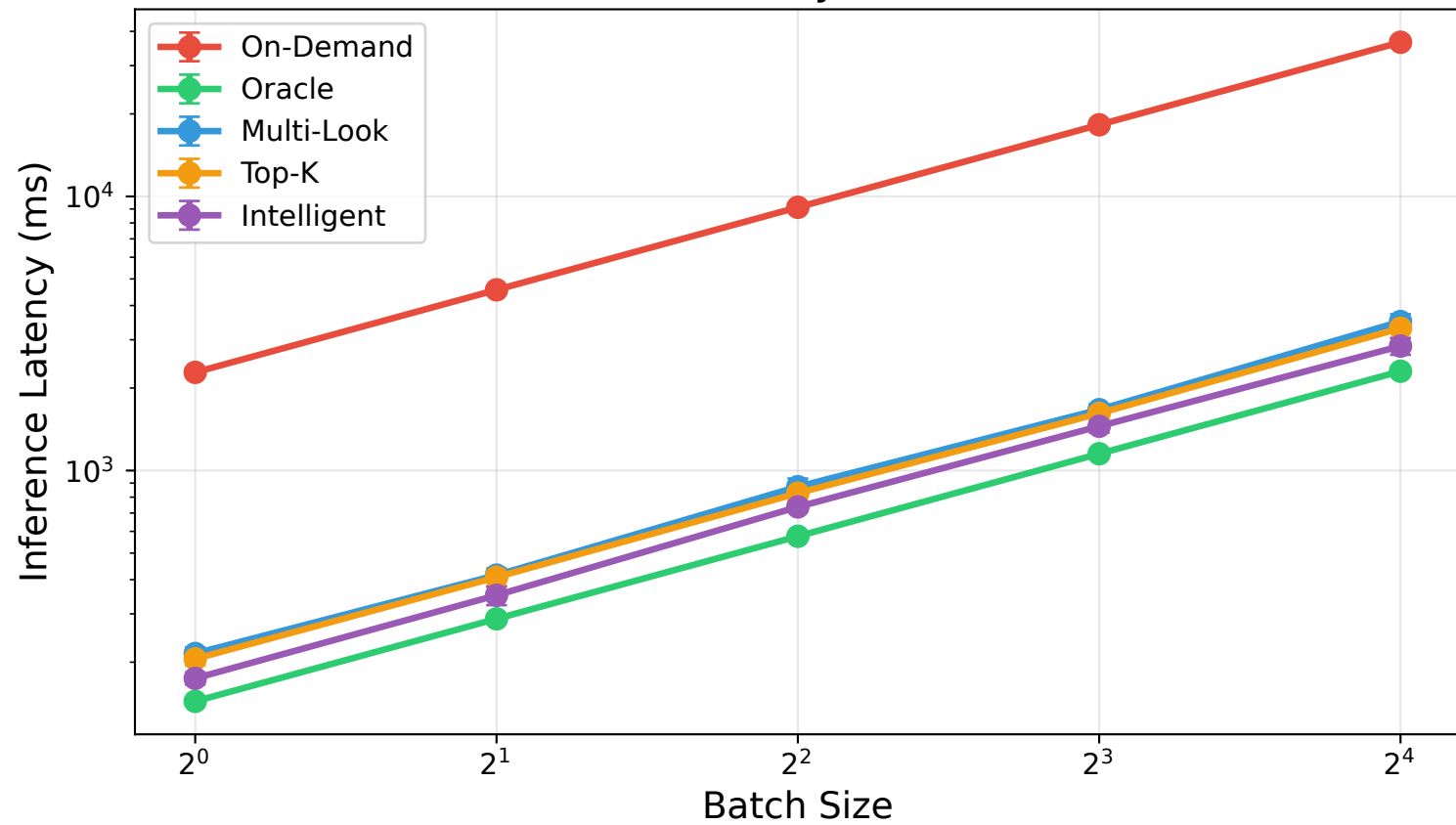
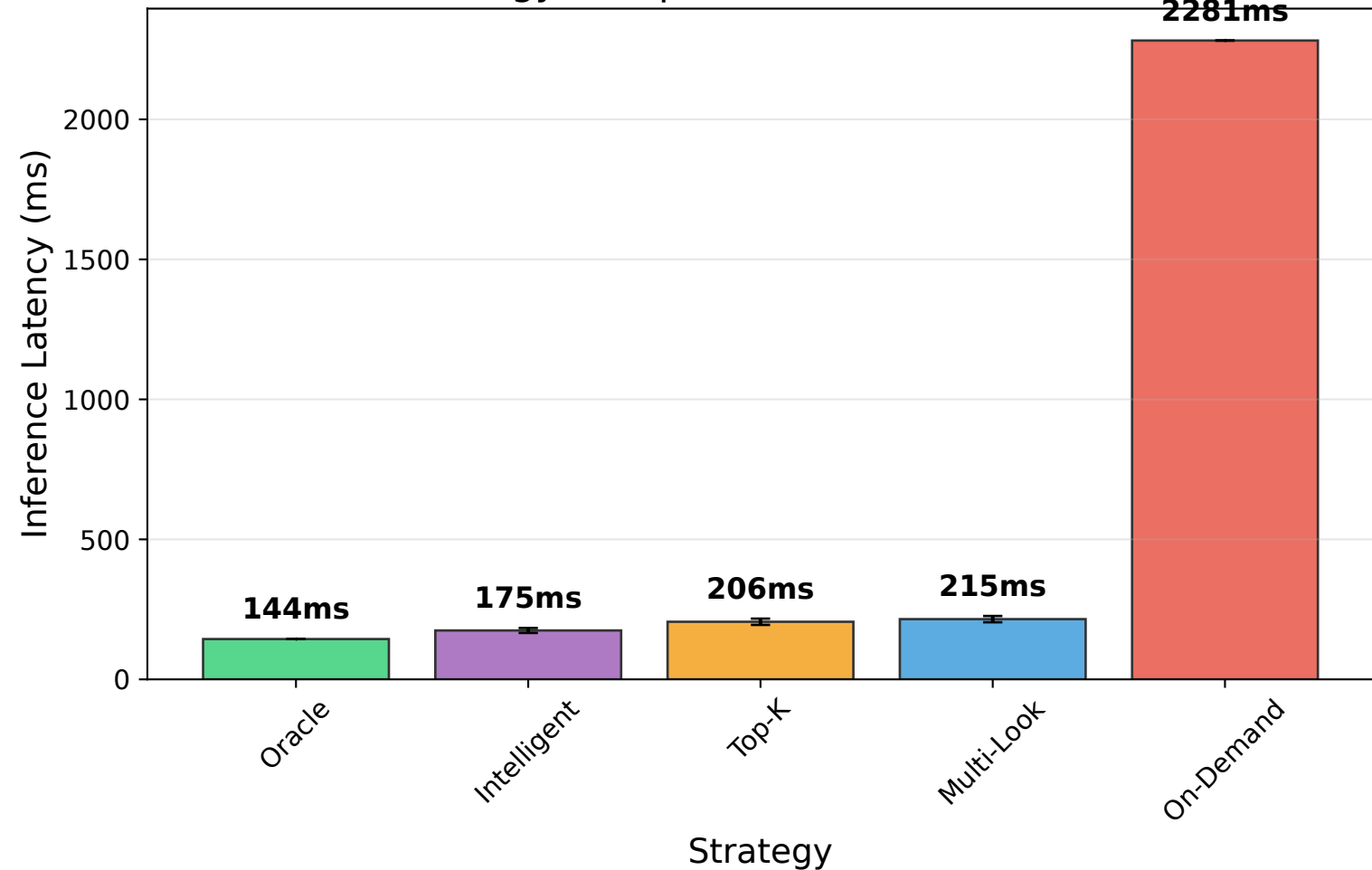


