# **S'MoRE**: Structural Mixture of Residual Experts for Parameter-Efficient LLM Fine-Tuning

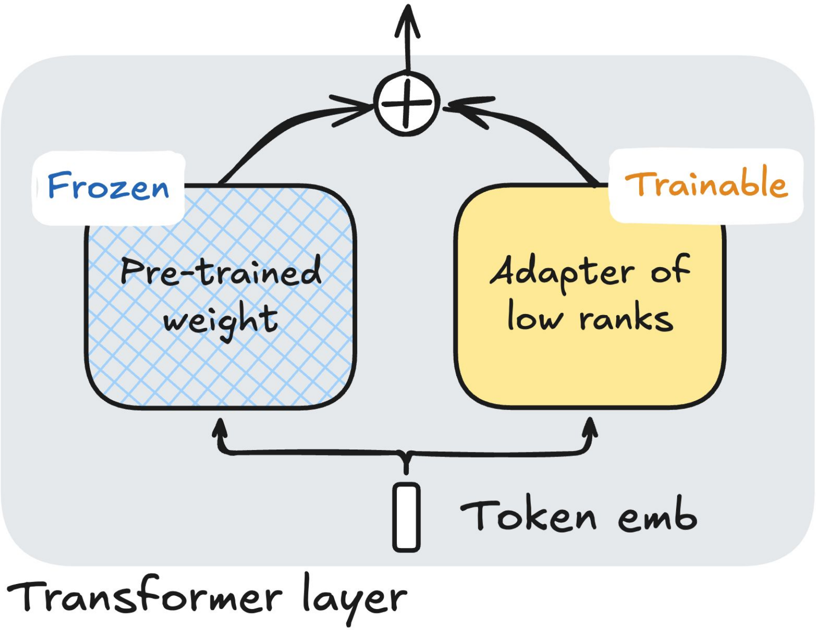
NeurIPS 2025

https://github.com/ZimpleX/SMoRE-LLM

# Problem Setup

Parameter-efficient fine-tuning (PEFT) on pre-trained LLM

- Adapt to downstream tasks
- Freeze pre-trained weight; Update low-rank adapter parameters



# Evolution of PEFT Adapters

## S'MoRE

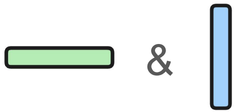
- Integrating & extending designs of both LoRA & MoE
- Exploiting **structural** relationship among residual experts
- Boosting MoE's model capacity while maintaining LoRA's parameter efficiency

| | **Technique** | | | **Benefit** | |
|---|---|---|---|---|---|
| | **Low-rank approximation** | **Conditional computation** | **Structural mixture** | **Parameter efficiency** | **Model capacity** |
| LoRA | ☑ | ☐ | ☐ | + | |
| MoE | ☐ | ☑ | ☐ | | + |
| Mixture of LoRA | ☑ | ☑ | ☐ | + | + |
| S'MoRE | ☑ | ☑ | ☑ | + | ++ |

