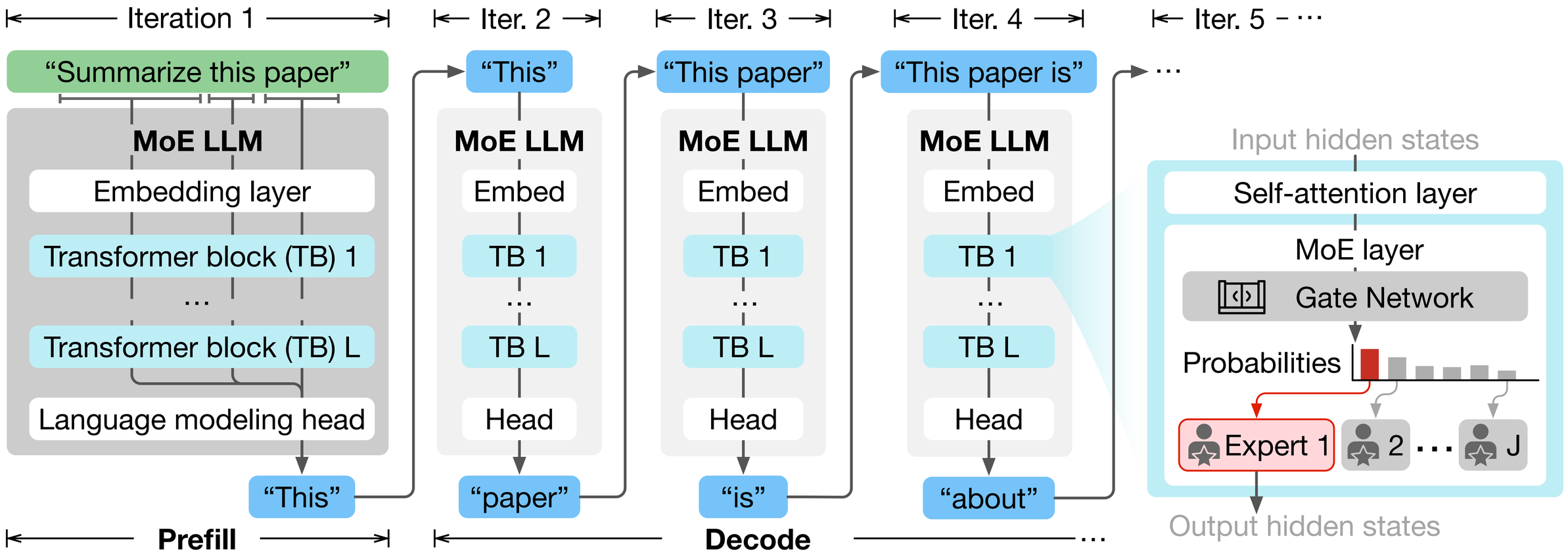
Taming Latency-Memory Trade-Off in MoE-Based LLM Serving via Fine-Grained Expert Offloading

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

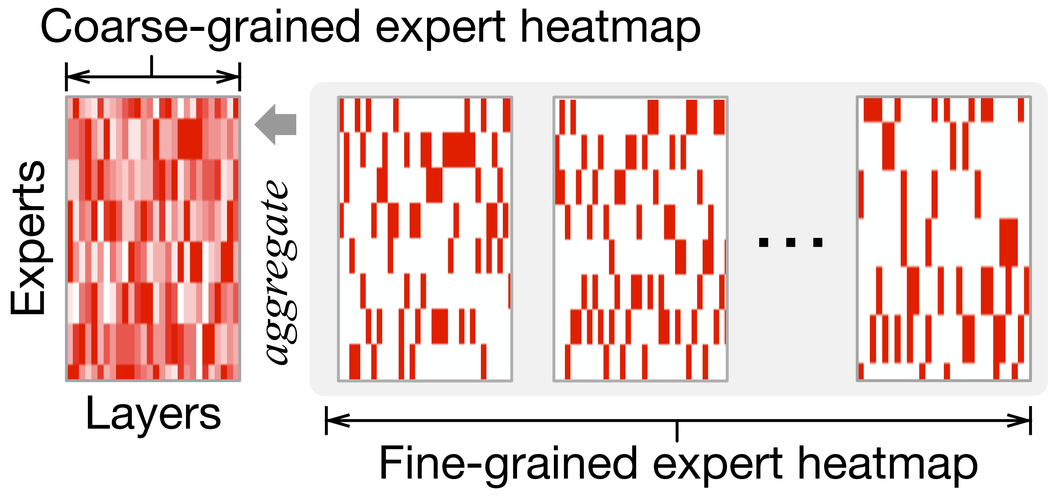
Stevens Institute of Technology[1], Rice University[2], University of Waterloo[3], Rutgers University[4]

