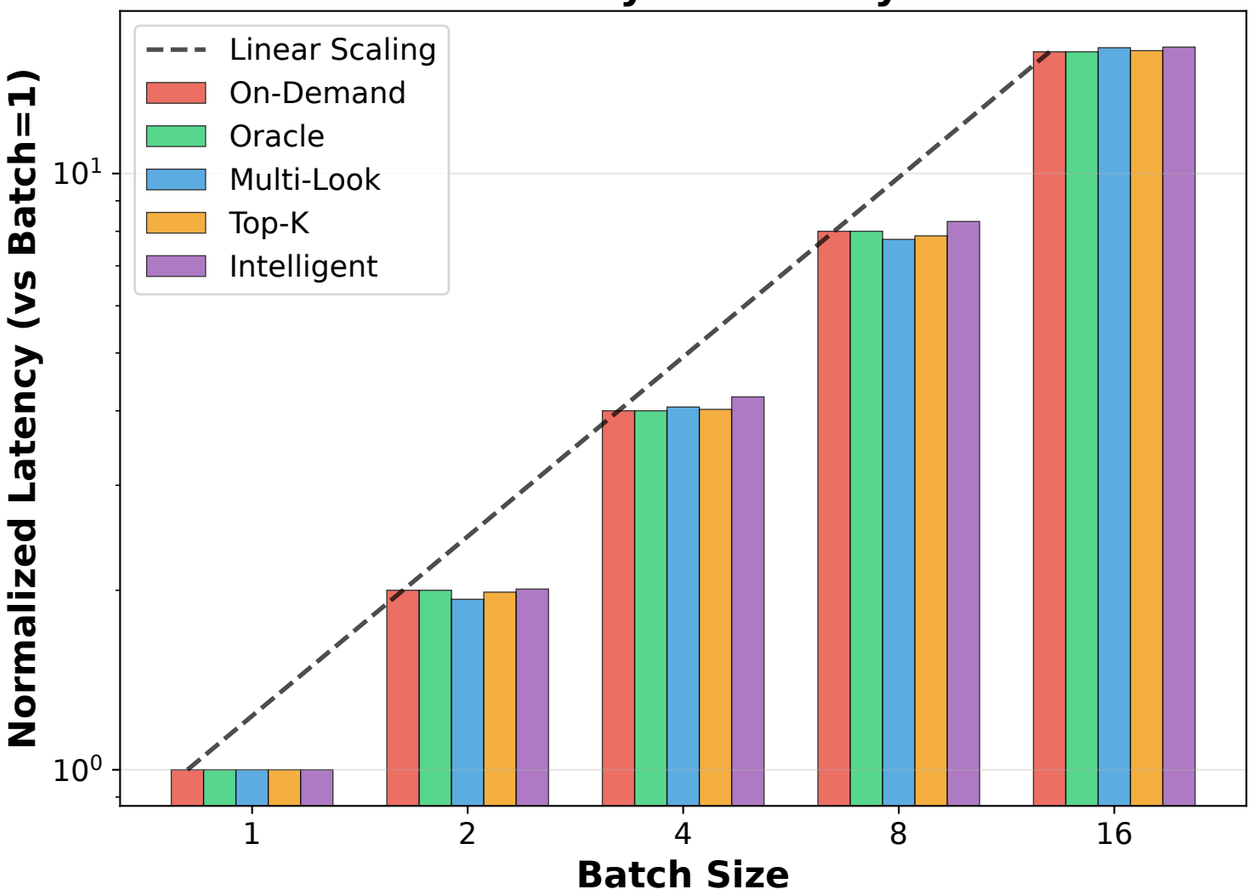
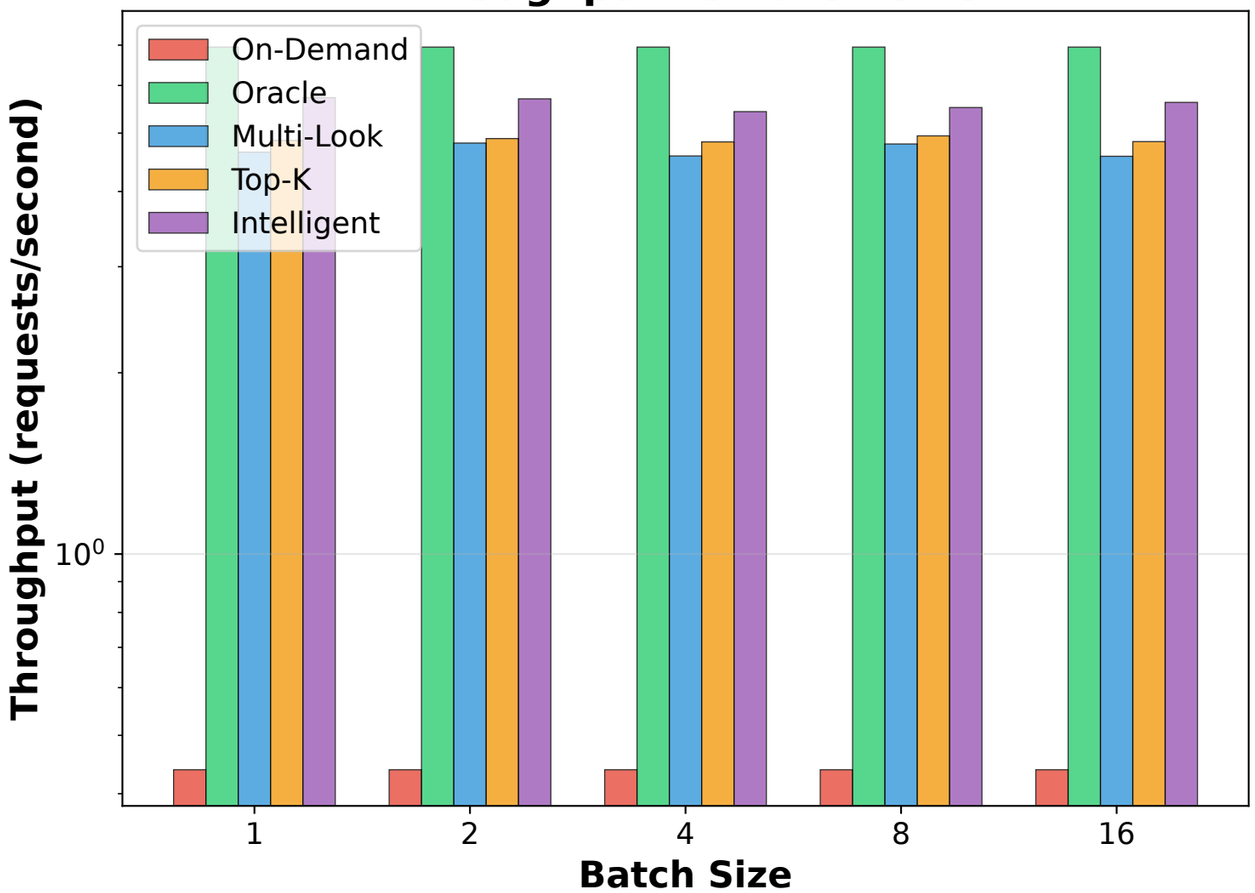
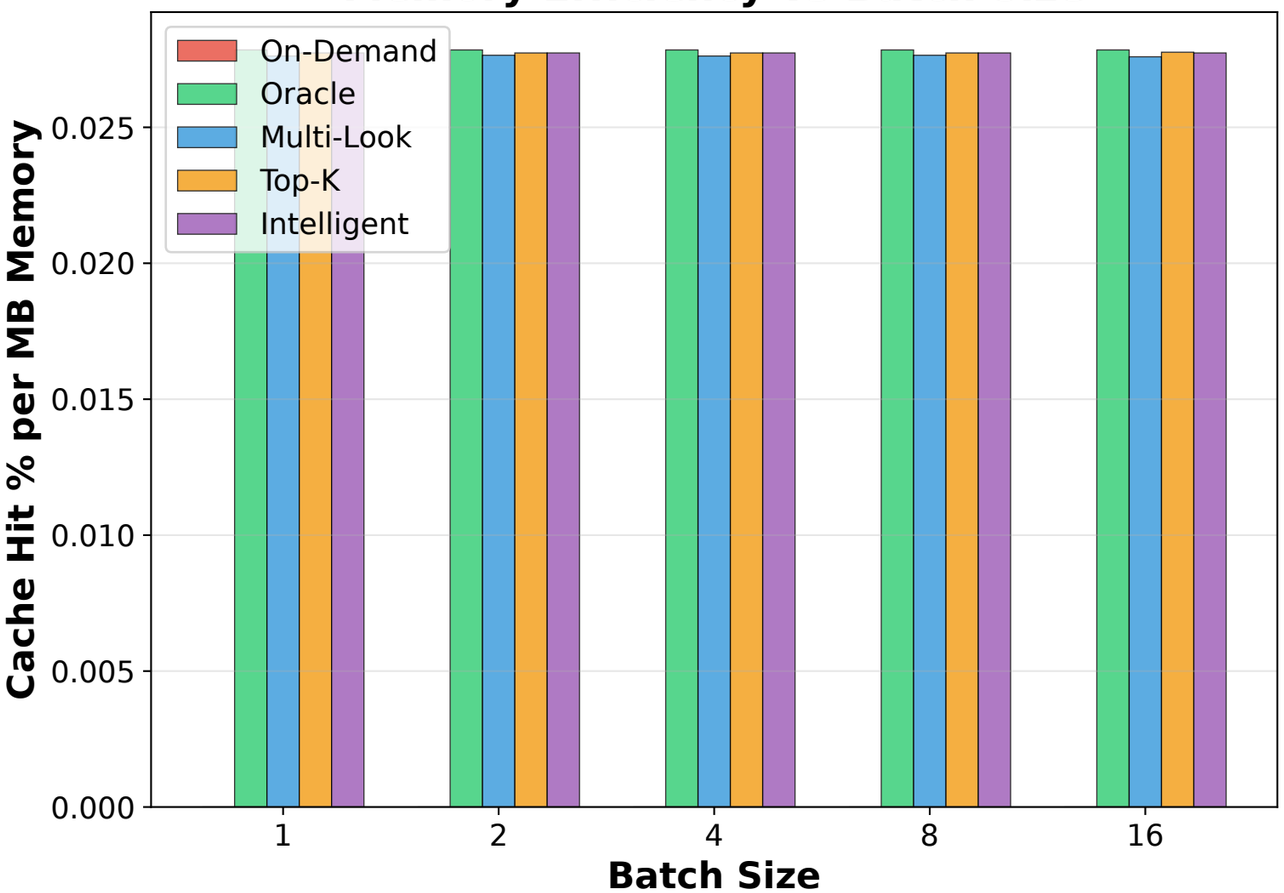
Latency Scalability



Throughput vs Batch Size



Memory Efficiency vs Batch Size



Expert Loading Efficiency