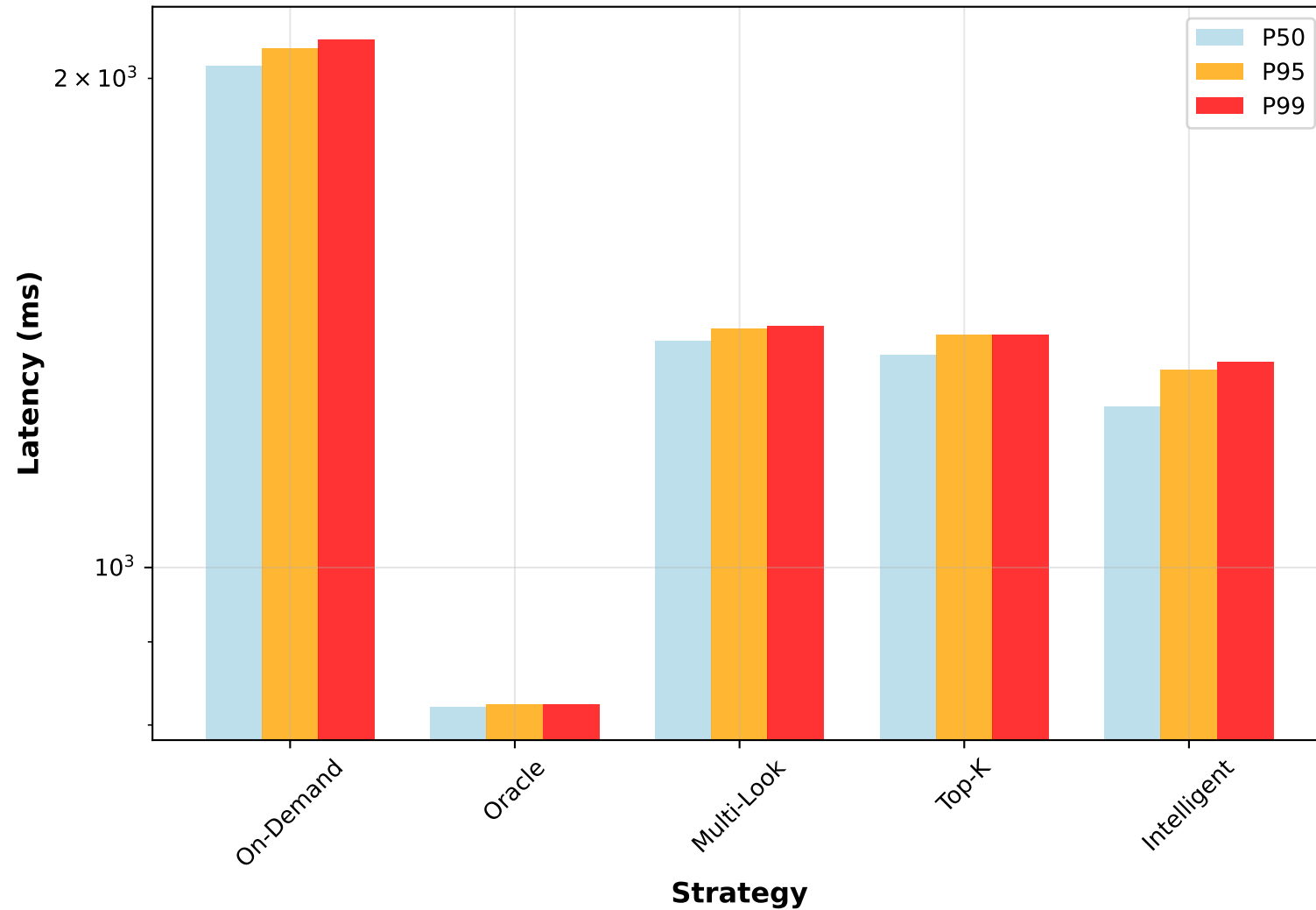
