QICCR Scalability Analysis

Near-linear scaling

- QICCR (4x A100 GPUs)
- LLM Beam Search
- Sequential Search

Throughput (paths/second) vs Concurrent Reasoning Paths