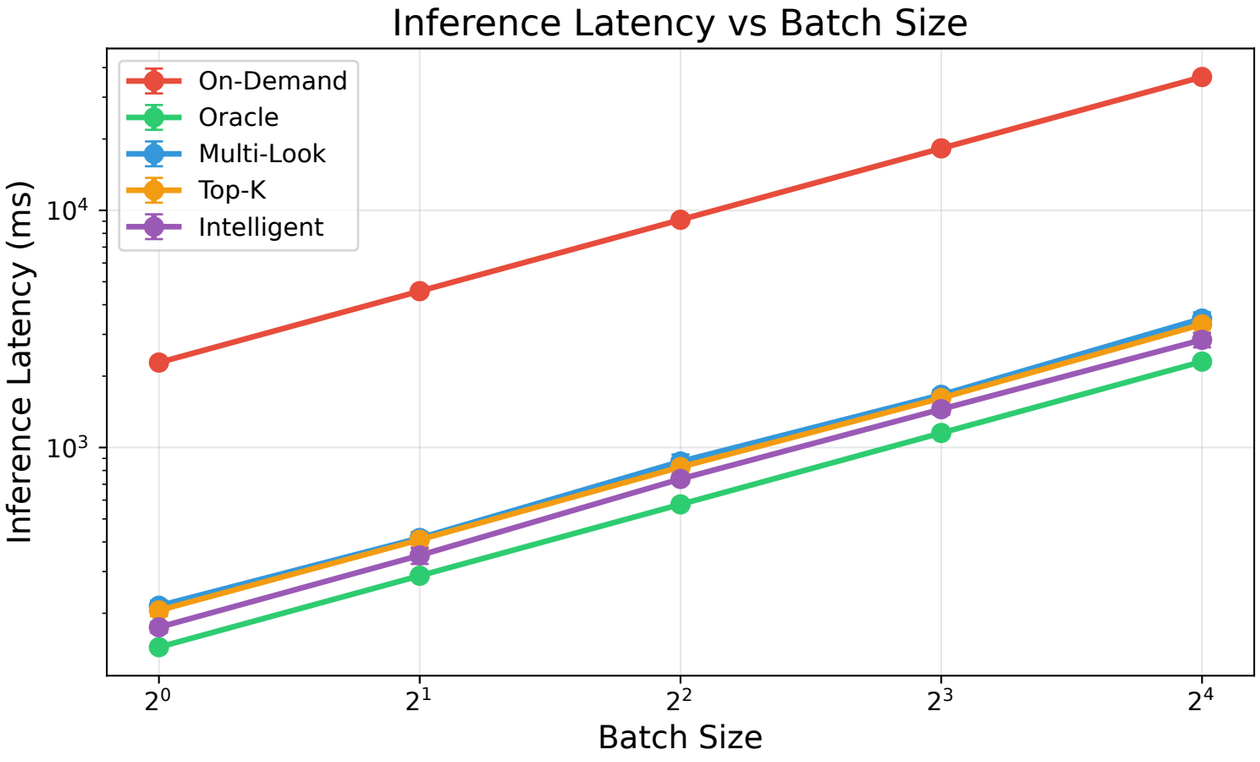
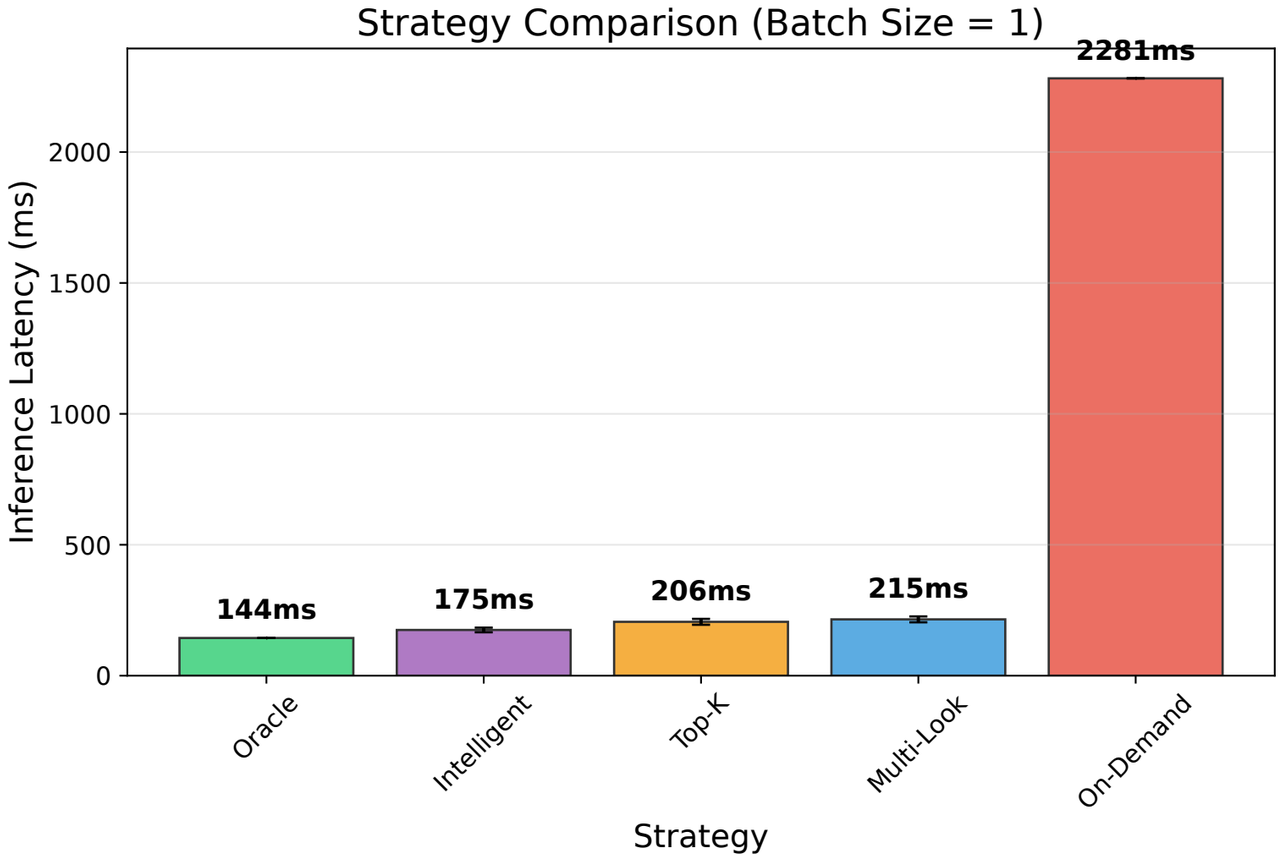
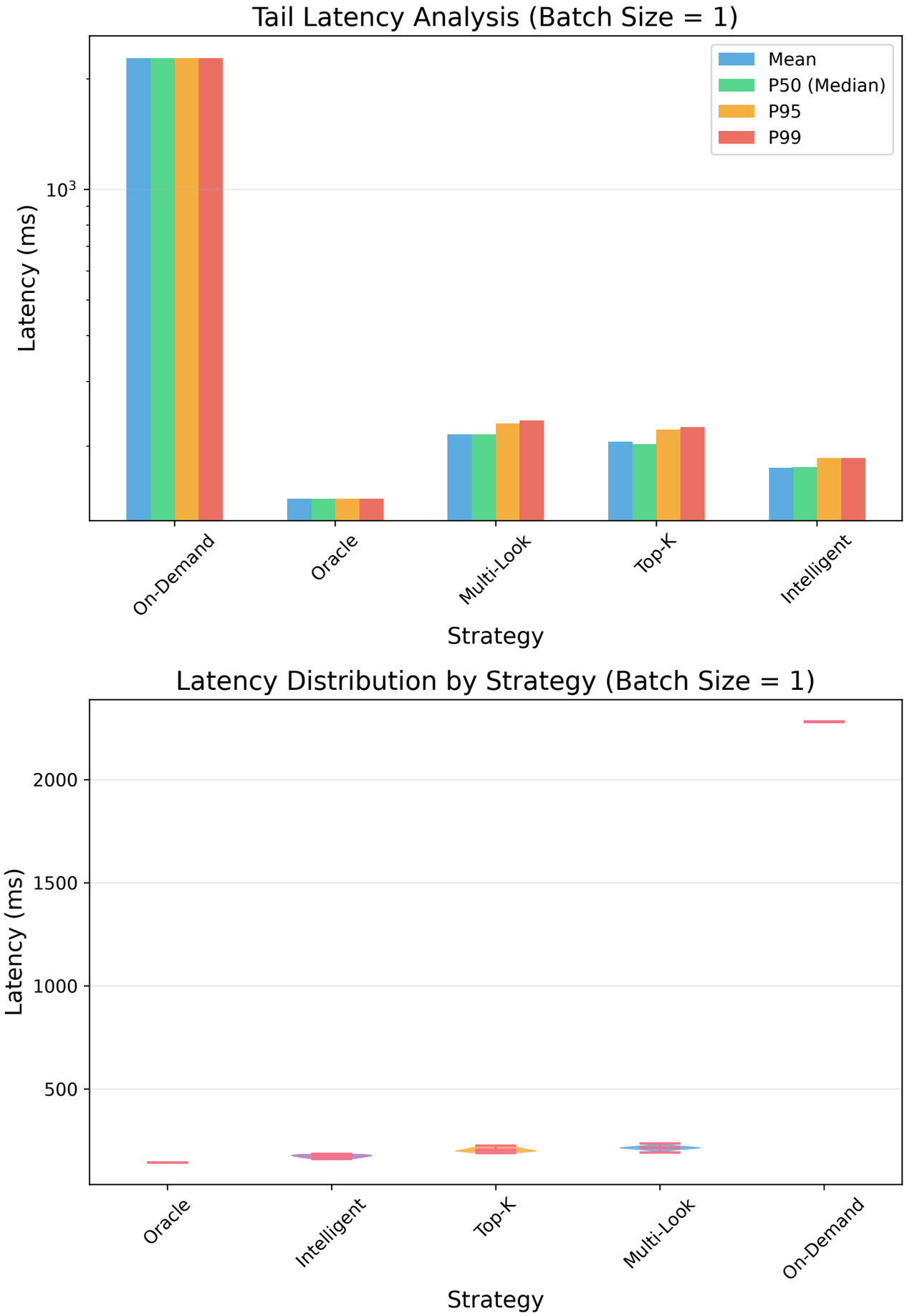