EuroSys 2026 | IntelliSys Lab

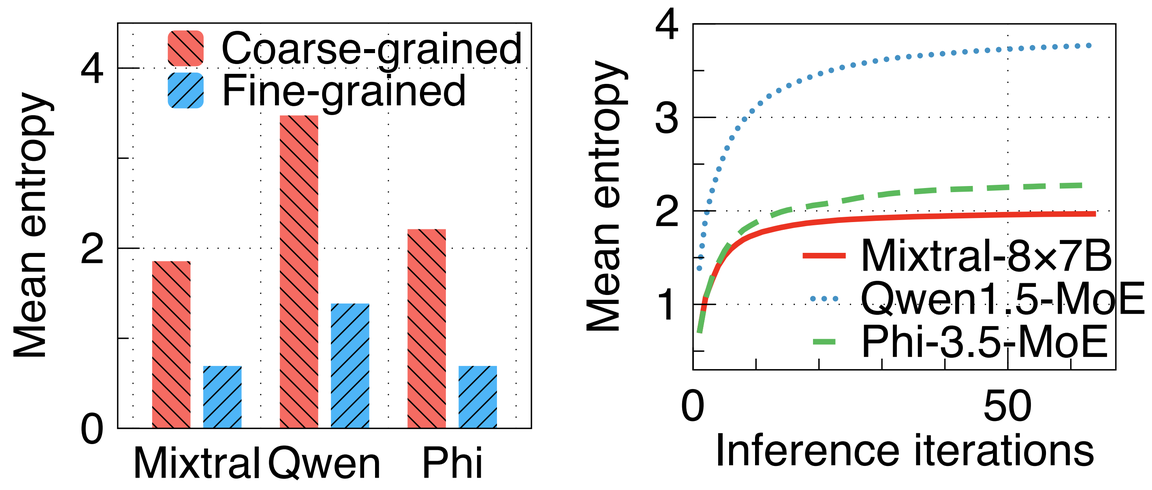
0. Mixture-of-Experts (MoE) Based Large Language Model (LLM) Serving



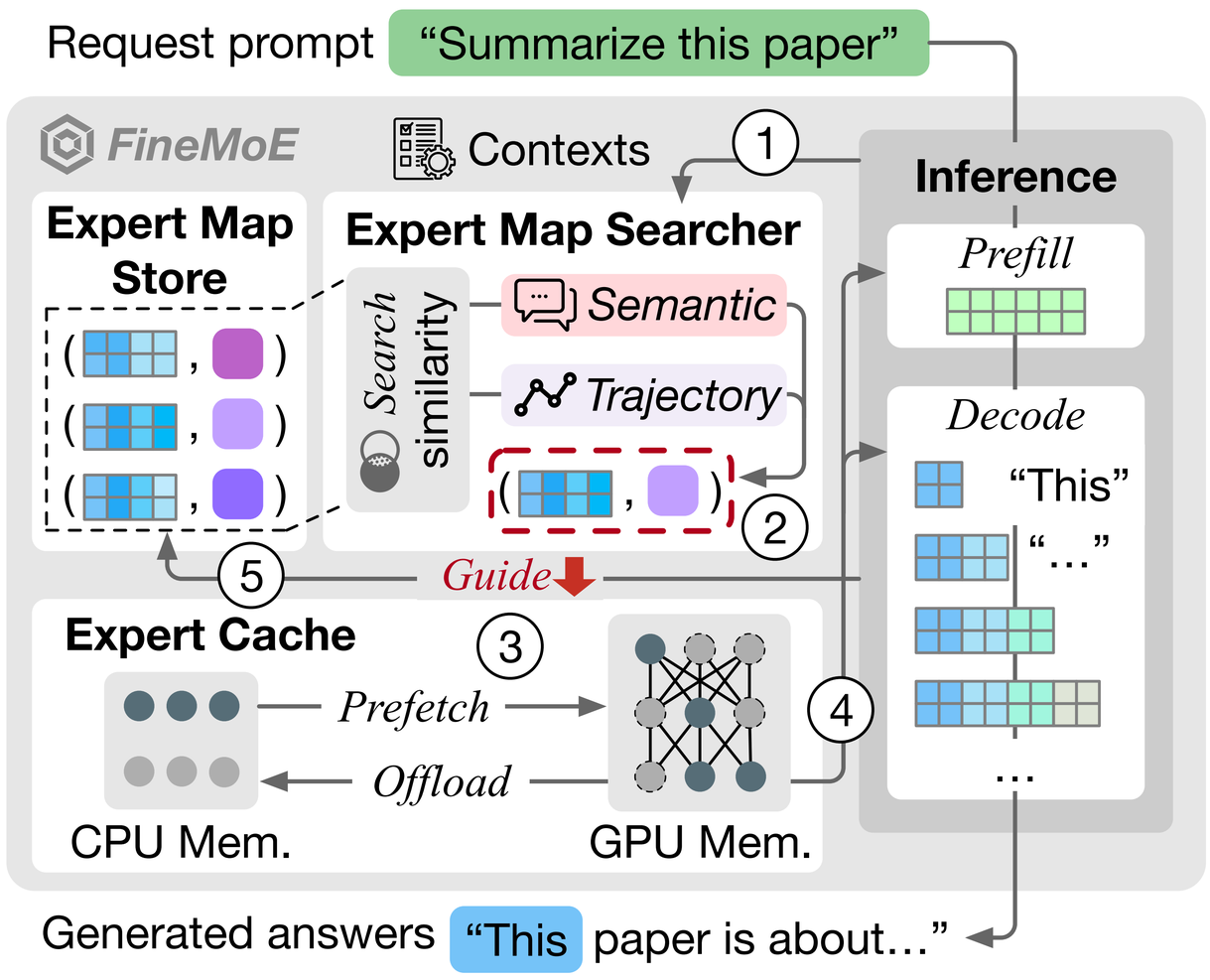
1. Coarse- vs. Fine-grained Expert Offloading



