**DeepSeek R1 Throughput Comparison
(Baseline vs Speculative Decoding)**

Speedup: 1.20×

Total token throughput (tok/s)

10567.2

12679.9

Baseline (No Speculative Decoding)

With Speculative Decoding