EuroSys '26, April 27–30, 2026, Edinburgh, Scotland UK

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Taming Latency-Memory Trade-Off in MoE-Based LLM Serving via *Fine-Grained* Expert Offloading

Hanfei Yu[1], Xingqi Cui[2], Hong Zhang[3], Hao Wang[4], Hao Wang[1]

Stevens Institute of Technology[1], Rice University[2], University of Waterloo[3], Rutgers University[4]

IntelliSys Lab

Extending Scaling Law: Mixture-of-Experts (MoE)

Meta
Open Pre-Trained Transformers (OPT) Library

175B

LLAMA 3

70B

deepseek

1.6T

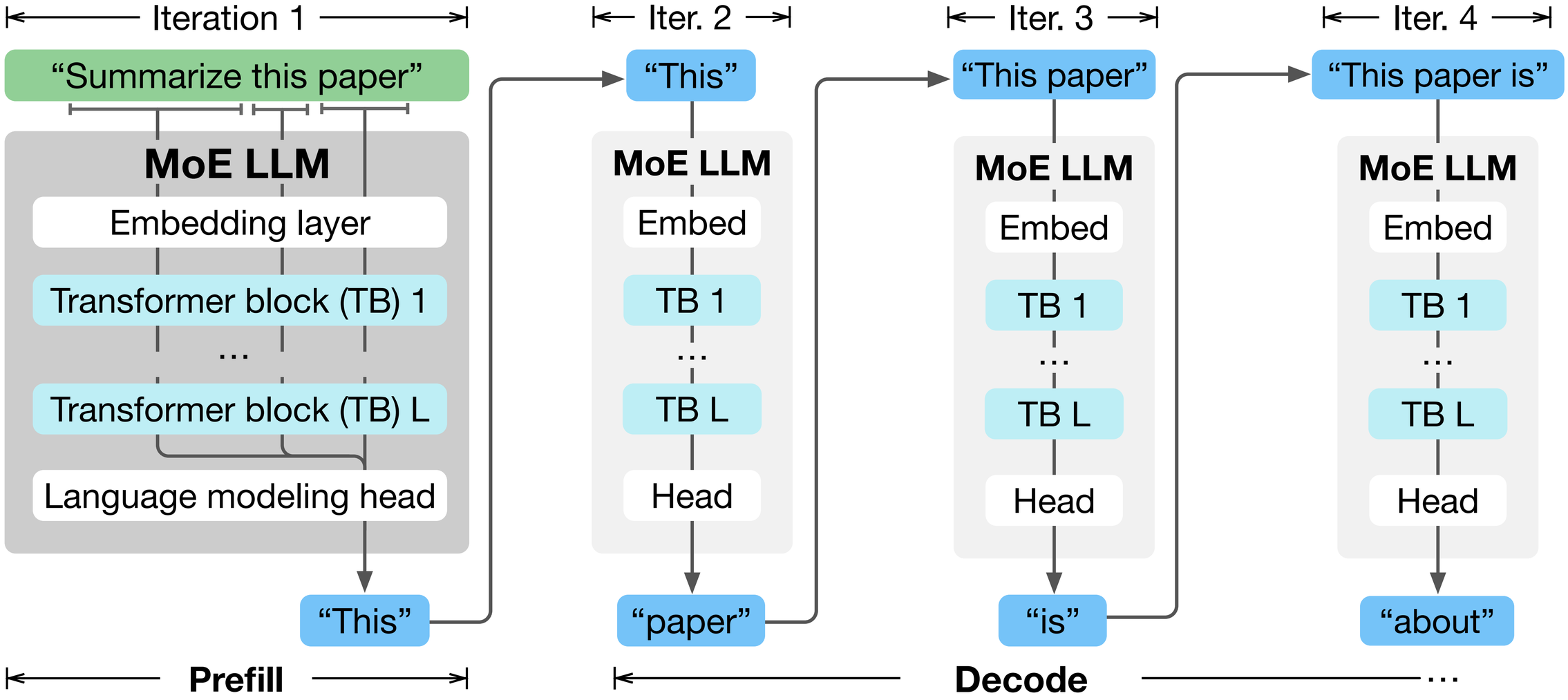
Llama 4

~2T

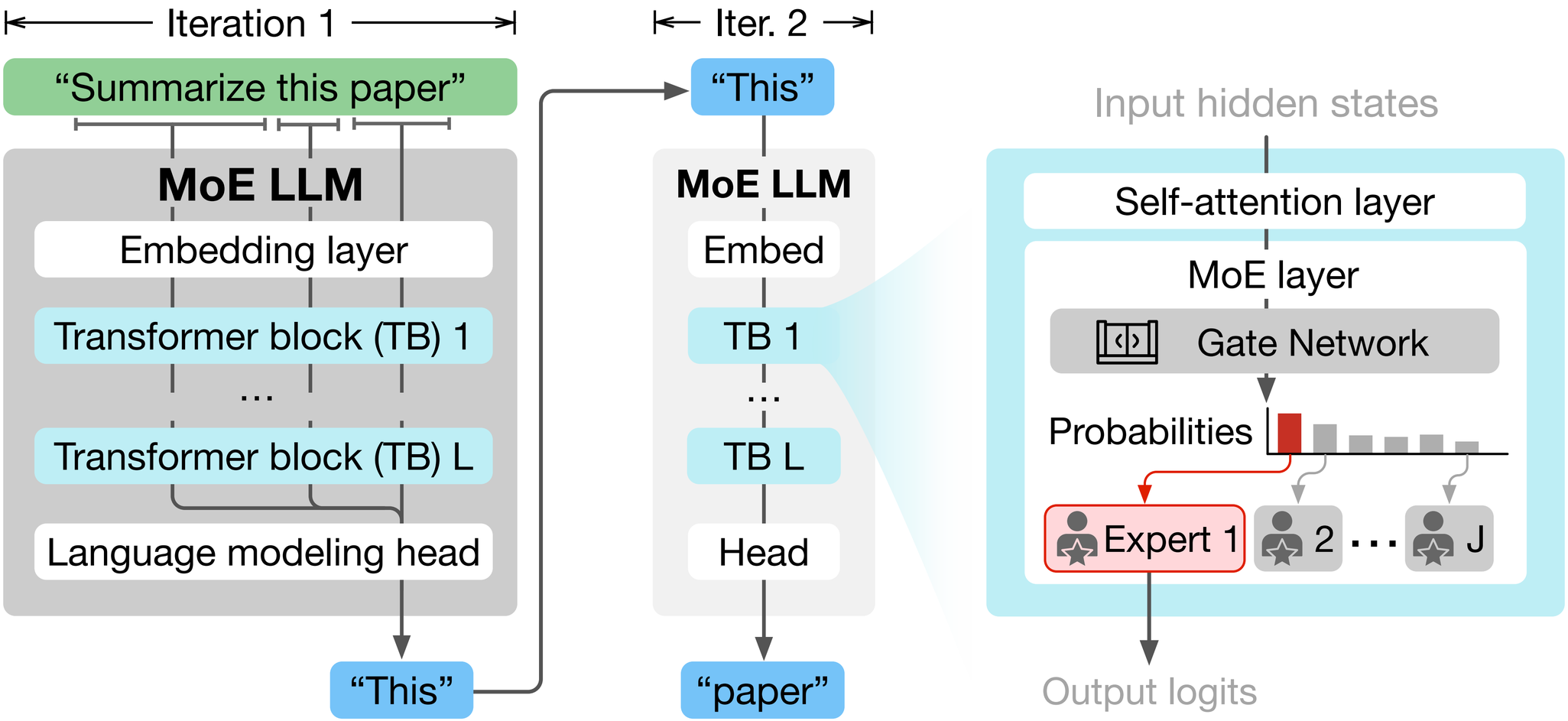
Grok 3

~1T

Large-Language Models (LLMs)

