.063	71.20	0.031	50.59	0.008	59.15	0.022	
	Dense	HydraLoRA (4)	35.93	0.006	73.54	0.023	66.34	0.002	71.60	0.023	50.75	0.012	59.63	0.013
		HydraLoRA (8)	35.93	0.012	72.31	0.007	62.08	0.042	71.60	0.012	50.99	0.012	58.58	0.017
		MixLoRA (4)	39.66	0.021	72.84	0.134	65.44	0.134	70.40	0.134	51.30	0.007	59.93	0.086
		MixLoRA (8)	39.32	0.021	74.78	0.270	66.42	0.069	69.60	0.134	51.14	0.037	60.25	0.106
		S'MoRE (2-2)	40.00	0.017	75.31	0.085	66.99	0.037	72.20	0.085	52.01	0.015	61.30	0.048
		S'MoRE (4-4)	39.66	0.017	74.43	0.085	67.32	0.045	72.80	0.202	52.01	0.168	61.24	0.103
	Noisy top-k	MixLoRA (4)	39.32	0.037	71.96	0.069	64.70	0.134	70.00	0.134	51.46	0.069	59.49	0.089
		MixLoRA (8)	37.97	0.069	72.84	0.270	65.03	0.134	70.80	0.270	51.46	0.069	59.62	0.162
		S'MoRE (2-2)	39.66	0.029	73.19	0.135	64.95	0.135	70.00	0.102	51.54	0.029	59.87	0.086
		S'MoRE (4-4)	39.66	0.037	74.96	0.135	66.26	0.102	71.40	0.135	52.17	0.273	60.89	0.136
	Switch	MixLoRA (4)	38.98	0.021	73.37	0.134	66.42	0.069	72.00	0.134	51.22	0.009	60.40	0.073
		MixLoRA (8)	39.32	0.021	73.72	0.069	65.85	0.134	71.80	0.134	51.30	0.021	60.40	0.076
		S'MoRE (2-2)	39.66	0.029	74.78	0.135	66.75	0.069	71.40	0.102	52.25	0.045	60.97	0.076
		S'MoRE (4-4)	40.34	0.021	74.78	0.168	67.16	0.202	72.40	0.085	52.09	0.021	61.35	0.099
LLaMA 3 8B	Base	80.34	0	89.77	0	70.35	0	73.80	0	59.91	0	74.83	0	
	LoRA	81.69	0.028	91.36	0.028	81.00	0.028	87.00	0.028	81.77	0.028	84.56	0.028	
	Dense	HydraLoRA (4)	83.39	0.013	91.53	0.160	81.82	0.013	88.20	0.082	83.82	0.160	85.75	0.086
		HydraLoRA (8)	81.69	0.079	91.53	0.015	81.49	0.024	86.60	0.015	84.14	0.297	85.09	0.086
		MixLoRA (4)	81.69	0.026	92.24	0.247	81.24	0.033	89.40	0.478	84.06	0.247	85.73	0.206
		MixLoRA (8)	82.37	0.132	91.71	0.247	81.00	0.033	88.60	0.075	85.40	0.478	85.82	0.193
		S'MoRE (2-2)	82.37	0.090	92.24	0.190	81.90	0.037	89.40	0.054	88.24	0.480	86.83	0.170
		S'MoRE (4-4)	82.71	0.190	91.89	0.247	81.90	0.033	90.00	0.076	85.48	0.247	86.40	0.157
	Noisy top-k	MixLoRA (4)	82.37	0.075	91.53	0.247	80.75	0.075	87.80	0.075	82.00	0.478	84.89	0.190
		MixLoRA (8)	83.39	0.950	91.53	0.247	80.67	0.075	88.40	0.247	83.19	0.478	85.44	0.399
		S'MoRE (2-2)	82.37	0.305	91.36	0.090	81.82	0.104	88.20	0.047	83.27	0.190	85.40	0.147
		S'MoRE (4-4)	82.37	0.104	91.71	0.305	82.06	0.047	90.00	0.480	85.48	0.714	86.32	0.330
	Switch	MixLoRA (4)	82.37	0.132	92.95	0.478	81.08	0.047	88.80	0.478	84.53	0.247	85.95	0.276
		MixLoRA (8)	82.03	0.033	91.71	0.132	81.24	0.047	88.60	0.247	85.95	0.950	85.91	0.282
		S'MoRE (2-2)	83.05	0.133	92.24	0.061	81.82	0.029	89.80	0.076	86.42	0.247	86.67	0.109
		S'MoRE (4-4)	83.39	0.076	92.42	0.305	82.15	0.047	89.80	0.305	85.87	0.305	86.73	0.208

necessity of more advanced PEFT adapters of higher model capacity. 2. For models using dense gates, HydraLoRA achieves the lowest parameter count. However, its average accuracy is notably lower than both the 1-layer MoE model MixLoRA and the 2-layer S’MoRE. Since for all models, we set the same parameter budget for hyperparameter tuning, this means that HydraLoRA cannot effectively utilize more parameters to boost its accuracy (see also Fig. 4). 3. On all gate types, S’MoRE achieves significantly higher average accuracy than all baselines. In particular, MixLoRA belongs to the MoLRE family (§3.1) whose layer operation can be categorized by Eq. 1. Thus, it can be seen as a single-layer S’MoRE. Clearly, building a two-layer structure (“2-2” or “4-4”) from a flat layer of experts (“4” or “8”) boosts the accuracy without requiring additional trainable parameters. 4. Finally, the comparable parameter counts of MixLoRA and S’MoRE implies that our multi-layer design introduces low parameter overhead, which is consistent with our analysis in §3.4.

Table 3: LLaMA 3-8B: model Accuracy / Pass@1, and the best-performing models’ trainable parameters (B).

Gate	Method	GSM8K		HumanEval	
		Accuracy	Param. (B)	Pass@1	Param. (B)
	Base model	55.95	0	26.22	0
	LoRA	59.97	0.014	43.29	0.014
Dense	HydraLoRA (4)	62.47	0.317	40.85	0.082
	HydraLoRA (8)	62.24	0.297	44.51	0.079
	MixLoRA (4)	61.11	0.132	39.02	0.026
	MixLoRA (8)	59.36	0.132	40.85	0.033
	S’MoRE (2-2)	62.40	0.104	42.07	0.090
	S’MoRE (4-4)	65.20	0.957	43.90	0.104
Switch	MixLoRA (4)	59.67	0.047	42.68	0.075
	MixLoRA (8)	61.56	0.247	39.63	0.247
	S’MoRE (2-2)	62.47	0.133	45.73	0.190
	S’MoRE (4-4)	63.91	0.957	42.07	0.090

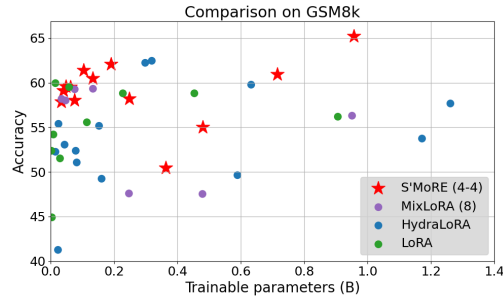


Figure 4: Change of accuracy w.r.t. trainable parameters, corresponding to models in Table 3.

Table 4: Results on Gemma 2-9B. We evaluate on representative benchmarks due to limited resources.

Method	ARC-e		CSQA		Winogrande		HumanEval		Avg	
	Accuracy	Param. (B)	Accuracy	Param. (B)	Accuracy	Param. (B)	Pass@1	Param. (B)	Acc. / Pass@1	Avg Param. (B)
LoRA	79.72	0.289	85.91	0.145	87.06	0.145	43.29	0.072	74.00	0.163
MixLoRA (4)	85.54	0.059	85.83	0.096	88.79	0.169	43.29	0.096	75.86	0.105
MixLoRA (8)	83.07	0.168	85.83	0.096	89.19	0.315	44.51	0.168	75.65	0.187
S'MoRE (2-2)	86.24	0.042	86.40	0.169	90.13	0.169	44.51	0.096	76.82	0.119
S'MoRE (4-4)	86.60	0.169	86.32	0.060	90.13	0.315	46.34	0.060	77.35	0.151

Table 5: S'MoRE on LLaMA 3.2-1B with more layers. We follow a simple hyperparameter tuning strategy, ensuring the same design space sizes and parameter budgets for the 2- and 3-layer variants.

Layer sizes	ARC-c		ARC-e		Commonsense QA		OpenBook QA		Winogrande	
	Accuracy	Param. (B)	Accuracy	Param. (B)	Accuracy	Param. (B)	Accuracy	Param. (B)	Accuracy	Param. (B)
2-2	40.00	0.017	75.31	0.085	66.99	0.037	72.20	0.085	52.01	0.011
2-2-2	39.32	0.017	74.25	0.102	67.40	0.053	72.60	0.205	52.88	0.011
4-4	39.66	0.017	74.43	0.085	67.32	0.045	72.80	0.202	52.01	0.168
4-4-4	40.34	0.029	73.90	0.205	67.32	0.053	73.60	0.202	52.09	0.013

4.3 Results on GSM8K & HumanEval

We evaluate on GSM8K and HumanEval using LLaMA 3-8B. The observations on accuracy / Pass@1 and parameter efficiency from Table 3 is consistent with those from Table 2: S'MoRE achieves significant accuracy improvement while maintaining parameter efficiency. Fig. 4 helps us better understand how the model accuracy scales with the amount of trainable parameters. 1) For S'MoRE, the accuracy consistently increases with parameters in the low-parameter region (less than 0.2B). Then the accuracy drops when we keep increasing the parameters. Interestingly, in the region from 0.4B to 1B, we see an almost linear increase of accuracy w.r.t. parameters – the accuracy eventually surpasses that of all other models with a large margin at around 1B. 2) For HydraLoRA, its accuracy peaks at around 0.3B. Unlike S'MoRE, keeping increasing the parameters does not help with HydraLoRA's accuracy improvement. This observation is consistent with Table 2. 3) Similar to HydraLoRA, the 1-layer MixLoRA does not show good scaling of accuracy w.r.t. parameters. S'MoRE may discover good structures among experts, which in turn helps experts better utilize their parameters.

