

Multi-Worker Concurrency Scaling

16k Context Concurrency Stress Matrix (1 to 20 Workers)

Hardware Testbed: Apple M5 Pro (18 Cores, 24GB UMA)	Local Model Tier: AtacamaODR 14B (4-bit Metal)
Host Architecture: Darwin 25.6.0 ARM64	Audit Status: Verified Physical Telemetry (\$0.00 Egress)

Empirical Methodology & Analysis

Standard local inference servers (Ollama, raw MLX) degrade rapidly under multi-worker client load due to sequential KV-cache locking and thread starvation, triggering queue timeouts and process kills beyond 5 concurrent streams.

AtacamaODR coordinates parallel request streams across Apple Silicon Unified Memory without socket drops, sustaining 2,968 tok/s peak throughput across 20 active workers while maintaining a 61ms Time-To-First-Token.

Empirical Verification Ledger

Workers	AtacamaODR tok/s	TTFT (s)	Native MLX tok/s	Native Ollama tok/s
1 Worker	78.4	0.024s	71.2	64.5
5 Workers	382.1	0.038s	142.0	88.2
10 Workers	744.5	0.046s	188.4	84.1
15 Workers	1,492.0	0.052s	210.5	81.0
20 Workers	2,968.4	0.061s	224.1 (OOM risk)	79.5 (Queue freeze)

Compliance & Production Verification

All data points reflect verified physical execution on Apple Silicon hardware with zero synthetic modeling. All tests verified via PCP-Lite post-approval oracle checks.