Cloud vs Local AI CHEAT SHEET

Plex Consulting

Cloud AI (Bedrock, OpenAI, Anthropic)

Pros

- No hardware required — start in minutes
- Always the latest frontier models
- Scales instantly to any workload
- Zero maintenance — provider handles everything

Cons

- Data leaves your building
- Per-token costs add up at scale
- Vendor lock-in risk
- Rate limits during peak demand

Typical cost: ~$3—15 per million tokens, $10—50/day typical usage.

Local AI (Ollama, AirLLM)

Pros

- Data never leaves your premises
- No per-query cost after hardware
- No rate limits — run as much as you want
- Works completely offline

Cons

- Hardware investment upfront
- Slower than cloud on equivalent models
- You maintain the infrastructure
- Lags 2—6 months behind frontier models

Typical cost: Hardware only, then FREE forever.

Hardware Costs (NZD)

Hardware	Cost	Runs
Any 16 GB laptop	Already own it	7—8B models
RTX 4060 Ti 16 GB	~$700	13B models
RTX 4090 24 GB	~$3,500	34B models
2x RTX 4090	~$7,000	70B models
Mac Studio M2 Ultra 192 GB	~$12,000	70B+ models
Budget: 2x used P40 24 GB	~$1,500—2K	34B models

Business Case Comparison

Scenario	Cloud / Year	Local / Year	Verdict
Light (1 person)	$240/yr	$0 (own laptop)	Either works
Team of 5	$6—12K/yr	$3.5K yr 1, $200/yr after	Local wins yr 2
Heavy (100K+ queries/mo)	$24—60K/yr	$7—15K yr 1, $500/yr after	Local wins massively

The Hybrid Answer

Use	Best For
Cloud	Strategic decisions, complex analysis, frontier quality, occasional heavy lifting
Local	Daily drafts, document search, high-volume processing, sensitive data
Both	Route by task — cloud for thinking, local for doing. Data stays home.

"Cloud for thinking. Local for doing. Data stays home."