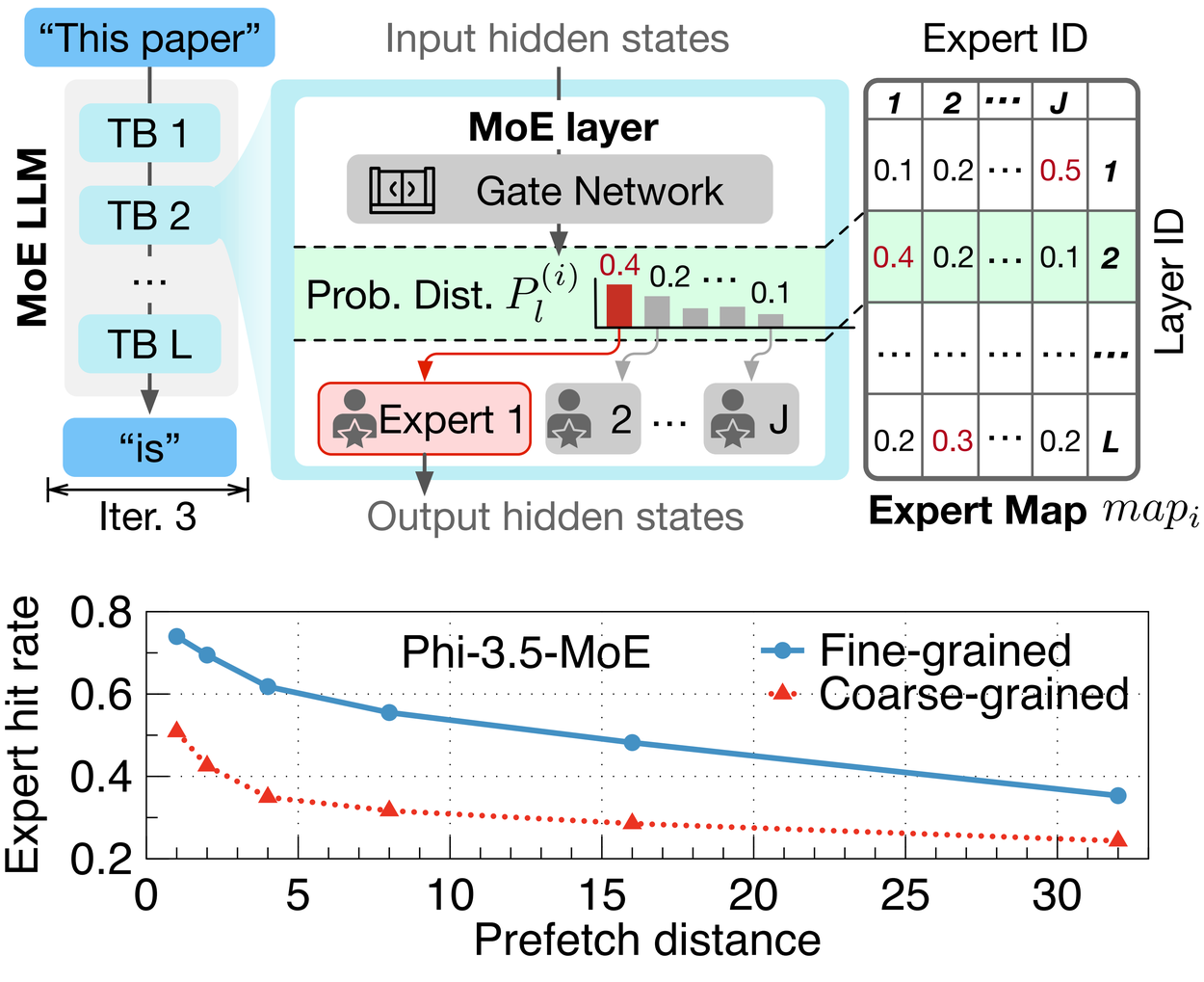
2. Entropy of Expert Activation Patterns



3. Workflow of FineMoE



4. Design of Expert Probability Map



We thank anonymous reviewers and our shepherd, Dr. Yaniv David, for their valuable feedback. This project is supported by U.S. National Science Foundation and AWS Cloud Credit for Research.