4.4 Evaluation on Gemma

We extend our evaluation to the Gemma model family. Table 4 shows the comparison with representative baselines. Consistent with the observations on the LLaMA family, S'MoRE achieves significant boost in accuracy / Pass@1 with comparable or fewer parameters (see "MixLoRA (4) vs. S'MoRE (2-2)" and "MixLoRA (8) vs. S'MoRE (4-4)"). The performance gains across multiple model scales (1B, 7B, 9B) and model families (LLaMA, Gemma) reaffirm the benefits from structural mixture.

4.5 Scaling up with Layers

We evaluate if increasing the number of S'MoRE layers can further improve accuracy. We follow a simple hyperparameter tuning strategy: for all the 2-layer S'MoRE under consideration, we add a 3rd layer with identical configuration (w.r.t. number of experts s , fanout f , expert dimension r , etc.) as the 2nd layer. Thus, the sizes of the design spaces for the 3-layer and 2-layer S'MoRE are equal. We also enforce the same parameter budget for the 2- and 3-layer models. Table 5 summarizes the comparison. Adding one more layer improves accuracy significantly in many cases. The accuracy improvements do not necessarily come at the cost of more parameters. For example, for Winogrande, "2-2-2" structure improves the accuracy of "2-2" by 0.87 with the same parameter count.

5 Conclusion

We introduced S'MoRE, a novel Structural Mixture of Residual Experts framework that jointly achieves the efficiency of low-rank adaptation (LoRA) with the flexibility of Mixture-of-Experts (MoE), and further boosts MoE's model capacity by exploiting experts' inherent structure. By applying hierarchical residual decomposition and tree-based routing, S'MoRE effectively emulates exponentially more experts without instantiating additional expert instances, and achieves similar computation and parameter efficiency as the vanilla LoRA. We further propose a structural flexibility metric to quantify the model capacity, and theoretically show that S'MoRE's unique model architecture design is the key to boost structural flexibility compared with various LoRA-MoE hybrids. On extensive experiments, we confirm S'MoRE's state-of-the-art fine-tuning performance.

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.
- AF Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.
- Joshua Ainslie, Tao Lei, Michiel de Jong, Santiago Ontañón, Siddhartha Brahma, Yury Zemlyanskiy, David Uthus, Mandy Guo, James Lee-Thorp, Yi Tay, et al. Colt5: Faster long-range transformers with conditional computation. *arXiv preprint arXiv:2303.09752*, 2023.
- Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL <https://api.semanticscholar.org/CorpusID:268232499>.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024.
- Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. <https://github.com/sahil280114/codealpaca>, 2023.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. URL <https://arxiv.org/abs/2107.03374>.
- Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024.
- 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.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL <https://arxiv.org/abs/2110.14168>.
- LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. <https://github.com/InternLM/lmdeploy>, 2023a.
- OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023b.
- Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. URL <https://arxiv.org/abs/2401.06066>.
- DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd*

- Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.106. URL <https://aclanthology.org/2024.acl-long.106>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning, 2022a. URL <https://arxiv.org/abs/2209.01667>.
- 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, 2022b.
- Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. Mixture-of-loras: An efficient multitask tuning for large language models. *arXiv preprint arXiv:2403.03432*, 2024.
- Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023.
- Xu Owen He. Mixture of a million experts, 2024. URL <https://arxiv.org/abs/2407.04153>.
- K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989. ISSN 0893-6080.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Ningyuan Teresa Huang and Soledad Villar. A short tutorial on the weisfeiler-lehman test and its variants. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 8533–8537. IEEE, June 2021. doi: 10.1109/icassp39728.2021.9413523. URL <http://dx.doi.org/10.1109/ICASSP39728.2021.9413523>.
- Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=NjNfLdxr3A>.
- 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.
- Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*, 2024a.
- Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient SMoe with hints from its routing policy. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=eFWG9Cy3WK>.
- Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*, 2023.

- 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*, 2024a.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022. URL <https://arxiv.org/abs/2205.05638>.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32100–32121. PMLR, 21–27 Jul 2024b. URL <https://proceedings.mlr.press/v235/liu24bn.html>.
- Jan Ludziejewski, Jakub Krajewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, Marek Cygan, and Sebastian Jaszczur. Scaling laws for fine-grained mixture of experts. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33270–33288. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/ludziejewski24a.html>.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7):197605, 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 *EMNLP*, 2018.
- Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=jxpsAj7ltE>.
- 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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18332–18346. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/rajbhandari22a.html>.
- David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models, 2024. URL <https://arxiv.org/abs/2404.02258>.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- Noam Shazeer, Azalia Mirhoseini*, Krzysztof Maziarczyk*, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=B1ckMDqlg>.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024b.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Chaoqi Wang, Zhuokai Zhao, Chen Zhu, Karthik Abinav Sankararaman, Michal Valko, Xuefei Cao, Zhaorun Chen, Madian Khabza, Yuxin Chen, Hao Ma, et al. Preference optimization with multi-sample comparisons. *arXiv preprint arXiv:2410.12138*, 2024a.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large models: A survey of methodologies. *arXiv preprint arXiv:2410.19878*, 2024b.

Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7880–7899, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.450. URL <https://aclanthology.org/2024.emnlp-main.450/>.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=uWvKBCYh4S>.

- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=ryGs6iA5Km>.
- Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=EvDeiLv7qc>.
- Hanqing Zeng, Hanjia Lyu, Diyi Hu, Yinglong Xia, and Jiebo Luo. Mixture of weak and strong experts on graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=wYvuY60SdD>.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.
- Ziyu Zhao, Yixiao Zhou, Didi Zhu, Tao Shen, Xuwu Wang, Jing Su, Kun Kuang, Zhongyu Wei, Fei Wu, and Yu Cheng. Each rank could be an expert: Single-ranked mixture of experts lora for multi-task learning, 2025. URL <https://arxiv.org/abs/2501.15103>.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.
- 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15913–15923, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state the claims made.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: See Appendix E.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The full set of assumptions and a complete proof are provided in either the main paper or its appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is released at: <https://github.com/ZimpleX/SMoRE-LLM>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We release all the training and test details in our experiment section and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Error bars and other statistical significance tests are conducted.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide sufficient information on the computational resources we used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed, since it focuses on foundational research in machine learning.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Table of Contents (Appendix)

A More Related Works	23
B Details of Model Design	24
B.1 Three types of gates implemented in practice	24
B.2 Alternative design: bottom-up routing	24
C Details of Theoretical Analysis	25
C.1 Derivation of parameter & computation costs	25
C.2 Proof of model capacity	26
C.2.1 Proof for two special S'MoRE configurations	26
C.2.2 Proof of Theorem 3.3	29
C.2.3 Proof of Theorem 3.4	30
C.2.4 Proof of Corollary 3.5	35
C.2.5 Proof of Corollary 3.6	36
D Additional Experimental Results	36
D.1 Dataset details	36
D.2 More details on experimental setup	36
D.3 Wall-clock time & potential system optimizations	37
D.4 Routing cost	37
E Limitations and Broader Impact	38

A More Related Works

More related works on MoE. In addition to the works mentioned in §2, we list a few additional works that explore MoE designs in LLM. LLaMA-MoE [Zhu et al., 2024] applies Mixture-of-Experts under the Continual Pre-training (CPT) setting. It breaks down LLaMA’s pre-trained weight matrices into sub-matrices and use them to initialize the experts’ parameters. MC-SMoE [Li et al., 2024b] discovers that there exists high redundancy among experts, and correspondingly proposes algorithms to cluster, merge and then compress a model with many experts. MoA [Feng et al., 2024] explores Mixture-of-LoRAs for the multi-task tuning scenarios, where it trains a LoRA for each task separately, and then assembles the multiple LoRAs into an MoE where a learnable router selects the most suitable expert based on the task category.

Heterogeneous experts. In most MoE models, experts are homogeneous and may have an identical model architecture. Heterogeneous experts have been recently explored in language modeling [Raposo et al., 2024, Ainslie et al., 2023] and graph learning [Zeng et al., 2024]. MoD [Raposo et al., 2024] and CoLT5 [Ainslie et al., 2023] consider the combination of light and heavy experts so that tokens through the light path can be processed faster and more cheaply. Mowst [Zeng et al., 2024] further discovers that mixture of weak and strong experts can enhance MoE’s capacity beyond that of a strong expert alone. However, existing heterogeneous MoE designs consider a *horizontal* stacking of different types of experts, where the weak/light branch operates in parallel to the strong/heavy one. However, in S'MoRE, the residuals of different orders can also be interpreted as experts of different strength, with the 1st-order residuals being the strongest. And we explore a *vertical* stacking design where the higher-order residuals transform and propagate to the lower-order ones. The heterogeneous expert design in S'MoRE encodes additional structural information than existing models.

Scaling behavior. There is an emerging trend to research the unique scaling behaviors of MoE systems compared with dense models. Ludziejewski et al. [2024] summarizes a scaling law indicating that finer expert granularity may improve model capacity, and He [2024] provides a practical implementation for very large number of experts. Zhao et al. [2025] applies such fine-grained experts in the fine-tuning tasks. Through analysis and experiments, we hypothesize that “structural flexibility” may be a neglected yet critical factor impacting MoE scaling. S’MoRE has the potential to scale better than regular MoEs, for fine-tuning tasks and beyond. We leave the application of S’MoRE on other types of tasks as future work.

