Qwen MoE: Comprehensive Tail Latency Analysis Tail Latency Percentiles (Batch Size 1) P99 Latency Scaling P50 On-Demand P95 Oracle 2×10^3 2×10^{3} P99 Multi-Look Top-K Intelligent P99 Latency (ms) Latency (ms) 10^{3} 10^{3} 400°+ 16 2 10 12 14 6 8 **Batch Size** Strategy **Performance Consistency Tail Latency Characteristics** 1.07 6 1.04 1.03 1.02 1.00 1.0 4.8% Coefficient of Variation (%) **Tail Ratio (P99/P50)** 1.6% 1.2% 0.2 0.1% 0.0 ntelligent 400x oracle 10P.T oracle **Strategy** Strategy