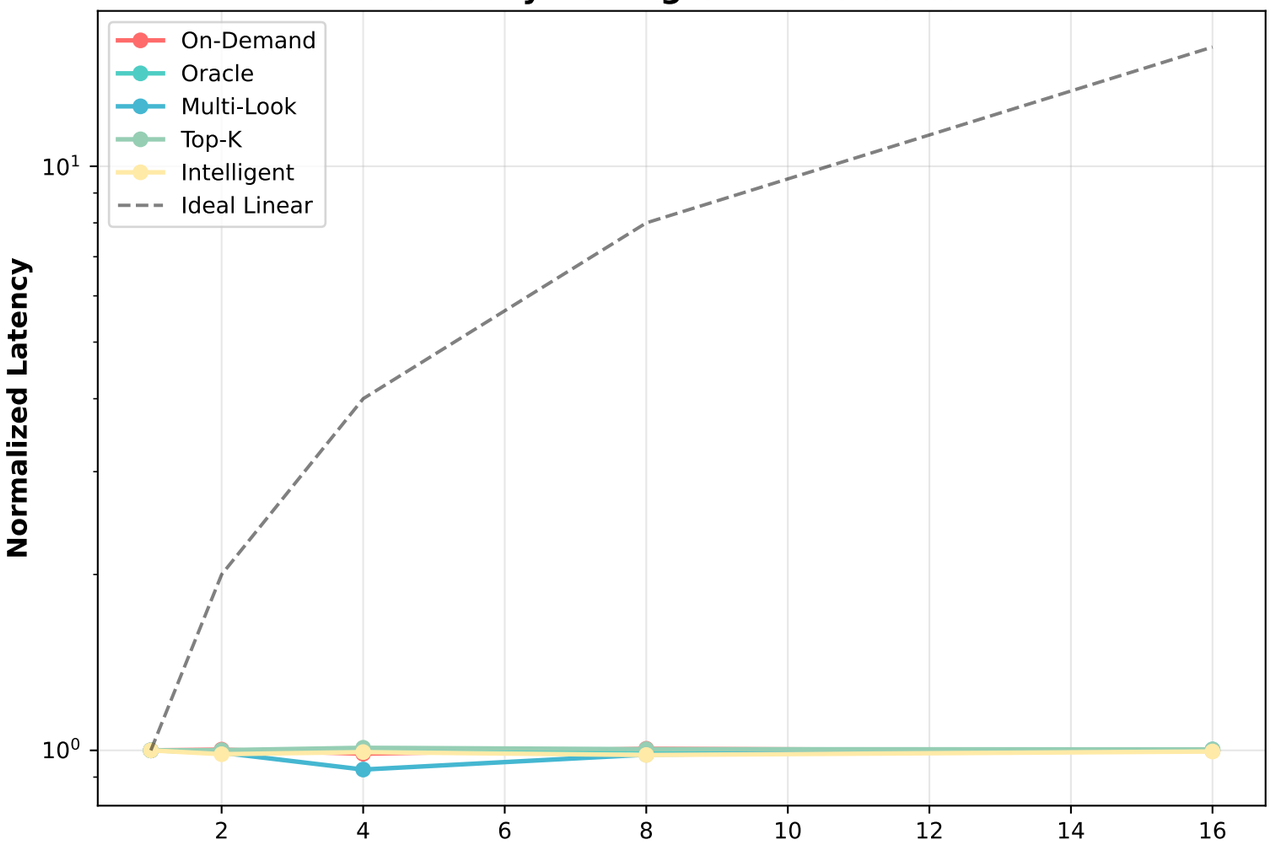
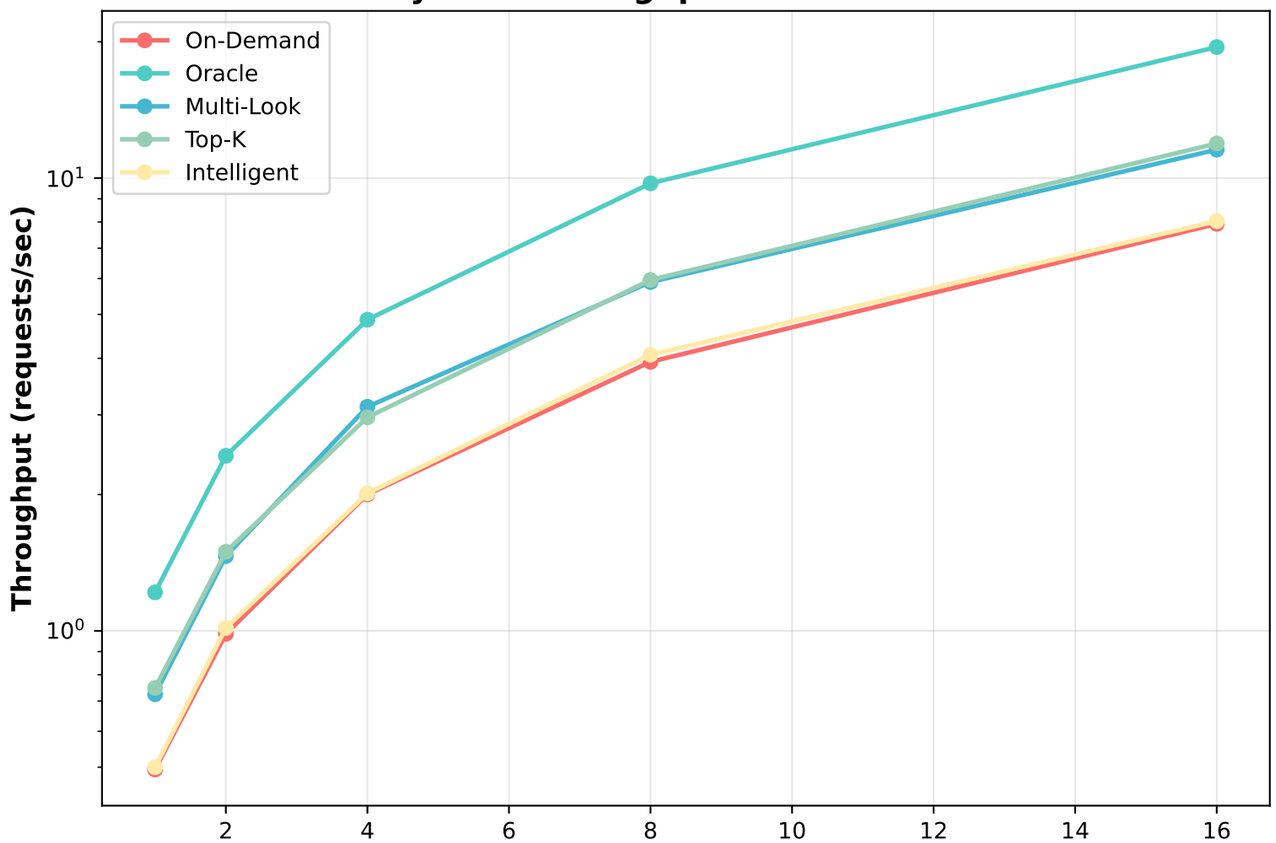
Qwen MoE: System Scalability Analysis