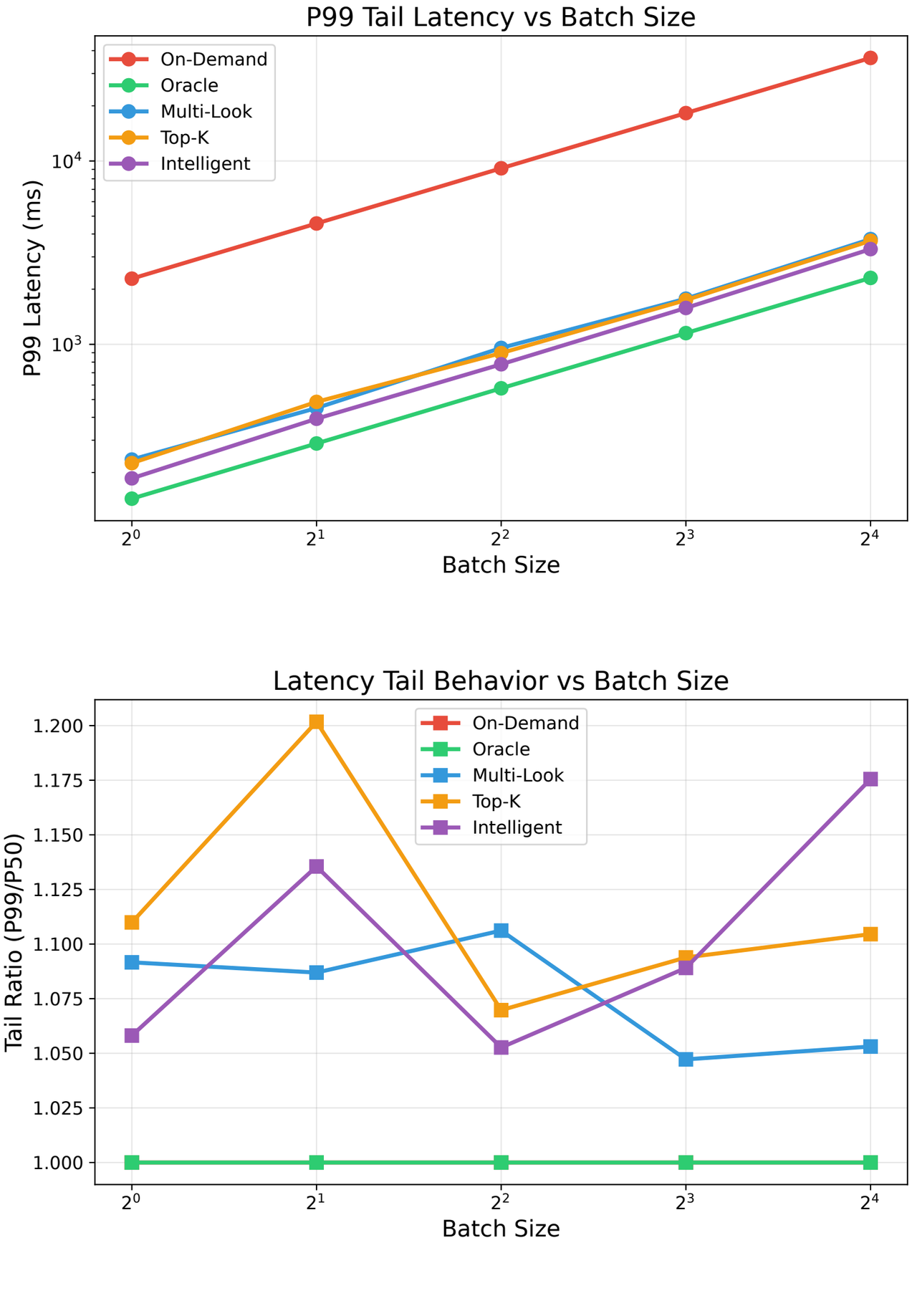
Switch Transformer Prefetching Performance Analysis