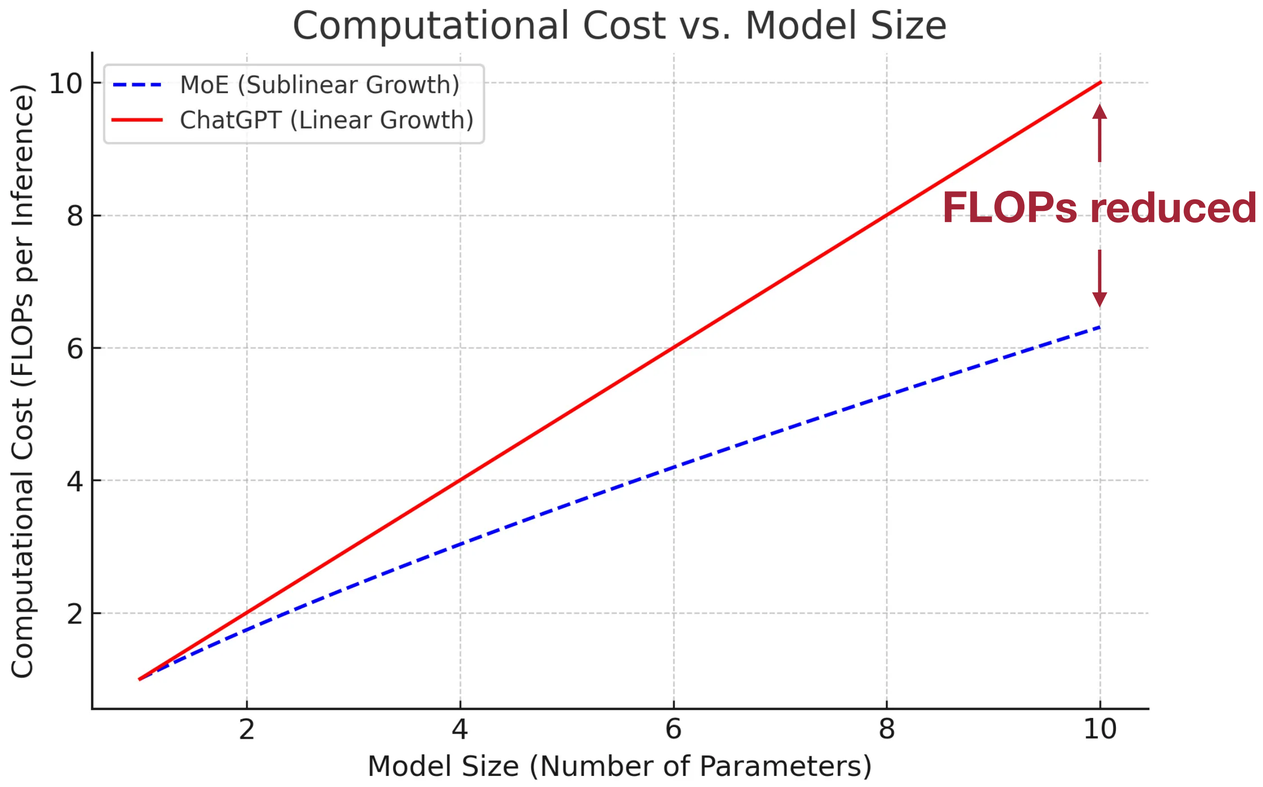


Mixture-of-Experts (MoE) Based LLMs



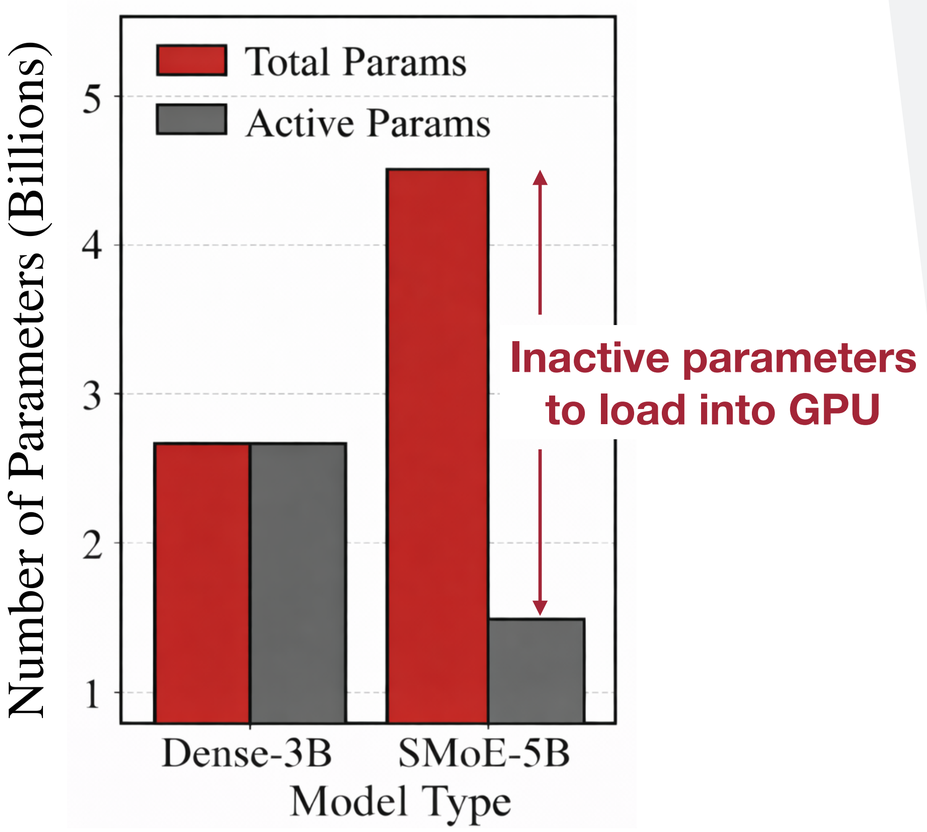
Pros & Cons of MoE-based LLMs

Compute Efficiency



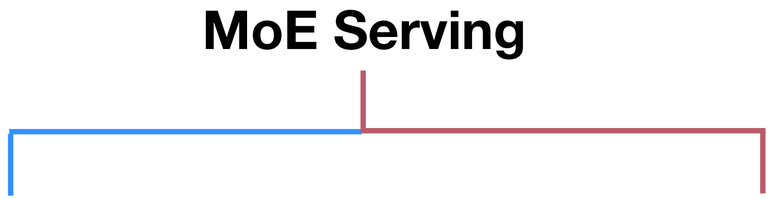
https://medium.com/chivian-technology/mixture-of-experts-moe-vs-chatgpts-algorithm-a-technical-comparison-a692d533bbc4

Serving Memory Inefficiency

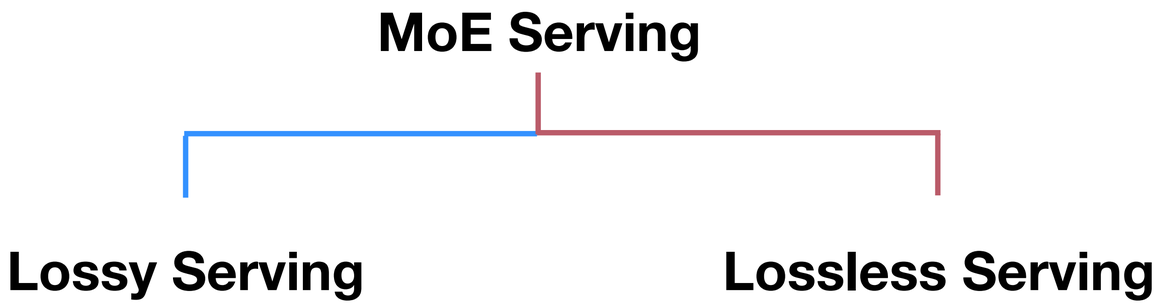


Pan, Bowen, et al. "Dense training, sparse inference: Rethinking training of mixture-of-experts language models." arXiv preprint arXiv:2404.05567 (2024).

