Classification Latency Distribution

Median: 2350ms | Throughput: 0.41 QPS | Tokens/query: 485

Mean: 2469ms

P95: 3254ms

Density

Latency (ms)