Switch Transformer Prefetching Performance Analysis

Inference Latency vs Batch Size

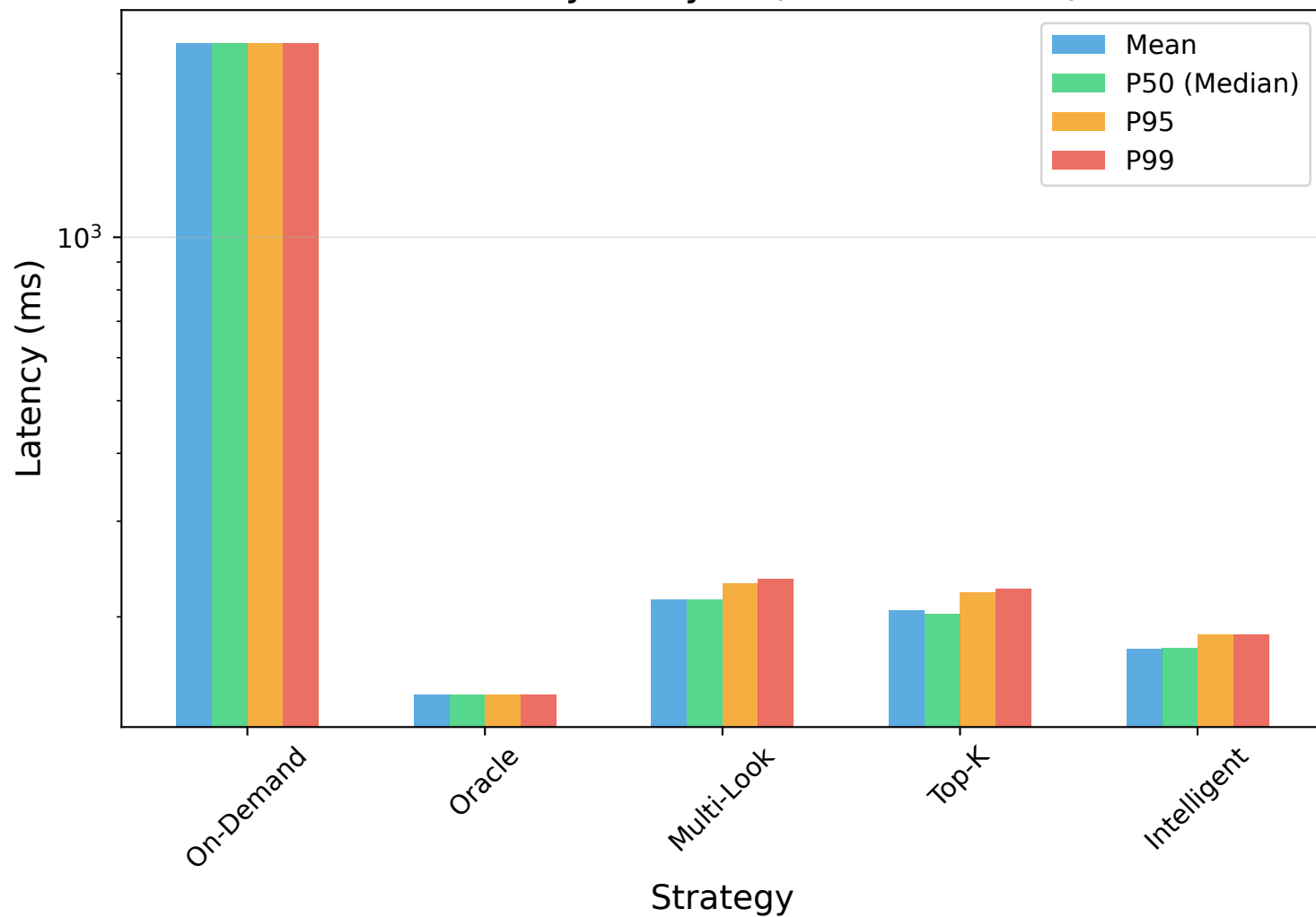


Strategy Comparison (Batch Size = 1)