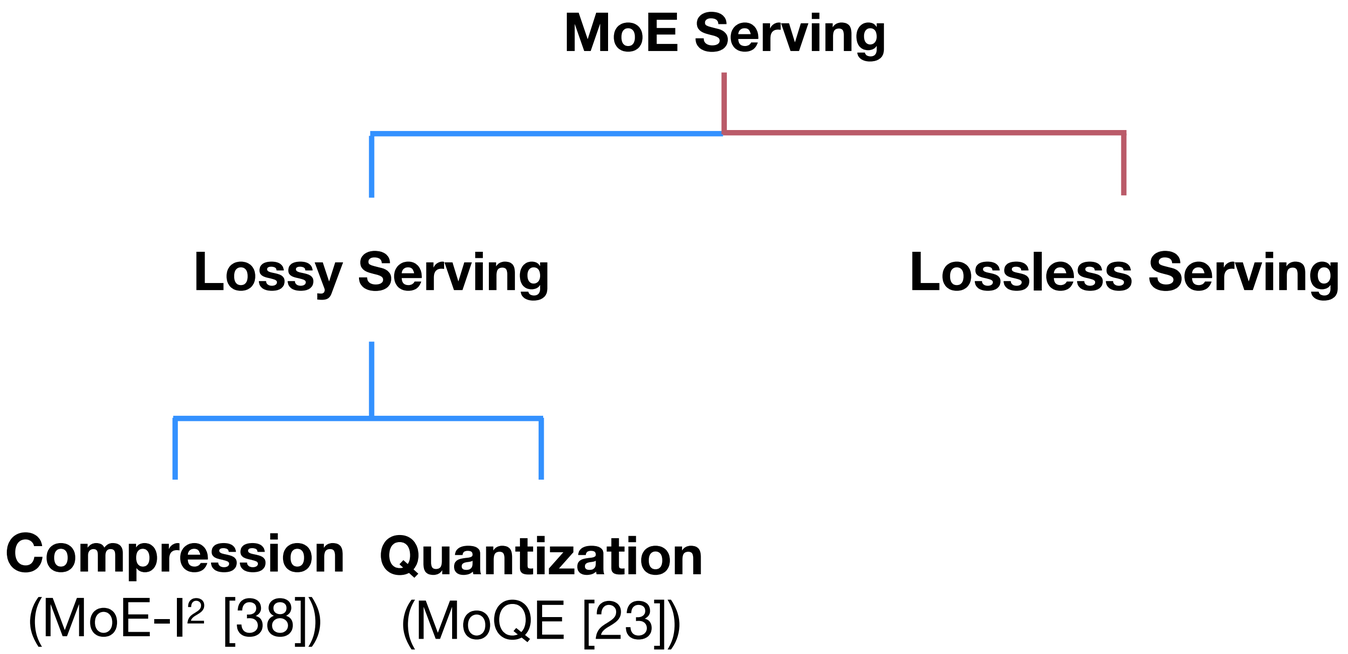
Existing Works on Efficient MoE Serving



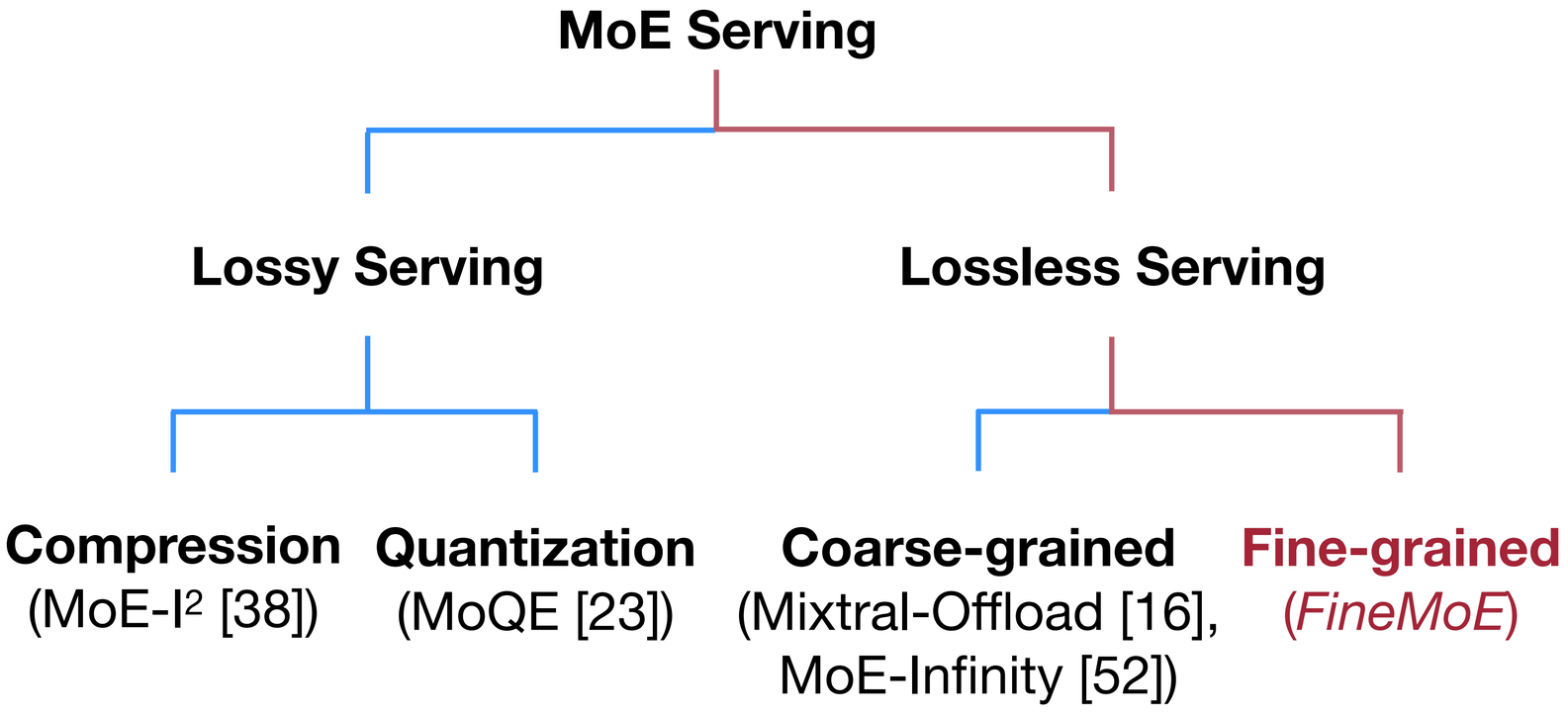
Existing Works on Efficient MoE Serving

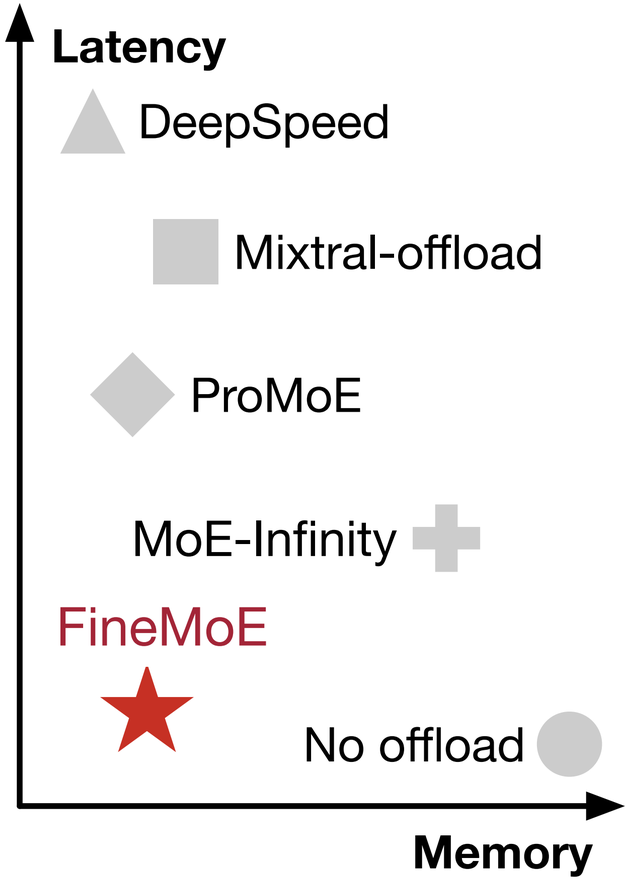


Existing Works on Efficient MoE Serving

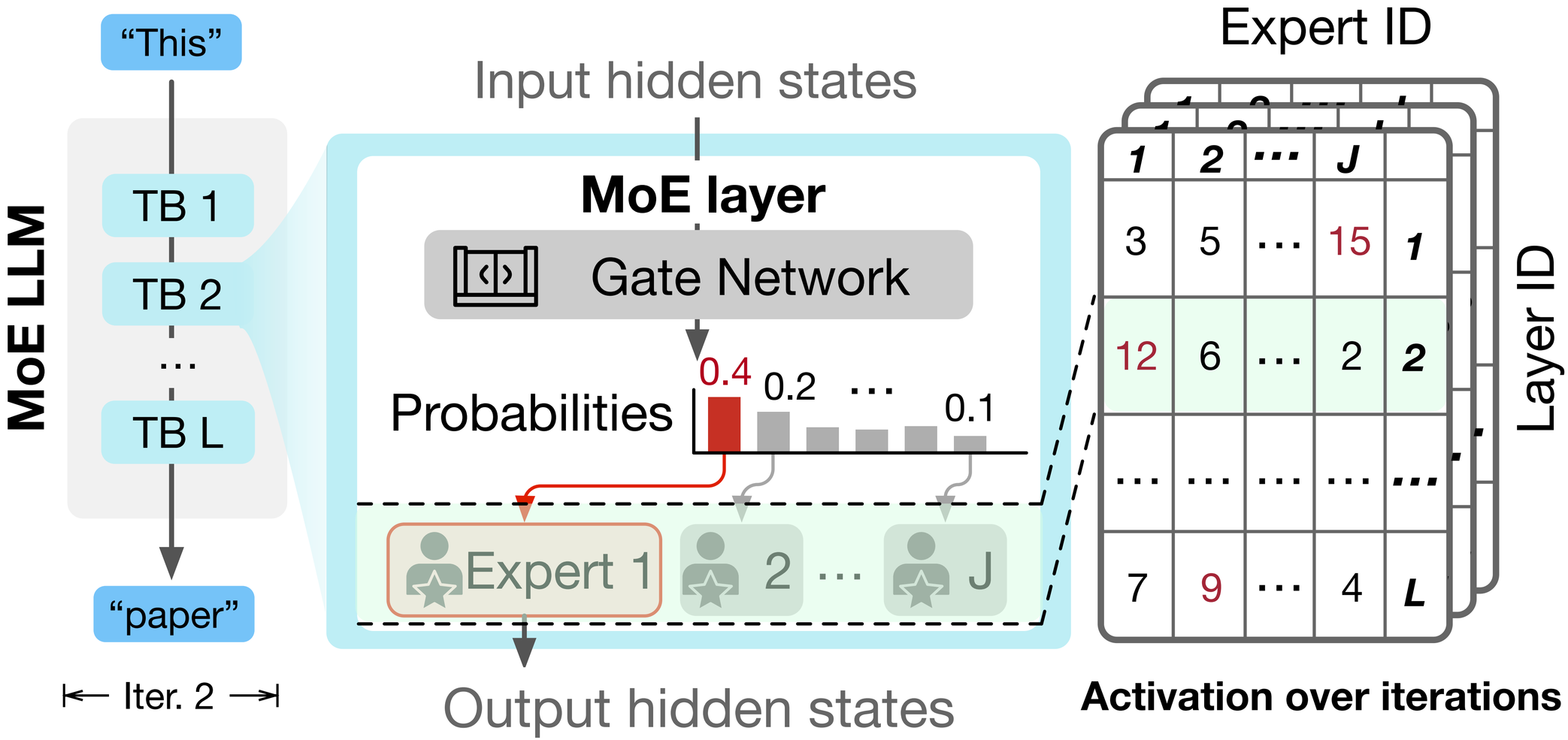


Existing Works on Efficient MoE Serving



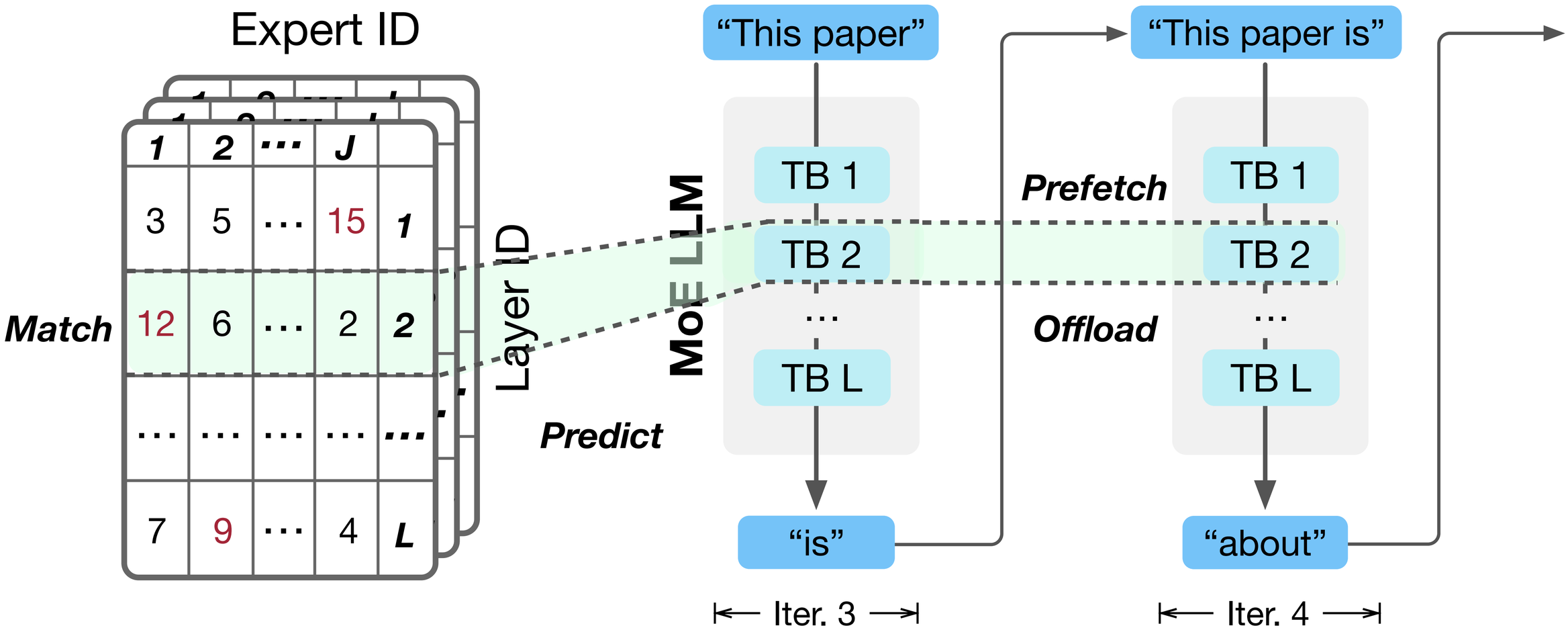


Expert Tracing and Offloading

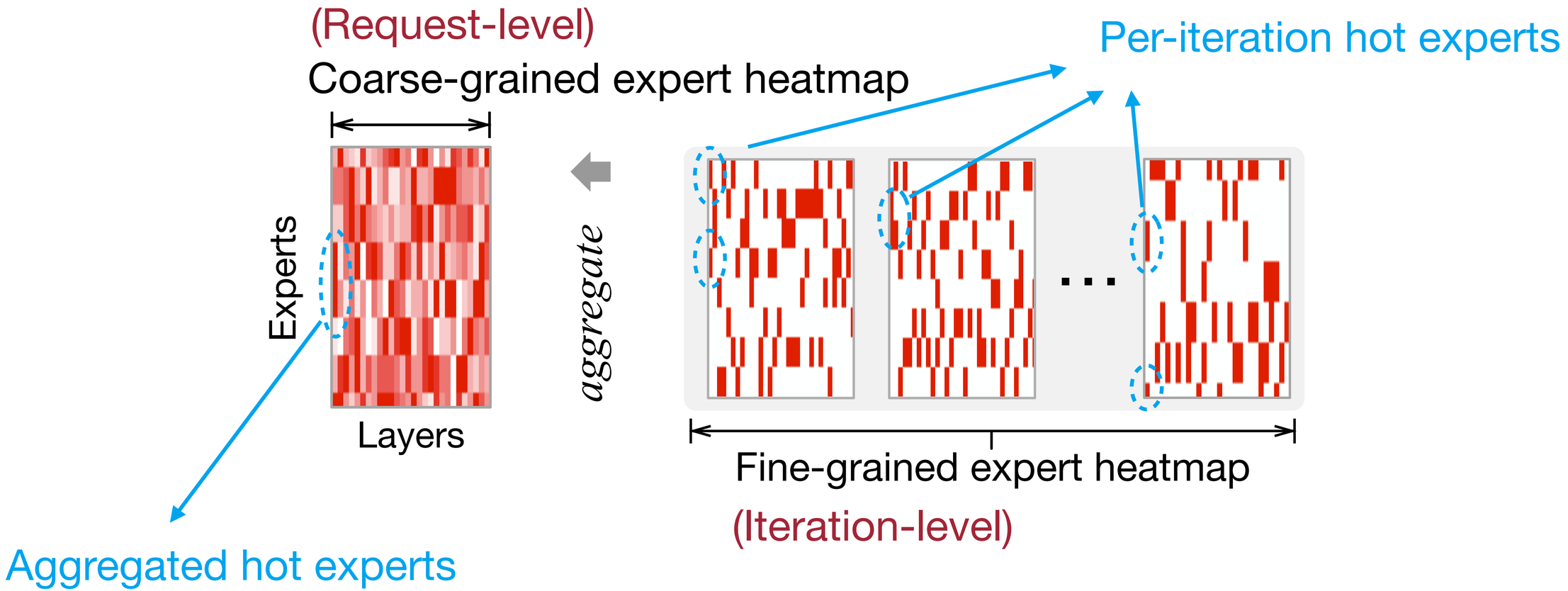


Leyang Xue; et al. MoE-Infinity: Efficient MoE Inference on Personal Machines with Sparsity-Aware Expert Cache. arXiv:2401.14361v3