B Details of Model Design

B.1 Three types of gates implemented in practice

In our experiments, we implement S’MoRE and the baselines with various types of dense and sparse gates, which we describe in more details here:

- *Dense gate:* Here we activate all children experts, meaning that $f_\ell = s_\ell$. For each children, the expert score α is still generated by the gating neural net. So the dense gate can be understood as a soft version of the sparse gates below. Now since all experts are activated, we do not include additional auxiliary loss for load-balance of all the experts. The same dense gate design is also used by HydraLoRA [Tian et al., 2024] which we include as one of the experimental baselines (see §4).
- *Sparse noisy top- k gate* [Shazeer et al., 2017]: For each layer ℓ , the router selects the top $k = f_\ell$ children residuals with the highest gating scores (generated by Eq. 6). However, a common issue with sparse expert selection is expert under-utilization, where certain experts are overused while others remain idle, resulting in inefficient training dynamics. To address this, following [Shazeer et al., 2017], we first add a learnable noise term on top of gating score α , to encourage exploration of different expert choices. Then we implement a layer-wise importance loss plus balance loss, computed separately for each set of experts $\mathcal{R}_\ell$ based on gating score distribution and activation frequency. We sum up the two auxiliary losses of each S’MoRE layer and each transformer layer, and add it to the model prediction cross-entropy (CE) loss, i.e.,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \gamma \cdot \sum_{\ell_{\text{trans}}=1}^{L_{\text{trans}}} \sum_{\ell=1}^L \left(\mathcal{L}_{\ell_{\text{trans}}, \ell}^{\text{importance}} + \mathcal{L}_{\ell_{\text{trans}}, \ell}^{\text{balance}} \right), \quad (11)$$

where L_{trans} denotes the total number of transformer layers and γ is a coefficient ($\ll 1$) controlling the strength of the load-balance constraint. The importance and balance loss computation exactly follows the original paper (see Section 4 and Appendix A of Shazeer et al. [2017]).

- *Sparse switch-transformer gate* [Fedus et al., 2022b]: This is another popular sparse gate design. There are two main differences from the noisy top- k gate above. First, the switch gate implements an optional jitter noise that is applied to the gate’s input embedding rather than the final gating score. In addition, the switch gate integrates a different way to compute the load-balance auxiliary loss. The auxiliary losses of all transformer and S’MoRE layers are still added back to the CE loss of the main task, just like Eq. 11. To ensure fair comparison in experiments, we exactly follow the balance loss implementation in the MixLoRA codebase³.

B.2 Alternative design: bottom-up routing

We propose the following bottom-up router fully compatible with S’MoRE’s structural mixing design.

In bottom-up routing, our goal remains the same as the top-down router of §3.3: to customize different children experts for different parents. Yet, under bottom-up routing, the parent index is unknown when we decide the children. So in our design, we will still include a key vector $\mathbf{k}$ (see Eq. 6). Yet this $\mathbf{k}$ is not directly associated with any specific parent expert. It represents a node position in the routing tree.

To avoid the discussion being overwhelmed by notations, we use the following example to illustrate the core idea. Consider a 3-layer S’MoRE, where each layer has a fanout of $f = 2$, and has $s = 6$ experts. Like our original design, the router still has the following:

³<https://github.com/TUDB-Labs/MixLoRA>

- a down-projection matrix that projects the original token x to d_{down} -dimensional x_{down} (e.g., $d_{\text{down}} = 16$);
- a small MLP in each layer (except layer 1).

In addition, we instantiate a learnable “key” tensor K in each layer. K ’s last dimension is $d_{\text{down}} = 16$, and K ’s leading dimensions depend on the fanout in parent layers (except that for layer 1, K ’s dimension additionally depends on the number of experts). For example, in layer 1, K has shape $(2, 2, 6, 16)$. In layer 2 and 3, K ’s shapes are $(2, 2, 16)$ and $(2, 16)$.

The routing proceeds as follows:

1. At layer 1, we compute dot product between K and x_{down} along the last dimension, generating a score tensor of shape $(2, 2, 6)$. Taking the top-2 along the last dimension, we determine the $2 \times 2 \times 2$ children experts for all the 2×2 parents.
2. Then we follow Eq. 3 to aggregate the children experts’ outputs, generating 2×2 different output embeddings.
3. In layer 2, we concatenate the 2×2 layer-1 output embeddings and the $(2, 2, 16)$ -shaped K , along the last dimension. Then we feed the concatenated tensor into layer 2’s router MLP to generate the score distribution over all 6 candidate experts. Taking the top-2 along the last dimension of the MLP output, we now select the 2×2 layer-2 experts, conditioned on the layer-1 children.
4. Layer 3 operates similarly.

Comparison. From the above, it is clear that the bottom-up router is still computationally efficient (similar to the original top-down router). Both the bottom-up and top-down designs can model the interaction between parent and children experts. The main difference is that the bottom-up router makes routing decisions based on the children’s aggregated embedding, while the top-down router directly consumes the token embedding. So in the bottom-up design, the gradient can flow back to the lower-layer experts from the router. This may lead to interesting training behaviors that differ from the top-down case.

Due to limited GPUs resources, we are unable to run ablation with the bottom-up router. We leave such evaluation as an important future work.

C Details of Theoretical Analysis

C.1 Derivation of parameter & computation costs

Here we provide additional algorithmic details for the parameter and computation efficiency calculation in §3.4.

Parameter efficiency. From Eq. 8, we have

$$P_{\ell+1} = s_{\ell} \cdot d \cdot r_{\ell} + d_{\ell+1}^2 \quad (12)$$

Then Eq. 9 is derived as

$$\begin{aligned}
P_{\text{proj}} + \sum_{\ell=1}^L P_{\ell} &= d \cdot d_L + \sum_{\ell=0}^{L-1} P_{\ell+1} \\
&= d \cdot d_L + \sum_{\ell=0}^{L-1} s_{\ell} \cdot d \cdot r_{\ell} + \sum_{\ell=0}^{L-1} d_{\ell+1}^2 \\
&= d \cdot d_L + d \cdot \left(\sum_{\ell=0}^{L-1} s_{\ell} \cdot r_{\ell} \right) + \sum_{\ell=0}^{L-1} d_{\ell+1}^2 \\
&= d \cdot d_L + d \cdot d_L + \sum_{\ell=0}^{L-1} d_{\ell+1}^2 \\
&= 2 \cdot d \cdot d_L + \Delta
\end{aligned} \tag{13}$$

$$\text{where} \quad \Delta = \sum_{\ell=0}^{L-1} d_{\ell+1}^2 \ll 2 \cdot d \cdot d_L$$

Computation cost. In Eq. 3, each $B_{\ell}^n \cdot A_{\ell}^n \cdot x$ requires $C' = d \cdot r_{\ell} + r_{\ell} \cdot d_{\ell+1}$ operations. Each $W_{\ell} \cdot x_{\ell}^n$ requires $C'' = d_{\ell} \cdot d_{\ell+1}$. Consider all activated experts i in layer $\ell + 1$, there can be at most $N' = \min\{s_{\ell}, F_{\ell}\}$ distinct $B_{\ell}^n \cdot A_{\ell}^n$ terms, incurring C' cost N' times. There are F_{ℓ} different x_{ℓ}^n inputs, each incurring C'' cost. Ignoring the element-wise addition “ $\sum_{n \in N_{\ell}^i}$ ” and multiplication $\alpha_{\ell}^{i,n}$, the total cost of layer $\ell + 1$ equals (where d_{ℓ} and F_{ℓ} follow Eq. 4 and Eq. 5):

$$\begin{aligned}
C_{\ell+1} &\leq \min\{s_{\ell}, F_{\ell}\} \cdot r_{\ell} \cdot (d + d_{\ell+1}) + F_{\ell} \cdot d_{\ell} \cdot d_{\ell+1} \\
&= \min\{s_{\ell}, F_{\ell}\} \cdot r_{\ell} \cdot d + \min\{s_{\ell}, F_{\ell}\} \cdot r_{\ell} \cdot d_{\ell+1} + F_{\ell} \cdot d_{\ell} \cdot d_{\ell+1} \\
&\leq s_{\ell} \cdot r_{\ell} \cdot d + F_{\ell} \cdot r_{\ell} \cdot d_{\ell+1} + F_{\ell} \cdot d_{\ell} \cdot d_{\ell+1} \\
&= s_{\ell} \cdot r_{\ell} \cdot d + F_{\ell} \cdot d_{\ell+1} \cdot (d_{\ell} + r_{\ell})
\end{aligned} \tag{14}$$

The cost of the final projection equals $C_{\text{proj}} = d \cdot d_L$. So the overall computation cost is:

$$C_{\text{proj}} + \sum_{\ell=1}^L C_{\ell} = d \cdot d_L + d \cdot \left(\sum_{\ell=0}^{L-1} s_{\ell} \cdot r_{\ell} \right) + \Delta' \stackrel{\text{(b)}}{=} 2 \cdot d \cdot d_L + \Delta' \stackrel{\text{(c)}}{\approx} 2 \cdot d \cdot d_L \tag{15}$$

where $\Delta' \leq \sum_{\ell=0}^{L-1} F_{\ell} \cdot d_{\ell+1} \cdot (d_{\ell} + r_{\ell})$. Steps “(b)” and “(c)” follow similar reasoning to Eq. 9.

Under practical values of $d_{\ell} + r_{\ell} \leq d_{\ell+1} \ll d$, the overhead term Δ' is small or negligible compared to the main cost $2 \cdot d \cdot d_L$. In Table 6, we empirically calculate the value of $2 \cdot d \cdot d_L$, the overhead Δ' , and their ratio. We take a representative configuration with $f_{\ell} = 2$, $s_{\ell} = 4$, and $r_{\ell} = 8$ or 16 for all layers ℓ (consistent with the experiments in §4). For 2 layers, the overhead Δ' is just **1.2%** (or **2.3%**) of the cost of vanilla LoRA with rank $d_L = 64$ (or $d_L = 128$).