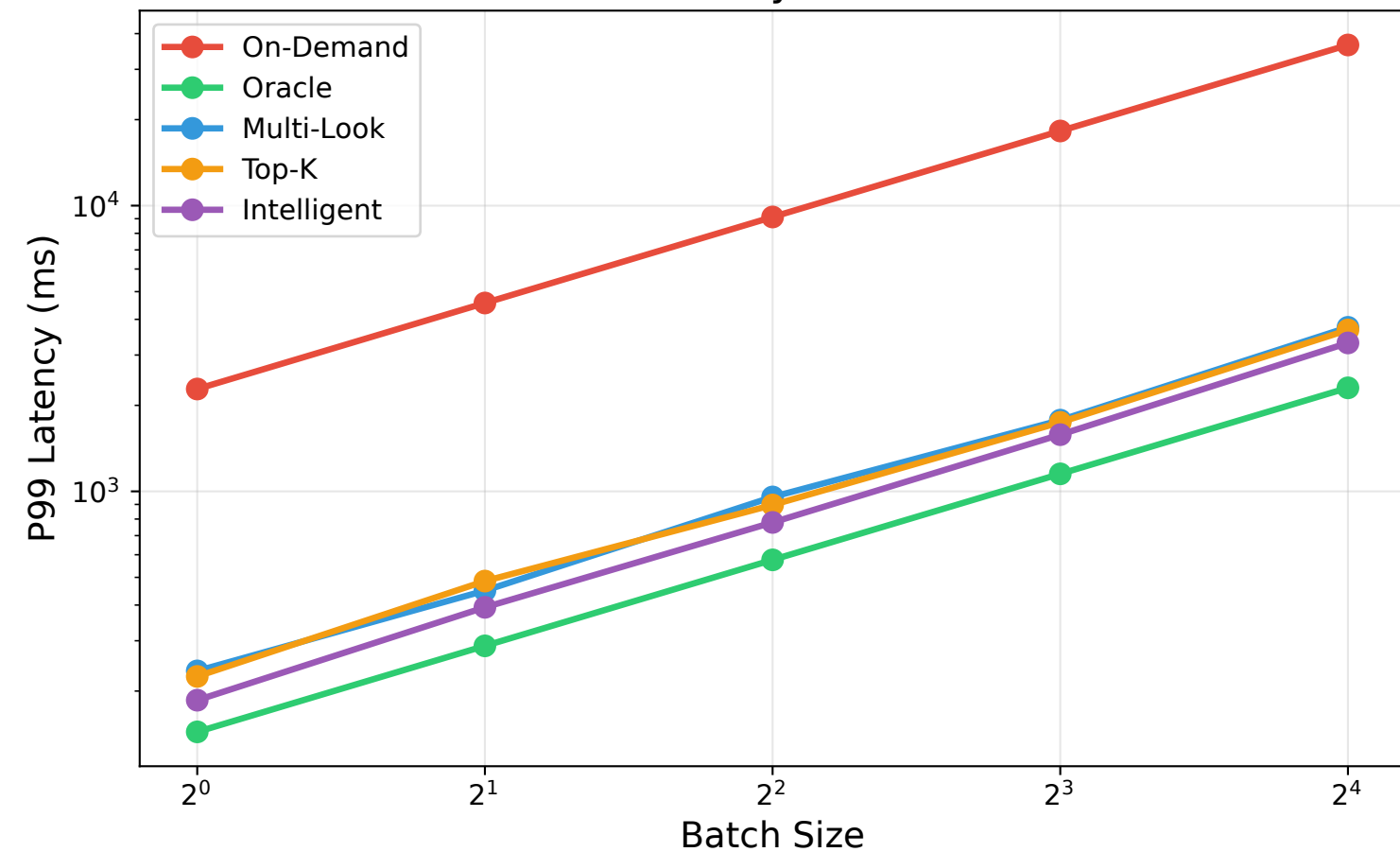


Comprehensive Tail Latency Analysis

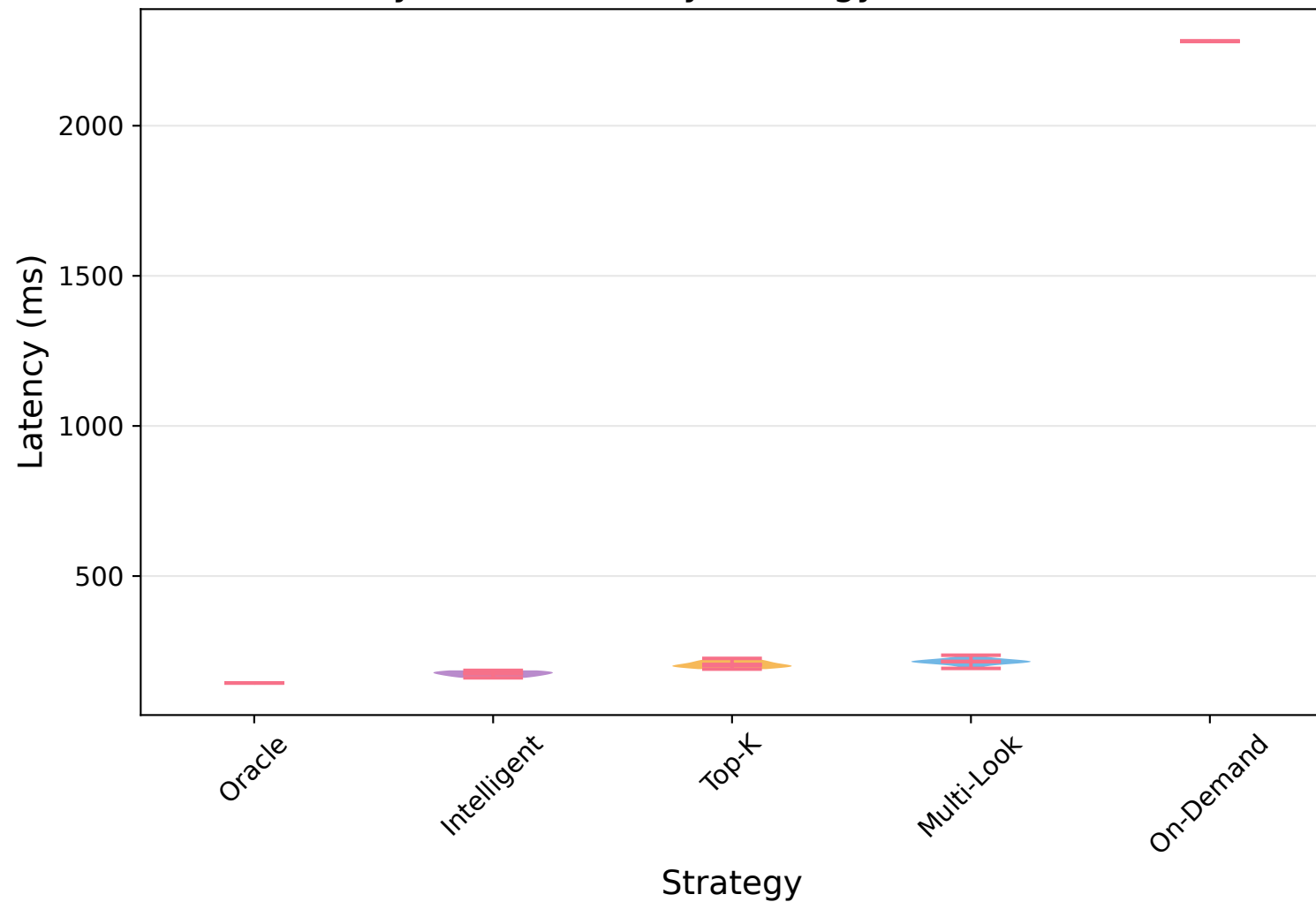
Tail Latency Analysis (Batch Size = 1)



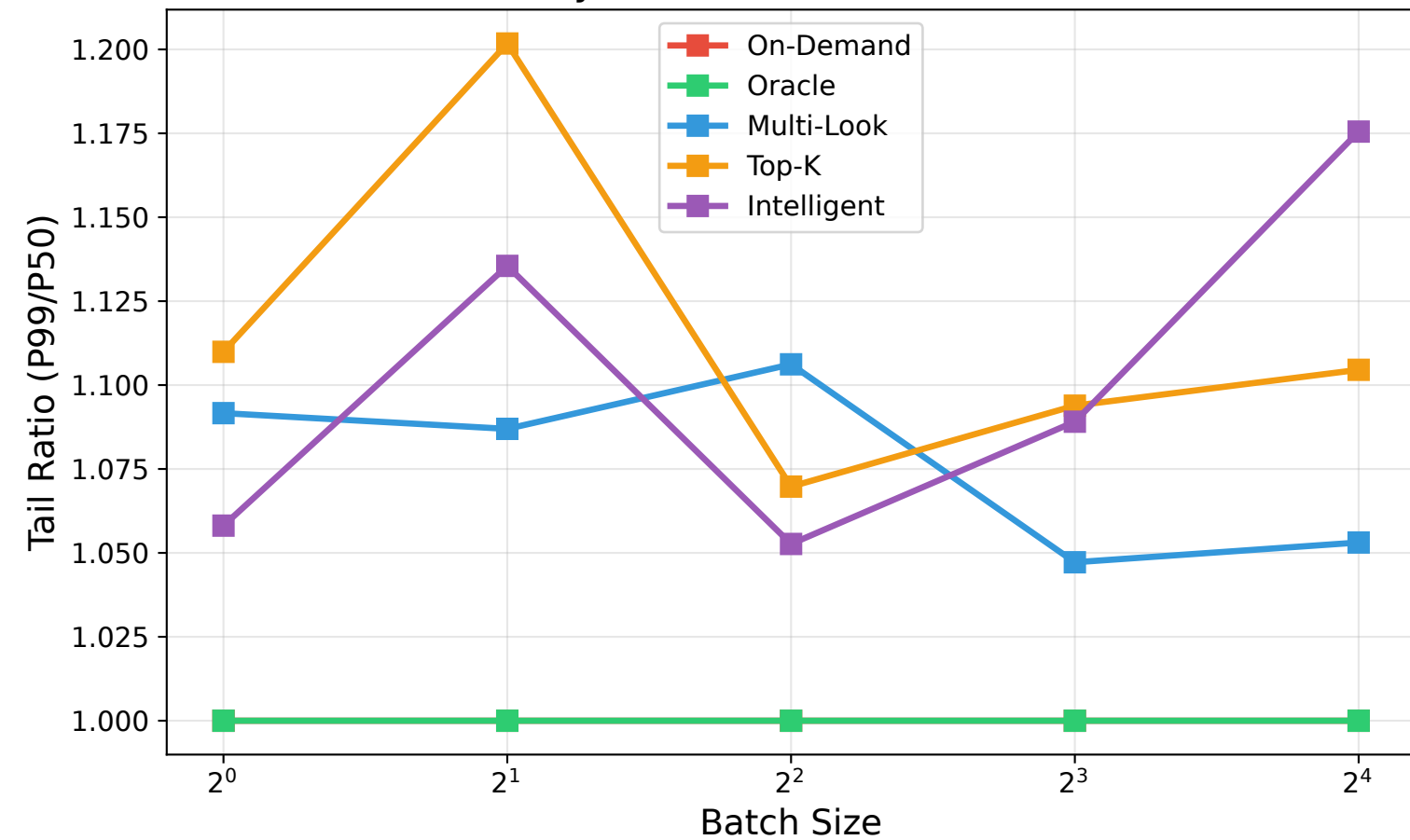
P99 Tail Latency vs Batch Size



Latency Distribution by Strategy (Batch Size = 1)