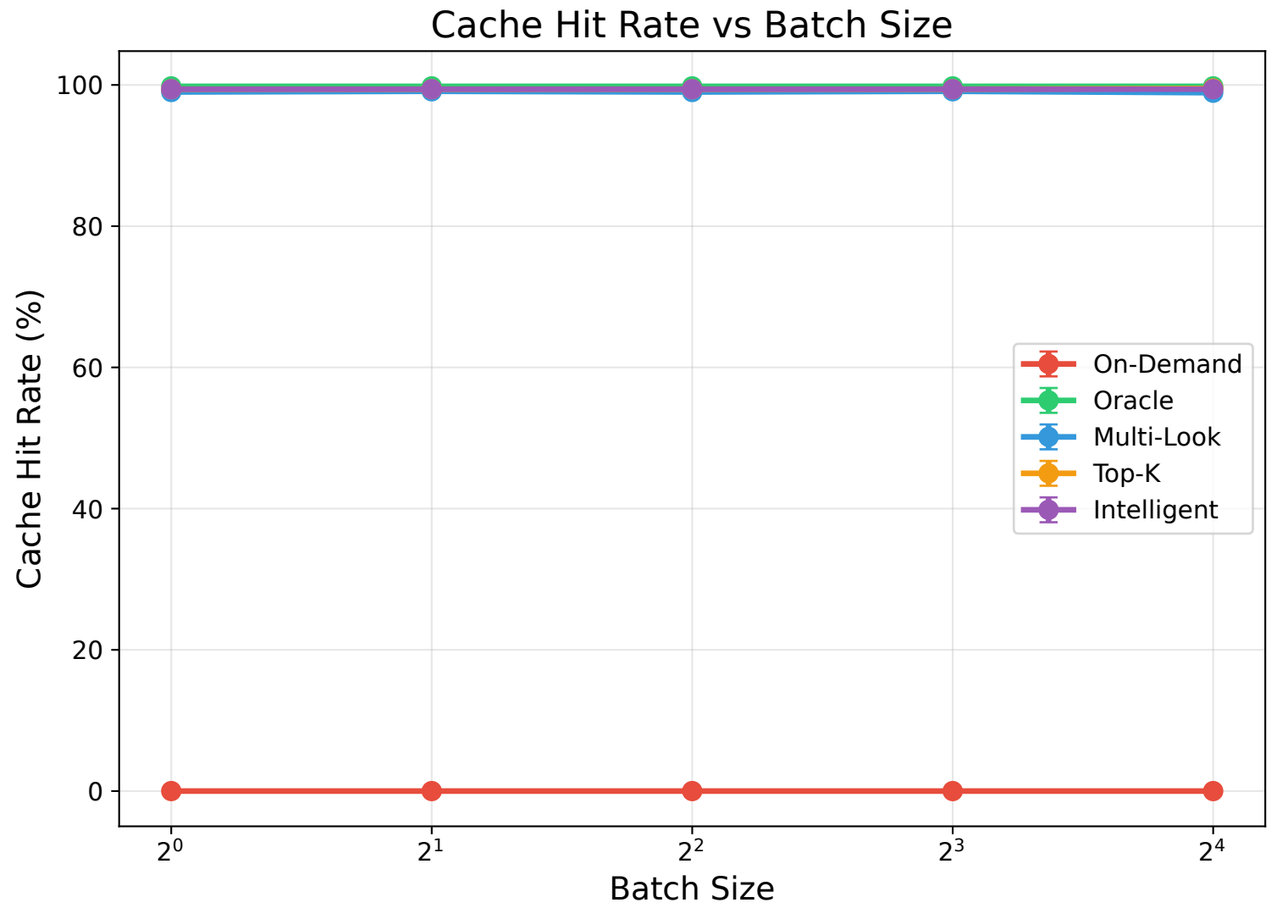


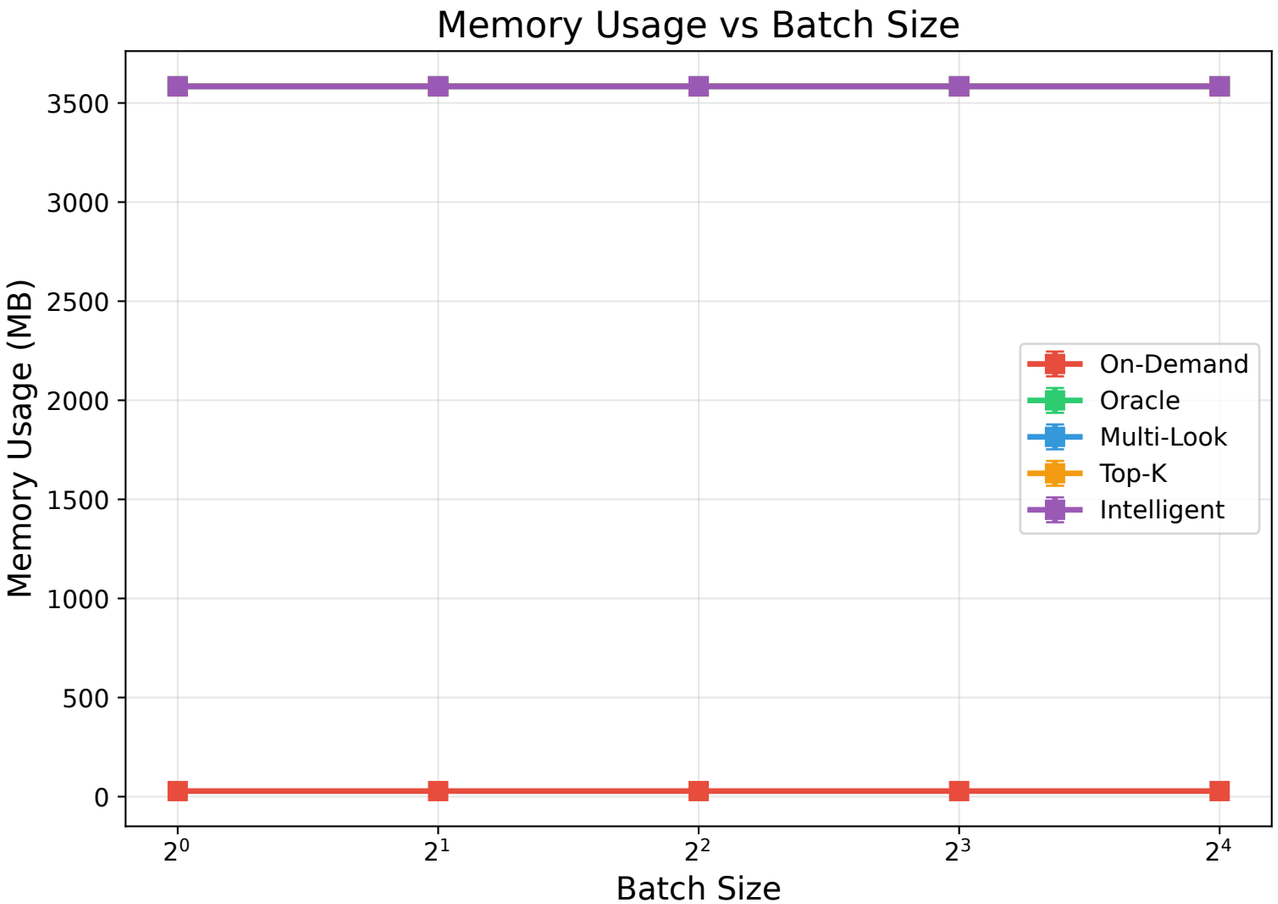