# Evolution of PEFT Adapters

**Similar efficiency**:

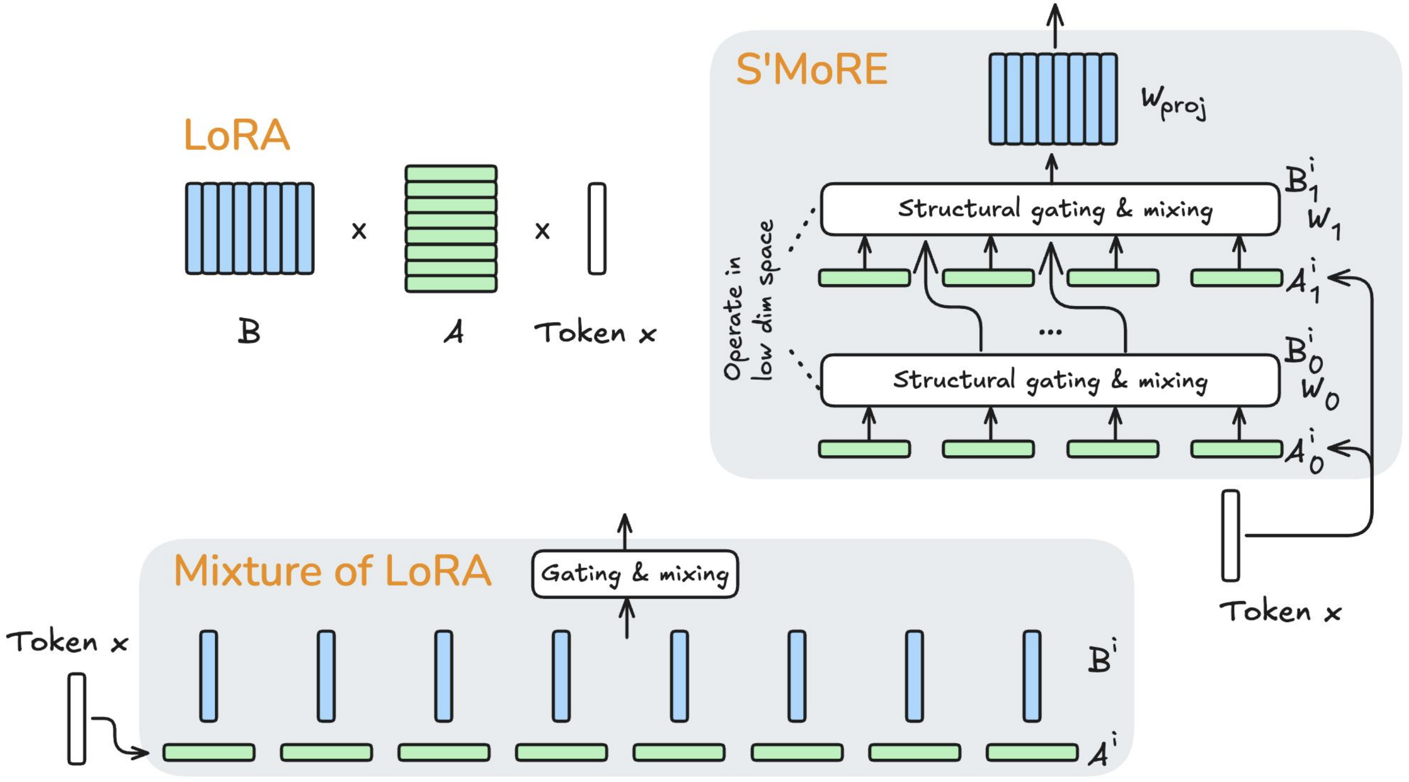
Same parameters




**Higher capacity**:

LoRA < Mixture of LoRA < S'MoRE

(measure via *structural flexibility*)