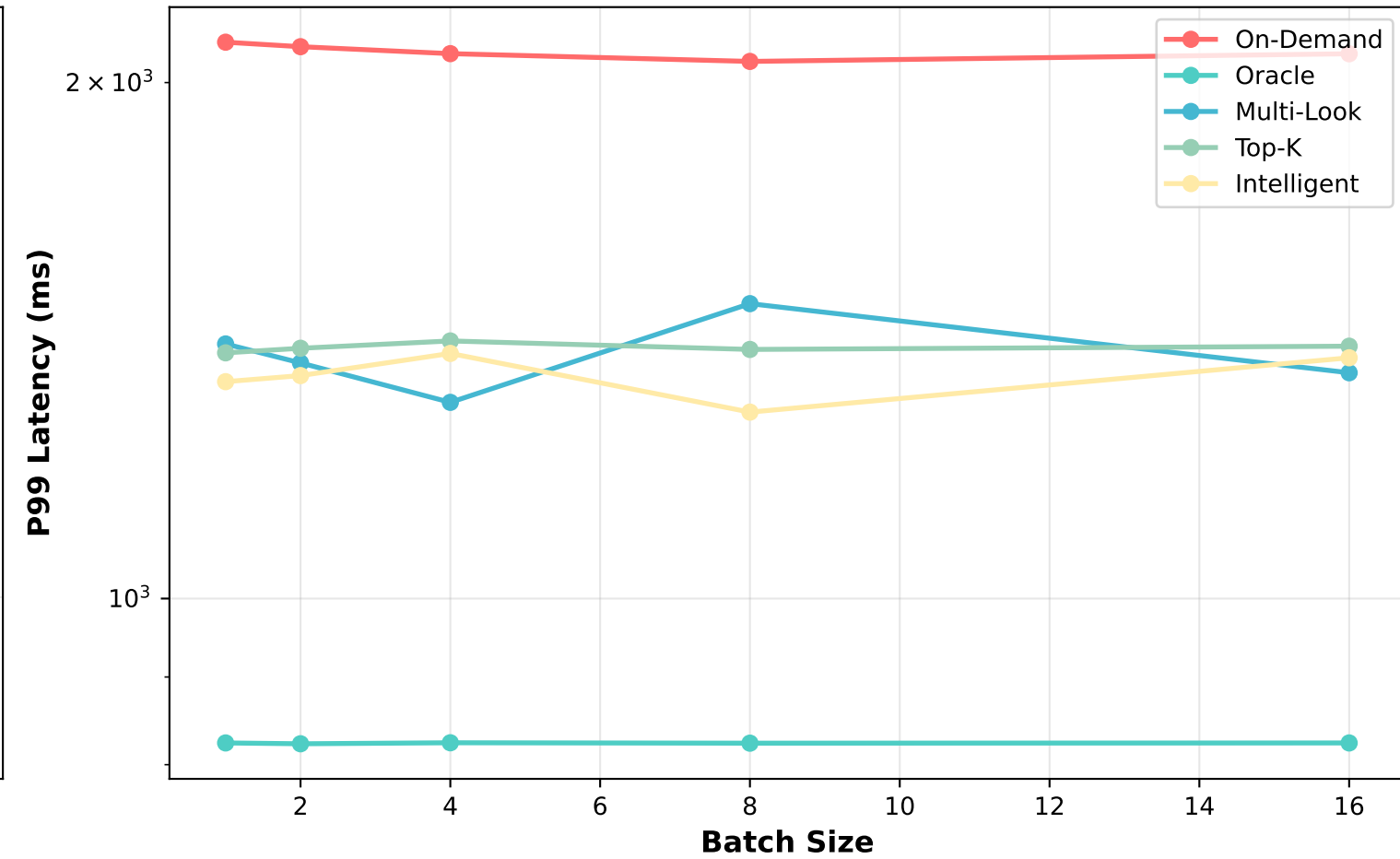


