# Running AI on Your Own Computer — A Plain-English Start

*A Beginner's Reference for Legal Professionals*

---

A starting guide for anyone who has never run an AI model on their own computer. It explains the words you'll hear, points you to the free apps that make it easy, and shows how people go from chatting with an AI to building one that can do tasks for them. No technical background needed.

## First, a few words explained

- **AI model** — the actual "brain" you talk to. ChatGPT is one example, but there are many, and some are free to download and run yourself.
- **Open model** — a model whose maker has released it publicly so you can download it and run it on your own machine, instead of only using it over the internet through someone else's website.
- **Running it locally** — the model lives and works on your own computer. Nothing you type gets sent out to another company. This is the part that matters most for us, and we'll come back to it.
- **Memory ("GB" / "VRAM")** — bigger, smarter models need more of your computer's working memory to run. When you see "8 GB" below, that's roughly how much room a model needs. A newer laptop handles small ones; a computer with a strong graphics card handles large ones.
- **License** — the legal terms for using a model. Some are free to use for business without strings; a few have restrictions. Worth a quick check before you rely on one for work.

## 1. The apps that make this easy

You don't install a model by hand. You install one simple app, and it downloads and runs the model for you — like an app store for AI brains. These are free.

- **LM Studio** — the friendliest place to start. A normal app with buttons and menus: browse a list, click to download a model, and start chatting. No commands. *If you try one thing, try this.*
- **Ollama** — popular, but you run it by typing short commands in a text window (for example, `ollama run qwen3`). Powerful, slightly more technical.
- **Jan / GPT4All** — other free, button-and-menu apps similar to LM Studio.
- **Hugging Face** — think of this as the giant public library where the models themselves are stored. The apps above quietly pull from here; you rarely need to visit it directly at first.

*Why do this on your own computer instead of a website? Because the model runs entirely on your machine, nothing you type ever leaves your control — which is exactly what our duty of confidentiality calls for. Using an AI website that sends client information to another company is a different question and shouldn't be treated as the same thing.*

### Prefer typing? Downloading through the terminal (optional)

The "terminal" is just a text window where you type instructions instead of clicking buttons — it's called **Command Prompt** or **PowerShell** on Windows, and **Terminal** on a Mac. It looks intimidating, but here you only ever type one short line. This is how **Ollama** downloads models:

- `ollama pull qwen3` — downloads the model and stops.
- `ollama run qwen3` — downloads it (if needed) and immediately starts a chat.
- `ollama list` — shows the models you've already downloaded.
- `ollama rm qwen3` — deletes one to free up space.

Type the line, press Enter, and a progress bar appears (models are a few gigabytes, so the first download takes a few minutes; after that it's instant). For a different model, swap the name — e.g. `ollama pull phi4`.

*Worth saying plainly: the terminal route and the button route do the same thing. LM Studio simply puts a friendly window over the exact same download. Nobody needs the terminal — use whichever you prefer.*

**Skip the terminal entirely: Hermes Desktop** (from Nous Research — the tool I use) is a normal app for Windows, Mac, and Linux that gives you all of this through buttons and menus, no commands required. On first launch it walks you through picking a model and sets things up for you; it uses Ollama in the background, so you get the same local, private setup without touching a text window. It's also the natural on-ramp to the "agents" in Section 3.

## 2. Which model to pick — by what your computer can handle

There are many good free models. The two things that decide which one is right for you are (1) whether your computer has enough memory to run it, and (2) the license. You don't need to memorize these — the apps above show you what will run. Here are sensible starting points:

- **A great all-purpose first pick: Qwen3**. It comes in several sizes; the small size needs only about 4.6 GB of memory, so it runs on a fairly ordinary newer laptop or Mac. Free to use for business.
- **If you have a gaming-style computer with a strong graphics card: DeepSeek R1** (good at step-by-step reasoning) or larger versions of Qwen3 will run and feel noticeably smarter.
- **For writing or reviewing code: Qwen coder** versions or **Devstral**.
- **If your computer is older or modest (about 8 GB): Phi-4** or **Llama 3.1 (8B)** are lightweight and still capable.

*The honest truth: there is no single "best" model. There's the best one for your computer and your task — and because new ones come out constantly, the answer changes. Start with one that runs well, and don't overthink it.*

## 3. Going further: "AI agents"

Everything above is about chatting with an AI — you ask, it answers. An **agent** is the next step up: instead of just answering, it carries out a task across several steps — looks things up, uses tools, checks its own work — with you supervising. Think of the difference between asking someone a question and handing them an assignment.

You don't build the machinery yourself. The simplest path is **Hermes Desktop** (from Section 1): it already *is* an agent. You just type what you want done in the chat — in plain English — and it plans the task, uses its tools, and carries it out. Nothing to wire up. It even remembers across sessions and gets better at your recurring tasks over time.

If you want to see what else is out there, other tools let you build agents too. They fall into two kinds, depending on whether you want to write code:

### If you don't code (start here):

- **n8n** — you build the agent by dragging boxes and connecting them on screen, like a flowchart. The most approachable option, and it includes safety controls so a human can review before anything important happens.
- **Dify, Flowise, Lindy AI, Relevance AI** — similar tools where you describe the agent in plain English and it's built for you.
- **Zapier, Make** — familiar automation tools (some firms already use them) that now have AI-agent features built in.

### If you (or someone you hire) can code:

- **CrewAI** — the easiest coding option; you set up a small "team" of agents with roles.
- **LangGraph** — the most powerful and reliable, but the hardest to learn. Best when an agent needs to run dependably for real work.

- **Microsoft Agent Framework, OpenAI Agents SDK** — other well-supported professional options.

*One honest note: often you don't need a separate agent at all. A single good AI request, or Hermes handling it directly in the chat, is usually enough.*

---

**Questions are welcome. — Travis M. Keil** • Keil Defense • 651.315.3097