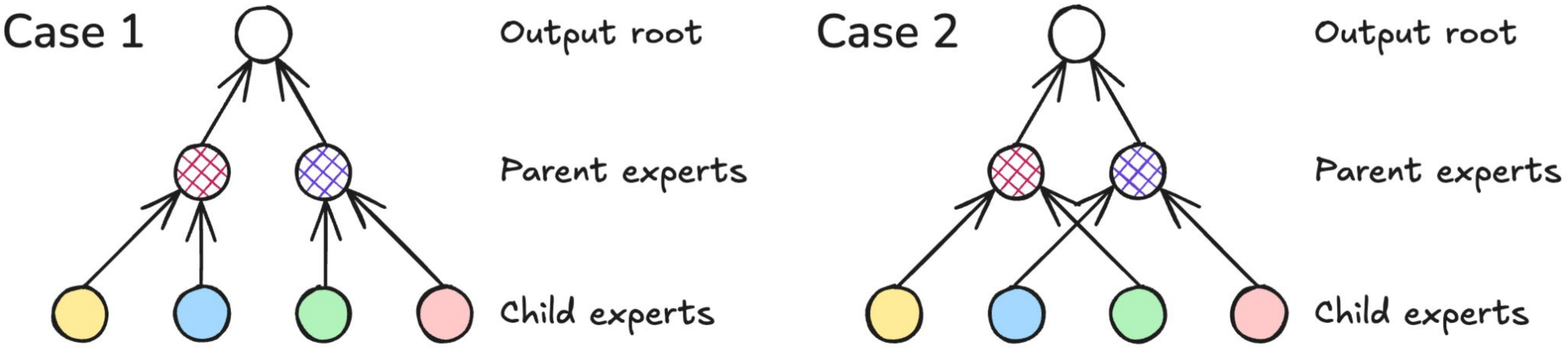


# How Does Structure Help?

MoE routing problem

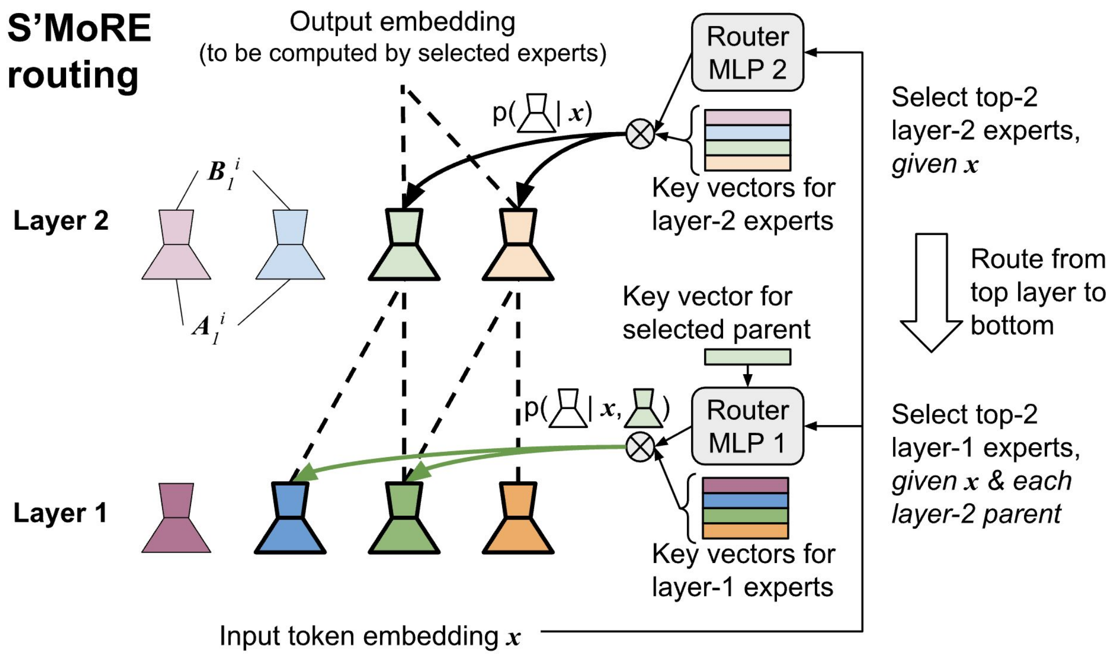
- What experts to activate? ⇒ Existing works
- How to connect activated experts? ⇒ S'MoRE & structural scaling

The **same** set of experts can form **exponentially many** different structures!



# S’MoRE Routing

- Hierarchical routing (top-down)
- Router computes conditional probability by
  - active ancestors
  - input token
- “Token-expert” similarity based on key-query dot product
- *Query* embedding: by Router’s compact MLP
- *Key* embedding: learnable for each residual expert



# S’MoRE Layer Propagation

Current layer output

Token emb (skip connect)

Previous layer output

$$\boldsymbol{x}^i_{\ell+1} = \sum_{n\in\mathcal{N}^i_\ell} \alpha^{i,n}_\ell \cdot \sigma\left(\boxed{\boldsymbol{B}^n_\ell \cdot \boldsymbol{A}^n_\ell} \cdot \boldsymbol{x} + \boxed{\boldsymbol{W}_\ell} \cdot \boldsymbol{x}^n_\ell\right)$$