Pattern-Based Expert Predictions



Key Problem: Aggregated Hot is Misleading



Serving Mixtral-8x7B

Problems of Coarse-Grained Offloading

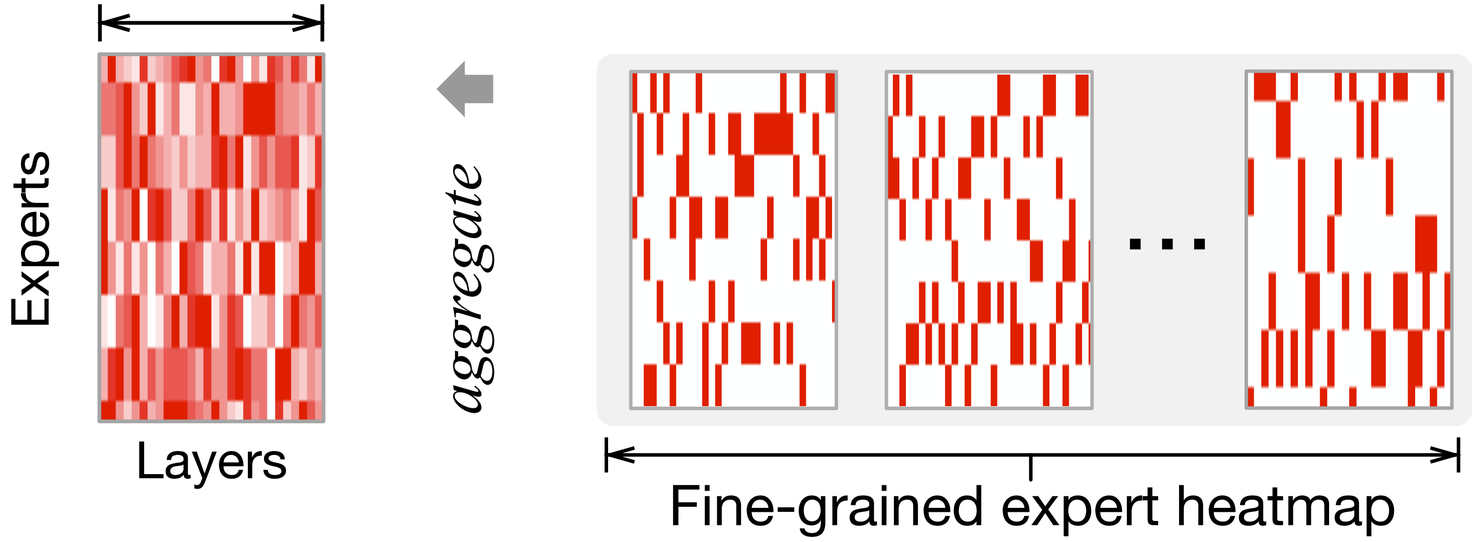
Coarse-grained expert pattern tracking

Low expert hit rates

Insufficient latency-memory trade-off

(Request-level)

Coarse-grained expert heatmap **diminish** predictability!

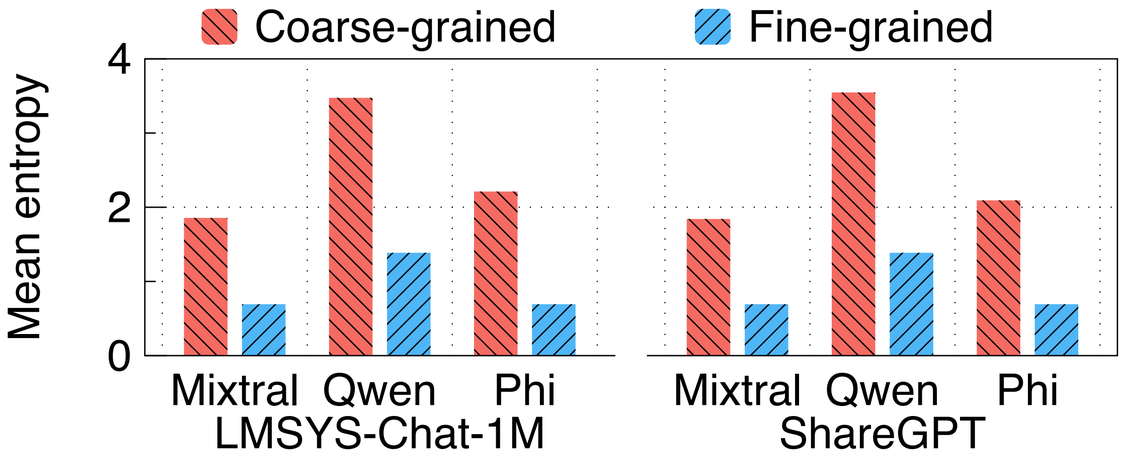


(Iteration-level)

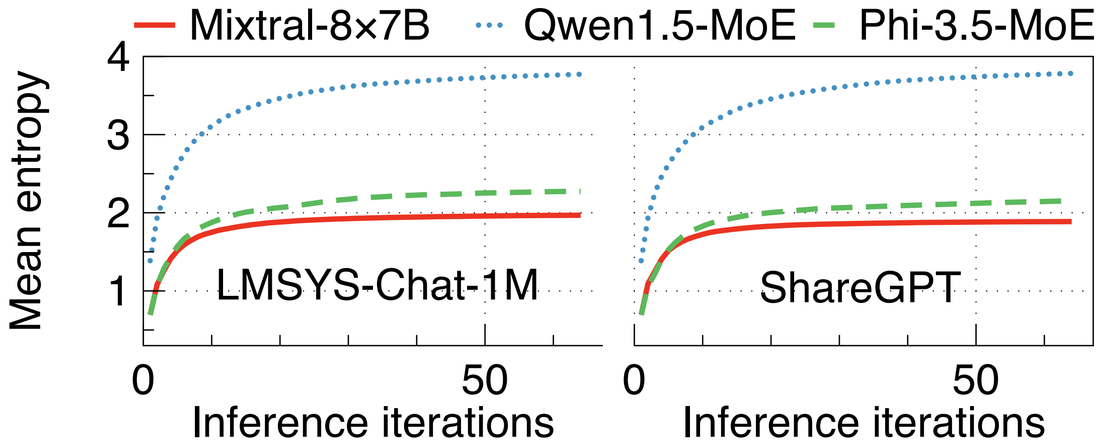
Serving Mixtral-8x7B

Problems of Coarse-Grained Offloading

Coarse-grained (request-level) expert patterns **diminish** predictability!