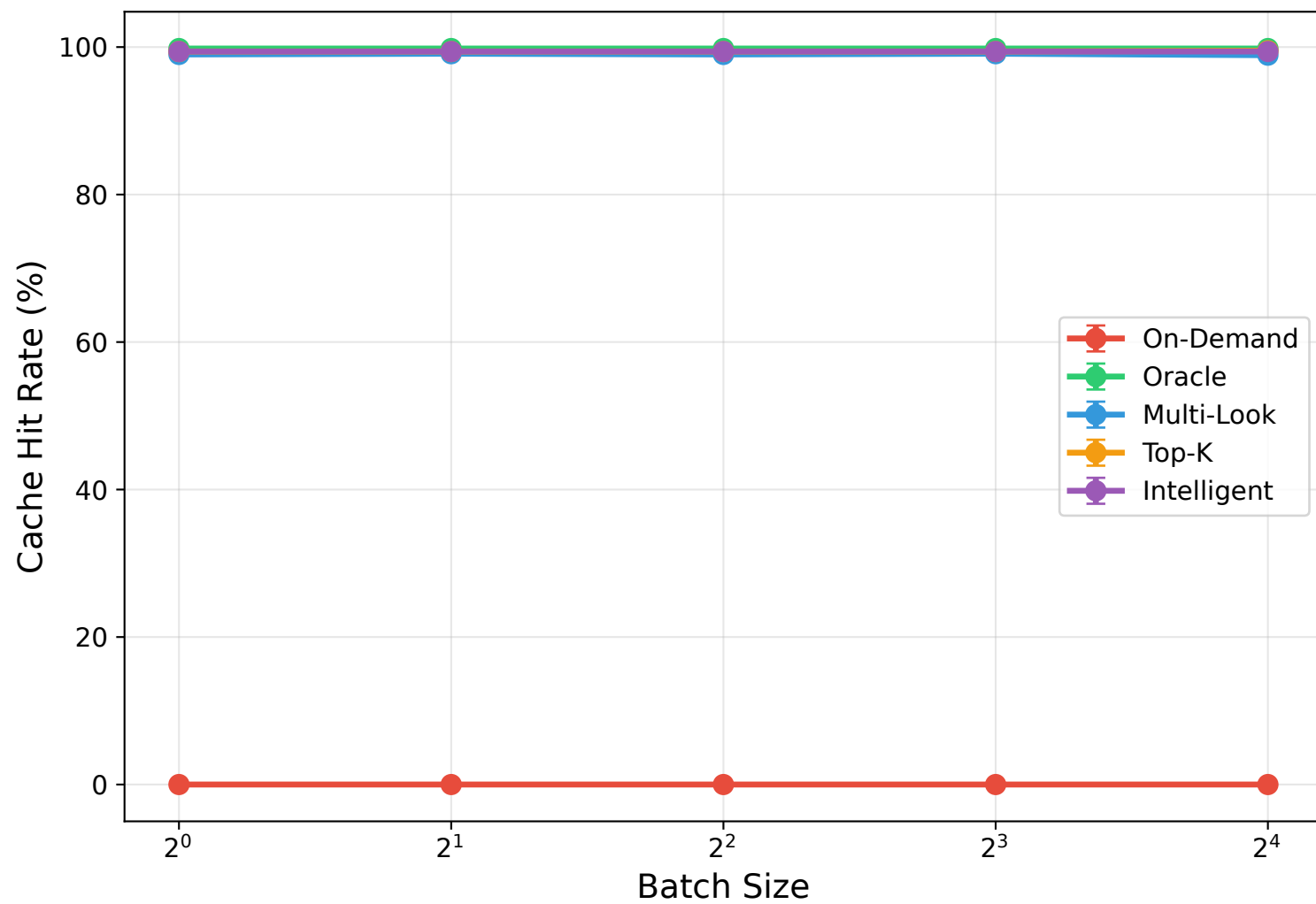


Latency Tail Behavior vs Batch Size

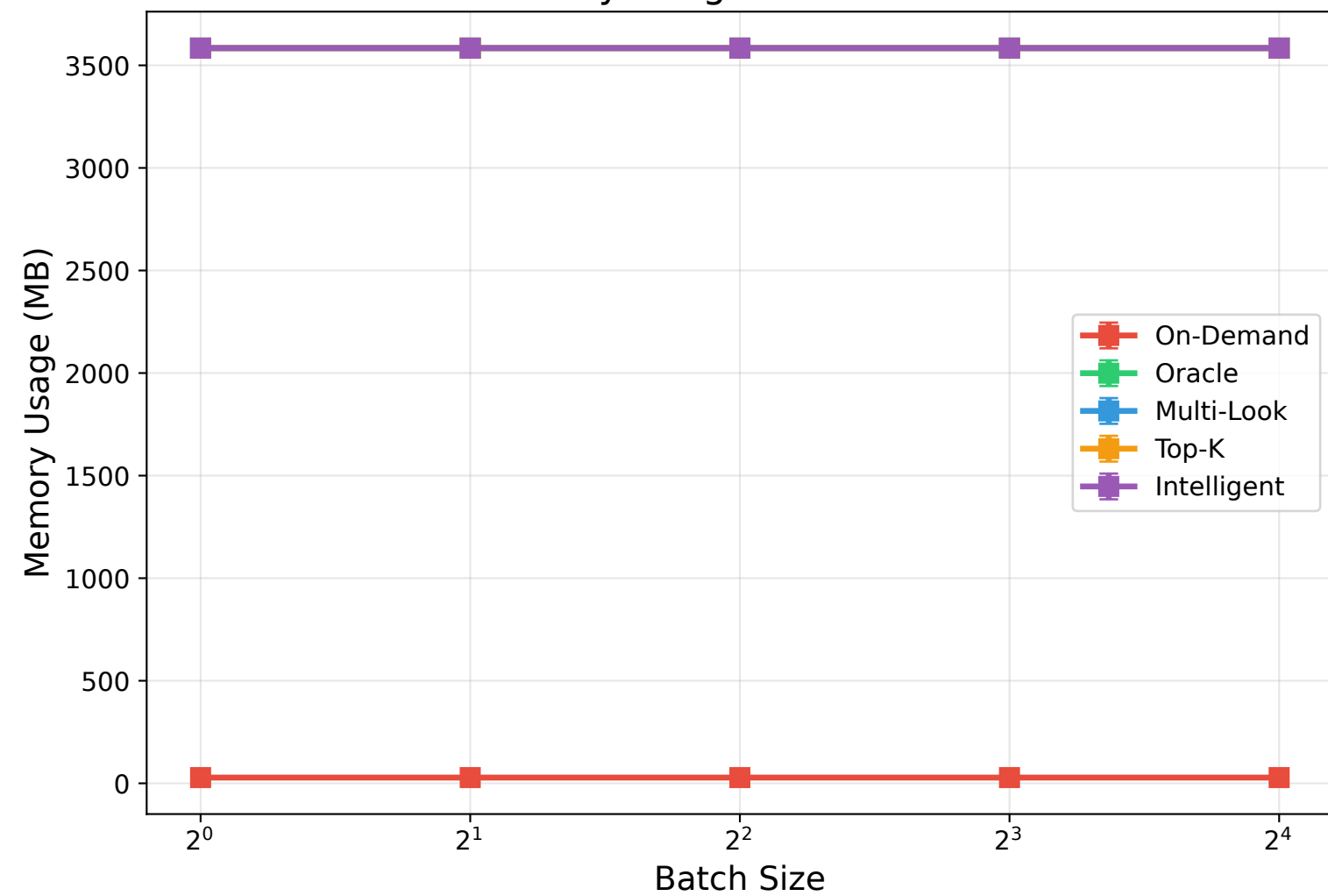


Cache Performance and Memory Analysis

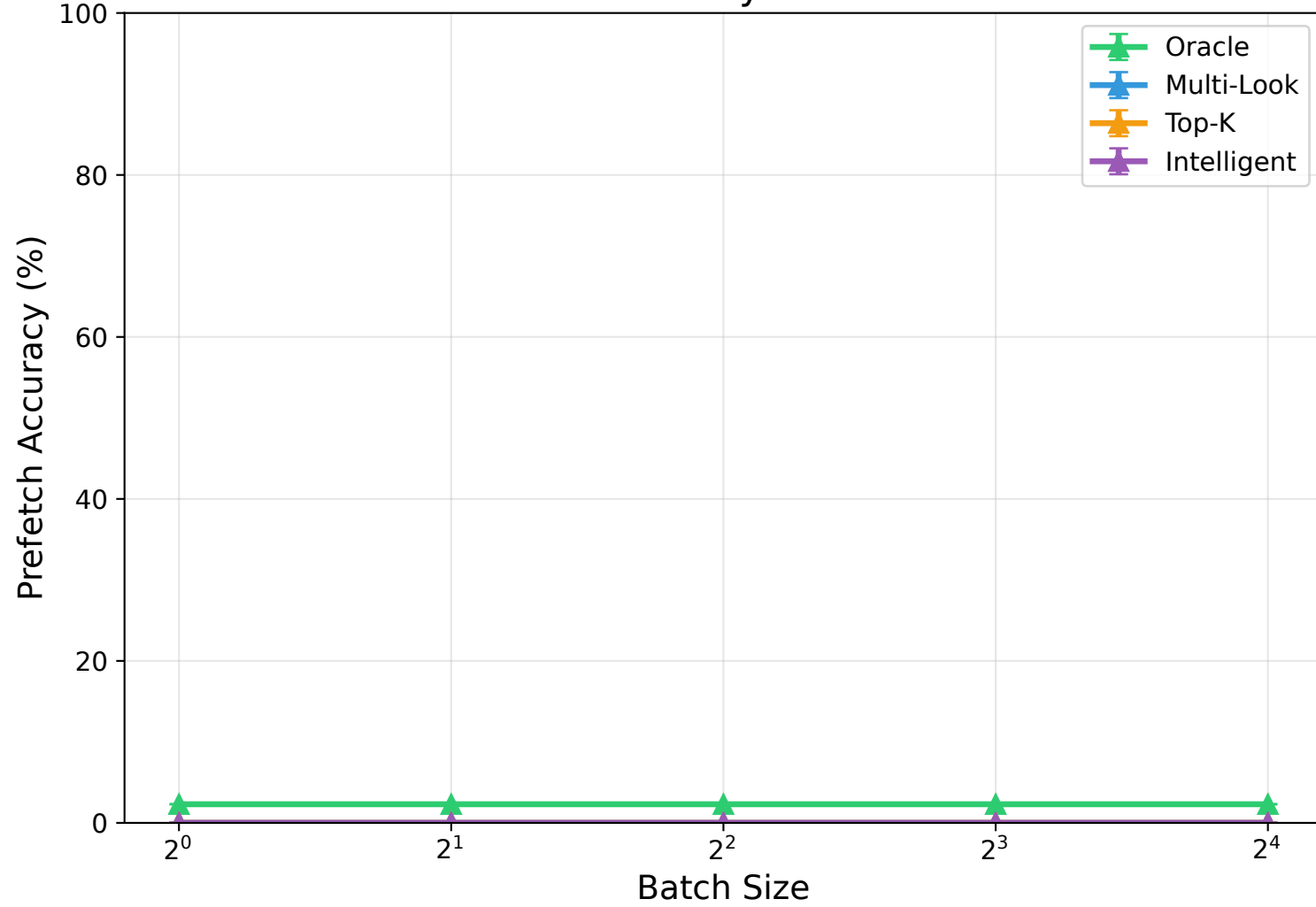
