Running Your Own AI in the Building

A Plain Guide to Local AI for Small Business Owners

John Hodsdon

Agentic Solutions

A free resource from agenticsolutions.biz

October 2026

Why I Wrote This

Most owners I talk to have tried ChatGPT, Grok, or something like them. Many have the same two worries. First, “Where is my information going when I paste it in there?” Second, “What is this going to cost me once my whole team is using it?”

There is another option that gets less attention: running an AI model on a computer you own, sitting in your own office. This is called a local AI, or a local LLM. I am setting one up myself right now, and this guide explains what I have learned about what it is, what it is good for, what it is not good for, and what equipment you actually need.

I will be honest about the limits. A local AI is a useful tool for some jobs. It is not a replacement for everything, and it is not right for every business.

Part 1: What a Local AI Is

The plain version

An **LLM** (large language model) is the engine behind tools like ChatGPT. It is a very large file of numbers that has been trained on huge amounts of text. When you type a question, the model predicts a useful answer, one word piece at a time.

When you use **cloud AI** (ChatGPT, Grok, Claude, Gemini, and others), that model runs on the provider’s computers in a data center. Your question travels over the internet to them, their computers work out the answer, and it travels back. You pay with a subscription, a per-use fee, or by letting them use your data, depending on the plan.

A **local AI** is the same kind of engine, but the model file sits on your own computer and your own computer does the work. Your question never leaves the building. Think of it as the difference between sending your books out to an outside firm and having a bookkeeper on staff. The outside firm may be more experienced, but the staff bookkeeper never takes your files out of the office.

Where these models come from

Several companies publish “open” models that anyone can download for free. Google publishes Gemma. Alibaba publishes Qwen. Meta publishes Llama. Mistral, a French company, publishes Mistral models. DeepSeek, a Chinese company, publishes DeepSeek models. OpenAI has also released an open model called gpt-oss. These are not the exact same models that power the top paid cloud services, but some of them are surprisingly good.

You download one of these models (usually a file of 2 to 40 gigabytes), load it into a free program on your computer, and start typing to it, much like you would with ChatGPT.

Part 2: The Benefits

1. Your data stays in the building. This is the main reason most owners look at local AI. Payroll, employee reviews, customer lists, contracts, bank statements, medical or legal details: none of it goes to an outside company. You do not have to read a privacy policy and hope. If the computer is unplugged from the internet, nothing can leave.

2. No per-use fees. Once you own the computer, running the model costs you electricity and nothing else. Five people can use it all day and the bill does not change. Cloud AI for a team usually runs a monthly fee per person, and heavy use of the paid business tiers can add up.

3. Works offline. If the internet goes down, the local AI keeps working. That matters less for most businesses than the first two points, but it is real.

4. Control. Cloud providers change their models, prices, and rules whenever they decide to. A model you downloaded stays exactly the same until you choose to change it. If it does a task well today, it will do it the same way next year.

5. Customization. You can give a local model standing instructions ("You write in our company's voice. Always use our job numbering format."), connect it to a folder of your own documents, and pick different models for different jobs. You can do some of this with cloud AI too, but locally you are not limited by what a provider allows.

Part 3: The Limitations, Honestly

1. Smaller models are less capable than the best cloud models. This is the big one. The top cloud models run on rooms full of hardware that cost millions of dollars. A model that fits on a business computer is smaller. For everyday writing and summarizing, the gap is often small. For hard reasoning, tricky math, complex analysis, or long multi-step problems, the top cloud models are clearly better. Local models also make things up (the industry calls this "hallucinating") somewhat more often, so you check their work.

2. Setup and upkeep are on you. No one at a help desk is going to fix it. The software has gotten much easier, and a first install takes an afternoon, but someone in your business has to be comfortable installing programs and following instructions.

3. Hardware costs money up front. You can try it on a computer you already own for free. A computer built to run better models costs roughly $1,500 to $3,500 or more, and graphics card prices are high right now (see Part 4).

4. Speed. On a good machine, a local model types out answers about as fast as you can read, or faster. On an older laptop it may be slow, a few words per second. Big models on modest hardware can be painfully slow.

5. No built-in web access. A local model only knows what it learned when it was built, which may be a year or more out of date. It cannot look up today's prices or news unless you set up extra tools. Cloud services usually search the web for you.

6. Power and heat. A computer with a strong graphics card draws real power when it is working, and puts out heat and some fan noise. It is not a big deal for one machine in an office, but it is not a laptop sitting quietly on a desk either.

7. Keeping models current. New and better open models come out every few months. That is good news, but someone has to notice, download the new one, and test it. Old models do not update themselves.

Part 4: The Equipment

Three terms you need to know

RAM (random access memory) is the computer's short-term working space. When you load a model, it has to fit into memory. More RAM means bigger models can be loaded. Most business laptops have 8 or 16 GB. A computer meant for local AI should have 32 GB or more.

GPU and VRAM. The **GPU** (graphics processing unit, or graphics card) was built for video games, but it turns out to be very good at the kind of math AI models do. It has its own separate, very fast memory called **VRAM** (video RAM). This is the single most important number for local AI on a Windows computer. If the whole model fits in VRAM, it runs fast. If it does not fit, part of it spills into regular RAM and it slows down a lot. NVIDIA graphics cards are the easiest choice because nearly all the AI software supports them well.