Comprehensive Tail Latency Analysis

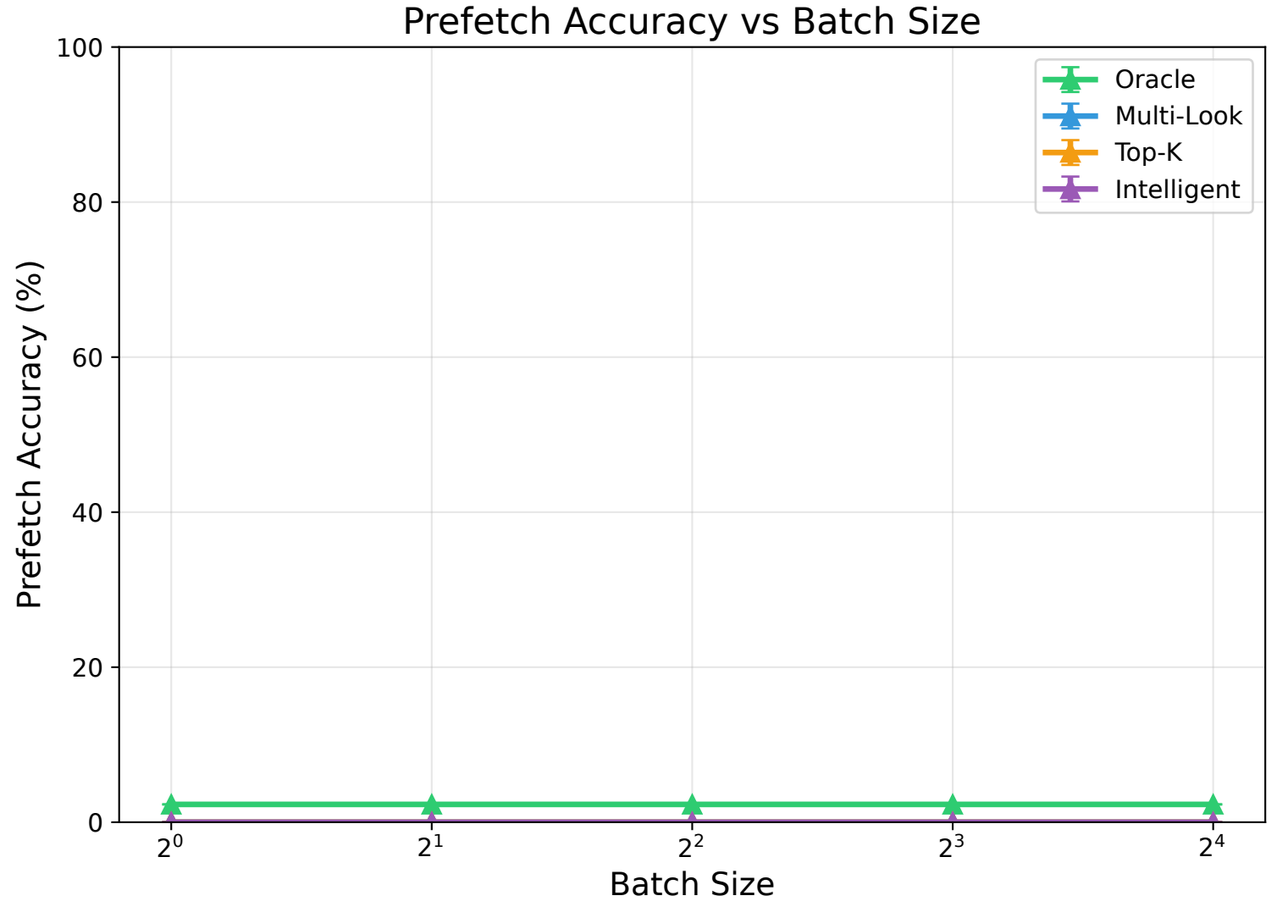


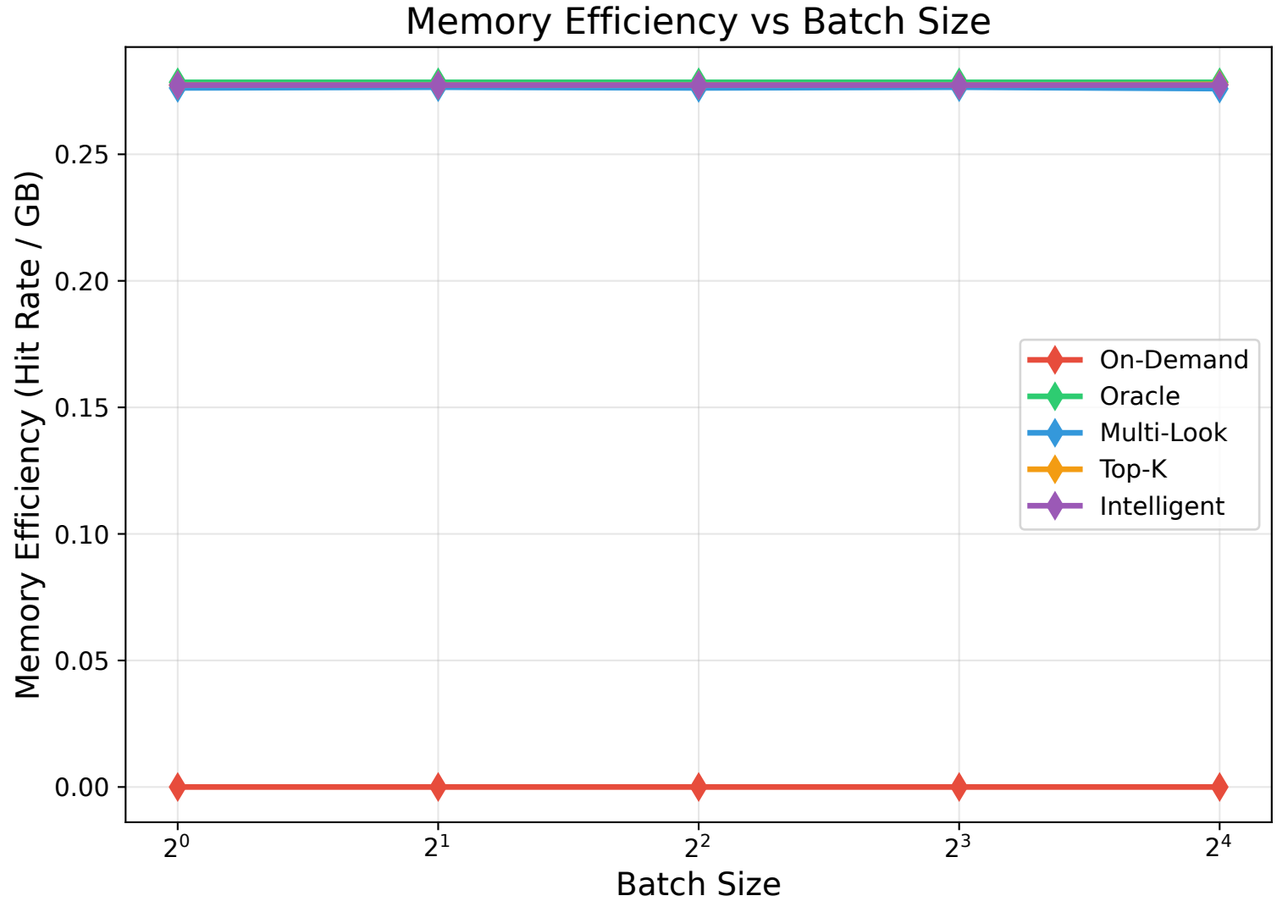


