ROUNDUP

Best Local AI Software Compared: Ollama, LM Studio, Jan & More

Six apps for running LLMs on your own hardware, compared on interface, local API, document chat and what the paid tiers cost.

Why run AI locally

Running a language model on your own machine gets you three things a hosted chatbot can't. Your prompts and files stay on your disk, there is no per-token bill, and the setup still works with the Wi-Fi off. The catch is hardware. Model size is limited by your RAM or VRAM, and a laptop won't run a model that needs a data centre to serve it.

The software layer has settled into a few clear types. Some tools are engines that serve models over an API. Some are desktop chat apps with a model browser built in. Others sit on top of an engine and add document chat, agents and workspaces. This roundup covers six of the most common options and sorts them by the job they do.

About scores

scored.tools rates tools on five criteria: quality of output, ease of use, pricing value, API and integration quality, and how well the tool solves the problem it claims to solve. None of the six tools below has a published scored.tools rating yet, so this piece gives no scores. The entries are grouped by type rather than ranked, and nothing here should be read as a ranking. Pricing and feature details come from each vendor's own pages as of November 2026.

The tools

Ollama

Score: not yet rated

Ollama is the engine most other local tools assume you have installed. You pull a model with one command, and it serves that model over a local REST API that other apps, scripts and editors can call. It runs on macOS, Windows and Linux, and it now ships a desktop chat window, though the command line and the API are where most people use it. Its pricing page lists a $0 pay-as-you-go tier that includes running models locally with no subscription. The paid plans are for Ollama's cloud models: Pro is $20/mo ($200/yr billed annually) with $60 of usage credits per month, and Max is

$100/mo with $300 of credits. If all you want is local inference, the free tier covers it. The cloud plans make sense when you want bigger open models than your hardware can hold while keeping the same API.

Best for: developers who want a local model endpoint to build against, or a backend for the GUI apps further down this list.

Pricing: local use is free. Cloud plans are Pro at $20/mo, Max at $100/mo, Team at $500/mo (early access) and Enterprise on custom terms.

Lm Studio

Score: not yet rated

LM Studio is a desktop app for finding, downloading and chatting with open models, and it is the easiest way here to get a model running without a terminal. It has a model browser, chat interface and local server in one window, plus a command-line tool for scripting. Its pricing page lists a Free plan covering local LLMs, offline voice transcription, limited web search and LM Link for up to 5 devices. Bionic+ costs $20 USD/month and adds US-hosted open-source models, web search with page extraction and LM Link for up to 20 devices. Pro is $100 USD/month with 5× usage limits and included usage of some hosted frontier models. The paid tiers push LM Studio toward hybrid local and hosted use. For purely local work the free plan is complete, and the app is closed source, which matters if you want to audit or fork it.

Best for: people who want a polished GUI for trying lots of models and serving them locally without touching config files.

Pricing: Free at $0. Bionic+ at $20 USD/month, Pro at $100 USD/month, and organisation plans with pooled credits or seats.

Jan

Score: not yet rated

Jan sells itself as an open-source ChatGPT replacement. It is a desktop chat app that runs open models and can also connect to hosted models such as ChatGPT, Claude and Gemini, so local and cloud chats sit in one interface. Its homepage says it is free and open source, with the code on GitHub, and lists cross-conversation memory as coming soon. Jan suits people who like the LM Studio approach but want source code they can inspect. The feature set is narrower than the workspace-style apps below, so if document chat across large folders is your main use, look at AnythingLLM or GPT4All first.

Best for: privacy-minded users who want an open-source desktop chat app that mixes local and hosted models.

Pricing: free and open source.

Gpt4all

Score: not yet rated

GPT4All is Nomic's desktop chatbot. Nomic says it needs "no cloud required, no data leaves your machine." It has installers for Windows (including Windows ARM), macOS and Ubuntu, and its headline feature is LocalDocs, which lets you chat with files stored on your machine. Nomic's page says it supports "thousands of models" and lets you customise assistants and workflows. The page lists no price for GPT4All itself. The $20/month figure shown there is for Nomic's separate Individual plans, so check which product you're paying for. GPT4All is a good fit for someone who wants offline document Q&A in a simple installer and doesn't need an API to build on.