Table 6: Overhead Δ' compared with the main computation cost $2 \cdot d \cdot d_L$

r_{ℓ}	L	d_L	$2 \cdot d \cdot d_L$	Δ'	Overhead ratio
8	2	64	0.5M	0.006M	1.2%
	3	96	0.8M	0.026M	3.3%
	4	128	1.0M	0.079M	7.5%
16	2	128	1.0M	0.025M	2.3%
	3	192	1.6M	0.104M	6.6%
	4	256	2.1M	0.315M	15.0%

Similar to the analysis in the “Parameter efficiency” paragraph, the gating MLPs are lightweight compared to the main cost $2 \cdot d \cdot d_L$, due to the small dimensionalities.

Thus, the total computation cost of S’MoRE is approximately $2 \cdot d \cdot d_L$, which is *the same as the cost of a vanilla LoRA* with rank d_L . This proves the both parameter and the computation efficiency of S’MoRE.

C.2 Proof of model capacity

C.2.1 Proof for two special S’MoRE configurations

Proposition C.1. (Proposition 3.1) *S’MoRE can express MoLRE, when $L = 1$ and $\sigma(\cdot)$ is the identity mapping.*

Proof. When $L = 1$, there is only a single layer propagation. When we set σ as the identity mapping, Eq. 3 becomes

$$\mathbf{x}_1 = \sum_{n \in \mathcal{N}_0} \alpha_\ell^n \cdot \mathbf{B}_0^n \cdot \mathbf{A}_0^n \cdot \mathbf{x} \quad (16)$$

where we omit the superscript i since there is just one parent node (the root of all all experts in a flat layer).

Combined with the final projection (see end of §3.2), the final output is computed by

$$\begin{aligned} \mathbf{x}' &= \mathbf{W}_{\text{proj}} \cdot \sum_{n \in \mathcal{N}_0} \alpha_\ell^n \cdot \mathbf{B}_0^n \cdot \mathbf{A}_0^n \cdot \mathbf{x} \\ &= \sum_{n \in \mathcal{N}_0} \alpha_\ell^n \cdot (\mathbf{W}_{\text{proj}} \cdot \mathbf{B}_0^n) \cdot \mathbf{A}_0^n \cdot \mathbf{x} \end{aligned} \quad (17)$$

where $\mathbf{A}_0^n \in \mathbb{R}^{r_0 \times d}$, $\mathbf{B}_0^n \in \mathbb{R}^{(s_0 \cdot r_0) \times r_0}$ and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d \times (s_0 \cdot r_0)}$.

For MoLRE with s_0 rank- r_0 experts, according to the definition in §3.1, we can express its layer operation as

$$\bar{\mathbf{x}}' = \sum_{n \in \mathcal{N}} \text{ROUTE}(\mathbf{x})^n \cdot \bar{\mathbf{B}}^n \cdot \bar{\mathbf{A}}^n \cdot \mathbf{x} \quad (18)$$

where we use overhead “bar” to distinguish variables of MoLRE from those of 1-layer S’MoRE. Here $\bar{\mathbf{A}}^n \in \mathbb{R}^{r_0 \times d}$ and $\bar{\mathbf{B}}^n \in \mathbb{R}^{d \times r_0}$.

To make Eq. 17 and Eq. 18 equivalent, we can have

- S’MoRE’s router implementing as $\text{ROUTE}(\mathbf{x})^n$
- $\mathbf{A}_0^n = \bar{\mathbf{A}}^n$ (by definition, both matrices have the same shape)
- $\mathbf{B}_0^n = \begin{bmatrix} \mathbf{0}_{r_0} \\ \vdots \\ \mathbf{0}_{r_0} \\ \mathbf{I}_{r_0} \\ \mathbf{0}_{r_0} \\ \vdots \\ \mathbf{0}_{r_0} \end{bmatrix}$, which is a binary matrix by vertically stacking s_0 square blocks of $r_0 \times r_0$ sub-matrices. The n -th block is a $r_0 \times r_0$ identity matrix, $\mathbf{I}_{r_0}$, while all the other blocks are 0 (denoted as $\mathbf{0}_{r_0}$).
- $\mathbf{W}_{\text{proj}} = [\bar{\mathbf{B}}^1, \dots, \bar{\mathbf{B}}^{s_0}]$.

Then $\mathbf{W}_{\text{proj}} \cdot \mathbf{B}_0^n = \bar{\mathbf{B}}^n$. And Eq. 17 becomes identical to Eq. 18, completing the proof. $\square$

Proposition C.2. (Proposition 3.2) *S’MoRE can express MoMOR, when setting $\sigma(\cdot)$ as the identity mapping.*

Proof. Without σ , we can collapse a multi-layer S’MoRE into a single-layer equivalent. For $L = 2$, following Eq. 3, we have

$$\begin{aligned}
x_2 &= \sum_{n \in \mathcal{N}_1} \alpha_1^n \cdot (B_1^n \cdot A_1^n \cdot x + W_1 \cdot x_1^n) \\
&= \sum_{n \in \mathcal{N}_1} \alpha_1^n \cdot B_1^n \cdot A_1^n \cdot x + W_1 \sum_{n \in \mathcal{N}_1} \alpha_1^n \cdot \left(\sum_{m \in \mathcal{N}_0^n} \alpha_0^{n,m} \cdot B_0^m \cdot A_0^m \cdot x \right) \\
&= \sum_{n \in \mathcal{N}_1} \hat{\alpha}_1^n \cdot B_1^n \cdot A_1^n \cdot x + \sum_{m \in \mathcal{N}_0} \hat{\alpha}_0^m \cdot (W_1 \cdot B_0^m \cdot A_0^m) \cdot x
\end{aligned} \tag{19}$$

where we define $\hat{\alpha}_1^n = \alpha_1^n$ and $\hat{\alpha}_0^m = \sum_{n \in \mathcal{N}_1 \text{ and } m \in \mathcal{N}_0^n} (\alpha_1^n \cdot \alpha_0^{n,m})$.

In general, for L layers and with the final projection step W_{proj} , we can summarize the propagation equation as

$$x' = \sum_{\ell=0}^{L-1} \sum_{i=1}^{s_\ell} \hat{\alpha}_\ell^i \cdot \left(\prod_{k=\ell+1}^L W_k \right) \cdot B_\ell^i \cdot A_\ell^i \cdot x \tag{20}$$

where we define $W_L = W_{\text{proj}} \in \mathbb{R}^{d \times d_L}$ and $\hat{\alpha}_\ell^i$ is a scalar coefficient by aggregating the router weights along all paths that end at the layer- $(\ell+1)$ expert i ⁴. In other words, $\hat{\alpha}_\ell^i$ generalizes the definition of $\hat{\alpha}_0^m$ above. The “path” here refers to the “ancestral path” (Definition C.5) ending at i . See more discussion on the routing tree in Appendix C.2.3. Also, if an expert is never selected, we let its $\hat{\alpha}_\ell^i = 0$. This way, we can replace the summation over $\mathcal{N}_\ell$ in Eq. 19 with the summation over $1 \leq i \leq s_\ell$ in Eq. 20.

For MoMOR model, following Eq. 2, we write its layer propagation as

$$x' = \sum_{\ell=1}^{L-1} \sum_{i=1}^{s_\ell} \text{ROUTE}_\ell(x)^i \cdot \bar{B}_\ell^i \cdot \bar{A}_\ell^i \cdot x \tag{21}$$

We can make Eq. 20 and Eq. 21 equivalent by a similar construction as the proof for Proposition 3.1. First, define a special binary projection matrix $P_{a \times b} \in \{0, 1\}^{a \times b}$ (where $a > b$) as

$$P_{a \times b} = \begin{bmatrix} \mathbf{0}_{(a-b) \times b} \\ \mathbf{I}_{b \times b} \end{bmatrix} \tag{22}$$

meaning that the first $a - b$ rows of $P_{a \times b}$ are all 0, and the bottom b rows are an identity matrix. It is easy to verify that for $a > b > c$:

$$P_{a \times b} \cdot P_{b \times c} = P_{a \times c} \tag{23}$$

Then we can set all parameters of S²MoRE as follows:

- Let the S²MoRE router implement $\text{ROUTE}_\ell(x)^i$.
- Let $A_\ell^i = \bar{A}_\ell^i$.
- Let B_ℓ^i be a $d_{\ell+1} \times r_\ell$ binary matrix, where its row $(i-1) \cdot r_\ell + 1$ to row $i \cdot r_\ell$ is a $r_\ell \times r_\ell$ identity matrix, and its all other rows are all 0. Here we let both i and the row index start from 1.
- Let $W_L = W_{\text{proj}} = [\bar{B}_0^1, \dots, \bar{B}_0^{s_0}, \dots, \bar{B}_{L-1}^1, \dots, \bar{B}_{L-1}^{s_{L-1}}]$ as the horizontal concatenation of all MoMOR’s up-projection matrices $\bar{B}_\ell^i$.

⁴The same expert i of layer $\ell+1$ may be selected multiple times, corresponding to different parents or ancestors. Thus, there can be multiple paths ending at the layer- $(\ell+1)$ expert i .

- Each $\mathbf{W}_k$ has shape $d_{k+1} \times d_k$ where $d_{k+1} = d_k + s_k \cdot r_k$. We set it as $\mathbf{W}_k = \mathbf{P}_{d_{k+1} \times d_k}$. Then it follows that

$$\prod_{k=\ell+1}^{L-1} \mathbf{W}_k = \mathbf{P}_{d_L \times d_{L-1}} \cdot \mathbf{P}_{d_{L-1} \times d_{L-2}} \cdots \mathbf{P}_{d_{\ell+2} \times d_{\ell+1}} = \mathbf{P}_{d_L \times d_{\ell+1}} \quad (24)$$

$$\begin{aligned} \Rightarrow \left(\prod_{k=\ell+1}^L \mathbf{W}_k \right) \cdot \mathbf{B}_\ell^i &= \mathbf{W}_{\text{proj}} \cdot \left(\prod_{k=\ell+1}^{L-1} \mathbf{W}_k \right) \cdot \mathbf{B}_\ell^i \\ &= \mathbf{W}_{\text{proj}} \cdot \mathbf{P}_{d_L \times d_{\ell+1}} \cdot \mathbf{B}_\ell^i \\ &= \bar{\mathbf{B}}_\ell^i \end{aligned} \quad (25)$$

Under the above construction, it is clear that Eq. 21 and Eq. 20 are exactly the same. Thus, S'MoRE can express MoMOR, concluding the proof.

