[Grab the markdown version for your agent to read]

LOCAL AI FOR BUILDERS

A founder's guide to models, tools, and local AI businesses

Now is the window to build local AI products. Founders are missing it because they're comparing benchmarks instead of looking for businesses.

One question that matters: is this model good enough for the job, and does running it locally make the product better?

This guide gives you the setup, the first workflow to copy, and the filter for spotting where local AI belongs.

When to Use Local vs Cloud

Local: Private files, sensitive customer data, offline work, fieldwork, low latency, repeated internal workflows, high API costs.

Cloud: Deep research, hard reasoning, tasks that need the strongest possible model.

Hybrid: Local does the private first pass. Cloud does the heavy reasoning. A human approves anything important. This is where most real products will land.

The Setup

Every local AI stack has four pieces:

Piece	What it is	Where to start
Model	The brain file	Gemma 4 (E4B for laptops, E2B for phones)
Warehouse	Where you find models	Hugging Face
Software	What runs the model	LM Studio (visual) or Ollama (API)
Workflow	The product you build	Start with the first workflow below

LM Studio is the friendliest starting point. Download it, search for Gemma, click download, start chatting. Ollama is more builder-oriented - install it, run a command, and you have a local API your apps can talk to. Google AI Edge is the path when you're ready to ship on-device products.

On Hugging Face, model cards can be overwhelming. Focus on five things: what is the model for, how big is it, what license, what hardware do people run it on, and are there quantized files available. Ignore everything else.

Hardware

RAM	What you can run
8 GB	Small models only. Stay under 4B parameters.
16 GB	Useful experiments. 4B-12B quantized.
32 GB	Larger workflows. 12B+ comfortably.
Strong GPU	27B+ becomes realistic.

You don't need a workstation. A spare laptop from 2021 runs a 4B model fine. Start with Q4 quantization. It's the easiest to run. Q8 keeps more quality but needs more memory.

First Workflow to Copy

Make a folder called `Customer Notes`. Put 10 real files in it — support tickets, sales calls, meeting notes, customer feedback.

Run Gemma against the folder. Produce one file: `whatcustomersaretellingus.md`

The file should cover:

- Repeated complaints
- Exact customer language
- Likely root cause
- The broken part of the business
- One thing to test this week

Run it again next week with fresh notes. That's the loop. A chat answer is nice. A reusable file changes the workflow.

The Eval

Before you decide what stays local and what needs cloud, test both.

Run the same 10 notes through local Gemma and a frontier cloud model. Compare the outputs. You're answering two questions: where does local already work, and where does cloud still earn its place?

Most founders skip this and guess. The eval takes 30 minutes and saves you from building on the wrong architecture.

Startup Idea Filter

Look for a customer with:

- Sensitive data that can't leave the building
- Repeated review work
- Dated software (early 2000s tools still in use)

- Expensive mistakes from missed details
- A workflow close to the device

Three or more of these? Worth exploring. All five? Strong signal.

---

3 Ideas That Pass the Filter

1. Home Health QA Reviewer
Home health agencies submit documentation before billing. Gaps and inconsistencies mean rejected claims. Build a local reviewer that flags missing fields, inconsistent notes, and compliance risks before anything gets submitted.

2. Offline Field Report Copilot
Restoration companies inspect damage and write reports from memory back at the office. Build a mobile app that drafts the report on-site from photos and voice notes. Flags missing pieces while the technician is still there.

3. Pre-Send Reviewer for Professional Services
Every firm has someone check drafts before they go out. Build a local desktop app that reviews outbound emails against the firm's own rules. Catches guaranteed-return language for wealth advisors, over-definitive sentences for lawyers, sensitive info leaks for HR.

Each one starts as a service with 5-10 customers. The checklist you build by hand becomes the product.

Final Note

Run Gemma. Try LM Studio. Copy the first workflow. Run the eval. Then look for one boring, repeated, sensitive workflow that nobody's automating because the data can't leave the building. That's where the product is.