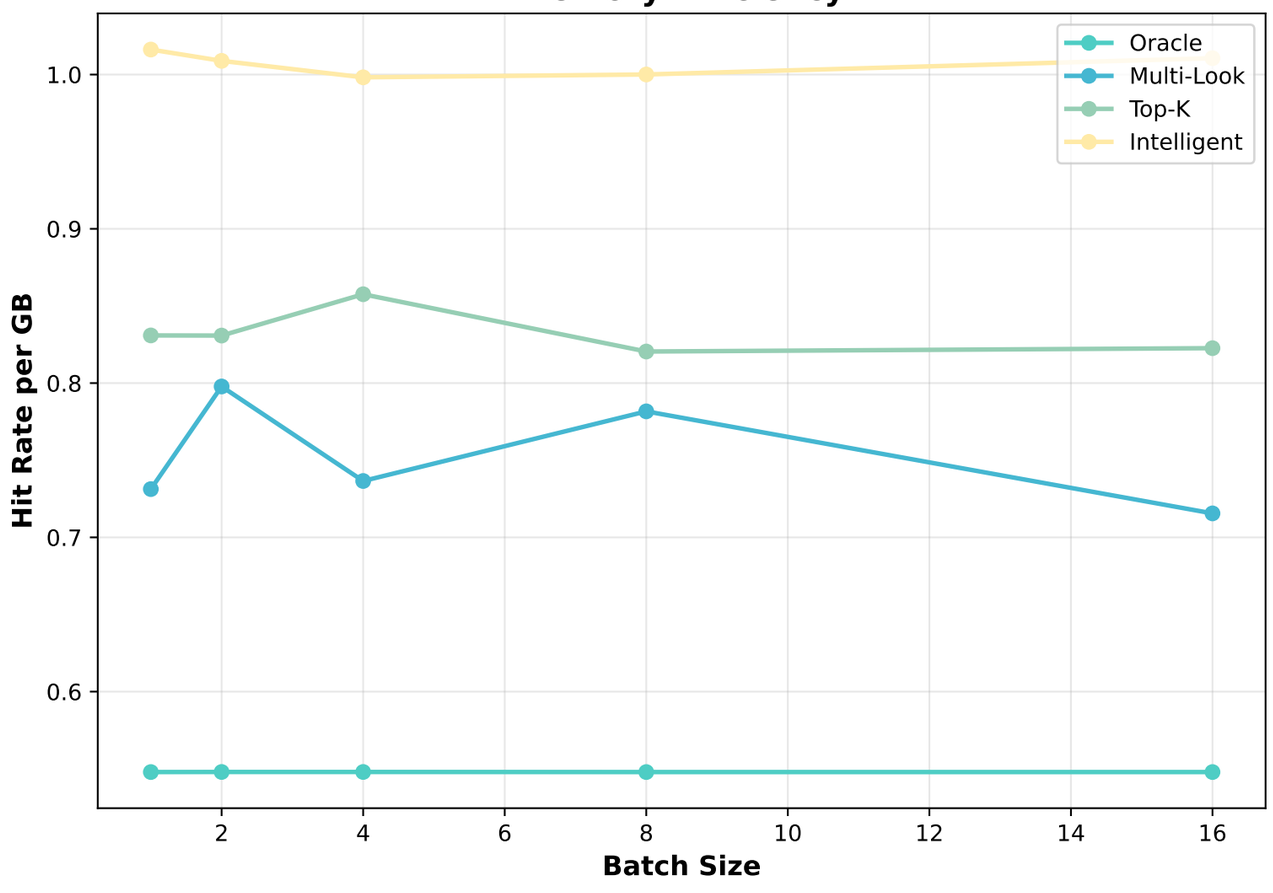
Latency Scaling Characteristics



System Throughput vs Batch Size



Memory Efficiency



Expert Loading Efficiency