Cache Hit Rate vs Batch Size



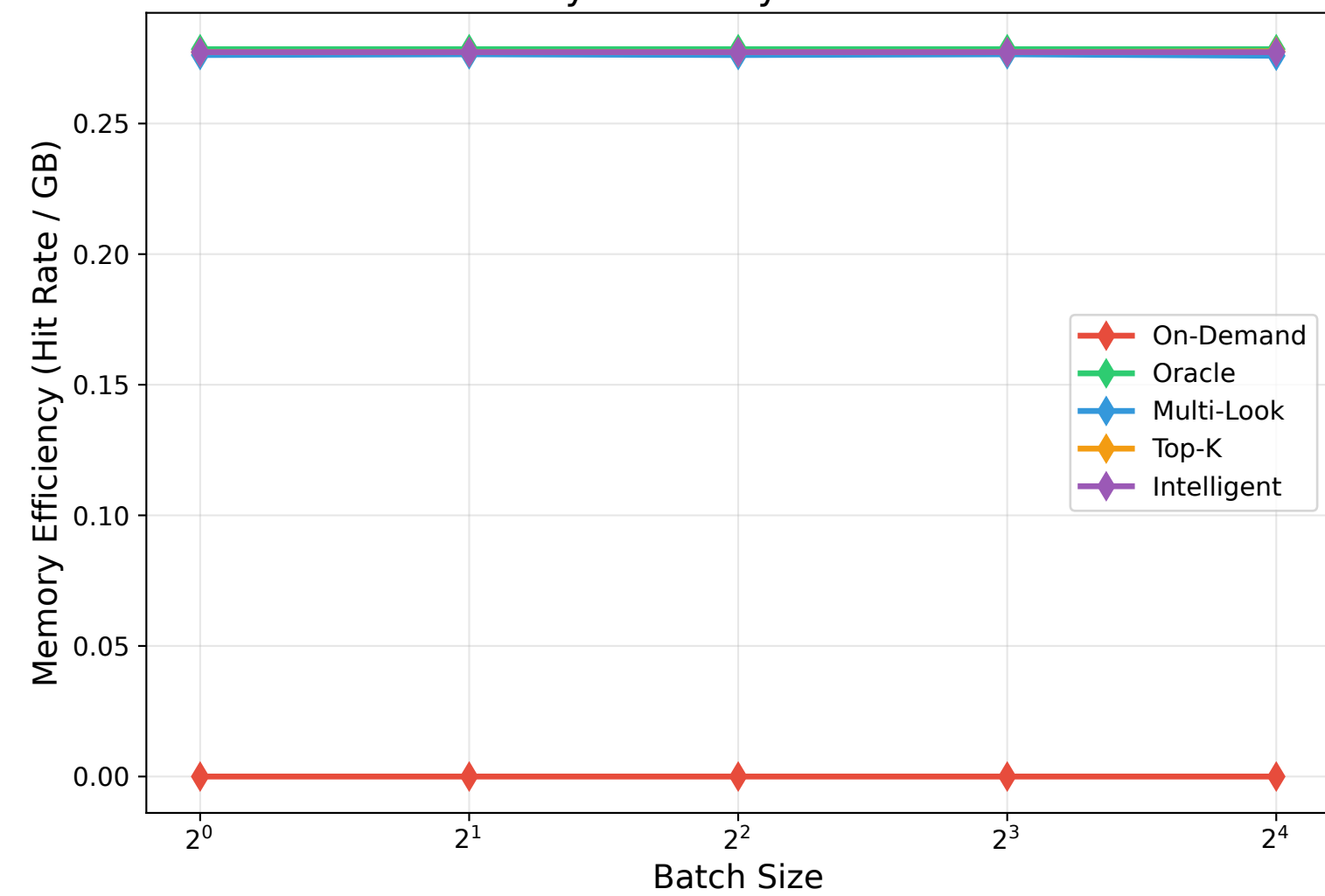
Memory Usage vs Batch Size



Prefetch Accuracy vs Batch Size

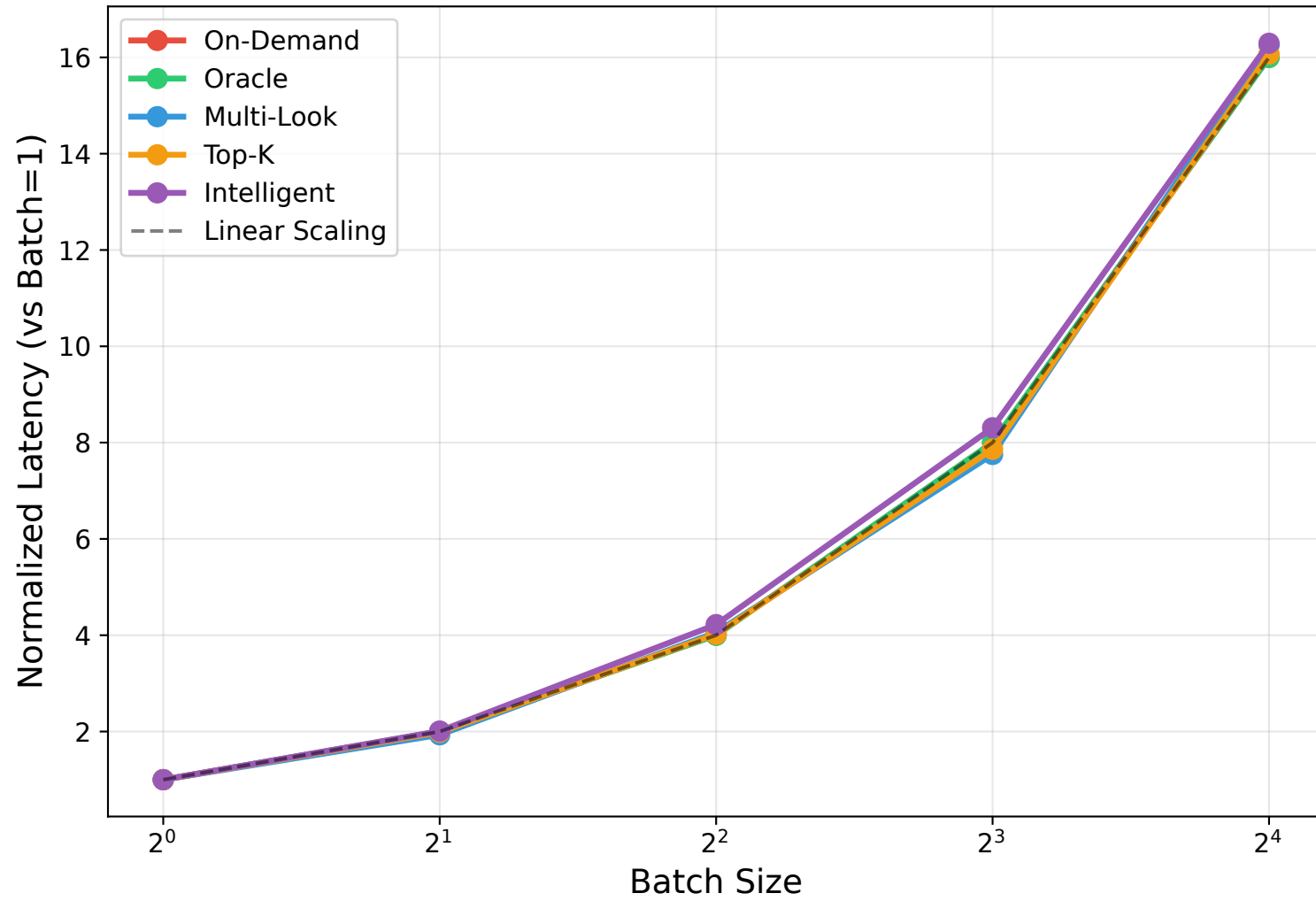


Memory Efficiency vs Batch Size

