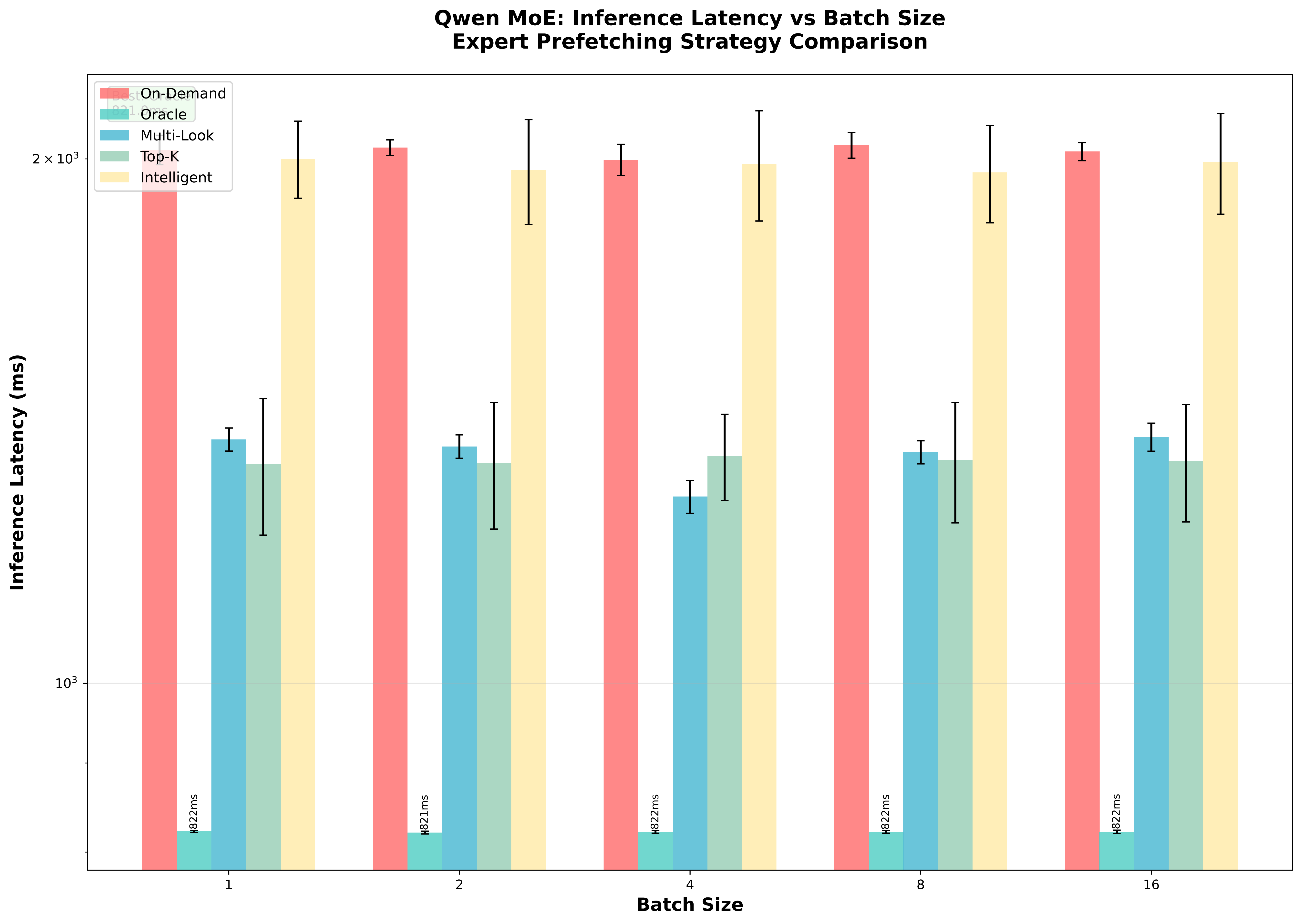

Qwen MoE: Inference Latency vs Batch Size
Expert Prefetching Strategy Comparison