(a) Mean entropy per layer

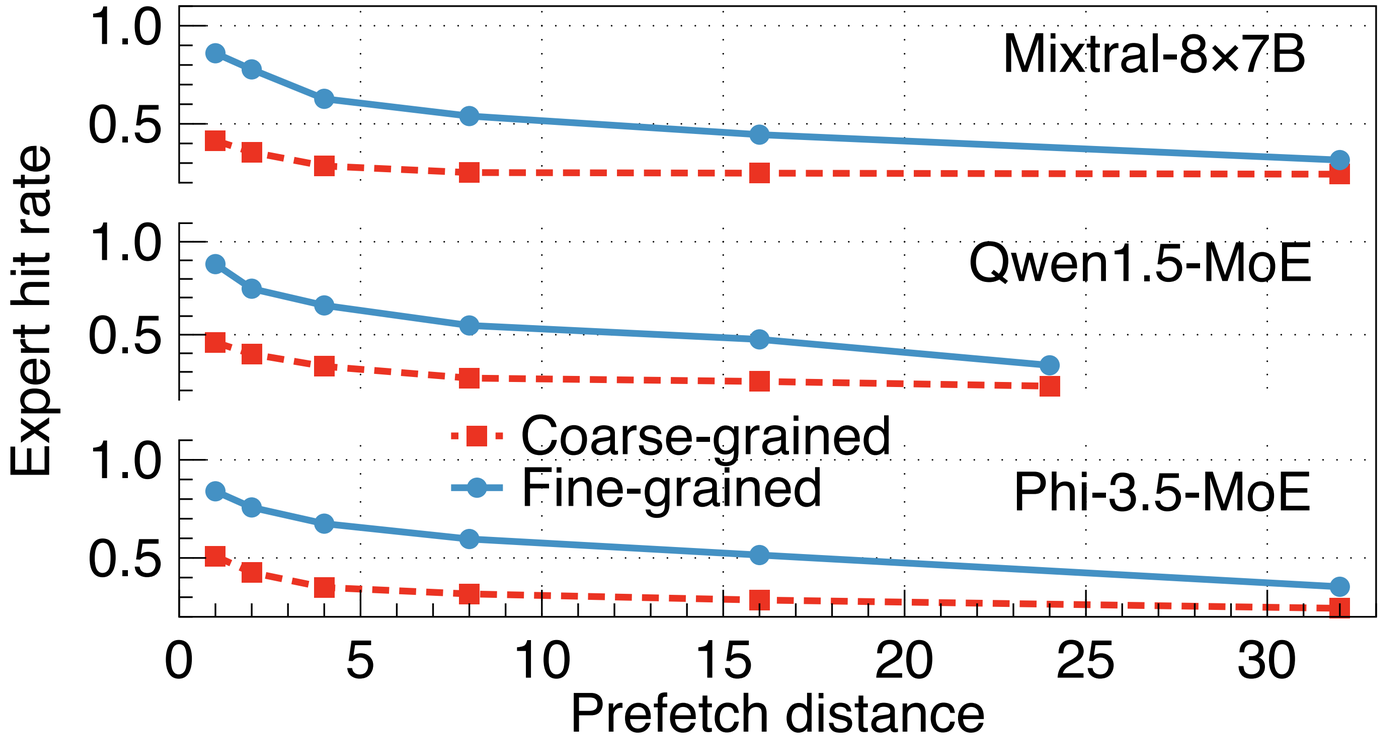


(b) Mean entropy per layer through iterations

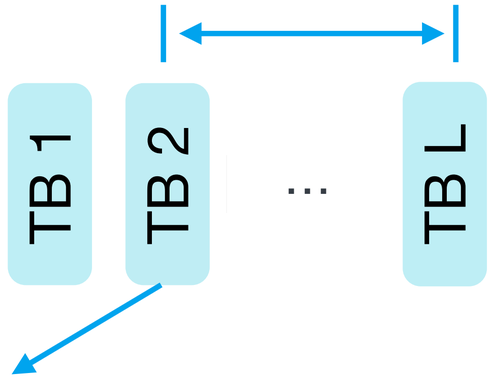
Higher entropy means **lower predictability**

Serving Mixtral-8x7B, Qwen1.5-MoE, and Phi-3.5-MoE

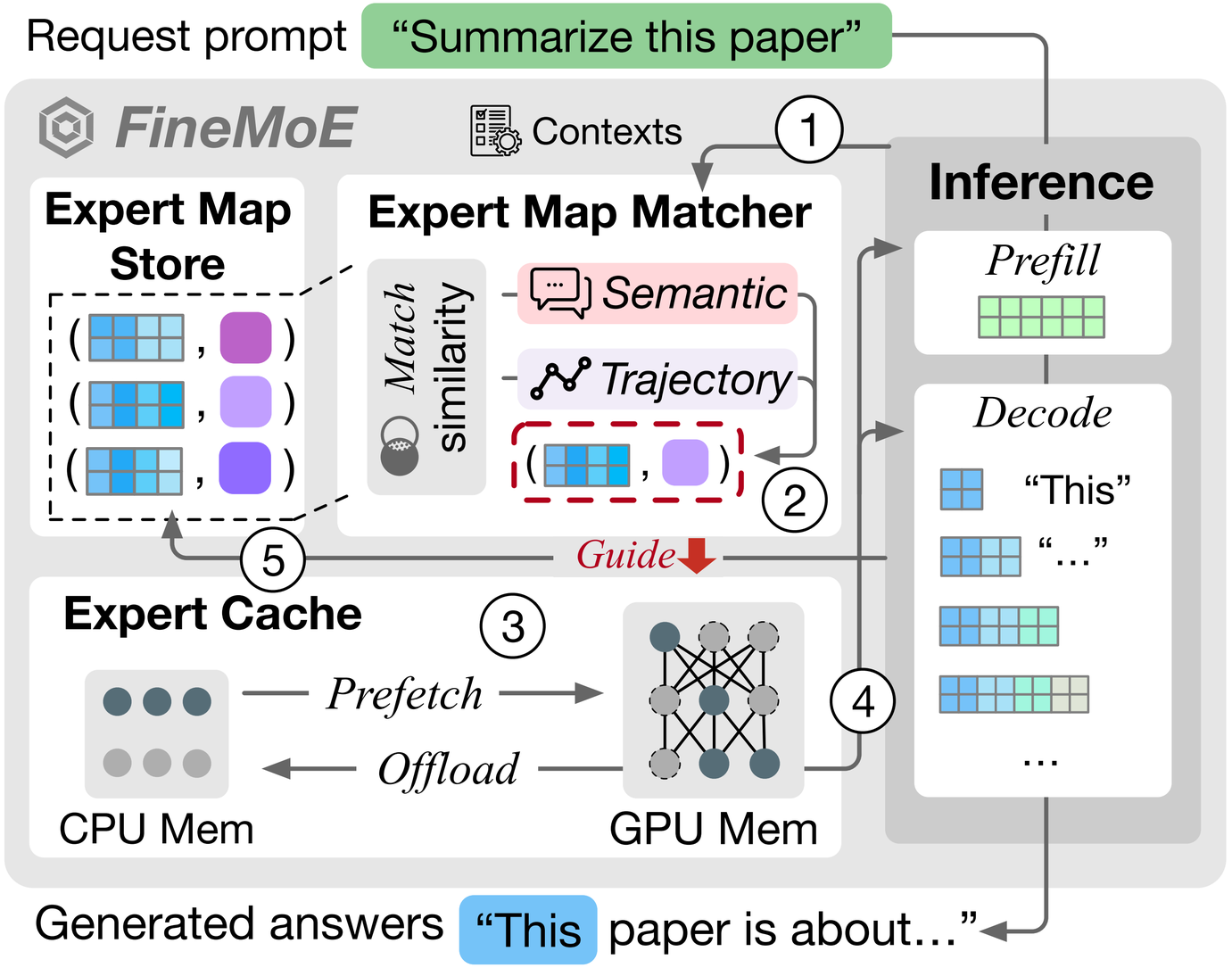
Coarse-Grained *vs.* Fine-Grained Offloading



Prefetch distance: # of layers ahead that the prefetch is issued before the target layer activates its experts to overlap loading overhead