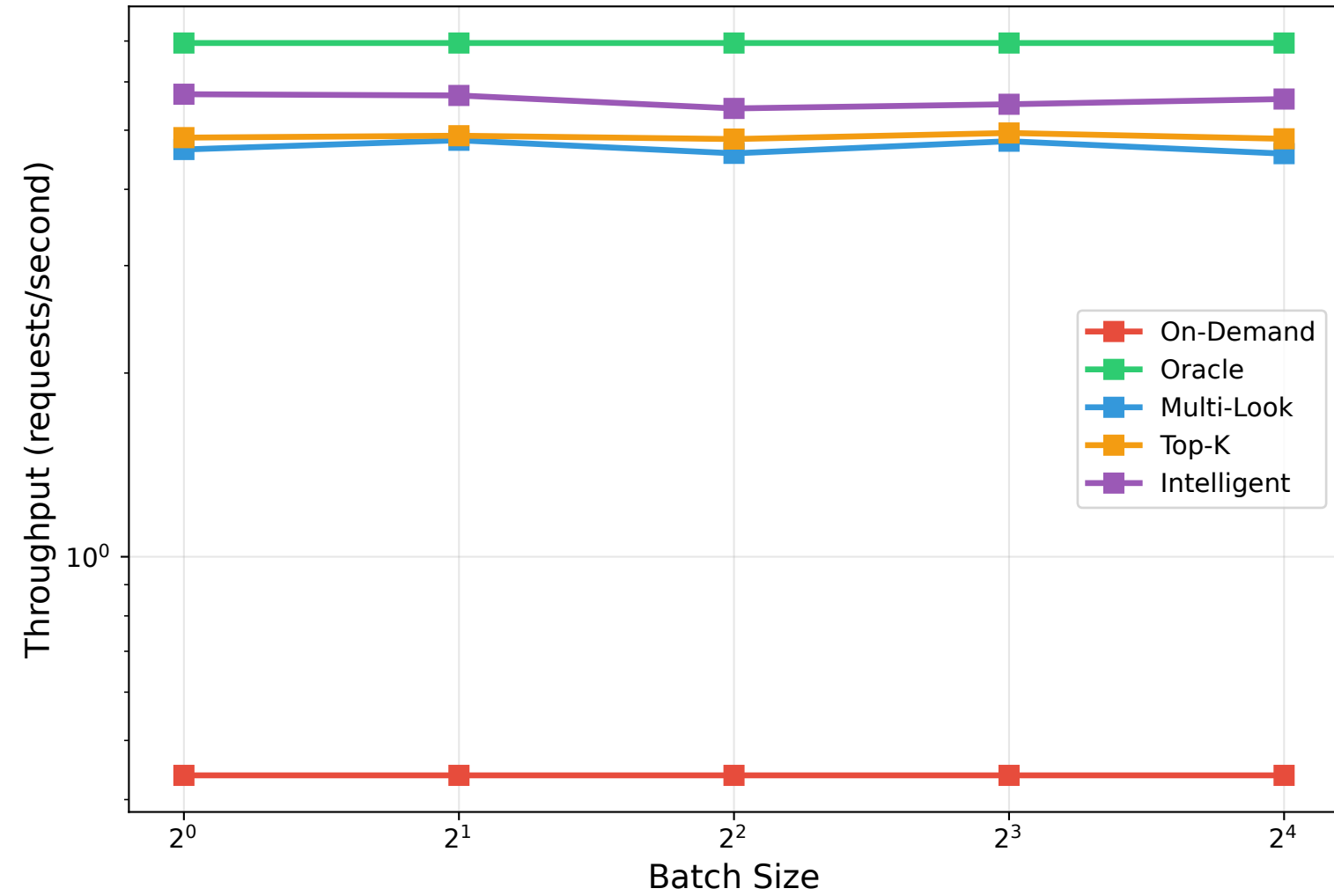


Scalability and Efficiency Analysis

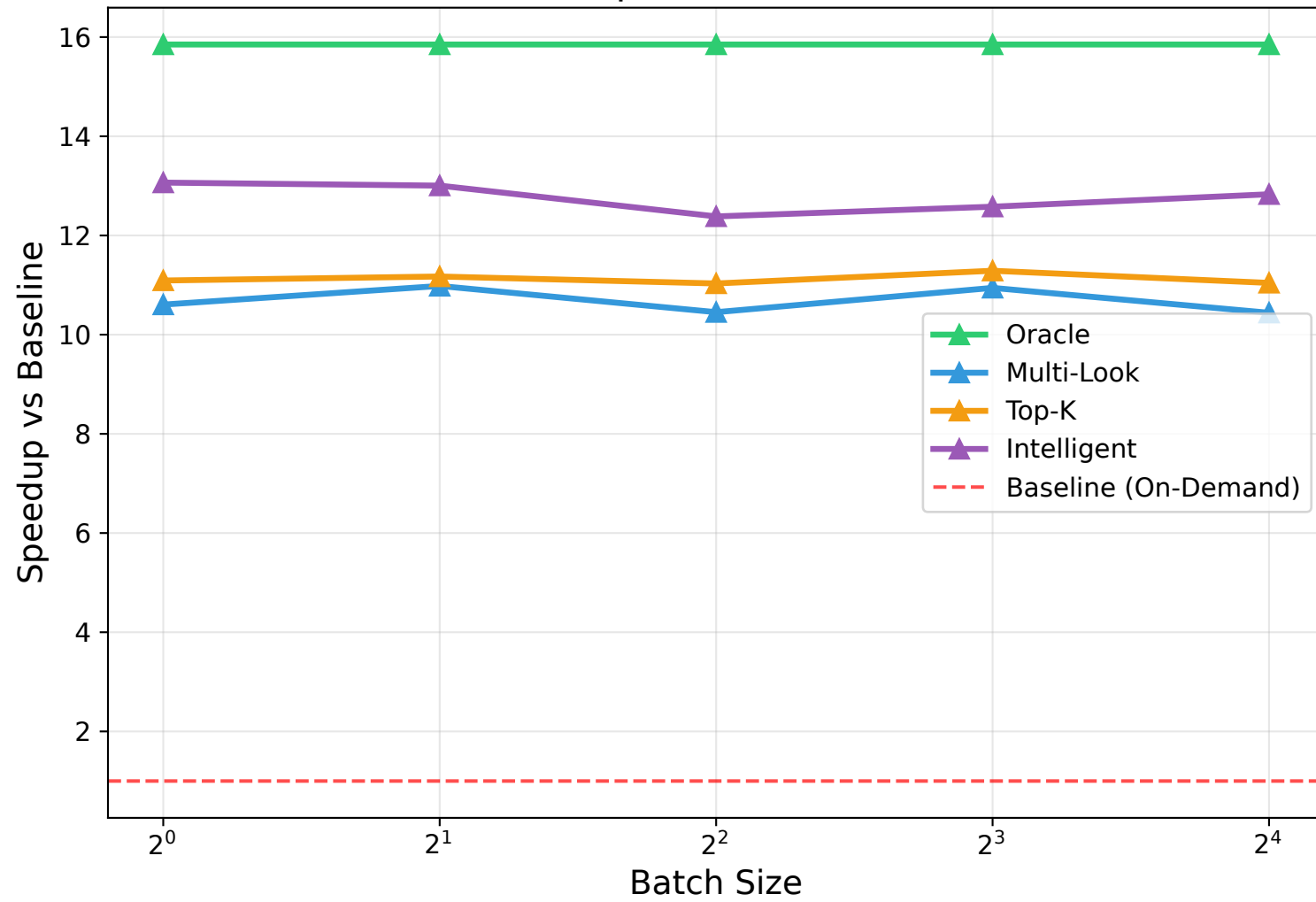
Latency Scalability



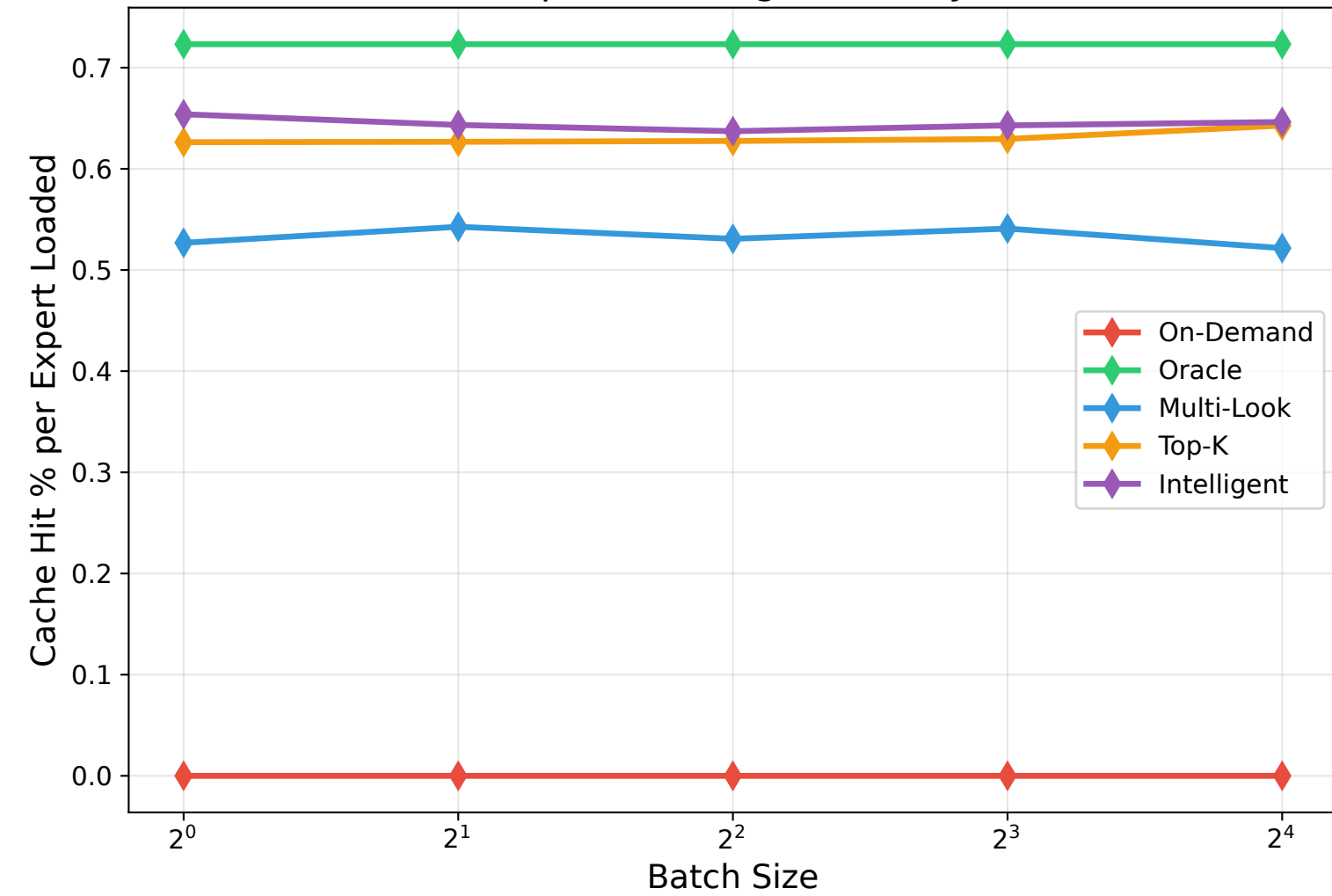