Remark. Note that the equivalence between S'MoRE and MoMOR can only be established when we set the layer dimension d_ℓ according to Eq. 4. This can be seen from the “minimum dimensionality” discussion in §3.2.

□

C.2.2 Proof of Theorem 3.3

Theorem C.3. (Theorem 3.3) *The structural flexibility of MoMOR is upper-bounded by $\Gamma_{\text{MoMOR}} = \max_{\mathbf{x}, \Theta} \text{dist}(\mathbf{x}; \Theta) \leq \binom{s_{L-1}}{f_{L-1}} \cdot \prod_{\ell=0}^{L-2} \left(\sum_{i=f_\ell}^{\min\{F_\ell, s_\ell\}} \binom{s_\ell}{i} \right)$.*

Proof. The upper bound of Γ_{MoMOR} basically quantifies the total number of combinations to select experts from each residual pool.

Assumption. We first simplify Eq. 2 that the router-generated coefficient $\text{ROUTE}_\ell(\mathbf{x})^i$ is just a binary mask. i.e., for a selected expert i , we have $\text{ROUTE}_\ell(\mathbf{x})^i = 1$. Otherwise, $\text{ROUTE}_\ell(\mathbf{x})^i = 0$. Such an assumption is just to ease the calculation of Γ_{MoMOR} and $\Gamma_{\text{S'MoRE}}$. It does not affect our fundamental conclusion that S'MoRE yields exponentially higher structural flexibility than MoMOR.

Based on Eq. 2, the MoMOR output is generated by a flat summation of different-order residues. Given any input $\mathbf{x}$, the number of distinct outputs cannot exceed the number of distinct ways to select residues from the pools $\mathcal{R}_0, \dots, \mathcal{R}_{L-1}$. Here we show some examples to illustrate the meaning of “distinct expert selection”.

- “Selecting experts 1,2,3 from $\mathcal{R}_0$ ” and “selecting experts 1,3,4 from $\mathcal{R}_0$ ” correspond to 2 distinct ways.
- “Selecting experts 1,2,3 from $\mathcal{R}_0$ ” and “selecting experts 3,2,1 from $\mathcal{R}_0$ ” correspond to the same way, because there is no ordering among the selected experts⁵.
- “Selecting experts 1,1,3 from $\mathcal{R}_0$ ” and “selecting experts 1,3,3 from $\mathcal{R}_0$ ”⁶ correspond to the same way due to our assumption of making $\text{ROUTE}_\ell(\mathbf{x})^i$ a binary mask. Basically we only care about whether an expert is selected or not. It does not matter how many times an expert is selected.

Remark. Distinct expert selections do not guarantee distinct outputs. For example, consider “selecting 1,2,3 from $\mathcal{R}_0$ ” and “selecting 1,3,4” from $\mathcal{R}_0$. Following the notation of Eq. 2, if the experts’ weights satisfy $\Delta \mathbf{W}_0^1 + \Delta \mathbf{W}_0^2 + \Delta \mathbf{W}_0^3 = \Delta \mathbf{W}_0^1 + \Delta \mathbf{W}_0^3 + \Delta \mathbf{W}_0^4$, then the two case generates the same output for all input $\mathbf{x}$:

⁵The order among selected experts does not matter because the sum aggregation of Eq. 2 is *permutation invariant*

⁶If we follow S'MoRE’s recursive expert selection process described in §3.3, the same expert of higher-order may be selected multiple times, from different lower-order parents.

$$\sum_{i \in \{1,2,3\}} \Delta \mathbf{W}_0^i \cdot \mathbf{x} = \sum_{j \in \{1,3,4\}} \Delta \mathbf{W}_0^j \cdot \mathbf{x} \quad (26)$$

Hence, counting the number of distinct ways of expert selection just gives an upper bound of Γ_{MoMOR} , because Γ_{MoMOR} is defined on the number of distinct outputs.

Counting the combinations. For the $\mathcal{R}_{L-1}$ pool with size s_{L-1} , there are $\binom{s_{L-1}}{f_{L-1}}$ ways to pick f_{L-1} residues. For $\mathcal{R}_{L-2}$ with $\ell \leq L-2$, there are $F_{\ell+1}$ parents, each picking f_ℓ children in the pool. Different parents can pick the same children. The number of distinct children selected by all parents ranges from f_ℓ to $\min\{F_\ell, s_\ell\}$. This makes the total count $\sum_{i=f_\ell}^{\min\{F_\ell, s_\ell\}} \binom{s_\ell}{i}$. From basic Combinatorics, each layer ℓ contributes to a multiplicative factor in the total count. Thus, the final upper bound is:

$$\Gamma_{\text{MoMOR}} \leq \binom{s_{L-1}}{f_{L-1}} \cdot \prod_{\ell=0}^{L-2} \left(\sum_{i=f_\ell}^{\min\{F_\ell, s_\ell\}} \binom{s_\ell}{i} \right) \quad (27)$$

□

C.2.3 Proof of Theorem 3.4

Theorem C.4. (Theorem 3.4) Setting $\sigma(\cdot)$ as an MLP, there exists some Θ' such that the structural flexibility of S'MoRE is $\Gamma_{\text{S'MoRE}} = \min_{\mathbf{x}} \text{dist}(\mathbf{x}; \Theta') = \prod_{\ell=0}^{L-1} \binom{s_\ell}{f_\ell}^{F_{\ell+1}}$, where $F_L := 1$.

Proof. We prove in two stages:

1. We show that following the routing process of S'MoRE, there can be $\Gamma_{\text{S'MoRE}}$ non-isomorphic depth- L trees, where each tree node is an expert residue.
2. We construct a S'MoRE instance where its L -layer propagation (Eq. 3) generates distinct outputs for all non-isomorphic trees above, regardless of input token embedding $\mathbf{x}$.

Both can be proven by induction.

Assumption. Similar to Theorem 3.3, we make simplification to the layer propagation Eq. 3, that the coefficient $\alpha_\ell^{i,n}$ is just a binary mask. i.e., for a selected children n , we have $\alpha_\ell^{i,n} = 1$. Otherwise, $\alpha_\ell^{i,n} = 0$.

Stage 1: Number of non-isomorphic trees. Recall the expert selection / tree construction process in §3.3: each active parent expert of layer $\ell+1$ (in $\mathcal{R}_\ell$) selects $f_{\ell-1}$ children out of all the $s_{\ell-1}$ experts of layer ℓ . So by traversing all the L layers, the router builds a depth- L balanced tree (which has $\prod_{\ell=0}^{L-1} f_\ell$ leaf nodes in total). Note that

1. For each parent, its f_ℓ selected children are distinct (i.e., the same parent cannot select the same child twice).
2. However, the same expert may appear in the same tree-level multiple times, corresponding to different parents or ancestors.
3. There is **no ordering** among the selected children, since Eq. 3 performs “sum” aggregation which is *permutation invariant*. e.g., it is equivalent to say that a parent of layer ℓ selects “children 1,3,4” and “children 4,3,1”.

Due to Point 2 above, we cannot uniquely identify a tree node by the its corresponding expert’s layer index and expert index. Yet, Points 1 and 3 ensure that any tree node n is *uniquely identifiable* by n ’s ancestral path $\mathcal{P}_n$.

Definition C.5. (Ancestral path $\mathcal{P}_n$) Let (ℓ, i) denote expert i of layer ℓ . Suppose a tree-node n at tree-level t corresponds to expert $(L - t + 1, i)$. Then n 's ancestral path, $\mathcal{P}_n = ((L - t + 1, i), (L - t + 2, i'), \dots, (L, i' \dots'))$, defines the unique path to traverse from n up to the tree root (where we treat the root as a *virtual* node that is the parent of all $(L, i' \dots')$, and we omit the root in the path).

Definition C.6. (Leaves' ancestral paths $\mathcal{T}$) Given a tree, define $\mathcal{T} = \{\mathcal{P}_n \mid n \text{ is a leaf node}\}$ as the set of ancestral paths of all leaf nodes, where there are $\prod_{\ell=0}^{L-1} f_\ell$ leaves, all at tree-level L .

Two trees are *isomorphic* if their structures are equivalent. That means, we can permute or swap the children (together with their corresponding descendant sub-tree) of some parent nodes to make the two trees look exactly the same. $\mathcal{T}$ enables us to define isomorphism. In our construction, the children are not ordered (Point 3 above), and so permuting or swapping children does not change $\mathcal{T}$. Thus, isomorphic trees have the same $\mathcal{T}$. On the other hand, we can show trees of the same $\mathcal{T}$ can be made equivalent by permutation or swapping, and thus are isomorphic. In sum, we can define tree isomorphism by $\mathcal{T}$ as follows:

Definition C.7. (Isomorphism) Given two trees, let their leaves' ancestral paths be $\mathcal{T}$ and $\mathcal{T}'$. The two trees are isomorphic if and only if $\mathcal{T} = \mathcal{T}'$.

We next derive the number of depth- L non-isomorphic trees by induction.

Imagine that we apply the top-down expert selection from layer ℓ down to layer 1 (with $\ell \geq 1$): at layer ℓ , we select $f_{\ell-1}$ experts from $s_{\ell-1}$ experts; at layer $\ell - 1$, for each of the selected parent of layer ℓ , we select $f_{\ell-2}$ from $s_{\ell-2}$ experts, and so on.

Induction hypothesis: the number of non-isomorphic trees yielded by such an expert-selection process equals:

$$\Gamma_{\text{S'MORE}}^\ell = \prod_{k=0}^{\ell-1} \binom{s_k}{f_k}^{F_{k+1}/F_\ell} \quad (28)$$

for some $1 \leq \ell < L$.