FineMoE: Fine-Grained Expert Offloading



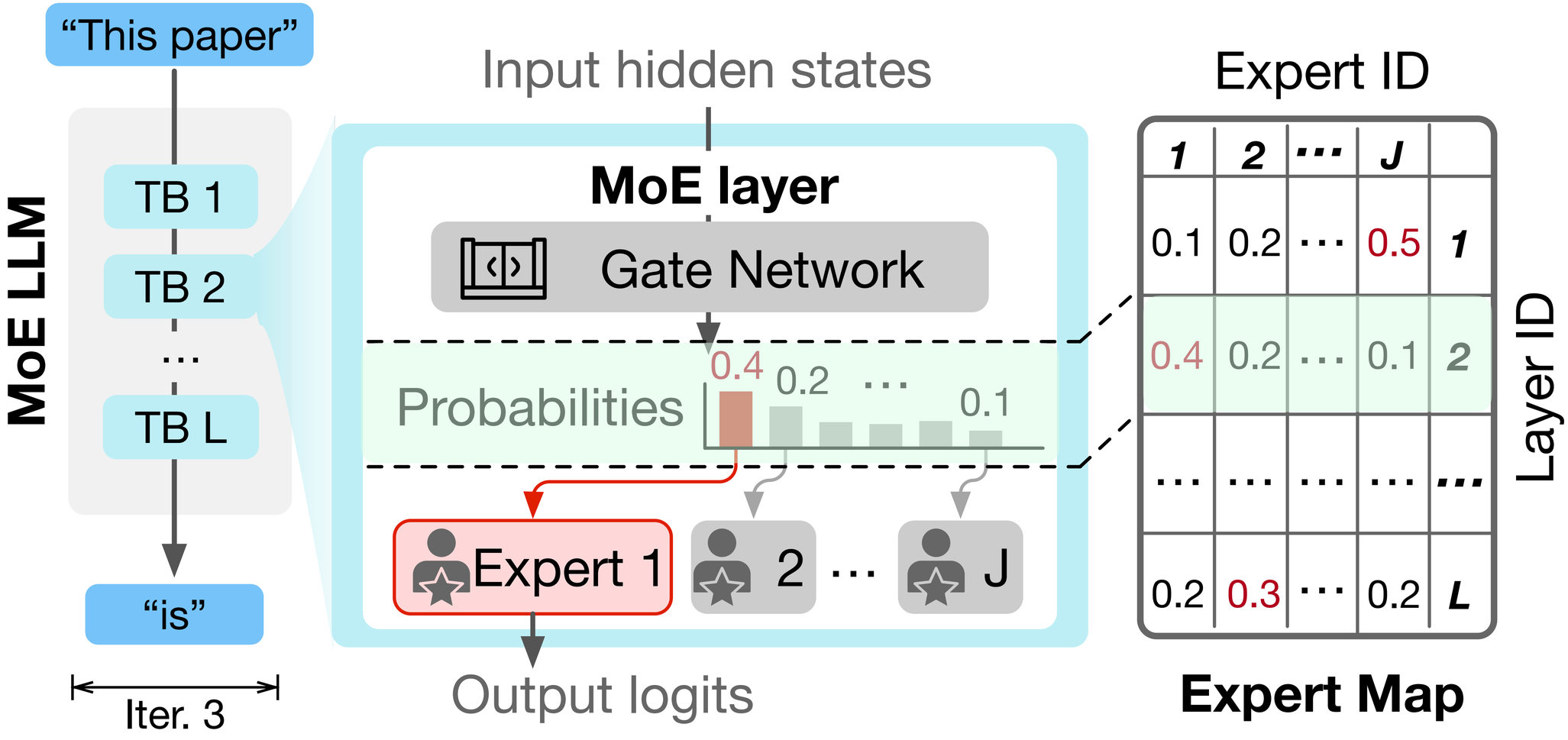
Fine-grained expert map

Fine-grained map search

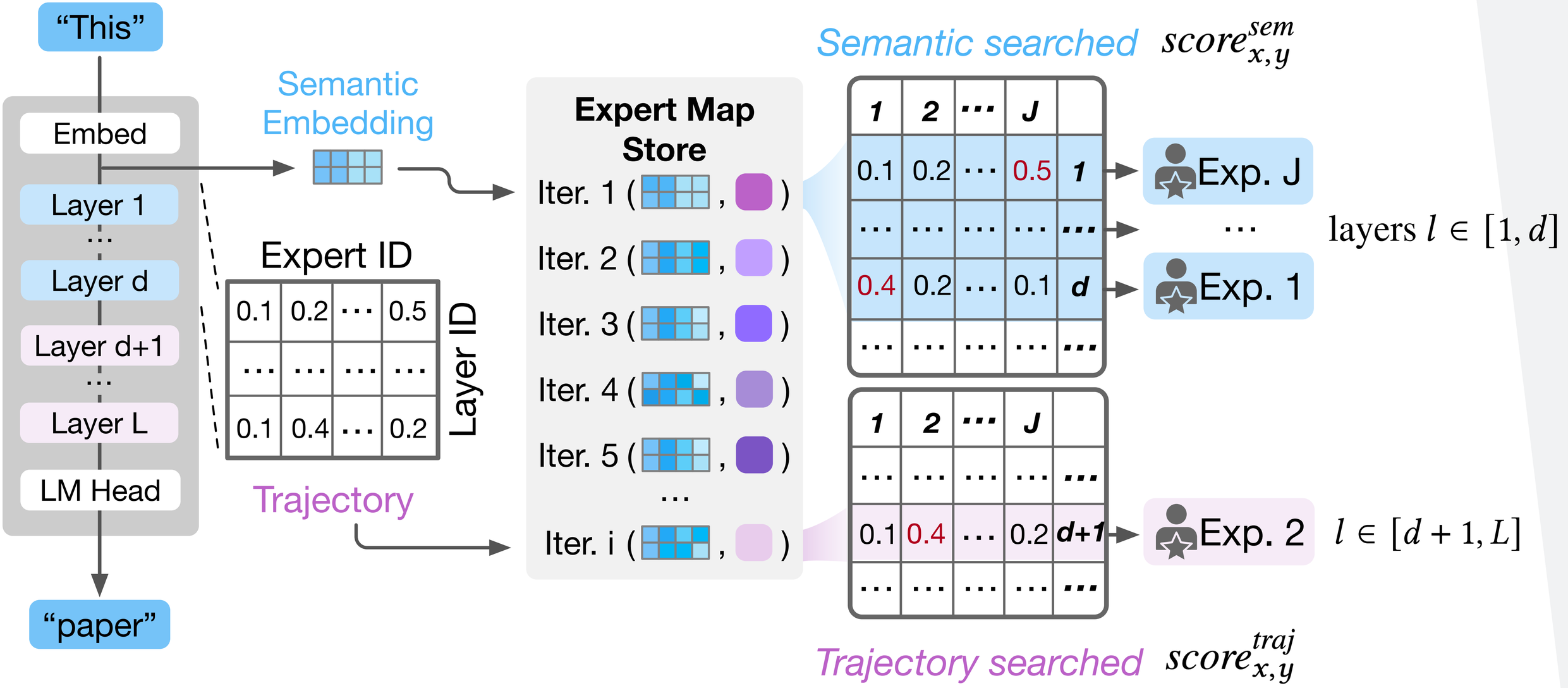
Fine-grained prefetch & cache management

Fine-Grained Expert Map

- Iteration-level expert selection
- Probability distributions rather than hit count

