ReMoE: Boosting Expert Reuse through Router Fine-Tuning in Memory-Constrained MoE LLM Inference

Xiongwei Zhu[1], Xiaojian Liao[1#], Tianyang Jiang[2], Yusen Zhang[1], Liang Wang[1], Limin Xiao[1]

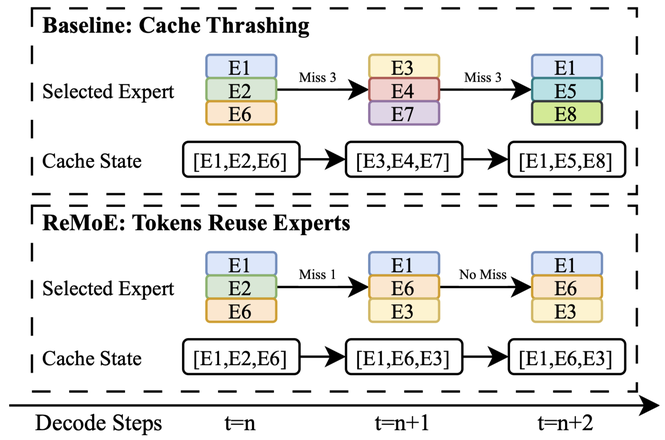
[1]Beihang University [2]Huawei Technologies Co., Ltd. [#]Corresponding Author

ICML International Conference On Machine Learning

Motivation

- Edge-side LLM deployment is emerging driven by on-device applications.
- MoE models sparsely activate only a subset of experts per token, reducing activated computation while maintaining high model capacity.
- Advances in mobile storage make local weight storage more feasible.

Challenge



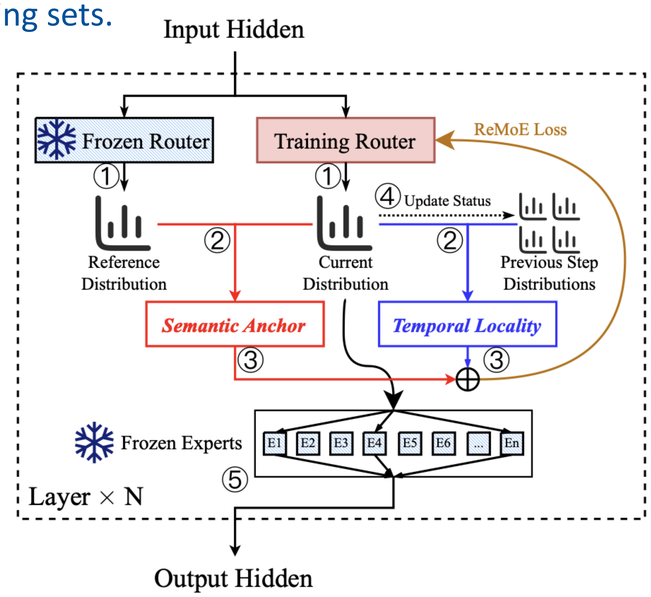
Memory-constrained MoE inference faces frequent expert switching.

- Only a small set of experts can stay in fast memory. Missed experts must be fetched from slow storage, leading to frequent evictions and I/O bottlenecks.
- Standard MoE routers (with load-balancing losses) tend to select disjoint experts across adjacent tokens, producing poor temporal locality and cache thrashing, which hurts latency in memory-constrained, single-request decoding on edge devices.

Method & Framework

Key Idea: Fine-tune only the router to increase expert reuse while limiting semantic shift.

- ***Semantic Anchor*** keeps routing close to the original pretrained router.
- ***Temporal Locality*** encourages reuse and stable local working sets.



Semantic Anchor:

$$L_{\text{Trust}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N_r} P_t^{(k)} \log \frac{P_t^{(k)}}{P_t^{\text{ref},(k)}}$$

Locality Loss:

$$L_{\text{Loc}} = \lambda_{\text{Reuse}} L_{\text{Reuse}} + \lambda_{\text{Smooth}} L_{\text{Smooth}} + \lambda_{\text{Lag}} L_{\text{Lag}} + \lambda_{\text{WS}} L_{\text{WS}}$$

① **Reuse Loss**

$$L_{\text{Reuse}} = -\log\left(\frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{K}\sum_{k\in\tilde{E}_{t-1}} P_t^{(k)}\right)$$

② **Smoothness Loss**

$$L_{\text{Smooth}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}[D_{\text{KL}}(P_t\|P_{t-1}) + D_{\text{KL}}(P_{t-1}\|P_t)]$$

③ **Lagged Inertia Loss**

$$L_{\text{Lag}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{|\mathcal{D}|}\sum_{\substack{d\in\mathcal{D}\\ t-d\geq 1}}\frac{1}{2}[D_{\text{KL}}(P_t\|P_{t-d}) + D_{\text{KL}}(P_{t-d}\|P_t)]$$

④ **Working-set compression loss**

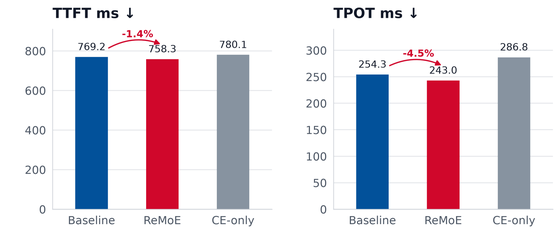
$$L_{\text{WS}} = -\frac{1}{n}\sum_{b=1}^{n}\sum_{k=1}^{N_r}\bar{P}_b^{(k)}\log\bar{P}_b^{(k)}, \quad \bar{P}_b = \frac{1}{W}\sum_{j=1}^{W} P_{(b-1)W+j}$$

Experiments

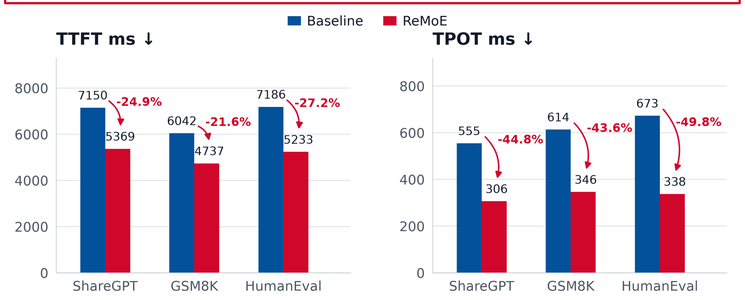
*CE-Only: router-only fine-tuning with only next-token cross-entropy, without trust or locality regularization.

Benchmark / Metric	Baseline	CE-only*	ReMoE
GSM8K (EM, strict)	38.89	36.92	38.13
GSM8K (EM, flex)	39.04	37.23	38.36
HumanEval (pass@1)	26.83	28.05	29.27
MMLU (acc)	57.72	57.44	57.81
IFEval (prompt loose)	17.93	–	17.93
IFEval (prompt strict)	17.19	–	16.08

vLLM GPU-CPU Expert Offloading



llama.cpp on Jetson Orin NX (SSD Offloading)



Contact & Cooperation

liaoxj@buaa.edu.cn