Base case $\ell = 1$: we are just sampling a single level. So the number of non-isomorphic trees equals the number of total ways to select f_0 experts from s_0 , which is $\binom{s_0}{f_0}$.

And

$$\begin{aligned} \Gamma_{\text{S'MORE}}^1 &= \prod_{k=0}^{1-1} \binom{s_k}{f_k}^{F_{k+1}/F_1} \\ &= \binom{s_0}{f_0} \end{aligned} \quad (29)$$

So the base case holds.

Induction from ℓ to $\ell + 1$: To construct a tree by selecting experts from layer $\ell + 1$ to 1, we follow two steps:

1. We select f_ℓ out of s_ℓ experts. Denote them as $\mathcal{E}_\ell = \{(\ell + 1, i_1), \dots, (\ell + 1, i_{f_\ell})\}$, where $i_a \neq i_b$ for all $a \neq b$.
2. We start from each $(\ell + 1, i_m)$ and recursively activate experts from layer ℓ down to 1 (where $1 \leq m \leq f_\ell$), following the procedure described above. Denote each such tree by its leaves' ancestral paths, $\mathcal{T}_{\ell, i_m}$.

Let $\mathbb{T}_\ell$ be the set of all possible $\mathcal{T}_{\ell, i_m}$ — note that $\mathbb{T}_\ell$ does not have subscript i_m , since an ancestral path ends at a virtual root node independent of i_m (see Definition C.5), and thus $\mathbb{T}_\ell$ is the same for all i_m . Based on the induction hypothesis, $|\mathbb{T}_\ell| = \Gamma_{\text{S'MORE}}^\ell$.

For such a tree constructed by the two steps above, let $\mathcal{T}_{\ell+1}$ be its leaves' ancestral paths:

$$\mathcal{T}_{\ell+1} = \bigcup_{k=1}^{f_\ell} \{p \oplus (\ell+1, i_k) \mid p \in \mathcal{T}_{\ell, i_k}\} \quad (30)$$

where “ $\oplus$ ” means appending $(\ell+1, i_k)$ to the end of the path p . By Definition C.7, the total number of non-isomorphic trees equals the number of distinct $\mathcal{T}_{\ell+1}$, which can be calculated with the following reasoning:

- There are $\binom{s_\ell}{f_\ell}$ distinct ways to choose $\mathcal{E}_\ell$ of Step 1.
- For each choice of $\mathcal{E}_\ell$, there are $|\mathbb{T}_\ell|$ choices of $\mathcal{T}_{\ell, i_k}$ for each i_k of $\mathcal{E}_\ell$, leading to $|\mathbb{T}_\ell|^{f_\ell}$ distinct combinations.

So the number of distinct $\mathcal{T}_{\ell+1}$ equals:

$$\begin{aligned} |\mathbb{T}_{\ell+1}| &= \binom{s_\ell}{f_\ell} \cdot |\mathbb{T}_\ell|^{f_\ell} \\ &= \binom{s_\ell}{f_\ell} \cdot (\Gamma_{\mathbf{S}, \text{MoRE}}^\ell)^{f_\ell} \\ &= \binom{s_\ell}{f_\ell} \cdot \left(\prod_{k=0}^{\ell-1} \binom{s_k}{f_k}^{F_{k+1}/F_\ell} \right)^{f_\ell} \\ &= \binom{s_\ell}{f_\ell} \cdot \prod_{k=0}^{\ell-1} \binom{s_k}{f_k}^{F_{k+1} \cdot \frac{f_\ell}{F_\ell}} \\ &= \binom{s_\ell}{f_\ell}^{F_{\ell+1}/F_{\ell+1}} \cdot \prod_{k=0}^{\ell-1} \binom{s_k}{f_k}^{F_{k+1}/F_{\ell+1}} \\ &= \prod_{k=0}^{\ell} \binom{s_k}{f_k}^{F_{k+1}/F_{\ell+1}} \\ &= \Gamma_{\mathbf{S}, \text{MoRE}}^{\ell+1} \end{aligned} \quad (31)$$

This completes the induction step. Thus, the total number of non-isomorphic trees for all L layers equals $\Gamma_{\mathbf{S}, \text{MoRE}}^L = \prod_{\ell=0}^{L-1} \binom{s_\ell}{f_\ell}^{F_{\ell+1}/F_L} = \prod_{\ell=0}^{L-1} \binom{s_\ell}{f_\ell}^{F_{\ell+1}}$ where $F_L := 1$.

Stage 2: Distinguishing non-isomorphic trees. We next show that there exists some parameters Θ' such that the layer propagation following Eq. 3 generates distinct output for non-isomorphic trees.

Notational correction to Eq. 3: In §3.2, we use $\mathbf{x}_\ell^i$ to denote the output embedding where i is the *expert* index. This notation is not precise since the same expert can appear as multiple tree nodes, as discussed in the Stage 1 proof above. To make the correction, we instead let $\mathbf{x}_\ell^i$ denote the embedding of *node* index i^7 for tree-level $L - \ell$, meaning that there can be $\mathbf{x}_\ell^i$ and $\mathbf{x}_\ell^{i'}$ mapped to the same expert, where $i \neq i'$.

Including the bias term: Our proof requires a minor modification of Eq. 3 to add a bias term $\mathbf{b}_\ell^n \in \mathbb{R}^{d_{k+1}}$ associated with each expert n . So the updated layer propagation equation, adapted from Eq. 3 now becomes:

$$\mathbf{x}_{\ell+1}^i = \sum_{n \in \mathcal{N}_\ell^i} \sigma(\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \mathbf{W}_\ell \cdot \mathbf{x}_\ell^{i \rightarrow n} + \mathbf{b}_\ell^n) \quad (32)$$

⁷In our terminology above, this means that each (ℓ, i) now corresponds to a *distinct* ancestral path.

where ℓ is the *layer* index; i is the index of a *tree node*, while n is still the index of an *expert*. $\mathcal{N}_\ell^i$ denotes the set of indices of the children experts selected by node i . Note, “ $i \rightarrow n$ ” means that a tree node i picks a previous-layer expert n as its child. So with a slight abuse of notation, we use superscript “ $i \rightarrow n$ ” to index such a child tree node. $\alpha_\ell^{i,n}$ of Eq. 3 is omitted since we simplify the expert weight as binary mask, as stated above.

We are now ready for the proof.

First, note that since the operations by Eq. 3 are permutation invariant, S’MoRE will generate the same output for all isomorphic trees.

Next, we consider non-isomorphic trees. Again we prove by induction.

Similar to the Stage 1 setting, we consider an expert-selection process from layer ℓ down to layer 1. After building such an ℓ -level tree, the model propagates the input token $\mathbf{x}$ from layer 1 up to layer ℓ to generate the output $\mathbf{x}_\ell$. Note, since in the induction step, the propagation terminates at $\mathbf{x}_\ell$, we do not need to superscript $\mathbf{x}_\ell$ with an additional node index i . In other words, $\mathbf{x}_\ell$ here is analogous to the *final* embedding $\mathbf{x}_L$ described in §3.2.

Induction hypothesis: For any ℓ -level non-isomorphic trees $\mathcal{T}_\ell \neq \mathcal{T}'_\ell$, we can set the layer 1 to ℓ parameters of S’MoRE such that $\mathbf{x}_\ell \neq \mathbf{x}'_\ell$.

Base case $\ell = 1$: For a single layer, the propagation simplifies to

$$\mathbf{x}_1 = \sum_{n \in \mathcal{N}_0} \sigma(B_0^n \cdot A_0^n \cdot \mathbf{x} + b_0^n) \quad (33)$$

where non-isomorphic trees under $\ell = 0$ degrades to distinct neighbor sets $\mathcal{N}_0$.

We want distinct outputs $\mathbf{x}_\ell \neq \mathbf{x}'_\ell$ for *all* inputs $\mathbf{x}$. So we have the following simple way to construct the parameters:

- $B_0^n = \mathbf{0}$ and $A_0^n = \mathbf{0}$, which leads to $B_0^n \cdot A_0^n \cdot \mathbf{x} + b_0^n = b_0^n$ for all input $\mathbf{x}$;
- Let the first element of b_0^n store the expert index (an integer from 1 to s_0), and the rest of the elements be 0.

We reuse the following lemma from Xu et al. [2019]:

Lemma C.8. (see Lemma 5 of Xu et al. [2019]) Assume a countable input feature space $\mathcal{X}$. There exists a function $f : \mathcal{X} \rightarrow \mathbb{R}^d$ so that $h(X) = \sum_{x \in X} f(x)$ is unique for each set $X \subset \mathcal{X}$ of bounded size.

In our case, σ of Eq. 32 corresponds to function f of Lemma C.8, and we treat b_0^n as the function’s input features. The “feature space” consisting of all possible b_0^n is clearly countable (since each element of b_0^n is either 0 or a bounded integer). The neighbor set $\mathcal{N}_0$ corresponds to X of Lemma C.8, which can be an arbitrary combination of the children experts.

Thus, due to the universal approximation theorem [Hornik et al., 1989], we can instantiate σ as an MLP to implement such a function f , to guarantee that all non-isomorphic trees get a unique output $\mathbf{x}_1$. This proves the base case.

Induction from ℓ to $\ell + 1$: Consider two trees constructed by recursive expert selection from layer $\ell + 1$ to 1. We use “prime” to denote quantities of the second tree. For example, their leaves’ ancestral paths are $\mathcal{T}_{\ell+1}$ and $\mathcal{T}'_{\ell+1}$. According to the analysis in the Stage 1 proof above, there are two possibilities to make the two trees non-isomorphic. i.e., $\mathcal{T}_{\ell+1} \neq \mathcal{T}'_{\ell+1}$:

1. The sets of level-1 nodes are different: $\mathcal{E}_\ell \neq \mathcal{E}'_\ell$;
2. Otherwise, let $\mathcal{E}_\ell = \mathcal{E}'_\ell = \{(\ell + 1, i_1), \dots, (\ell + 1, i_{f_\ell})\}$. There exists i_m such that $\mathcal{T}_{\ell, i_m} \neq \mathcal{T}'_{\ell, i_m}$ for some $1 \leq m \leq f_\ell$.