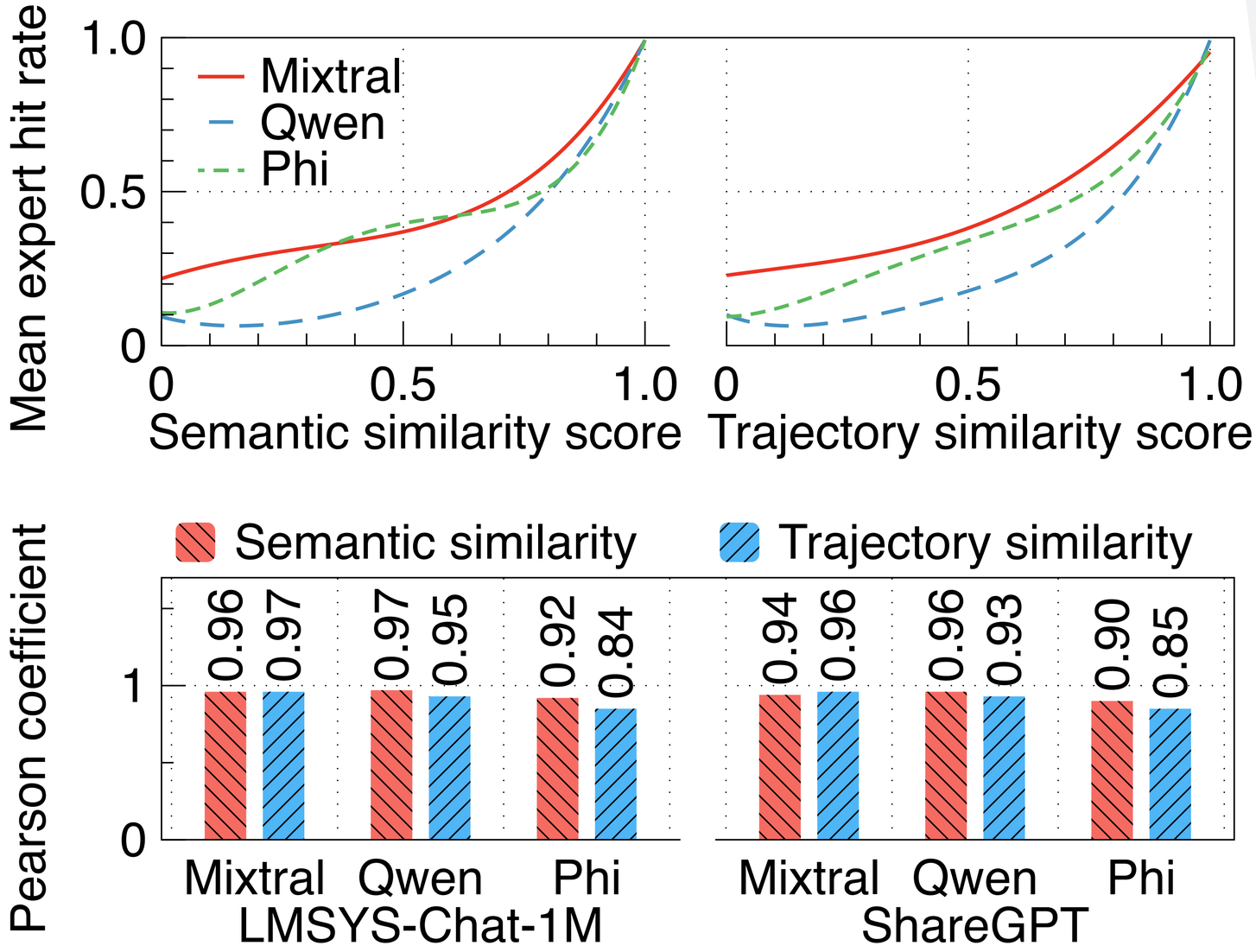


Fine-Grained Expert Map Search



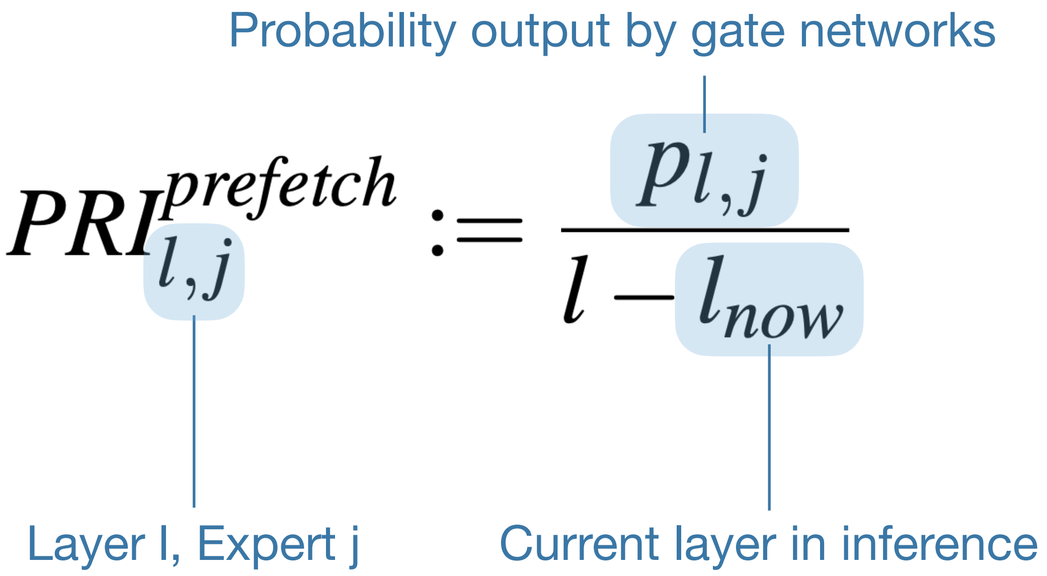
Expert Semantic & Trajectory Similarity

Expert semantic & trajectory similarity can effectively guide **expert usage**

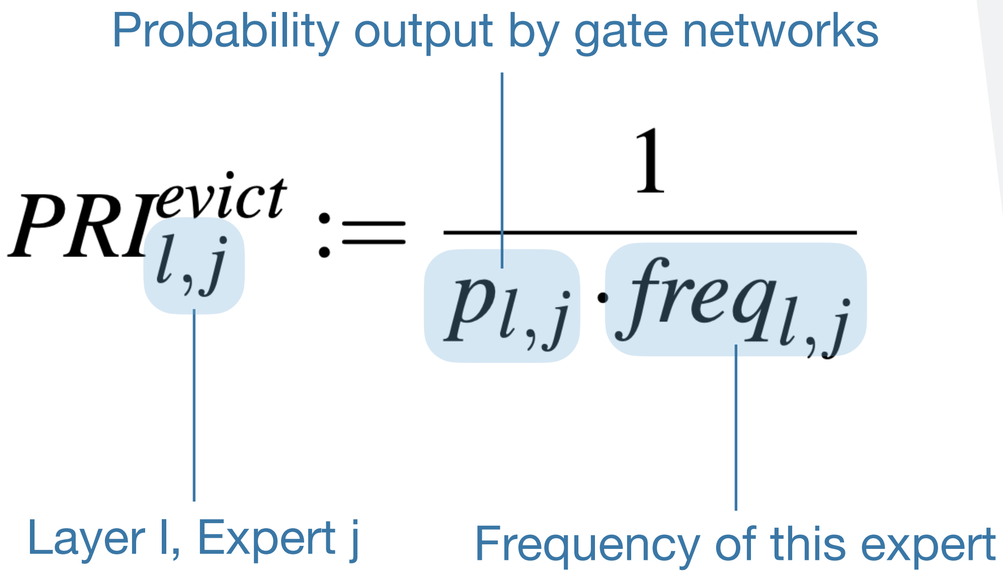


Fine-Grained Expert Prefetch & Cache

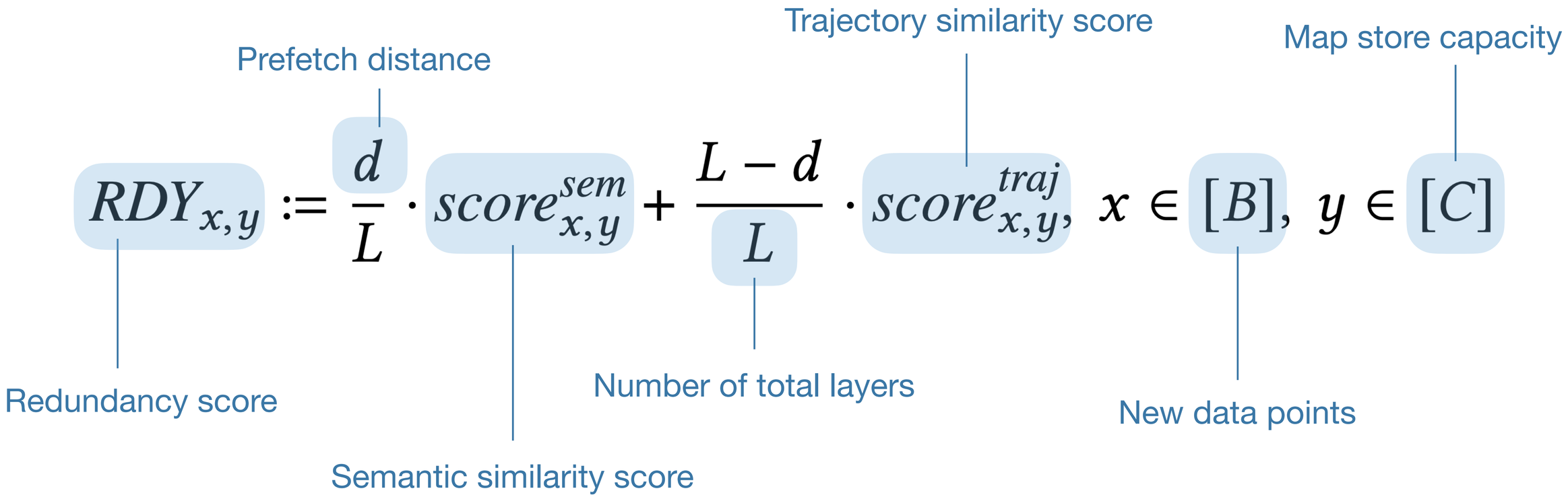
Prefetch Priority



Eviction Priority



Expert Map Store Deduplication



Leverage the contributions of two similarity scores

Evaluation

Implementation

HuggingFace Transformers
MoE-Infinity

Baselines

MoE-Infinity
ProMoE
Mixtral-Offload
DeepSpeed

Metrics

Time-To-First-Token (TTFT)
Time-Per-Output-Token (TPOT)
Expert Hit Rate

Datasets

ShareGPT

Prompts from
OpenAI ChatGPT

LMSYS-Chat1M

One-million
Chats & Conversations

MOE-INFINITY: Efficient MoE Inference on Personal Machines with Sparsity-Aware Expert Cache. arXiv
ProMoE: Fast MoE-based LLM Serving using Proactive Caching. arXiv
Fast Inference of Mixture-of-Experts Language Models with Offloading. arXiv
DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. SC'22
ShareGPT: Share Your Wildest ChatGPT Conversations. https://sharegpt.com/
LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. arXiv

Models and Testbed

MoE (LLM) Models

MoE Models	Parameters (active / total)	Experts Per Layer (active / total)	Num. of Layers
Mixtral-8×7B [23]	12.9B / 46.7B	2 / 8	32
Qwen1.5-MoE [60]	2.7B / 14.3B	4 / 60	24
Phi-3.5-MoE [1]	6.6B / 42B	2 / 16	32

GPU Testbed

6 NVIDIA RTX 3090

Pairwise NVLinks

144 GB GPU memory

<$10K USD

https://docmost.intellisys.synology.me/share/a782usgds1/p/6-gpu-machine-recipe-WVmaOsjzKE

