Choosing hardware for local AI: GPUs, NPUs and Apple Silicon

26 October 2026 · 6 min read · blogs.iictp.org/hardware-for-local-ai-gpus-npus-apple-silicon

For local models, the main constraint is memory, followed by memory speed, and only then raw compute. Match the hardware to the model you intend to run, not the other way round.

TL;DR The short version

- For local models, memory capacity matters most, then memory bandwidth, and only then raw compute.
- Discrete GPUs are fast with fixed memory; Apple Silicon's unified memory holds large models at lower speed.
- Size the hardware to the smallest model that passes your tests, at your chosen quantization, plus room for context.

Read online, listen to it, or change the text size: https://blogs.iictp.org/hardware-for-local-ai-gpus-npus-apple-silicon

What matters, in order

1. **Memory capacity:** the model has to fit, with room for context.
2. **Memory bandwidth:** generating each token reads the weights, so bandwidth limits speed.
3. **Compute:** matters most for processing long prompts and for serving many users.
4. **Software support:** the runtime must support your hardware.

The families

- **Discrete GPUs:** the usual choice for servers; fast, with a fixed memory size.
- **Apple Silicon:** unified memory shared by CPU and GPU lets a laptop or desktop hold large models, at lower speeds than server GPUs.
- **NPUs:** low-power accelerators in new laptops and phones, good for small models; check what your runtime supports.
- **CPU only:** workable for small or heavily quantized models and for light use.

Sizing steps

1. Choose the smallest model that passes your tests.
2. Work out its memory at your chosen quantization, plus context.
3. Pick hardware with comfortable headroom.
4. Measure speed on your own prompts before you buy in volume.

Related credential: Local LLM Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. IICTP. Local LLM Practitioner credential page. iictp.org/certifications/llmp.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/hardware-for-local-ai-gpus-npus-apple-silicon. More guides: blogs.iictp.org.