

The Practical AI Stack 2026

A plain-English system for choosing useful AI tools and running private AI on a mid-range Windows PC—without buying every subscription or guessing at model sizes.

What you get: a curated tool map, a realistic local setup, copy-ready workflows, a troubleshooting checklist, and a decision rule for what to run in the cloud versus on your own computer.

Version 1.0 · 4 October 2026

Prepared for practical builders, creators, students, and small teams.

This guide is educational. Tool names, pricing, model availability, and system requirements change. Check the official links before installing or paying.

START HERE

Do not collect tools. Build a small system.

Most people do not need twenty AI subscriptions. They need a repeatable loop: find information, think with it, make an output, and keep private work private when it matters.

The \$19 promise: by the end of this guide, you should know which tool to open for a job, which local model to try first, and what your PC can realistically handle.

The 4-part loop

01 • FIND

Search and verify

Use a web-aware tool for current facts, then open the source. AI can sound certain while being wrong.

02 • THINK

Draft and compare

Use a general assistant for outlining, rewriting, planning, and explaining. Keep a human review step.

03 • MAKE

Turn it into an asset

Use design, video, coding, or automation tools only after the job and audience are clear.

04 • KEEP

Protect and reuse

Keep sensitive documents local where practical, save winning prompts, and record which model produced what.

Mid-range PC baseline

Use this guide's "mid-range" label as a practical baseline, not a guarantee: Windows 10 22H2 or newer, a modern 6-core CPU, 16 GB RAM, an SSD, and ideally 4–8 GB dedicated GPU memory. LM Studio recommends at least 16 GB RAM and 4 GB dedicated VRAM; larger models may need more memory or slower CPU-only inference.

Rule of thumb: start with a 3B–8B quantized model. A model that runs comfortably and answers quickly is more useful than a larger model you avoid opening.

TOOL MAP

Popular tools, assigned to jobs

Pick one default tool per job. Free tiers and features change, so the links are the source of truth.

JOB	START WITH	WHY IT BELONGS	WATCH-OUT
General work	ChatGPT or Claude	Drafting, analysis, planning, coding help	Verify facts; limits and models change
Current research	Perplexity	Fast web-oriented answers with source links	Open and inspect important sources
Long material	NotebookLM	Ask questions against a source set	It is not a replacement for reading or citation checking
Visual posts	Canva AI	Quick layouts, copy, and creative variations	Review brand, rights, and factual claims
Short video	CapCut AI	Captions, edits, and social-ready assembly	Check music, likeness, and commercial rights
Code	Cursor or an AI-enabled editor	Useful inside a real codebase	Never accept unreviewed code or secrets
Private chat	LM Studio or Ollama	Local models, offline use, and repeatable workflows	Quality is model- and hardware-dependent
Local interface	Open WebUI + Ollama	Browser interface and multi-model organization	Secure local network access; do not expose it casually
Local images	ComfyUI	Node-based, controllable image workflows	GPU memory and model licensing matter

Simple choice: if the task depends on today’s information, use a web-connected assistant. If the task contains private notes or must work without internet, test a local model first.

LOCAL SYSTEM

Your first private AI setup

The least confusing path is to choose one desktop runner, one small model, and one interface. Do not install every tool at once.

1

Prepare the PC

Update Windows and GPU drivers. Keep at least 20–30 GB free on the SSD for a first model and cache. If you have only 8 GB RAM, use cloud tools or very small models first.

2

Choose one runner

LM Studio is the visual path: download a model from its Discover area, chat locally, and optionally run a local server. **Ollama** is the simpler command-line/API path: install it, then run a model such as `ollama run qwen3:8b` if your hardware can handle it.

3

Start small

Try a 3B or 4B model on a modest machine. Move to an 8B quantized model when memory and speed are comfortable. Qwen3 is available in several sizes; Llama 3.2 includes a 3B option. Model tags and sizes change, so use the current library pages.

4

Add a browser interface only if needed

Open WebUI can provide a self-hosted interface for Ollama and OpenAI-compatible APIs, and its docs describe Windows setup. Keep it bound to your local machine unless you understand authentication, network exposure, and updates.

Recommended first stack:

Windows PC → Ollama or LM Studio → Qwen3 4B/8B or Llama 3.2 3B
→ Open WebUI only after the first local chat works.

What “private” does and does not mean

Local inference can keep prompts and files on your computer, but downloads, extensions, telemetry, browser history, backups, and other connected services can still create data trails. Treat privacy as a system setting, not a marketing word.

