

Local-AI Buyer's Checklist

Six things to verify before you buy the box – and real, measured numbers to compare it against.

Every figure below is a measurement from robotnurse.io's own benchmark posts – Mac Studio M3 Ultra (256 GB) and DGX Spark GB10 (128 GB). No vendor spec sheets, no estimates.

CHECKLIST

Verify these six before you pay

- ☐ **01 Memory / VRAM capacity vs. model size**

WHY You must fit **weights + KV cache + working set**, not just the weights file. A 125B MoE is **192.2 GB** at 8-bit but **360 GB** at bf16 – the bf16 build does not fit a 256 GB box. At 4-bit mixed-precision the same model is **111.7 GB** (~30 tok/s peak). Buy the most memory you can afford; it is the one spec you usually cannot add later.
- ☐ **02 Memory bandwidth – the number that predicts throughput**

WHY Token generation is memory-bound, so **bandwidth $\approx$ speed**. A healthy M3 Ultra reads **538 GB/s** and prefills a 25,970-token prompt at **373.7 tok/s**; a “sick” box reading 94 GB/s does the same at **118.2 tok/s** – 4.7 $\times$ slower on identical silicon. Run a read-bandwidth probe before trusting any tok/s claim.
- ☐ **03 Quantization fit per model class**

WHY Do not blanket-apply “4-bit.” On a 125B MoE, uniform 4-bit saves only **4.3 GB** versus mixed 4/8-bit but costs **8.5% quality**. Attention Q/K/V, norms, routing, and MTP-sensitive tensors must stay at higher precision – a wrong NVFP4 input-scale degenerated output intermittently, a real shipped bug.
- ☐ **04 Inference engine – match it to the hardware (llama.cpp • MLX • vLLM)**

WHY vLLM + NVFP4 was “**the only combination that exists and fits**” GLM-5.3-Flash on 2 $\times$ DGX Spark; the Mac Studio serves it via MLX (mixed 4/8-bit); llama.cpp runs GGUF (Q2XL, Q4_K_M, MXFP4). The same AEON 8-bit model decodes at **42.9 tok/s** with an MLX draft model versus **40.6** plain – the draft model matters as much as the engine.
- ☐ **05 Power, cooling & steady state**

WHY A wedged clock governor cost **3.3 $\times$ steady** throughput, and a shared box drifts ± 20 –30%. Warm up before measuring: a healthy box holds **~ 350 tok/s** on a 2,560-token prompt; under **150 tok/s** means something is wrong before you have even loaded the model.
- ☐ **06 Upgrade path**

WHY Know which axis you can extend. **Mac Studio RAM is soldered** – choose 256 GB up front, you cannot add it later. A **DGX Spark GB10** is a fixed 128 GB node that scales by adding a second unit (2 $\times$ 128 GB). Decide the ceiling before the purchase, not after.

REFERENCE

Real numbers to compare against

BANDWIDTH → THROUGHPUT · MAC STUDIO M3 ULTRA, 80-CORE GPU, 256 GB				
STATE	READ BANDWIDTH	8-BIT MATMUL	PREFILL (25,970 TOK)	
Healthy (clean)	538 GB/s	22.9 TFLOP/s	373.7 tok/s	
Sick (throttled / dirty)	94 GB/s	4.83 TFLOP/s	118.2 tok/s	
Kernel note: warm 4-bit matrix-vector kernel runs 850–910 GB/s cache-resident (~600 GB/s cold); a decode-kernel retune added +7% throughput.				
MODEL FIT · QUANT VS. MEMORY ON A 125B MOE (6B ACTIVE, 512 EXPERTS)				
QUANTIZATION	MEMORY	NOTE		
bf16 (base)	360 GB	does not fit a 256 GB box		
8-bit	192.2 GB	native 262k context		
4-bit mixed-precision	111.7 GB	~30 tok/s peak		
plain 4-bit	111.6 GB	—		
text-only 4-bit	94.5 GB	speculative head kept		
2-bit	68–80 GB	exists, not run		
vendored single-file	111 GB peak	~385 tok/s prefill		
KV cache: ≈98 KB/token → a 128k context ≈ 13 GB.				
ENGINE × HARDWARE · WHAT ACTUALLY RUNS WHERE				
HARDWARE	ENGINE	QUANT	MODEL	RESULT
2× DGX Spark GB10 (128 GB ea.)	vLLM	NVFP4	GLM-5.3-Flash	only combination that fits (121.7 GiB/rank)
Mac Studio M3 Ultra 256 GB	MLX (mlx-vlm)	mixed 4/8-bit	GLM-5.3-Flash	4-bit = −8.5% quality for 4.3 GB
Mac Studio (studio1)	dflash-mlx + DFlash2	8-bit / 4-bit	AEON Qwen3.8-27B	42.9 / 60.6 tok/s decode · 77% accept
Mac Studio (studio2)	llama.cpp GGUF	MXFP4	gpt-oss-120b	69.5 GB RSS

Source: robotnurse.io benchmark posts – figures transcribed verbatim from measured runs, source-cited per row.

robotnurse.io · Local AI, measured.