# Qwen MoE: Comprehensive Tail Latency Analysis

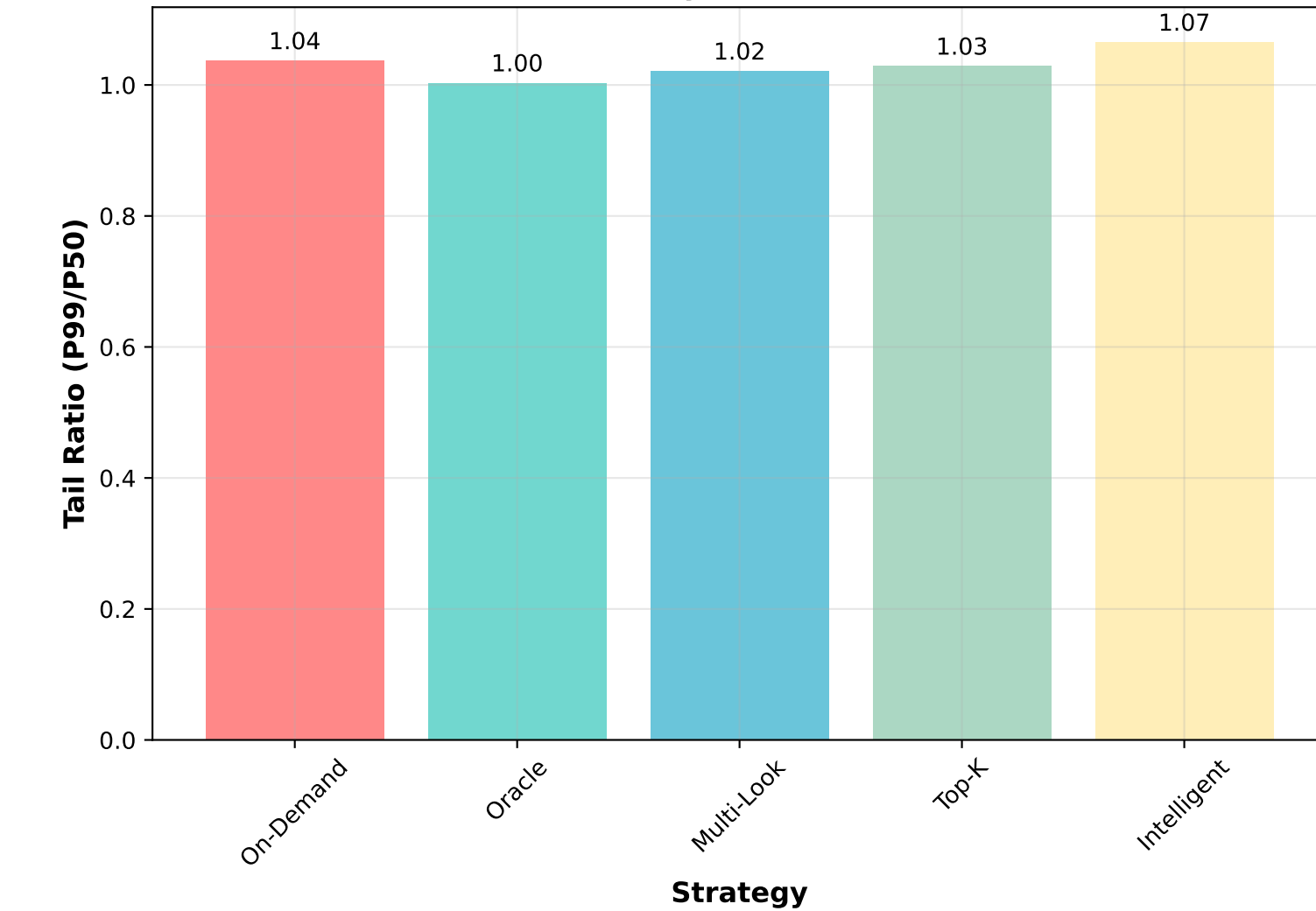
## Tail Latency Percentiles (Batch Size 1)



## P99 Latency Scaling



## Tail Latency Characteristics



## Performance Consistency