Active children experts

Router weights

Residual (low rank) experts

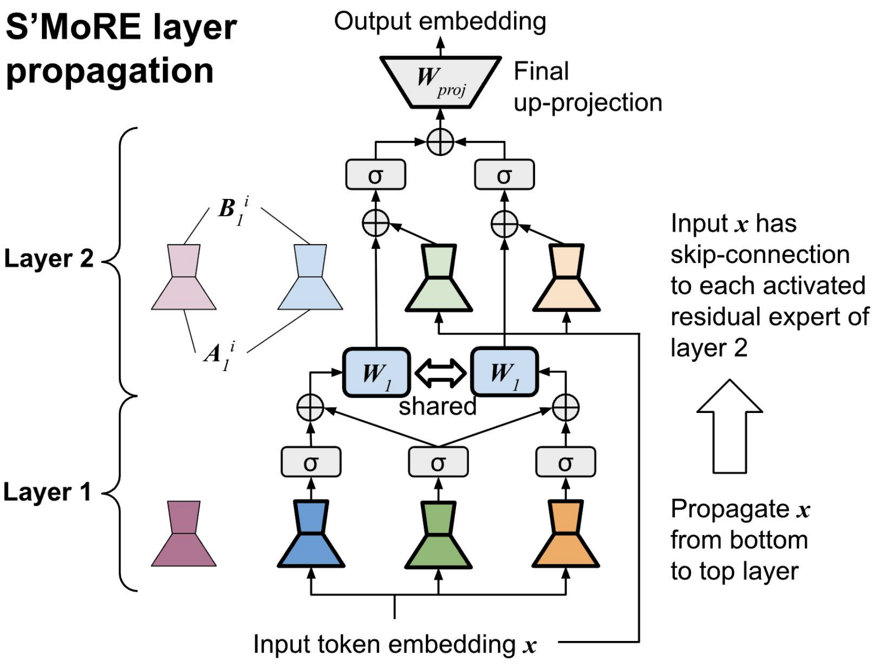
Shared proj. weights

Neighbor aggregation

Transformation

A special GNN

- Selected experts form a residual tree
- Token emb propagates from leaves to root
- Each layer: aggregation + transformation ⇒ **GNN**
- Craft each layer’s output dim for efficiency
- σ & W theoretically ensures expressive power



# Summary of Theoretical Properties

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

| $r_\ell$ | $L$ | $d_L$ | $2 \cdot d \cdot d_L$ | $\Delta$ | 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% |

## Parameter & computation efficiency

- Similar to vanilla LoRA

## Recovering 1-layer MoE

**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.*

## Expressive power w.r.t. "**structural flexibility**" $\Gamma \Rightarrow$ Graph isomorphism test

- Given token, $\Gamma$ = num distinct outputs that different expert structures can generate

(1-layer MoE)

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

**sum** over fanout

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

**exponent** over fanout

# Experiments

## Setup

- 7 benchmarks
- 2 model families & 3 model scales
- 3 gating types
- 4 or 8 total number of experts

## Main observations

- Accuracy boost due to **structural** mixture
- Comparable parameter size due to **residual** aggregation in low-dim space

