Ritesh Bhirud (Author), Rabab Alomairy (Advisor)

CUR-MoE: Portable Mixture-of-Experts with High-Ratio Compression

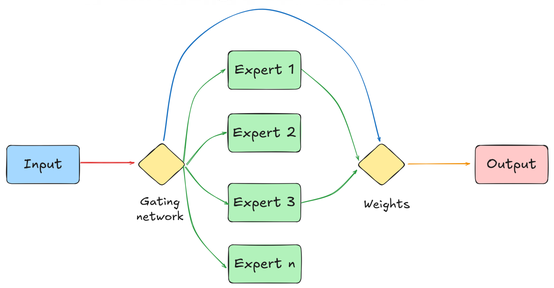
Objective

Challenge: MoE models face prohibitive memory scaling for HPC deployment [1]. While SVD-based compression achieves effective model compression [2], it uses abstract factorization that obscures expert behavior and limits interpretability of compressed models [3].

Our Innovation: We present the first Julia-based MoE framework using CUR decomposition—achieving comparable compression with interpretable experts, and portable multi-GPU execution across NVIDIA, AMD, Intel, and Apple.

Tools: Julia

Julia's LLVM pipeline enables portable MoE training across vendor GPUs



Julia combines high-level syntax with C-level speed, unifying prototyping and HPC-scale deployment—eliminating the Python/C++ split and accelerating research-to-production.

Benchmarking Hardware

CPU: AMD Ryzen 9 HX 370 (12 cores, 24 logical processors)
GPU: NVIDIA RTX 5060 Ti 16GB / H200 NVL 140 GB/ AMD Radeon RX 7600
RAM : 64GB / 512 GB

CUR-MoE: Practical Alternative with Preserved Interpretability

CUR-MoE Compression: Breakthrough Performance Analysis
Mixtral 8×7B Model Evaluation

Compression Method	Moderate Compression (20-40%)	High Compression (60%)	Interpretability	Memory at 60%	Throughput (tokens/sec) 20% / 40% / 60%
CUR-MoE *(Our approach)*	5.82-8.59 perplexity	31.64 perplexity	Interpretable decomposition	76.42 GB (~56% reduction)	100.3 / 104.7 / 143.9
MoE-SVD	5.94-8.66 perplexity[2]	33.24 perplexity[2]	Limited transparency	70.31 GB[2] (~60% reduction)	104.7 / 109.8 / 156.1[2]

Hardware Performance: Cross-Vendor GPU Compatibility vs CPU

Configuration:
512 → 2048 → 512 dims
8 experts, top-2 routing
Batch size: 64

CPU Baseline	AMD GPU	NVIDIA/CUDA GPU
8.276 ms	1.572 ms	1.312 ms
1.0× baseline	5.3× faster	6.3× faster

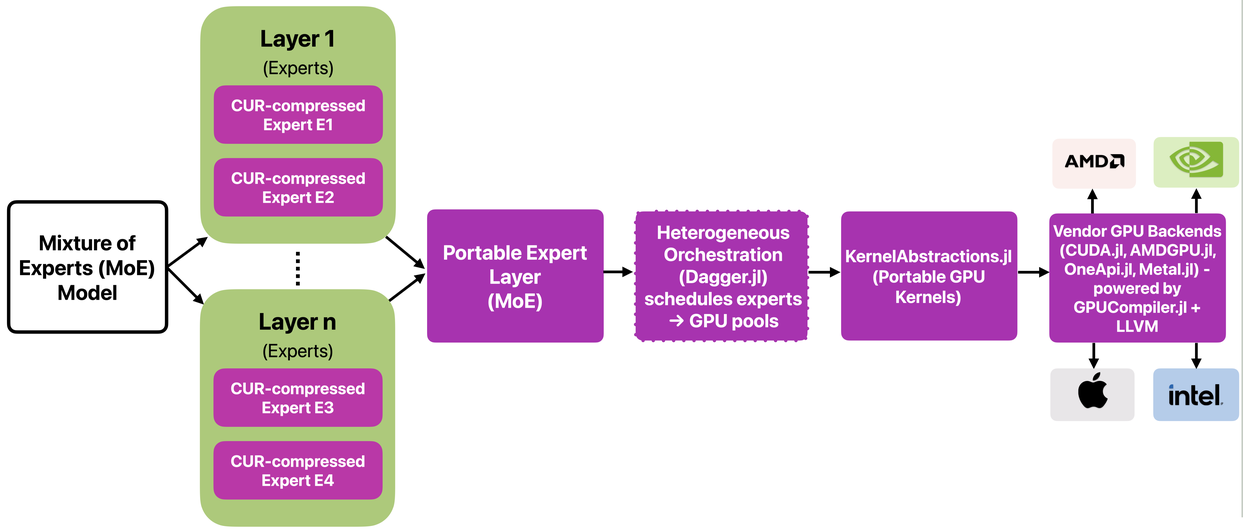
MoE Gating Mechanisms: Comprehensive Evaluation Across Routing Strategies