Our goal is to show that for each of the above cases, Eq. 32 can generate distinct outputs for $\mathcal{T}_{\ell+1}$ and $\mathcal{T}'_{\ell+1}$.

Similar to the construction in the $\ell = 1$ case, we set $B_\ell^n = \mathbf{0}$ and $A_\ell^n = \mathbf{0}$. And b_ℓ^n is a one-hot vector with the first element being the expert index (ranging from 1 to s_ℓ). Recall that $W_\ell \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ where $d_{\ell+1} = s_\ell \cdot r_\ell + d_\ell$ (see Eq. 4). We set

$$W_\ell = \begin{bmatrix} \mathbf{0}_{(s_\ell \cdot r_\ell) \times d_\ell} \\ I_{d_\ell \times d_\ell} \end{bmatrix} \quad (34)$$

where $\mathbf{0}_{(s_\ell \cdot r_\ell) \times d_\ell}$ is a $(s_\ell \cdot r_\ell) \times d_\ell$ all-0 matrix and $I_{d_\ell \times d_\ell}$ is a $d_\ell \times d_\ell$ identity matrix.

So Eq. 32 now becomes

$$x_{\ell+1}^i = \sum_{n \in \mathcal{N}_\ell^i} \sigma \left(\begin{bmatrix} \hat{b}_\ell^n \\ x_\ell^{i \rightarrow n} \end{bmatrix} \right) \quad (35)$$

where $\hat{b}_\ell^n$ is a length- $(s_\ell \cdot r_\ell)$ vector by discarding the trailing 0s of b_ℓ^n .

Since the layer- $(\ell + 1)$ output corresponds to the tree root, we can ignore the index i . Also note that $i \rightarrow n$ is essentially i_m of $\mathcal{E}_\ell$ above.

So we have

$$x_{\ell+1} = \sum_{n \in \mathcal{N}_\ell} \sigma \left(\begin{bmatrix} \hat{b}_\ell^n \\ x_\ell^{i_m} \end{bmatrix} \right) \quad (36)$$

Finally, we go back to the two cases above that makes two trees non-isomorphic. Clearly, for either case, the two non-isomorphic trees will have different sets of $\begin{bmatrix} \hat{b}_\ell^n \\ x_\ell^{i_m} \end{bmatrix}$. This allows us to apply Lemma C.8, and conclude that the outputs $x_{\ell+1}$ will also be different for the two non-isomorphic trees.

Note that 1. we are still dealing with a countable feature space, since there are finite number (i.e., $\Gamma_{\text{S'MoRE}}^\ell$) of distinct $x_\ell^{i_m}$; 2. Different sets of $\begin{bmatrix} \hat{b}_\ell^n \\ x_\ell^{i_m} \end{bmatrix}$ means different input “ X ” to function f in Lemma C.8.

This completes the induction step from ℓ to $\ell + 1$.

In sum, our layer propagation function in Eq. 32 ensures that we can find some S'MoRE parameters Θ' such that all depth- L non-isomorphic trees will lead to distinct outputs x_L .

Combining the proof for the two stages, we have shown that the “structural flexibility” of S'MoRE equals

$$\Gamma_{\text{S'MoRE}} = \prod_{\ell=0}^{L-1} \binom{s_\ell}{f_\ell}^{F_{\ell+1}}. \quad (37)$$

Final remark. In the proof, we require σ to be an MLP. In practice, we can implement σ simply as non-linear activation (e.g., ReLU). It is easy to see that setting σ as “an MLP with a *single* hidden layer of dimension $d_{\ell+1}$ ” is equivalent to setting σ simply as an activation function — For the single-layer MLP, the transformation matrix before the activation can be merged with $B_\ell^n \cdot A_\ell^n$ and W_ℓ of the S'MoRE layer. The transformation matrix after the activation can be merged with the next layer $W_{\ell+1}$.

Even if we implement σ as an MLP of at least 2 layers, it is still computation and parameter efficient. The input dimension to the MLP is d_ℓ , which is small (compared with the dimension of the token embeddings). Thus, it is reasonable to set the hidden dimension of the MLP layers also small. This makes the overall MLP very compact. We can follow similar reasoning as §3.4.

□

C.2.4 Proof of Corollary 3.5

Corollary C.9. (Corollary 3.5) Let $\Gamma_{S', \text{MoRE}^*}^\ell$ be the structural flexibility of ℓ -layer S' MoRE variant under Eq. 38. It satisfies the following recursion: $\Gamma_{S', \text{MoRE}^*}^\ell = \binom{s_{\ell-1}}{f_{\ell-1}} \cdot \binom{\Gamma_{S', \text{MoRE}^*}^{\ell-1} + f_{\ell-1} - 1}{f_{\ell-1}}$, where $\Gamma_{S', \text{MoRE}^*}^0 := 1$.

Proof. This proof utilizes the construction in proving Theorem 3.4.

First, we decompose Eq. 38 as (like before, we ignore router weight α for brevity):

$$\mathbf{x}_{\ell+1}^i = \sum_{n \in \mathcal{N}_\ell^i} (\mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} + \mathbf{W}_\ell \cdot \sigma(\mathbf{x}_\ell^n)) \quad (38)$$

$$= \underbrace{\left(\sum_{n \in \mathcal{N}_\ell^i} \mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} \right)}_{(a)} + \underbrace{\mathbf{W}_\ell \cdot \left(\sum_{n \in \mathcal{N}_\ell^i} \sigma(\mathbf{x}_\ell^n) \right)}_{(b)} \quad (39)$$

We consider how many distinct values (a) and (b) can take.

Term (b). Suppose an ℓ -layer S' MoRE $*$ can generate $\Gamma_{S', \text{MoRE}^*}^\ell$ distinct outputs, meaning that $\mathbf{x}_\ell^n$ can take $\Gamma_{S', \text{MoRE}^*}^\ell$ different values – This is as if we have a pool of $\Gamma_{S', \text{MoRE}^*}^\ell$ distinct elements.

The first question is, if we take $f_\ell = |\mathcal{N}_\ell^i|$ elements from this pool (where the same element can be taken multiple time, since different children n can have the same descendant sub-tree), how many unique multisets⁸ can we obtain. This is a classic “combination with replacement” problem, and the solution is $\binom{\Gamma_{S', \text{MoRE}^*}^\ell + f_\ell - 1}{f_\ell}$.

The second question is, can we encode each distinct multiset into distinct outputs via the form of $\sum \sigma(\cdot)$. Reusing Lemma C.8⁹, the answer is affirmative.

So term (b) can take $\Gamma_{S', \text{MoRE}^*}^\ell$ distinct values.

Term (a). Since the router takes top- f_ℓ experts, there are in total $\binom{s_\ell}{f_\ell}$ distinct $\mathcal{N}_\ell^i$. The key problem is if we perform the simple summation $\sum_{n \in \mathcal{N}_\ell^i}$ without the mapping σ , can we ensure distinct output for each distinct $\mathcal{N}_\ell^i$ (we cannot apply Lemma C.8 without σ)? i.e., for any $\mathcal{N}_\ell^i \neq \mathcal{N}_\ell^{i'}$, how can we ensure $\sum_{n \in \mathcal{N}_\ell^i} \mathbf{B}_\ell^n \cdot \mathbf{A}_\ell^n \cdot \mathbf{x} \neq \sum_{n' \in \mathcal{N}_\ell^{i'}} \mathbf{B}_\ell^{n'} \cdot \mathbf{A}_\ell^{n'} \cdot \mathbf{x}$. Setting a bias term encoding the expert index i , following Appendix C.2.3, does not work. A failure case is that $\mathcal{N}_\ell^i$ contains experts 1, 4 and $\mathcal{N}_\ell^{i'}$ contains experts 2, 3: $1 + 4 = 2 + 3$ even through $\{1, 4\} \neq \{2, 3\}$. Fortunately, there are existing encoding schemes that satisfies our requirement. For example, we can encode the s_ℓ experts into a “superincreasing sequence” where expert i is encoded into 2^i . In this case, it is guaranteed that $\sum_{n \in \mathcal{N}_\ell^i} 2^n \neq \sum_{n' \in \mathcal{N}_\ell^{i'}} 2^{n'}$ for any $\mathcal{N}_\ell^i \neq \mathcal{N}_\ell^{i'}$.

Combining (a) and (b). Finally, when we set $\mathbf{W}_\ell$ according to Eq. 34, we are guaranteed that any two different pairs of (a) and (b) will have different values of “(a) + (b)”. This means the total number of distinct $\mathbf{x}_{\ell+1}^i$ we can obtain from Eq. 38 equals:

$$\Gamma_{S', \text{MoRE}^*}^{\ell+1} = \binom{s_\ell}{f_\ell} \cdot \binom{\Gamma_{S', \text{MoRE}^*}^\ell + f_\ell - 1}{f_\ell} \quad (40)$$

Lastly, when $\ell = 1$, it is obvious that $\Gamma_{S', \text{MoRE}^*}^1$ should be $\binom{s_0}{f_0}$. If we define $\Gamma_{S', \text{MoRE}^*}^0 := 1$ and let $\ell = 0$, Eq. 40 becomes $\Gamma_{S', \text{MoRE}^*}^1 = \binom{s_0}{f_0} \cdot \binom{1+f_0-1}{f_0} = \binom{s_0}{f_0}$, which satisfies the initial condition.

This completes the proof.

⁸A multiset is a set where an element can appear multiple times.

⁹The original Lemma in Xu et al. [2019] is indeed derived on multisets.

□

C.2.5 Proof of Corollary 3.6

Corollary C.10. (Corollary 3.6) *The structural flexibility of S'MoRE[#] equals $\prod_{\ell=0}^{L-1} \binom{s}{f}^{F_{\ell+1}}$ where $F_L := 1$.*

Proof. This proof follows almost exactly as the the proof of Theorem 3.4 in Appendix C.2.3. The only difference is that now every layer has the same dimension d , rather than d being increased with larger layer index ℓ .

