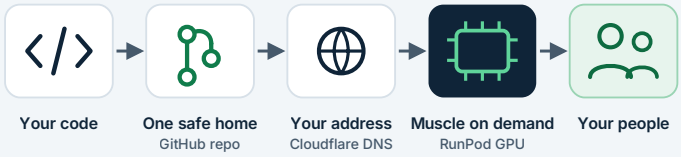
A PRACTICAL PATH FROM IDEA TO LIVE

What if going live were a checklist, not a leap of faith?

Most launches stall on plumbing — where the code lives, how your web address points to it, and what it costs to keep the lights on. There is a calm, repeatable path. Four steps. No surprises on the bill.



CODE → REPO → ADDRESS → RIGHT-SIZED COMPUTE → PEOPLE

THE IDEA, IN PLAIN TERMS

Going live is four decisions, made once — then repeated for every project.

Where does the code live? How do people find it? What runs the heavy lifting? And who is watching the meter? Answer those four questions deliberately and every launch after the first one gets faster, safer, and cheaper.

1 Give the code one safe home

GITHUB

- One repository per product — no copies on laptops or desktops.
- Protect the main branch: changes need a review before they land.
- Keep passwords and keys in a secrets vault, never in the code.
- Automatic checks run on every change before it can go live.

2 Point your name at it

CLOUDFLARE DNS

- Your domain becomes the front door; DNS records are the signposts.
- Turn on the proxy: encryption and a shield against bad traffic, included.
- Changes take minutes, not days — and roll back just as fast.
- One dashboard shows every address you own.

3 Rent muscle only when needed

RUNPOD GPU

- Graphics-card power by the hour — or by the second, serverless.
- Right-size the pod to the job; don't pay for idle horsepower.
- Auto-stop pods the moment work finishes.
- Queue jobs back-to-back so paid time is busy time.

4 Watch the meter, on purpose

TOKEN SPEND CONTROL

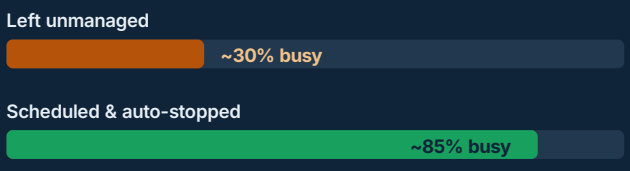
- Route routine tasks to small or local models; save the big AI for hard jobs.
- Set monthly caps and alerts per team and per project.
- Cache repeated answers instead of paying for them twice.
- A one-page report: spend, by task, by team, every month.

WHY UTILIZATION IS THE WHOLE GAME

A rented GPU earns its keep only while it's working.

An idle pod costs the same as a busy one. Scheduling jobs back-to-back and stopping pods automatically keeps paid hours in the productive band — the same machine, doing twice the work.

PAID GPU HOURS THAT DO REAL WORK



RETHINK THE REFLEX TO RENT

Before you send every job to the cloud, look at the computer on the desk.

Most offices buy machines with 8 or 16 GB of memory — fine for email, too small for AI. A workstation with an NVIDIA Blackwell-class graphics card and CUDA runs capable models **locally** — Qwen and Mistral through Ollama — privately, instantly, with no meter running. The cloud becomes the exception, not the default.

TODAY'S DEFAULT

The standard office machine

- 8–16 GB of memory; no real graphics horsepower.
- Every AI request leaves the building — and the meter runs.
- Sensitive documents travel to someone else's servers.
- Costs climb with every question your team asks.

THE SMARTER BUY

The AI-ready workstation

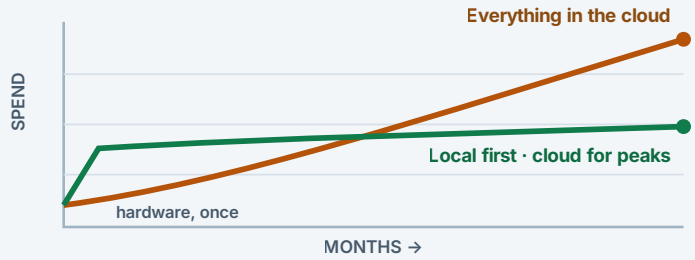
- NVIDIA Blackwell-class GPU with CUDA; 64 GB+ of memory.
- Everyday AI runs on the desk — instant, private, no per-question fee.
- Documents never leave your building unless you choose.
- Cloud GPUs reserved for the few genuinely heavy jobs.

Runs locally today: **Ollama** runtime · **Qwen** · **Mistral** — matched to the task, not the hype.

WHAT HAPPENS TO THE BILL

One purchase up front. A gentle line after.

Cloud-for-everything looks cheap in month one and steep by month twelve. Local-first starts with the hardware, then flattens — most questions never touch a meter. Tokens are spent where they earn their keep.



Your information stays home

Contracts, client files, and HR records are processed on a machine you own, inside walls you control.

HELPFUL FOR EVERYONE

Answers without the wait

Local models reply in the time it takes to read the question — no queue, no outage on someone else's end.

HELPFUL DAY TO DAY

Spend you can predict

Hardware is a line item you approve once. Token budgets cover only the hard jobs that truly need the big AI.

HELPFUL FOR LEADERS

Better results, task by task

Simple tasks go to simple tools; hard ones get the right model. Matching the tool to the job improves answers, not just bills.

HELPFUL FOR EVERYONE

Curious what this could look like for your team?

We'll map which of your tasks actually need the cloud — in plain language, no pressure.

EMAIL brent@llmadvisor.ai
WEB www.llmadvisor.ai
PHONE 617.595.8092