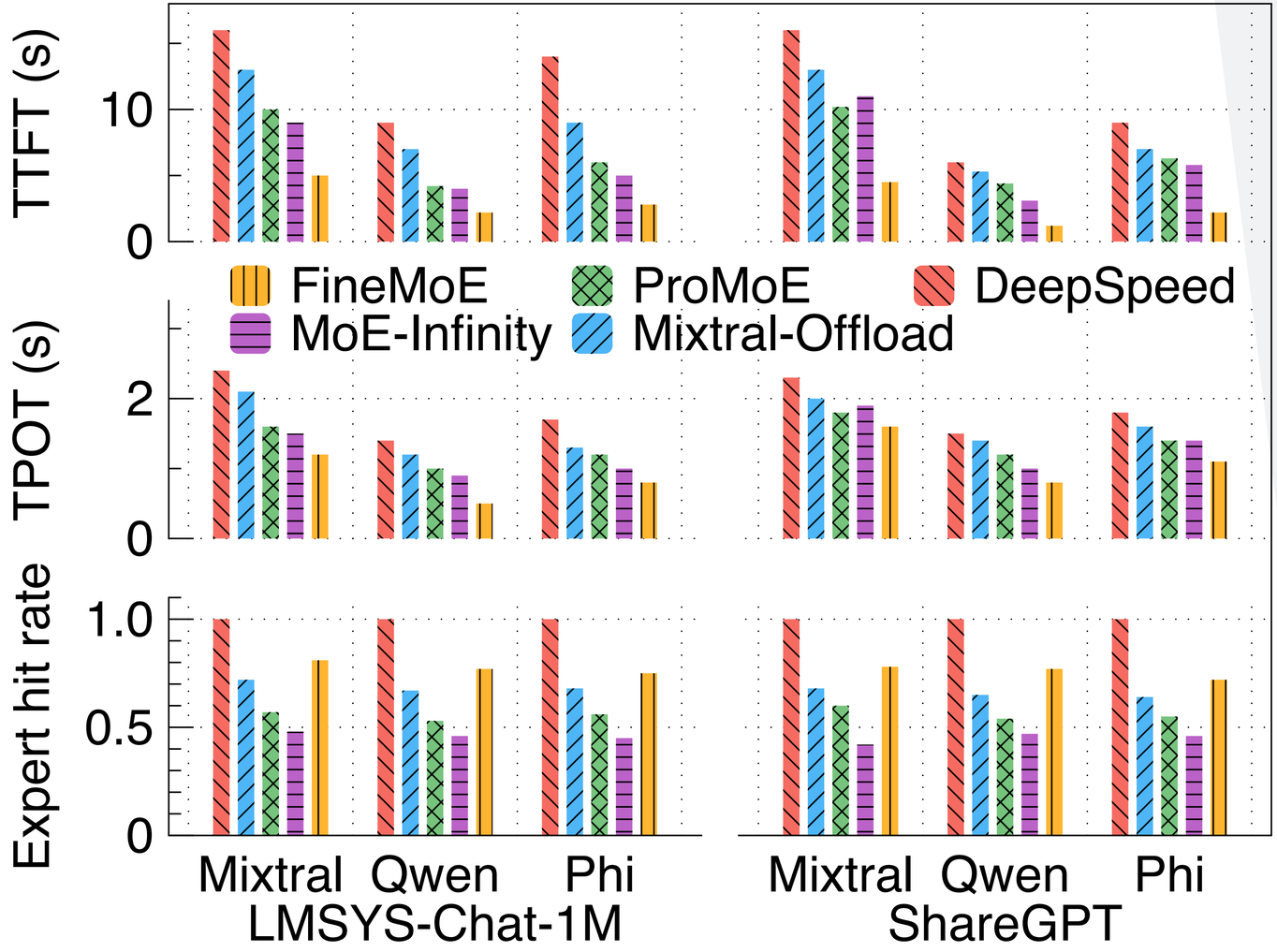
Serving Performance

Lower Inference Latency

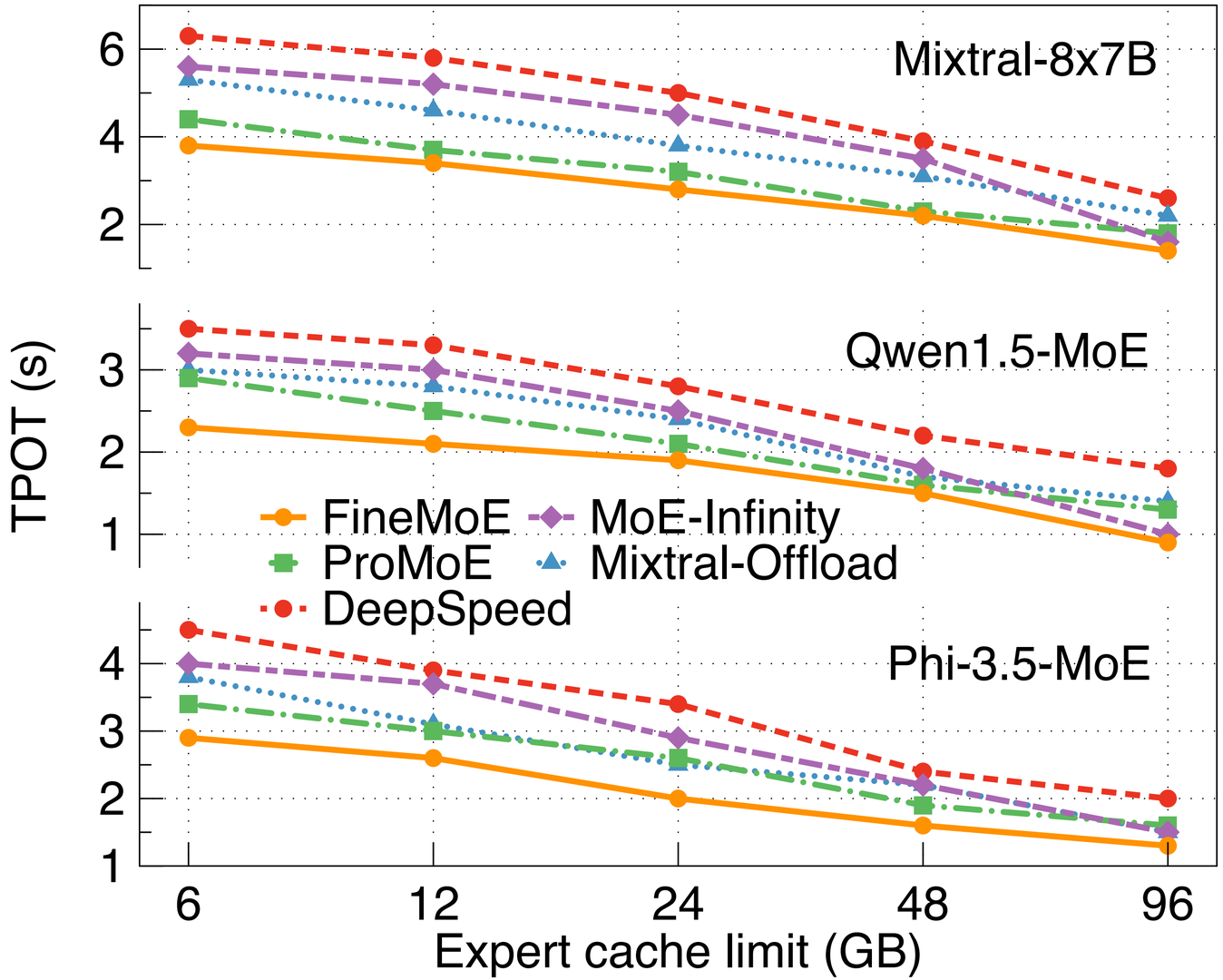
Higher Expert Hit Rate

47%

Inference Latency Reduction



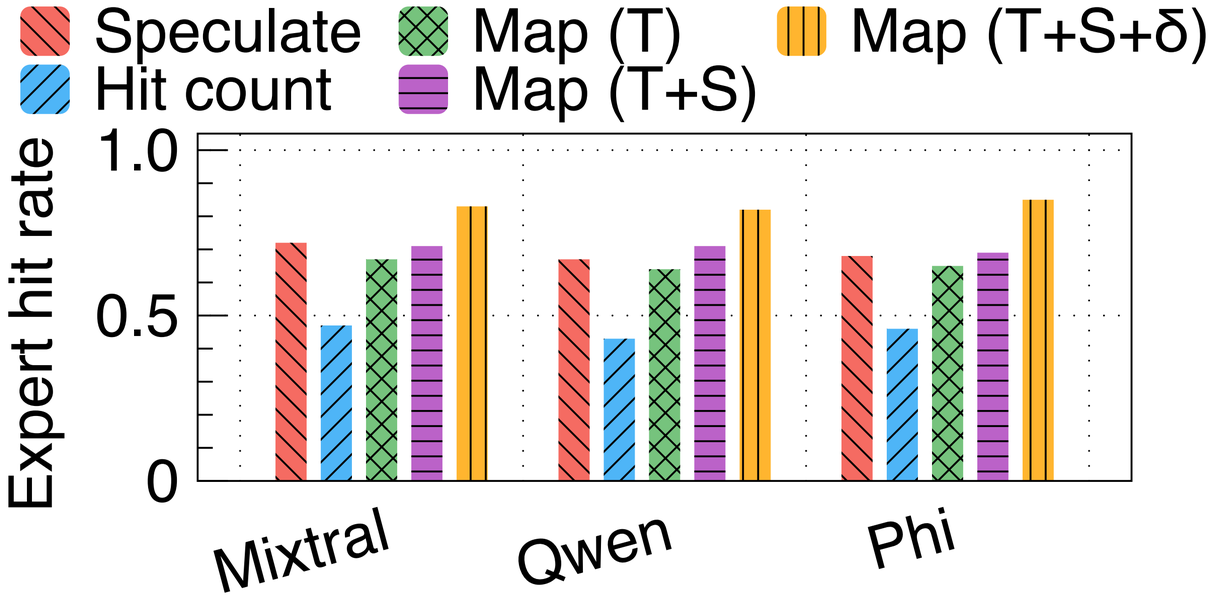
Limited Memory Budget

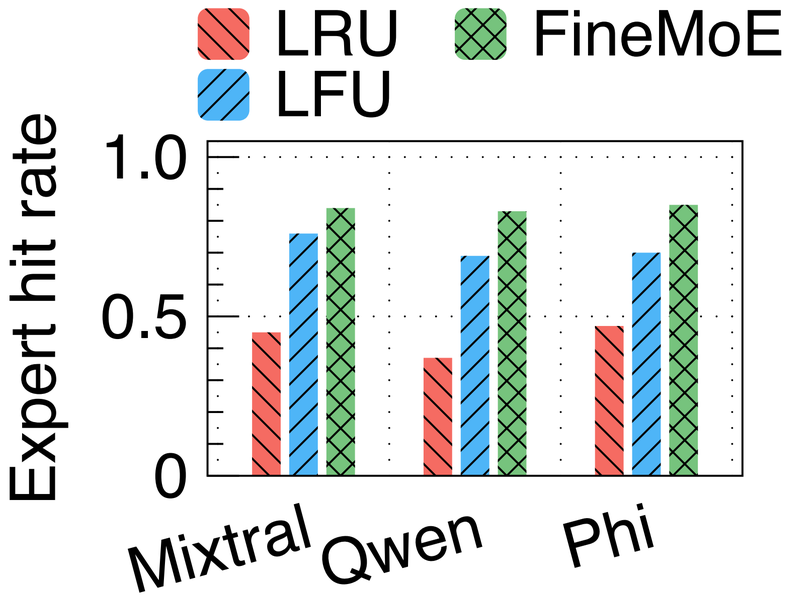


32%

TPOT reduction with 6GB memory

Expert Tracking, Prefetching, and Caching

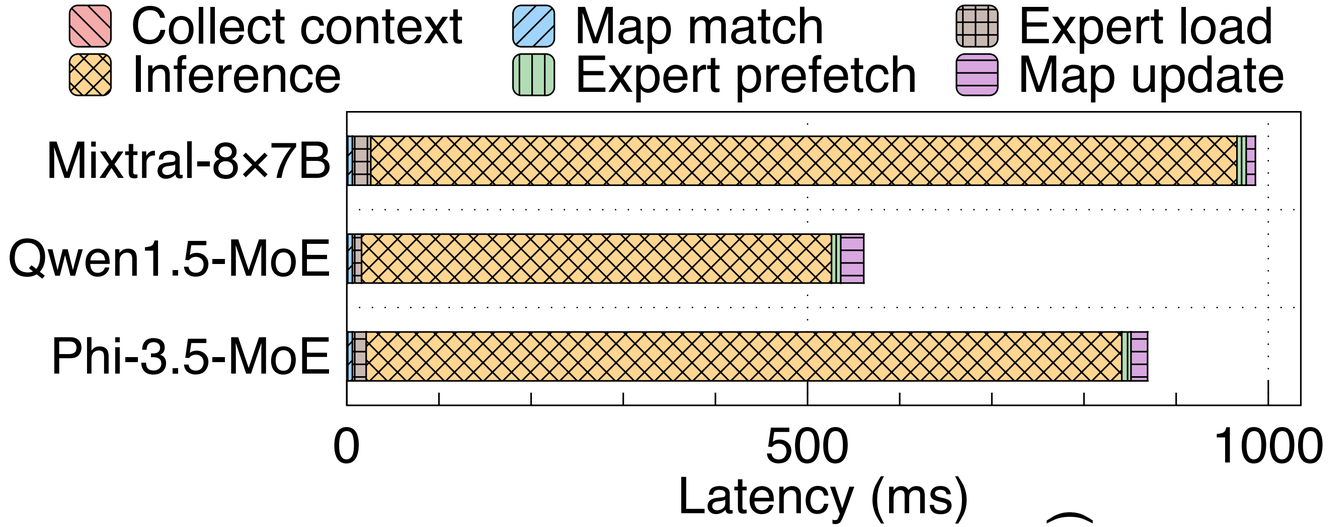


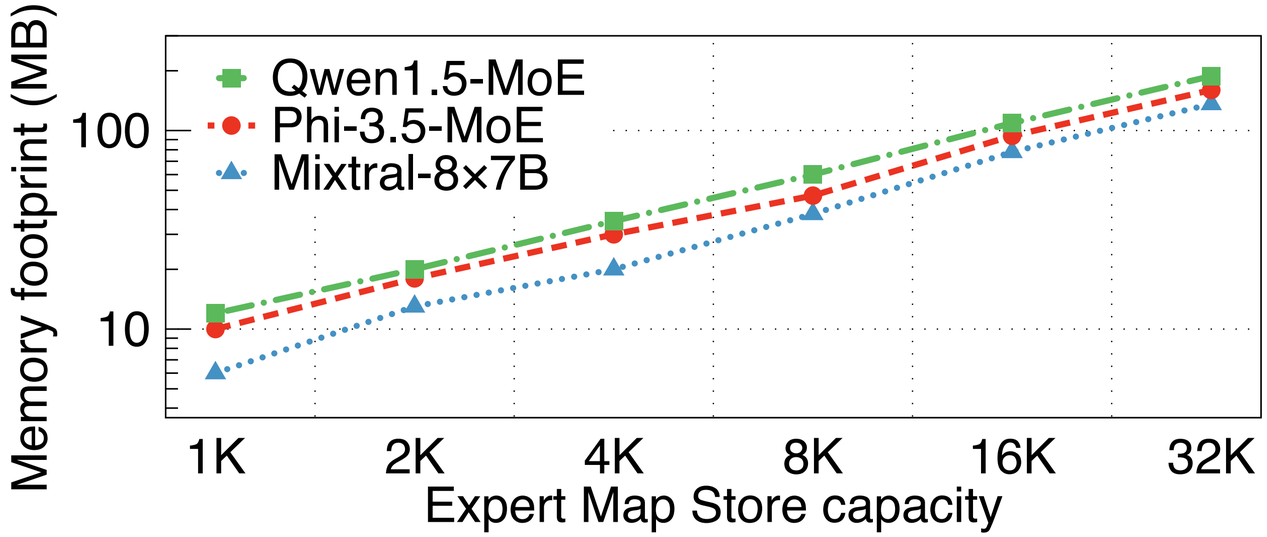


36%

Expert Hit Rate Improvement

System Overheads





Fine-grained
Expert Pattern Tracking

Fine-grained
Expert Map Search

Fine-grained
Expert Prefetch & Cache

FineMoE

47%

Inference Latency Reduction

36%

Expert Hit Rate Improvement

THANK YOU

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030

Fine-grained
Expert Pattern Tracking

Fine-grained
Expert Map Search

Fine-grained
Expert Prefetch & Cache

FineMoE

47% Inference Latency Reduction

36% Expert Hit Rate Improvement

Paper

Github