Throughput vs Batch Size



Performance Improvement vs On-Demand

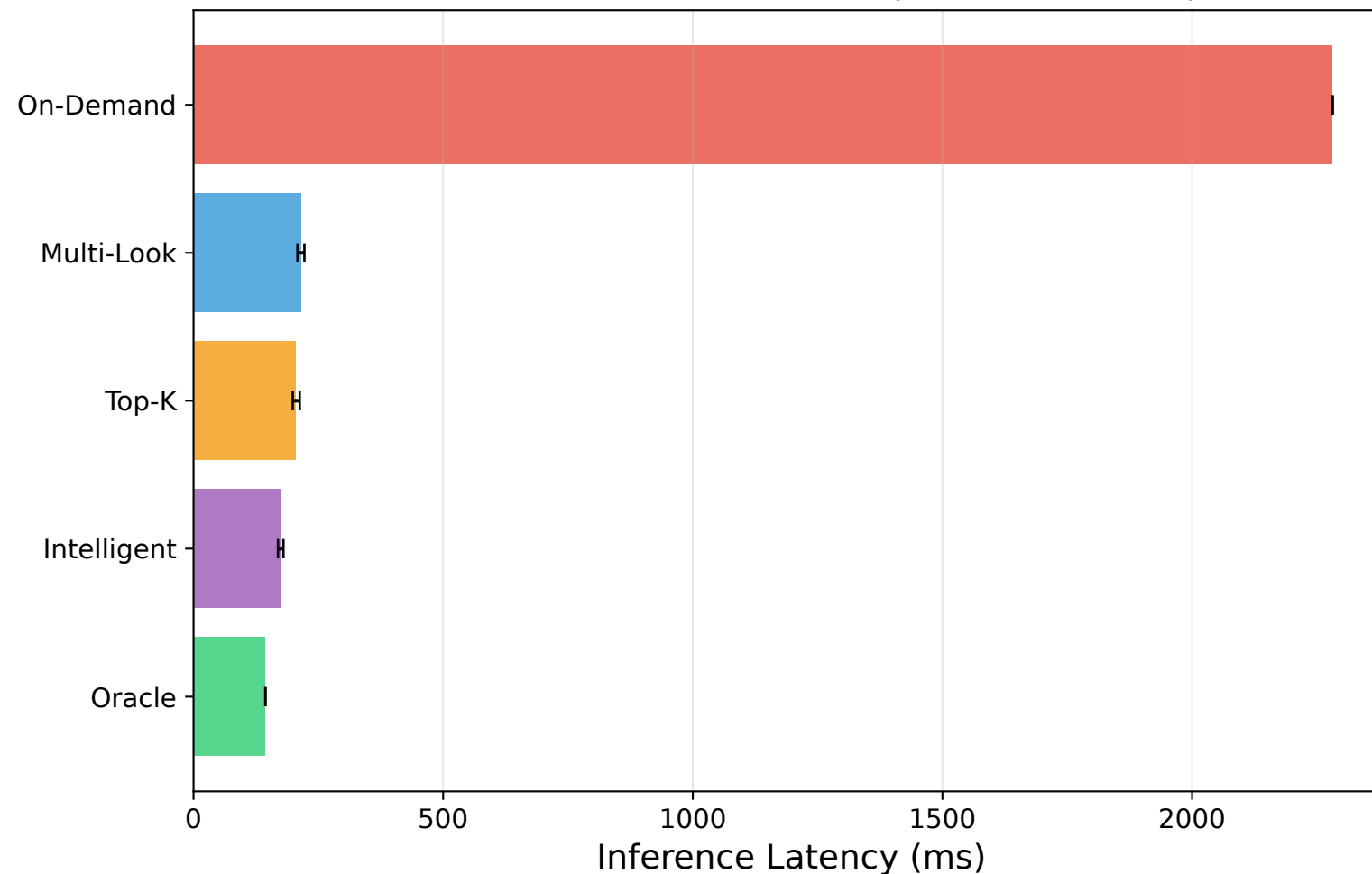


Expert Loading Efficiency

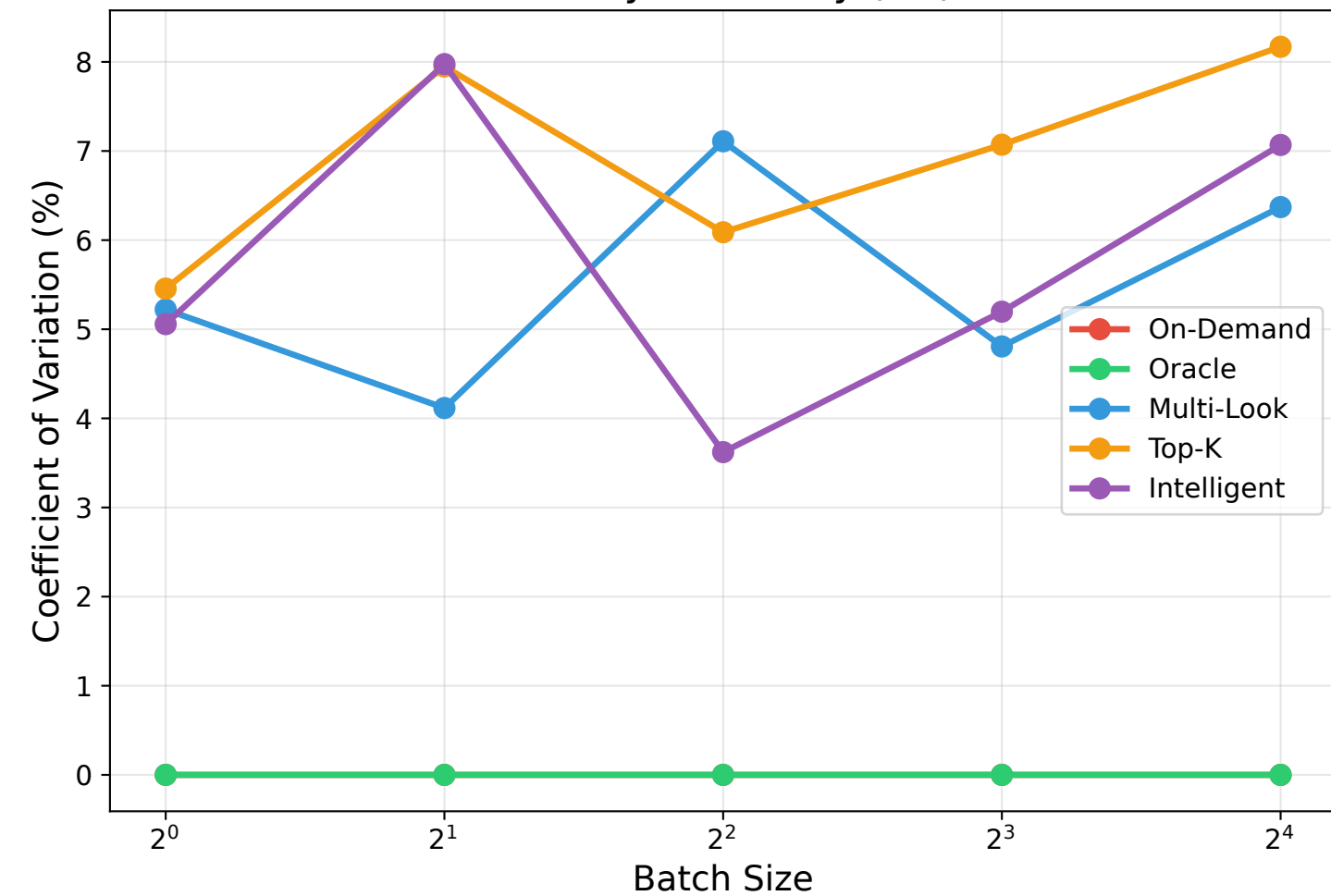


Statistical Analysis and Significance