Cache Performance and Memory Analysis



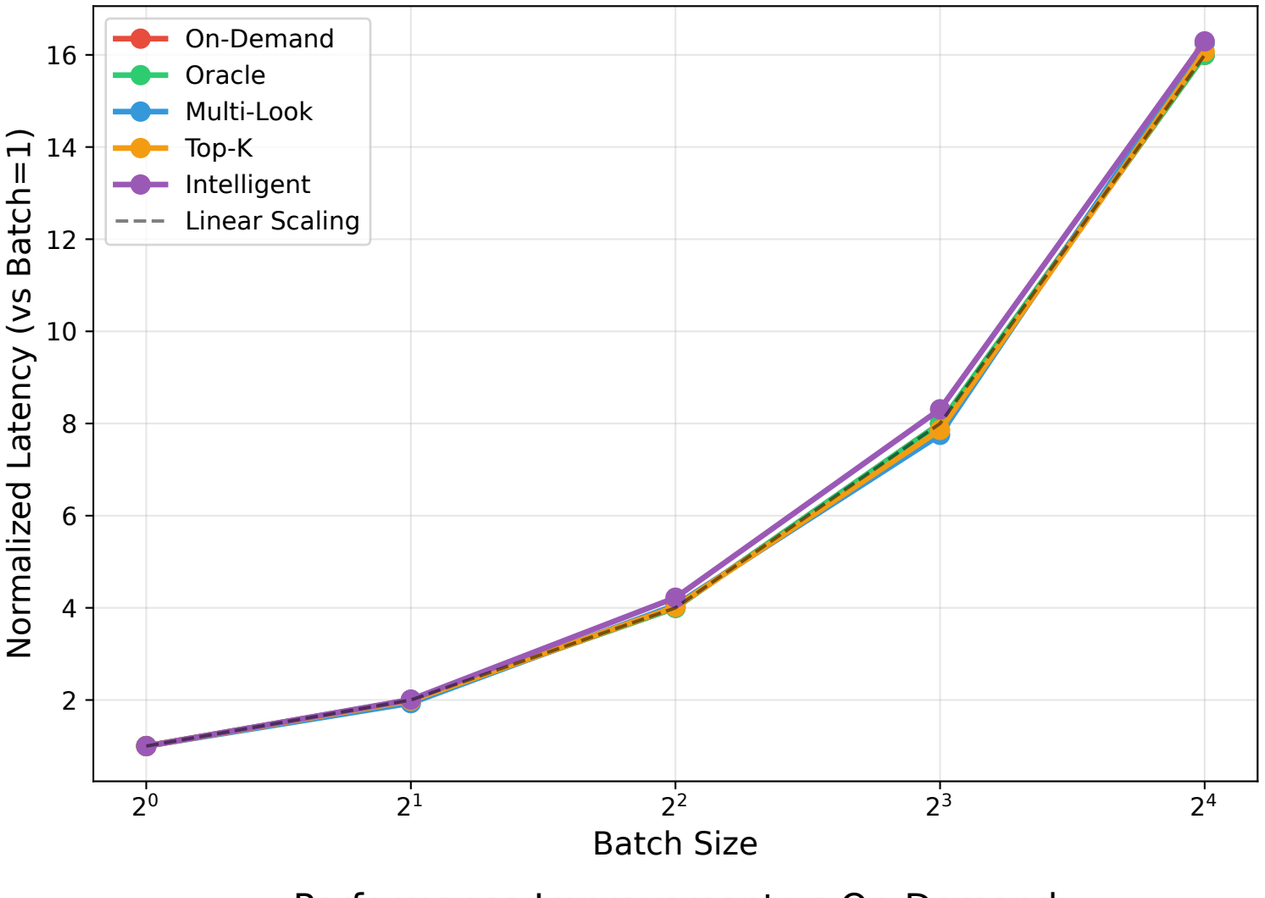




