Cortex vs. Mem0: Performance Benchmark

Executive Summary

We executed **10,000 real API operations** against a live Harper Fabric cluster running Cortex, including memory stores with server-side ONNX embeddings, semantic searches, document ingestion, and metadata lookups, with zero errors. We isolated the server-side CPU time of each operation (averaging ~29ms for searches and ~113ms for stores) and extrapolated how many operations fit within each tier's compute-hour budget. We then compared those numbers to Mem0's published retrieval limits.

Bottom line: Cortex's free tier delivers **23× more semantic retrievals** per month than Mem0 Free. At the $250/3-months tier, Cortex delivers **84× more retrievals** per month than Mem0 Pro ($249/month), with zero embedding API costs and complete data ownership.

23×
FREE TIER ADVANTAGE

84×
PAID TIER ADVANTAGE

How We Tested

We modeled a team of 5 AI-powered developers, each running 7 coding sessions/day (weekends at 30% volume). Each session averages ~12 operations, producing roughly 10,000 operations per month with the following mix:

Operation	What It Does	Count	Avg Server CPU
Memory Store	Save decision/context with ONNX embedding	2,033	~113ms
Memory Search	Semantic recall by meaning	5,283	~29ms
Count / Get	Stats or fetch by ID (no embedding)	2,033	~2ms
Synapse Ingest	Parse and embed team docs	244	~5ms
Synapse Search	Search ingested documents	407	~22ms

Server CPU is isolated by subtracting the measured ~45ms network baseline from wall-clock time. This is what counts toward your compute-hour budget. For this workload, database reads/writes and network I/O are not the binding constraint.

What Limits Retrievals

Harper Fabric's free tier includes 10M reads, 2M writes, and 5 GB storage, none of which come close to being exhausted. The binding constraint is **Application Compute Hours** (2 hours / 6 months on free, 180 hours / 3 months on $250).

Cortex embeds text server-side using all-MiniLM-L6-v2 (384-dim, ~23 MB ONNX model). A semantic search costs only ~29ms of server CPU. Stores cost more (~113ms avg) because the stored text is longer, but searches vastly outnumber stores in real usage.

Our 10,000-operation workload consumed just 399 seconds (~0.11 hours) of server CPU. That means the free tier's 2 compute hours would sustain this team for 18 months, three times longer than the 6-month billing period.

Results: Cortex vs Mem0

Free Tier Comparison	Cortex Free	Mem0 Free	Cortex Advantage
Retrievals / month	**~23,000**	1,000	**23×**
Max memories	**~250K+ (5 GB)**	10,000	**25×**

$250 Tier Comparison	Cortex @ $250	Mem0 Pro $249/mo	Cortex Advantage
Retrievals / month	**~4,200,000**	50,000	**84×**
Max memories	**~500K+ (10 GB)**	Unlimited	—
Cost	**$250 / 3-months**	$249/mo recurring	**$497 saved / quarter**

Methodology

- **Cluster:** All 10,000 operations ran on a Harper Fabric free-tier cluster (single colocated node).
- **Embedding:** all-MiniLM-L6-v2 (384-dim, ~23 MB ONNX). Runs server-side, no external API calls.
- **Operations:** 10,000 total across 5 simulated users, shuffled randomly.
- **CPU isolation:** 30 baseline measurements (~45ms network) subtracted to isolate server CPU.
- **Retrieval calc:** Compute-hour budget ÷ blended server CPU per operation, with read fraction applied.
- **Mem0 data:** mem0.ai/pricing (March 2026). Hobby: 1K retrievals/mo. Pro $249/mo: 50K retrievals/mo.
- **Error rate:** 0 out of 10,000 operations.

Conclusion

Cortex on Harper Fabric delivers dramatically more value at every price point. A team of 5 heavy AI users gets 23× more free retrievals than Mem0, with room to spare for the full 6-month billing period. At $250, teams get 84× the monthly retrieval capacity of Mem0 Pro, with zero embedding costs and full data ownership.