MoE Gating Mechanisms: Performance Comparison

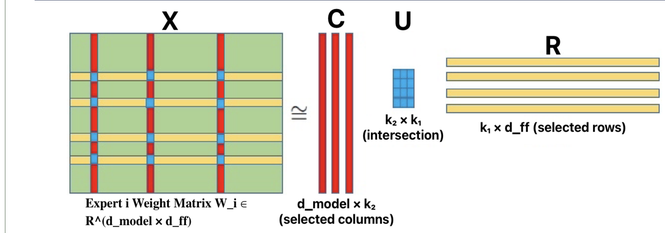
Gating Mechanism	Load Balance (higher = better)	Expert Utilization (higher = better)	Inference Speed (tokens/sec)	Text Diversity (0-1 scale)	Routing Stability (higher = better)
Random	0.955	1.008	134.6	0.809	0.870
TopK	0.713	0.849	135.4	0.845	0.924
Switch	0.672	0.749	170.5	0.810	0.873
StochasticTopK	0.739	0.853	153.1	0.829	0.842
ExpertChoice	0.815	0.954	167.9	0.770	0.899
SoftMoE	0.779	0.911	141.8	0.785	0.904
Hash	0.598	0.664	155.6	0.830	0.988
SharedExpert	0.810	0.907	143.2	0.791	0.908

Legend: Green highlighting = Top 3 performers in each metric category

Portable HPC Deployment: Julia-Enabled Cross-Vendor Compatibility



The algorithm: CUR



Selection Method: Leverage-based sampling

- Row selection: SVD-derived importance scores
- Column selection: Feature significance ranking
- Preserves interpretability (vs SVD abstractions)

Results

- **Moderate Compression (20–40%)**
  CUR-MoE and SVD-MoE perform equivalently (Original WikiText-2 perplexity: 3.92 → ~8.66), ensuring safe efficiency gains for HPC-scale models
- **Extreme Compression (70%)**
  SVD-MoE was not tested at this compression ,while CUR-MoE remains functional (35.29 perplexity, ~9× degradation)
- **HPC Scalability & Portability**
  MoE layers are embarrassingly parallel, inherently scalable across multi-GPU and mixed vendor clusters (NVIDIA, AMD, Intel, Apple) [1]. Julia's LLVM based compilation enables seamless portability across architectures.
- **Interpretability Breakthrough**
  CUR-MoE leverages CUR's inherent interpretability properties through preserved column/row structure [3], enabling analysis of expert specialization patterns and routing behavior—critical for debugging and optimizing large-scale MoE deployments.

References

[1] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang, "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models," arXiv preprint arXiv:2401.06066, 2024.
[2] Wei Li, Lujun Li, Hao Gu, You-Liang Huang, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. 2025. MoE-SVD: Structured Mixture-of-Experts LLMs Compression via Singular Value Decomposition. In *Proceedings of the Forty-second International Conference on Machine Learning* (ICML '25).
[3] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," Proc. Natl. Acad. Sci. U.S.A., vol. 106, no. 3, pp. 697–702, 2009.