WORKFLOWS

Five workflows that pay back quickly

1. Research brief

Ask a web-connected tool for a three-source brief. Open each source. Save the claim, source, date, and uncertainty. Use local AI to summarize private notes afterward.

2. Content repurposing

Give a local model a transcript or draft. Ask for five hooks, three titles, and a short version. Human-review tone, names, and claims before publishing.

3. Code companion

Ask the editor to explain a file before changing it. Request a small diff, tests, and a rollback plan. Never paste API keys or private credentials.

4. Document Q&A

Use NotebookLM for a selected source set or a local model for sensitive notes. Ask it to quote the section it used and say when the answer is not present.

5. Visual batch

Write the message first, then use Canva or CapCut for variations. Export one test asset, inspect it on mobile, and only then create a batch.

6. Weekly review

Ask AI to group tasks by outcome, identify repeated work, and suggest one automation. Keep the final priorities human-owned.

Copy-ready prompt: source-aware answer

“Answer the question in plain language. Separate what you know from what you infer. For every current or factual claim, provide the source title, URL, and access date. If the sources disagree, show the disagreement instead of smoothing it over.”

Copy-ready prompt: private local assistant

“You are a careful offline assistant. Use only the text I provide. Do not invent missing facts. Return: (1) a concise answer, (2) the exact evidence you used, and (3) one uncertainty or follow-up question.”

TROUBLESHOOTING

When the local model feels bad

SYMPTOM	TRY THIS FIRST
It is too slow	Use a smaller model, fewer context tokens, or GPU acceleration. Close memory-heavy apps.
It runs out of memory	Use a smaller quantized model, reduce context, and move model storage to a fast SSD with space.
Answers are weak	Use a better prompt, provide the needed context, or switch to a larger model/cloud assistant for reasoning-heavy work.
It hallucinates	Ask for evidence, constrain it to supplied text, and verify independently. Local does not mean correct.
Open WebUI cannot connect	Confirm Ollama is running, check the provider URL, and follow the current Open WebUI connection guide.
Image generation crashes	Lower resolution, use fewer steps, close other GPU apps, and check the model's VRAM expectations.
Installation feels risky	Use official download pages only, scan files, create a restore point, and avoid random "optimizers" or pirated model bundles.

Security checklist

- ☐ Download software from the official project page.
- ☐ Never paste passwords, API keys, wallet seed phrases, or private client data into a cloud assistant.
- ☐ Do not expose a local server to the public internet unless you understand authentication and network controls.
- ☐ Review model and asset licenses before commercial use.
- ☐ Keep a human approval step before sending, publishing, purchasing, or deleting.

Remember: the best AI system is the smallest system you will actually use every week.

LAUNCH CHECKLIST

Your 60-minute setup plan

00

Write one job

Example: "Summarize my private project notes and turn them into next actions."

15

Test one cloud tool

Run the same prompt in ChatGPT, Claude, or Gemini. Note quality, speed, and whether sources are needed.

30

Install one local runner

Choose LM Studio for a visual start or Ollama for a repeatable developer workflow.

45

Run one local model

Start with a small Qwen3 or Llama 3.2 variant. Test a private, non-sensitive sample first.

60

Save your operating rule

Cloud for current facts and frontier capability. Local for privacy, offline work, and experimentation. Human review for anything consequential.

Decision card

IF THE TASK IS...	DEFAULT ROUTE
Time-sensitive research	Web-connected assistant + source checking
Private notes or drafts	Local model when quality is sufficient
High-stakes advice	Qualified human + primary sources; AI only as an assistant
Code in a real project	AI editor + tests + review + version control
Creative production	Tool suited to the asset, with rights and brand review

Next move: choose one job, one runner, and one model. A small working system beats a giant unused toolbox.

SOURCES & UPDATE NOTES

Official links to check

This guide intentionally links to first-party pages because AI tools, prices, model names, and system requirements change quickly.

- [ChatGPT pricing and plan features](#)
- [Claude pricing](#)
- [Google Gemini updates](#)
- [Perplexity Help Center: Features](#)
- [Google NotebookLM](#)
- [Canva Magic Studio](#)
- [CapCut AI tools](#)
- [Cursor](#)
- [LM Studio system requirements](#)
- [LM Studio offline operation](#)
- [Ollama for Windows](#)
- [Ollama Qwen3 model library](#)
- [Ollama Llama 3.2 model library](#)
- [Open WebUI quick start](#)
- [Open WebUI + Ollama](#)
- [ComfyUI documentation](#)

SHIZU Growth Kits

This is a practical starting guide, not a guarantee that any tool will produce a particular result. Review the official documentation and licenses for your use case. Version 1.0 · 4 October 2026.

