FREE CHECKLIST · LOCAL AI

LOCAL LLM IN A BOX

The no-BS checklist: hardware, budget, and what you actually need to run your own AI on your own machine.

THGSN

01 Define the job (before any hardware)

The hardware decision follows from one question: what will this machine actually do? Pick your primary use — everything else is noise.

USE CASE	WHAT IT DEMANDS
Document Q&A / RAG over your files	Moderate model (13B–32B), large context window, long sessions
Code completion / pair programming	Strong mid-size model (13B–32B), latency matters more than size
Transcription (STT)	A mid GPU is plenty — the easiest job on this list
Image generation (ComfyUI)	Same GPU, different stack — 16–24 GB VRAM is the sweet spot
Long-form generation	Big model (70B+) — a different price bracket

Pick **one** primary job. A box tuned for three jobs is a box that is mediocre at all of them.

02 Model size to VRAM — the math that matters

Rule of thumb: a model in 4-bit quantization takes roughly **0.5 GB of VRAM per billion parameters**. Add 2–4 GB for the KV cache and headroom. These are the working numbers:

MODEL SIZE (Q4)	WEIGHTS IN VRAM	PRACTICAL MINIMUM VRAM
7B–8B	~4–5 GB	8 GB — comfortable, but limited ability
13B–14B	~8–10 GB	12 GB minimum, 16 GB comfortable
32B–34B	~20–22 GB	24 GB — the sweet spot for quality per dollar
70B	~40–42 GB	48 GB (dual GPU or Mac unified memory)

Q4 is where the quality-per-gigabyte curve is flattest. Q8 doubles VRAM for a small quality gain — do not plan hardware around Q8.

03 The four real hardware paths

PATH	WHAT YOU GET	WATCH OUT
Used 24 GB GPU (RTX 3090) + 32 GB RAM build	The best price/quality box for 8B–32B models. 32B Q4 fits and runs.	Buy a tested card; check ex-mining history; 350 W under load
RTX 4090 24 GB	Same sizes, roughly 2x faster inference	450 W, loud, price is 3–4x a used 3090 — pay only if speed is the point
Mac with large unified memory (64–128 GB)	Silent, efficient, runs 70B at low quant. No driver drama.	Slow for its size; unified memory is shared with the CPU; not for training
2x 24 GB GPUs on one board	48 GB → 70B Q4 on a PC	Needs a board with real x8/x8 PCIe lanes; PSU and cooling get complicated

The rest of the box: any modern 8+ core CPU is fine for models that fit in VRAM. 32 GB system RAM is the floor (64 GB if you plan to offload). Storage: 1 TB NVMe — models are 5–45 GB each and you will keep several.

04 The software stack (boring, and that is good)

You do not need a cluster or Kubernetes. The proven stack is three pieces:

- **Runtime** — Ollama (easiest, great for starting) or vLLM (faster, the standard for continuous serving).
- **Interface** — a web UI (Open WebUI or similar) on top of the runtime.
- **Your documents** — a simple RAG: chunked files + embeddings + the same model. A weekend project once the box works.

Total setup time on hardware that works: a few hours. If your setup takes a week, the hardware or the stack is wrong — not the model.

05 Do NOT buy

- **Datacenter GPUs (H100 and friends)** — 700 W to 1 kW each, server cooling, gray-market resale. A 3090 does 90% of what you imagine needing.
- **A brand-new $3,000 tower before testing the workload** — rent a cloud GPU for a week first; it costs less than one used card and tells you your real model size.
- **A Mac for training or fine-tuning** — unified memory is excellent for inference, weak for the training loop.
- **The second GPU on day one** — run the first box for a month, see which models you actually keep open, then add capacity.

06 Home vs. office reality check

If the box lives where people live: a single 3090/4090 is a 350–450 W heat source under load — plan airflow and check the noise curve at 50% load, not max. If it lives in a closet or office: the same card, a 750 W supply, a shelf and monitoring is a production system. Idle power is 15–25 W per GPU, so the electricity bill is not the risk — people keep believing it is.

07 The 60-second summary

CHECK	MINIMUM
[ ] Pick ONE primary job	Document Q&A, code, or STT
[ ] Choose model size from the job	13B–32B covers most real workloads
[ ] VRAM >= weights + 2–4 GB headroom	24 GB is the value sweet spot
[ ] 32 GB system RAM, 1 TB NVMe	Floor, not a recommendation
[ ] Ollama or vLLM + web UI	Hours to set up, not weeks
[ ] Test one month before adding hardware	The workload reveals the real need

The whole point of local AI is that the machine belongs to you: no per-token meter, no subscription, no data leaving the building. The hardware above makes that true for a few hundred to a couple of thousand dollars — not ten.