| Model | Gate | Method | ARC-c Acc. | ARC-c Param. | ARC-e Acc. | ARC-e Param. | CSQA Acc. | CSQA Param. | OBQA Acc. | OBQA Param. | Winogrande Acc. | Winogrande Param. | **Avg Acc.** | **Avg Param.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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 |
| LLaMA 3.2 1B | 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 |
| LLaMA 3.2 1B | Dense | 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 |
| LLaMA 3.2 1B | Dense | 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 |
| LLaMA 3.2 1B | Dense | 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 |
| LLaMA 3.2 1B | Dense | 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 |
| LLaMA 3.2 1B | Dense | 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 |
| LLaMA 3.2 1B | 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 |
| LLaMA 3.2 1B | Noisy top-$k$ | 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 |
| LLaMA 3.2 1B | Noisy top-$k$ | 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 |
| LLaMA 3.2 1B | Noisy top-$k$ | 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 |
| LLaMA 3.2 1B | 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 |
| LLaMA 3.2 1B | Switch | 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 |
| LLaMA 3.2 1B | Switch | 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 |
| LLaMA 3.2 1B | Switch | 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 |
| | | 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 |
| LLaMA 3 8B | 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 |
| LLaMA 3 8B | Dense | 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 |
| LLaMA 3 8B | Dense | 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 |
| LLaMA 3 8B | Dense | 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 |
| LLaMA 3 8B | Dense | 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 |
| LLaMA 3 8B | Dense | 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 |
| LLaMA 3 8B | 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 |
| LLaMA 3 8B | Noisy top-$k$ | 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 |
| LLaMA 3 8B | Noisy top-$k$ | 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 |
| LLaMA 3 8B | Noisy top-$k$ | 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 |
| LLaMA 3 8B | 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 |
| LLaMA 3 8B | Switch | 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 |
| LLaMA 3 8B | Switch | 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 |
| LLaMA 3 8B | Switch | 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 |

# Experiments

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

| Method | ARC-e Accuracy | ARC-e Param. (B) | CSQA Accuracy | CSQA Param. (B) | Winogrande Accuracy | Winogrande Param. (B) | HumanEval Pass@1 | HumanEval Param. (B) | **Avg 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 3: LLaMA 3-8B: model Accuracy / Pass@1, and the best-performing models' trainable parameters (B).

| Gate | Method | GSM8K Accuracy | GSM8K Param. (B) | HumanEval Pass@1 | HumanEval 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 |
| Dense | HydraLoRA (8) | 62.24 | 0.297 | **44.51** | 0.079 |
| Dense | MixLoRA (4) | 61.11 | 0.132 | 39.02 | 0.026 |
| Dense | MixLoRA (8) | 59.36 | 0.132 | 40.85 | 0.033 |
| Dense | S'MoRE (2-2) | 62.40 | 0.104 | 42.07 | 0.090 |
| Dense | S'MoRE (4-4) | **65.20** | 0.957 | 43.90 | 0.104 |
| Switch | MixLoRA (4) | 59.67 | 0.047 | 42.68 | 0.075 |
| Switch | MixLoRA (8) | 61.56 | 0.247 | 39.63 | 0.247 |
| Switch | S'MoRE (2-2) | 62.47 | 0.133 | **45.73** | 0.190 |
| Switch | S'MoRE (4-4) | **63.91** | 0.957 | 42.07 | 0.090 |

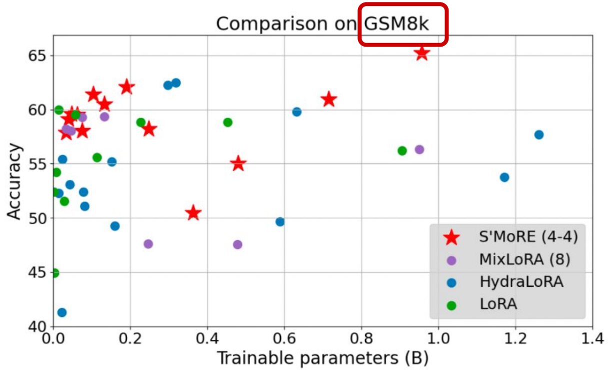



Figure 4: Change of accuracy w.r.t. trainable parameters, corresponding to models in Table 3.

- Consistent (or larger) gains across model families
- Structure improves scaling on math & coding

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 Accuracy | ARC-c Param. (B) | ARC-e Accuracy | ARC-e Param. (B) | Commonsense QA Accuracy | Commonsense QA Param. (B) | OpenBook QA Accuracy | OpenBook QA Param. (B) | Winogrande Accuracy | Winogrande 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 |

Increasing layers may further improve accuracy – with even **fewer** parameters

# Future Directions

Model scaling w.r.t. **STRUCTURE**!