How to Choose a Local AI Model (Step by Step)

AI Models & Technology · about 22 minutes · 2026-11-11

A fully illustrated 28-step guide to choosing a local AI model: define tasks and check real hardware, understand parameters and quantisation, vet licences, then install, benchmark on your own prompts and maintain the pick.

Phase 1 — Define the Job and Measure Your Hardware

1. **List the exact tasks it will handle**
   List the exact tasks it will handle.
2. **Prepare five to ten real test prompts**
   Prepare five to ten real test prompts.
3. **Decide where the model will run**
   Decide where the model will run.
4. **Fix your privacy and offline requirements**
   Fix your privacy and offline requirements.
5. **Read the available memory**
   Read the available memory.
6. **Check CPU, GPU and supported backends**
   Check CPU, GPU and supported backends.
7. **Set a realistic size budget**
   Set a realistic size budget.

Phase 2 — Read the Specs and Shortlist

8. **Decode parameter count and quantisation**
   Decode parameter count and quantisation.
9. **Match quantisation to available memory**
   Match quantisation to available memory.
10. **Choose a file format for your runtime**
    Choose a file format for your runtime.
11. **Check context length and special abilities**
    Check context length and special abilities.
12. **Read the licence before downloading**
    Read the licence before downloading.
13. **Check provenance and maintenance**
    Check provenance and maintenance.
14. **Shortlist two or three candidates**
    Shortlist two or three candidates.

Phase 3 — Install, Run and Benchmark

15. **Install a local inference runtime**

    Install a local inference runtime.

16. **Download weights from a trusted source**

    Download weights from a trusted source.

17. **Load the model and confirm it answers**

    Load the model and confirm it answers.

18. **Tune how many layers go to the GPU**

    Tune how many layers go to the GPU.

19. **Run the same prompt set on every model**

    Run the same prompt set on every model.

20. **Measure speed and time to first token**

    Measure speed and time to first token.

21. **Save the hardest real task for last**

    Save the hardest real task for last.

Phase 4 — Decide, Integrate and Maintain

22. **Score candidates on quality, speed and fit**

    Score candidates on quality, speed and fit.

23. **Set a default model and a fallback**

    Set a default model and a fallback.

24. **Connect it through a local API**

    Connect it through a local API.

25. **Pin the exact version and quantisation**

    Pin the exact version and quantisation.

26. **Watch memory and thermal headroom**

    Watch memory and thermal headroom.

27. **Keep a small regression prompt set**

    Keep a small regression prompt set.

28. **Re-evaluate every few months**

    Re-evaluate every few months.

Sources

1. Anthropic — AI research and documentation
   https://www.anthropic.com/
2. NIST — AI Risk Management Framework
   https://www.nist.gov/
3. Mozilla — web and browser resources
   https://www.mozilla.org/
4. Microsoft Learn
   https://learn.microsoft.com/

Written as a checklist from https://www.howcompass.com/how-to/How-to-Choose-a-Local-AI-Model. Corrections: https://www.howcompass.com