Qwen MoE: Comprehensive Tail Latency Analysis Tail Latency Percentiles (Batch Size 1) P99 Latency Scaling P50 2×10^3 P95 2×10^{3} P99 P99 Latency (ms) Latency (ms) 10^{3} 10^{3} On-Demand Oracle Multi-Look Top-K Intelligent 400°+ 2 10 12 14 16 6 8 **Batch Size** Strategy **Tail Latency Characteristics Performance Consistency** 1.04 9.0% 1.03 1.03 1.02 1.00 1.0 8 Coefficient of Variation (%) Tail Ratio (P99/P50) 5.1% 1.9% 1.5% 0.2 0.1% 0.0 ntelligent 400x oracle 10Px oracle **Strategy** Strategy