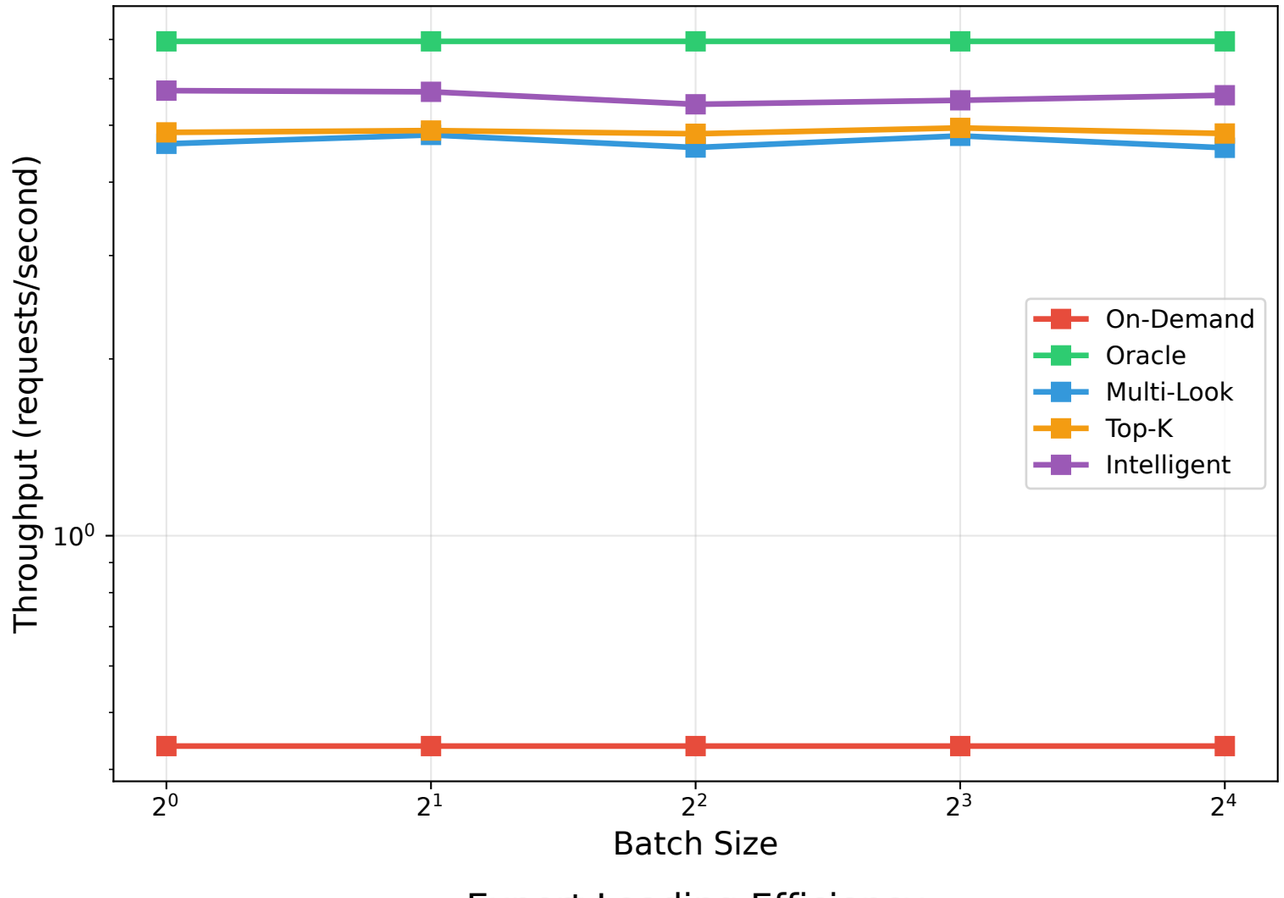


Scalability and Efficiency Analysis

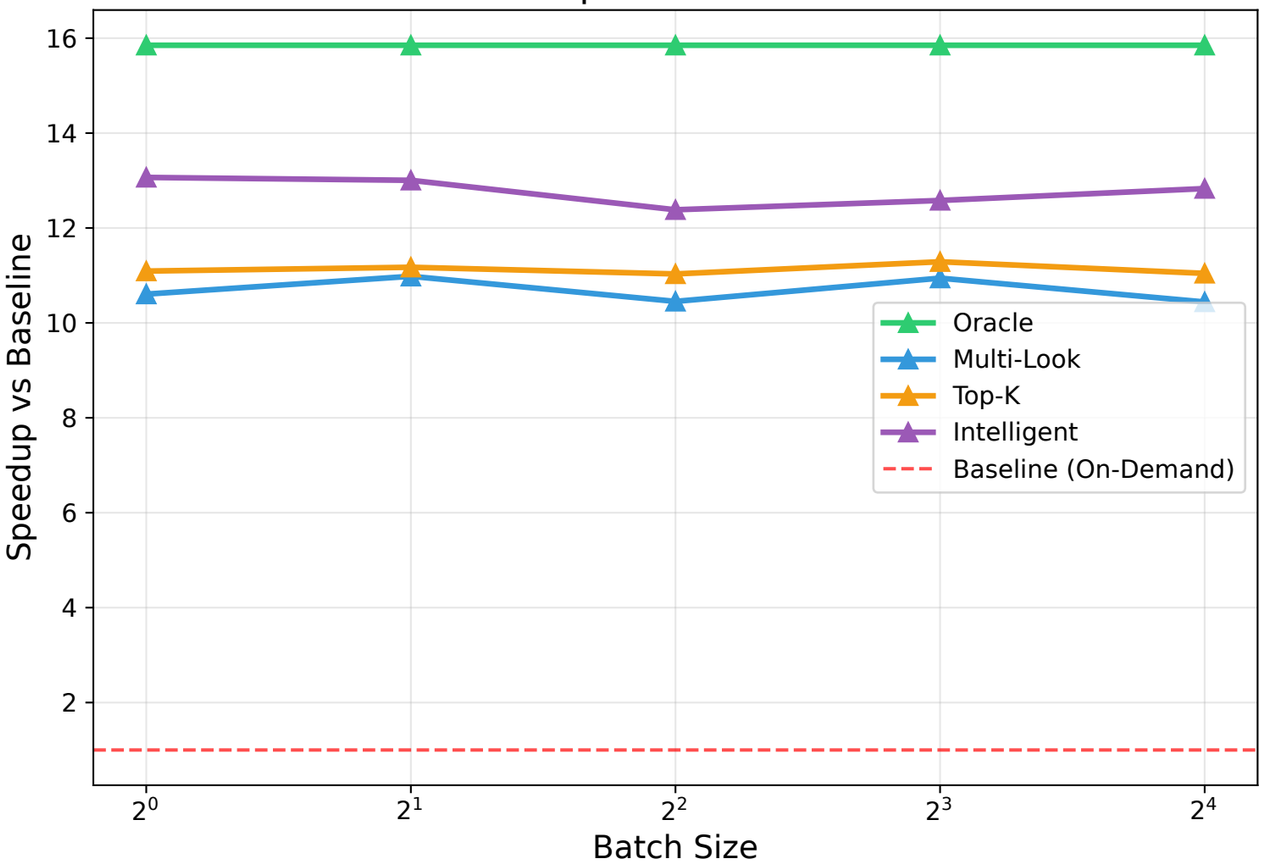