This just requires the following minor modification to the proof in Appendix C.2.3:

- When applying Lemma C.8, instead of constructing the mapping $\mathcal{X} \rightarrow \mathbb{R}^d$, we instead do the mapping $\mathcal{X} \rightarrow \mathbb{R}^{d'}$, with any $d' < d$ (Note that the Lemma does not have constraint on the output dimension d). So the output of $\sum \sigma(\cdot)$ is in a d' -dimensional subspace of $\mathbb{R}^d$.
- Updating Eq. 34, we set $\mathbf{W}_\ell$ to be a projection matrix with the first $d - d'$ rows being 0, and the rest d' rows being a projection from $\mathbb{R}^d$ to the $\mathbb{R}^{d'}$ that $\sum \sigma(\cdot)$ spans.

□

D Additional Experimental Results

D.1 Dataset details

We evaluate on a diverse set of benchmark consisting of 7 popular fine-tuning datasets. Specifically, ARC-c and ARC-e [Clark et al., 2018] evaluate logical reasoning and world knowledge through challenging multiple-choice questions. Commonsense QA [Talmor et al., 2018] assesses a model’s grasp of everyday knowledge and implicit relationships. OpenBook QA [Mihaylov et al., 2018] requires multi-step reasoning over scientific facts, while Winogrande [Sakaguchi et al., 2021] measures commonsense pronoun resolution. Accuracy is used as the evaluation metric for all above datasets. In addition, we evaluate the models on 2 more challenging datasets. GSM8K [Cobbe et al., 2021] contains 8.5k high-quality linguistically diverse grade school math word problems. Deriving the correct solution requires multi-step reasoning (2 to 8 steps) by the LLM model. CodeAlpaca [Chaudhary, 2023] contains 20k instruction-following data for fine-tuning LLM’s code generation capability. HumanEval [Chen et al., 2021] consists of 164 hand-written programming problems, to access the LLM’s capabilities in language comprehension, reasoning, algorithms, and simple mathematics. We train the LLM on CodeAlpaca and then evaluate the checkpoint on HumanEval. We measure the “Pass@1” metric, where we let the fine-tuned model to generate $k = 1$ solution for each problem, and evaluate whether it can pass the unit tests.

D.2 More details on experimental setup

For all models, we insert the adapters to the feed forward networks (FFN) of all transformer layers of the base models. Specifically, each FFN consists of an “up-projection” matrix, a “gate-projection” matrix and a “down-projection” matrix. We insert the adapter to each of the three matrices.

To ensure a fair comparison, we set an equal budget for trainable adapter parameters and compare different model architecture within this constraint. For LoRA [Hu et al., 2021], we vary the rank r in $\{2^k \mid 0 \leq k \leq 10\}$, and set the `lora_alpha` parameter as $2 \cdot r$ following standard practice. For MixLoRA [Li et al., 2024a], we adjust the number of experts within $\{4, 8\}$, keep the number of active experts within $\{1, 2, 4\}$ ¹⁰ (while ensuring that it does not exceed half of the total experts), and the expert dimension within $\{2^k \mid 0 \leq k \leq 6\}$. For HydraLoRA [Tian et al., 2024], we vary the number of heads in $\{4, 8\}$, and the rank r in $\{2^k \mid 0 \leq k \leq 8\}$. For S'MoRE, in most experiments (except the “scaling-up” study in Table 5), we limit S'MoRE to two layers due to resource constraints. We vary the number of experts (s_0, s_1) within $\{(2, 2), (4, 4)\}$: the fanout (f_0, f_1) is $(1, 1)$ when

¹⁰“Number of active experts” is only set for the sparse gates (“noisy top- k ” and “switch”). For dense gates, the number of active experts equals total number of experts.

Table 7: Wall-clock time (second) comparison

Method	ARC-c	ARC-e	CSQA	OBQA	Winogrande	Average
MixLoRA	426	794	3343	3539	3007	2222
S'MoRE	489 (1.15 $\times$)	957 (1.21 $\times$)	4289 (1.28 $\times$)	4406 (1.24 $\times$)	4014 (1.33 $\times$)	2831 (1.24 $\times$)

$(s_0, s_1) = (2, 2)$ and is $(2, 2)$ when $(s_0, s_1) = (4, 4)$ ¹¹. We vary the expert dimension (r_0, r_1) within $\{(2^k, 2^k) \mid 0 \leq k \leq 6\} \cup \{(2^k, 2^{k+1}) \mid 0 \leq k \leq 5\} \cup \{(2^k, 2^{k+2}) \mid 0 \leq k \leq 4\}$. All baselines and S'MoRE are trained with 2 epochs, with learning rate $1e-4$. The learning rate follows a cosine schedule.

Software & hardware. We implement S'MoRE by adding a customized adapter to the Hugging Face PEFT library [Mangrulkar et al., 2022]. All models are trained via the LLaMA-Factory [Zheng et al., 2024] SFT pipeline, ensuring a consistent execution environment. Similarly, all the evaluations are conducted through OpenCompass [Contributors, 2023b], which is a unified evaluation framework providing a standard API for all considered benchmarks. For the computation hardware, all experiments are run on a single node with 4 NVIDIA A100 80GB GPUs.

D.3 Wall-clock time & potential system optimizations

While §3.4 ensures that S'MoRE theoretically incurs negligible computation overhead, it is true that without system-level optimization, the multi-layer structure may increase the wall-clock time. Yet, such overhead is small.

Measurement. Table 7 shows the wall-clock time to finish training of MixLoRA and S'MoRE, measured on the same machine (with 4 NVIDIA A100 GPUs) and same software environment (based on LLaMA-Factory). The backbone model is LLaMA 3-8B. Trainable parameters of MixLoRA (8 rank-64 experts) and S'MoRE (2 layers, each with 4 rank-64 experts) are comparable.

On average, S'MoRE incurs 24% wall-clock time overhead, which is relatively small. The above measurement is based on S'MoRE under native PyTorch implementation, without any system optimization. It is reasonable to expect that the wall-clock time overhead can be further reduced by applying standard techniques, such as

- **CUDA kernel fusion**, which combines the operation of multiple S'MoRE layers into a single CUDA kernel. This can effectively reduce the “kernel launch” overhead associated with deeper S'MoRE (in native PyTorch, each layer may require its own “kernel launch”).
- **Token-level parallelism**, which interleaves the processing of different layers across different tokens. This is achievable by custom Triton kernels or `torch.compile(...)` optimization. Such parallelism addresses the load-balance between the router and expert layers (since the router is more lightweight than the expert propagation), which improves GPU utilization. Such parallelism can also break the dependency between the top-down routing and bottom-up propagation, as these two stages can be interleaved across tokens.

D.4 Routing cost

In Fig. 5, we visualize the router computation cost (Eq. 6) relative to that of the experts' layer propagation (Eq. 3), corresponding to the best-performing models in Table 2. The x -axis denotes the different S'MoRE structures (in Table 2, we do not include the results corresponding to the “4-2” and “8-8” S'MoRE architectures, due to space limit). The costs are measured by the total number of arithmetic operations performed by the routers versus by the experts. In general, when the residual rank r_ℓ is lower, the cost of routing becomes *relatively* higher (since the router operation is independent of the ranks). However, in all cases, the routing cost is insignificant compared to the cost of expert propagation (at most 26%). This is consistent with our theoretical complexity analysis in §3.4.

¹¹Same as above, the fanouts are only set for sparse gates. For dense gates, the fanout of layer ℓ equals the total number of experts in layer ℓ

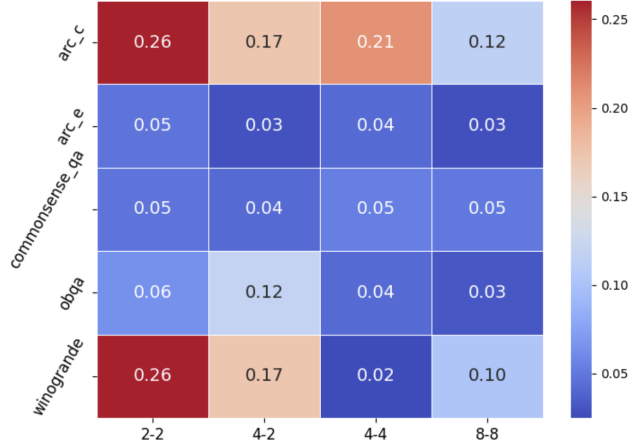


Figure 5: Cost of router (Eq. 6) relative to expert propagation (Eq. 3), measured by their number of arithmetic operations. Here we consider S’MoRE with noisy top- k gate on LLaMA 3.2-1B

E Limitations and Broader Impact

Limitations. This paper focuses on a novel model architecture design for parameter-efficient MoE, whose computation graph differs from those of standard LoRA and single-layer MoE. We do not focus on the corresponding system-level optimization, and thus our implementation of the S’MoRE layers is written in native PyTorch. To optimally utilize the GPU resources and further accelerate the model execution on commercial hardware, dedicated CUDA kernels may be developed and different levels of execution parallelism (e.g., data-, model- and pipeline-parallelism) may be explored. In addition, we may integrate S’MoRE into state-of-the-art LLM acceleration frameworks such as vLLM [Kwon et al., 2023] or LMDeploy [Contributors, 2023a] to boost the practical execution speed. We treat such system-level optimization as meaningful future work.

Broader impact. This work focuses on developing a new PEFT model for the general LLM fine-tuning tasks. It does not have any direct negative societal impact. In the future, S’MoRE may be extended to other tasks or models, to broaden its impact on the enhanced model capacity. For example, we may apply the hierarchical residual design to foundation models under pre-training. The dramatically improved “structural flexibility” under the same parameter and computation budget has the potential to break the ceiling of the current scaling law for both the dense and MoE LLMs.