Qwen MoE: Memory Efficiency vs Cache Effectiveness
Pareto Frontier Analysis

Oracle
(57.0 MB/expert)

Top-K
(36.2 MB/expert)

On-Demand
(No Caching)

90% Hit Rate Threshold

Cache Hit Rate (%)

Memory Efficiency (MB per Expert Cached)

Pareto Optimal
Dominated (Caching)
No Caching
Pareto Frontier