Latency Scalability



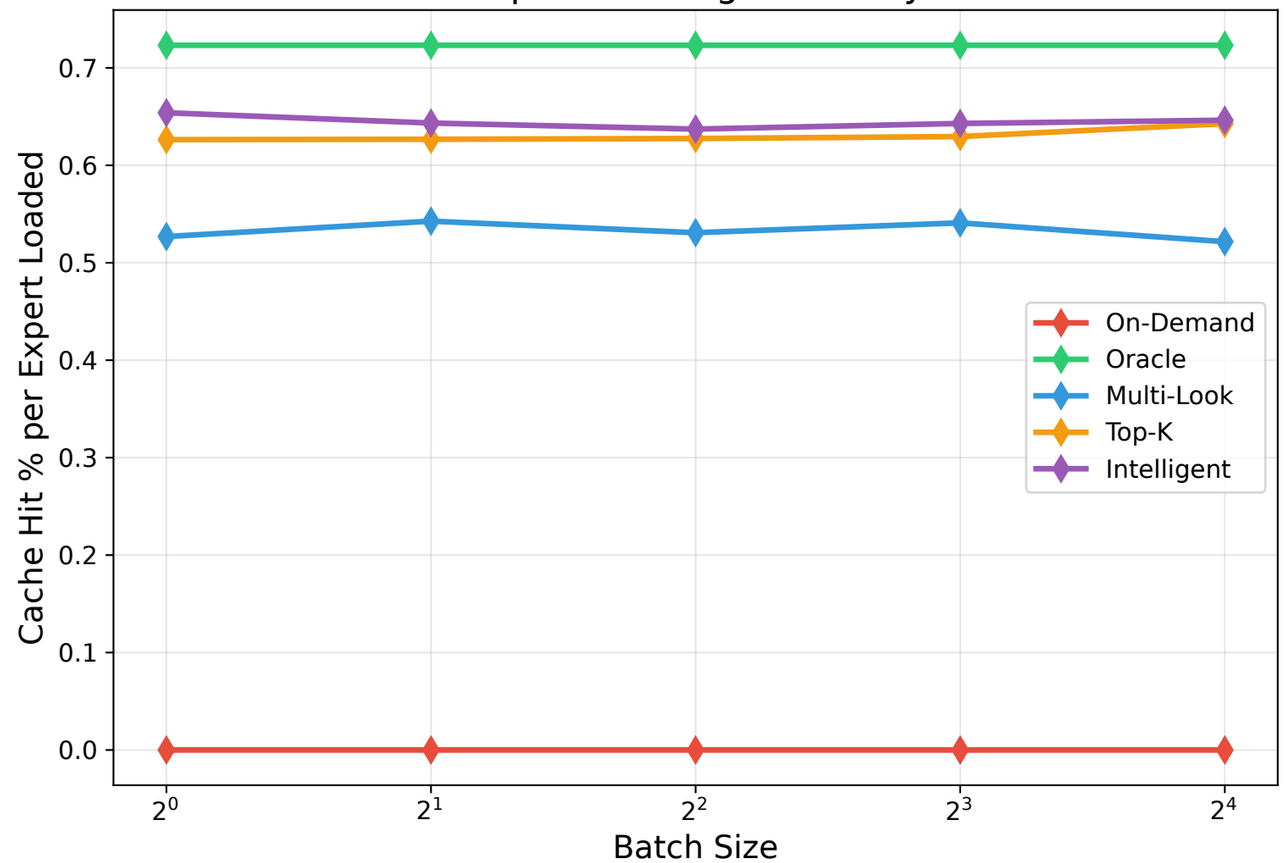
Throughput vs Batch Size



Performance Improvement vs On-Demand



Expert Loading Efficiency



Statistical Analysis and Significance