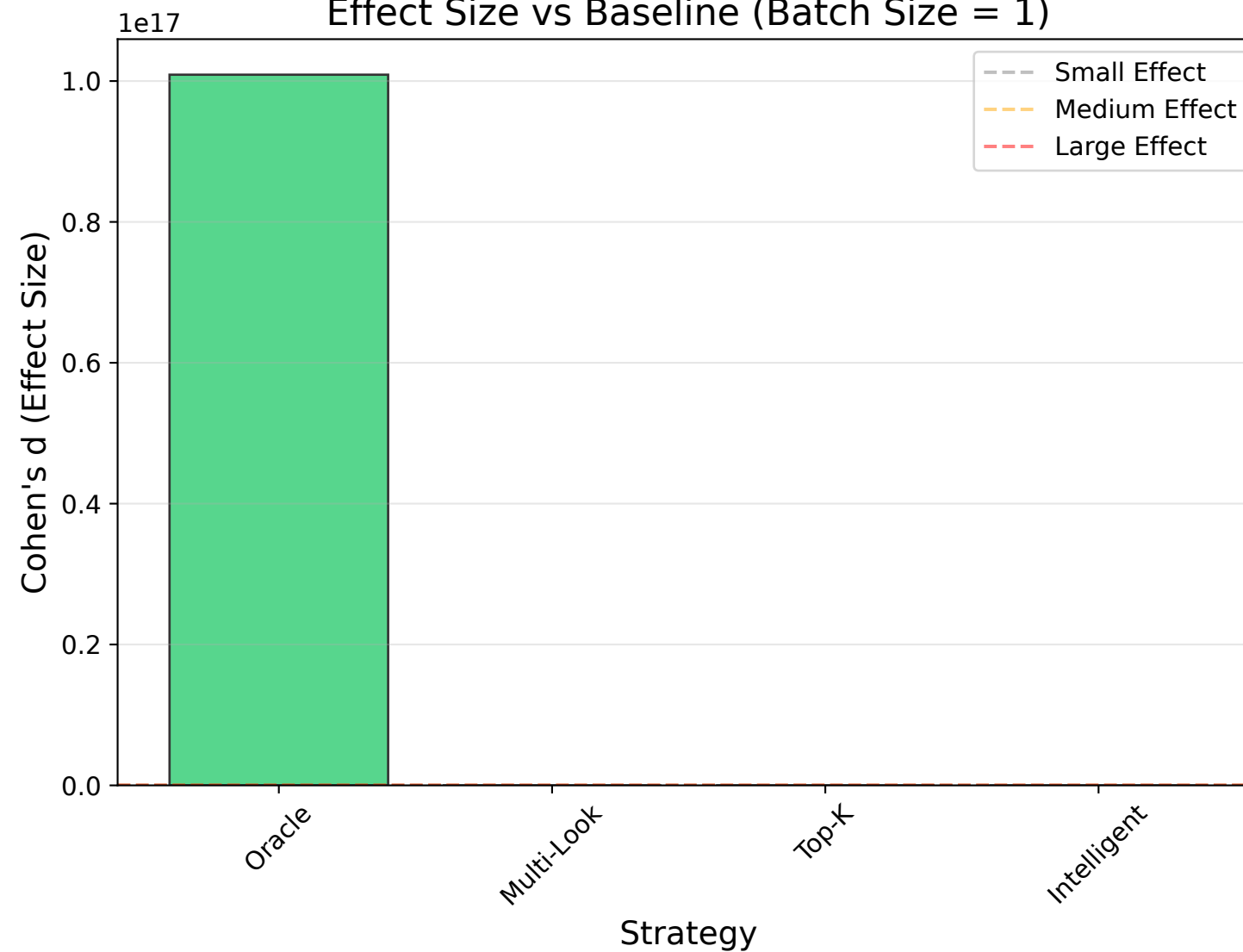
95% Confidence Intervals (Batch Size = 1)



Latency Variability (CV)



Effect Size vs Baseline (Batch Size = 1)



Performance Consistency