Best for: non-developers who want offline chat with local documents on ordinary hardware, including Windows ARM laptops.

Pricing: no price listed for the GPT4All desktop app. Nomic's separate Individual plans are $20/month.

Anythingllm

Score: not yet rated

AnythingLLM, from Mintplex Labs, is a workspace app built around retrieval (RAG) and agents. You create workspaces, load documents into them and chat against those documents using whichever model provider you connect. It can use local engines like Ollama and LM Studio as well as hosted APIs. Per its pricing page, the desktop app is free and self-hosting with Docker is also free. Hosted cloud instances cost $50/monthly for Basic (you bring your own LLM API key) and $99/monthly for Pro with a 72-hour support SLA. On-premise deployment with SSO and RBAC sits under Enterprise. This is the strongest option here for a small team that wants a shared, self-hosted document assistant, since the Docker route keeps everything on your own server at no licence cost.

Best for: teams and power users who want document workspaces and agents on top of a local model backend.

Pricing: desktop and Docker self-hosting are free. Cloud Basic is $50/monthly, Pro is $99/monthly, and Enterprise is quoted on request.

Msty

Score: not yet rated

Msty Studio is a desktop front end that handles local and hosted models in one place, with features aimed at people who chat with models all day and want more control than a basic chat window gives. Its pricing page lists a Free tier at $0 forever, Aurum at $149 USD per user/year, and Aurum Lifetime at $349 USD per user as a one-time payment. Enterprise and Teams plans need a minimum of 5 seats. Msty has also widened into a family of products (Go, Nexus and Stack), some still in beta or preview, which makes the pricing page busier than it needs to be. Of the six tools here, Msty is the

only one asking for a licence fee for advanced desktop features rather than for hosted compute. You pay for the software itself, and whether that's worth it depends on whether you use the extra features.

Best for: heavy chat users who want one polished app for local and cloud models and don't mind paying for extra features.

Pricing: Free at $0. Aurum at $149 USD per user/year or $349 USD lifetime. Teams and Enterprise on request.

Comparison table

Tool	Type	Open source	Local use cost	Paid tier from	Paid tier buys
Ollama	Engine + API, basic chat app	Yes	Free	$20/mo (Pro)	Cloud model credits
Lm Studio	Desktop GUI + local server	No	Free	$20 USD/month (Bionic+)	Hosted models, more LM Link devices
Jan	Desktop chat app	Yes	Free	None listed	n/a
Gpt4all	Desktop chat + LocalDocs	Yes	No price listed	None listed for the app	n/a
Anythingllm	RAG and agent workspaces	Yes	Free (desktop or Docker)	$50/monthly (Cloud Basic)	Managed hosting, support SLA
Msty	Desktop front end, local + cloud	No	Free	$149 USD per user/year (Aurum)	Advanced app features

Final picks

These picks are based on fit with the feature set and pricing each vendor publishes, since none of the tools carries a scored.tools rating yet.

- **Building apps on a local model** calls for Ollama. It's free for local use, has a simple API, and most other tools here can use it as a backend.
- **The easiest way to try models with a GUI** is Lm Studio. You get model discovery, chat and a local server in one free app.
- **Open-source desktop chat** goes to Jan, for people who want to read the code behind the app holding their conversations.

- **Offline document chat for non-developers** goes to Gpt4all, thanks to LocalDocs and installers for every major OS, including Windows ARM.
- **A shared document assistant for a team** is Anythingllm self-hosted with Docker, pointed at an Ollama backend.
- **A power-user chat front end** is Msty, if the Aurum features justify $149 a year to you. Try the free tier first.

A common stack pairs two of these, with Ollama or LM Studio as the engine and AnythingLLM or Msty on top for documents and workflow. Everything in that pairing can run at no cost on hardware you already own, and you only pay once you want hosted models or managed infrastructure.

Generated 2026-11-22 · Read the live article at scored.tools/blog/best-local-ai-software-comparison/