**CC Throughput Overhead (Random 4000 ⇒ 1000)**

Number of Concurrent Requests

CC Overhead (%)

LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4