Rarely Rare
Rare by Design

A practical buyer's guide · 2026

The Local AI Buyer's Guide

Which hardware runs which model sizes, when local beats cloud on cost, and how to choose — without overspending or guessing.

WHAT'S INSIDE

Prepared by Rarely Rare · Vendor-neutral AI infrastructure consulting

01 · WHY RUN AI LOCALLY

The case for keeping AI in your building

For the first time, a small team can put a "personal AI supercomputer" on a desk for the price of a high-end laptop. Here's why companies are doing exactly that.

57%
cite data privacy as the #1 barrier to cloud AI

1.6–4×
cheaper on-prem for sustained inference (up to 8.6× vs pay-per-use)

75%
of enterprises will run hybrid AI by 2027 (IDC)

Privacy & compliance

Prompts, responses, and fine-tuning data never leave your network — the strongest possible posture for finance, healthcare, legal, gov, and defense.

Predictable cost

A fixed hardware cost replaces unpredictable per-token cloud bills. For steady, high-volume workloads, on-prem wins on TCO.

Latency & control

No round-trips to a cloud API, no rate limits, no model deprecations forced on you. You own the stack.

02 · HOW MUCH MACHINE DO YOU NEED?

Match the model size to memory

The single most important spec for local AI is **memory** — it determines the largest model you can run. Quantization (compressing the model) shrinks the requirement dramatically. **Q4 is the sweet spot** (~1% quality loss).

Model size	Example models	Memory at Q4	What it's good for
7–8B	Llama 3 8B, Mistral 7B	**~4–8 GB**	Chat, drafting, classification, simple RAG
13–14B	Qwen 14B	**~12–16 GB**	Stronger reasoning, coding assistance
70B	Llama 3 70B	**~40–48 GB**	Near-frontier quality, complex agents
200B+	DeepSeek, large MoE	**~110–128 GB**	Frontier-scale local reasoning

Rule of thumb: memory (GB) ≈ model parameters (B) ÷ 8 at Q4 · ÷ 4 at Q8 · × 2 at FP16. Add ~20% headroom for context (KV cache).

03 · HARDWARE COMPARED

The four ways to run AI locally

There's no single "best" machine — it depends on the model size you need and whether you prioritize raw speed or large-model capacity.

Machine	Memory	Price	Strengths	Watch-outs
NVIDIA DGX Spark GB10 Grace Blackwell	128 GB unified	**~$4,699**	Purpose-built; full NVIDIA AI stack preinstalled; runs ~200B-param models; link two for ~400B	Lower memory bandwidth than a discrete GPU
Dell Pro Max with GB10 OEM GB10 system	128 GB unified	**~$3,699–$3,999**	Same GB10 silicon; ships with DGX OS; strong value entry point	Same bandwidth trade-off as DGX Spark
Apple Mac Studio M3 Ultra / M4 Max	up to 512 GB unified*	**~$4k–$10k+**	Runs very large models others can't fit; quiet, low-power; great dev experience	Slower tokens/sec; *512GB config discontinued Mar 2026
NVIDIA RTX 5090 discrete GPU build	32 GB VRAM	**~$2,000+****	Fastest tokens/sec (1,792 GB/s) for models under ~32 GB; great for 7–13B	Can't fit 70B+ in VRAM; **street prices often $3k–5k

The core trade-off: speed vs. capacity

A discrete GPU (RTX 5090) has ~3× the memory bandwidth but a quarter of the memory. If a model fits in its VRAM, it's the fastest option. If it doesn't, it either won't run or crawls. Unified-memory machines (GB10, Mac) hold the *whole* model in one pool — so they run the big models a 5090 simply can't.

Quick guide: ≤13B and need speed → RTX. Up to ~200B with a turnkey AI stack → DGX Spark / Dell GB10. The largest possible models on one box → Mac Studio.

04 · LOCAL VS. CLOUD: THE COST MATH

When local pays for itself

Cloud AI has near-zero upfront cost but charges per token forever. Local AI is the reverse: a fixed cost up front, then near-zero marginal cost. The more you use it, the more local wins.

Illustrative: a $4,000 local machine running steady daily inference often breaks even versus comparable cloud API spend within months — after which inference is effectively free. For sustained workloads, on-prem runs **1.6–4×** cheaper, and up to **8.6×** cheaper than pay-per-use pricing. (Your actual breakeven depends on volume, model size, and utilization — we calculate it for your specific case.)

05 · DECISION FRAMEWORK

Start here

If you...	Start with
Just need to know which machine to buy	**AI Workstation Selection** — $500–$2,000
Want private models running in production	**Local LLM Deployment** — $2,000–$10,000
Need a prioritized, costed AI roadmap	**AI Strategy Workshop** — $5,000–$25,000
Are a regulated org preparing for production AI	**Enterprise AI Readiness Assessment** — $10,000–$50,000+

Get a free, vendor-neutral recommendation

Tell us one AI use case you've been holding back on for privacy or cost reasons, and we'll tell you honestly whether running it locally makes sense — and on what hardware. No pitch, no obligation.

Book a scoping call: cal.com/rarelyrare
Web: rarelyrare.com · **Email:** zeronoisereport@gmail.com

Rarely Rare — Rare by Design. We are a vendor-neutral advisor and do not resell hardware. Specs and prices verified June 2026 and subject to change; treat sizing figures as planning estimates. Sources: Dell & NVIDIA product pages; Apple; independent local-LLM hardware and VRAM analyses (2026); IDC hybrid-AI projection.