Apple's unified memory. Apple's newer Macs (with M-series chips) do not have separate RAM and VRAM. They share one pool, called unified memory, between the main processor and the graphics part of the chip. That means a Mac with 64 GB of unified memory can load models that would need a very expensive graphics card on a Windows computer. The tradeoff is that Macs are generally slower at producing answers than a high-end NVIDIA card, and you cannot upgrade the memory later.

Quantization is a way of shrinking a model so it fits on smaller hardware. Think of it like saving a photo at a lower quality setting. The file gets much smaller, and the picture still looks almost the same unless you look closely. A model stored at "4-bit" quantization is about a quarter the size of the original and loses only a little quality. This is why a model with 27 billion "parameters" (the number usually printed in the name, like 27B) can sometimes fit on a 12 GB card. Shrink it too far, though, and quality drops noticeably.

Model size. You will see models labeled 3B, 8B, 14B, 27B, 70B. The B stands for billions of parameters. Bigger generally means smarter and slower and needing more memory. As a rough rule of thumb, a model at 4-bit quantization needs a little over half a gigabyte of memory per billion parameters, plus some extra room for the conversation itself. So an 8B model needs around 5 to 6 GB, and a 70B model needs over 40 GB.

One newer twist: some models are built as "mixture of experts." Only a small part of the model works on each word, so they run faster than their size suggests. You may see names like 35B-A3B, meaning 35 billion total but only about 3 billion active at a time. These can run reasonably on a mid-range computer if it has plenty of regular RAM.

The three tiers

Prices below are what I found in October 2026. Graphics card prices in particular are well above their original list prices right now. Trackers show the RTX 5070 at roughly $770 to $890 new against a $549 launch price, and the RTX 5090 several times its $1,999 launch price. Check current prices before you buy.

Tier	Rough cost	Typical equipment	Models that fit	Good for
Entry	$0 (use what you have)	Existing laptop or desktop, 16 GB RAM, no special graphics card, or a Mac with	Small models, 3B to 9B, such as Llama 3.2 3B, Qwen 3.5 9B, Gemma 4 E4B	Trying it out, short drafts, simple summaries, rewording emails.

Tier	Rough cost	Typical equipment	Models that fit	Good for
		16 GB		Can be slow.
Mid	About $1,800 to $2,500 for a new desktop	32 to 64 GB RAM, NVIDIA card with 12 to 16 GB VRAM (RTX 5070 12 GB, RTX 5060 Ti 16 GB)	8B to 14B models comfortably (Gemma 4 12B, Qwen 3.5 9B, Qwen3 14B). Some larger models squeezed down with quantization or split with system RAM	Daily drafting, document summaries, internal Q&A over company files, private data work
High	About $2,500 to $4,000 and up	NVIDIA card with 24 GB+ VRAM (used RTX 3090, RTX 4090, RTX 5090), or a Mac with 64 GB+ unified memory (Mac mini M4 Pro 64 GB, Mac Studio)	27B to 35B models comfortably (Qwen 3.8 27B, Gemma 4 26B, Qwen 3.6 35B-A3B). 70B models such as Llama 3.3 70B at 4-bit on 48 GB+ of memory, often slowly	Heavier analysis, longer documents, better quality writing, coding help, several users at once

A few notes on the table:

- **Entry tier.** If your laptop has 16 GB of RAM, you can try local AI this week for free. It will work. It will not impress you on hard questions, and it may be slow, but it will show you what the experience is like before you spend anything.
- **Mid tier.** This is where most small businesses should land if they decide to go further. A 16 GB card gives more breathing room than a 12 GB card. One price tracker in October 2026 listed a new RTX 5060 Ti 16 GB below a new RTX 5070 12 GB, so for AI use alone, the cheaper card with more memory can be the better buy. The RTX 5070 is faster, though, for models that fit in it.
- **High tier.** A Mac mini M4 Pro with 64 GB of unified memory was listed around $2,500 new in October 2026, and a Mac Studio M4 Max with 64 GB around $3,600. Per estimates from hardware tracking sites, the Mac mini can load a 70B model but produces answers at only about 5 words per second on it, which is slow for chat. On the Windows side, a top card like the RTX 5090 (32 GB) is fast but was selling for several thousand dollars in October 2026, far above its launch price.

What else matters

- **Storage.** Models are big files. A 1 TB solid state drive (SSD) is a sensible minimum if you plan to keep several models around.
- **Power supply.** A desktop with a mid or high tier graphics card needs a power supply sized for it. Follow the card maker's recommendation.
- **Operating system.** Windows 11, macOS, and Linux all work. This guide's walkthrough uses Windows 11 because that is what most businesses run.

Part 5: The Software

You need two things: a program that runs the model (the "engine"), and something to type into (the "front end"). All three tools below are free.

Ollama (my main recommendation). Ollama is a free program that downloads models and runs them for you. You tell it which model you want by name, and it handles the rest. It runs on Windows, Mac, and Linux, it has a large library of models at ollama.com, and many other tools connect to it. It started as a command line tool (meaning you type commands in a text window), and it now also includes a simple chat window on Windows and Mac.

LM Studio. LM Studio is a free desktop program with a friendlier, point-and-click look. You browse models, click download, and chat. If the idea of typing commands puts you off, start here. It is a good way to try several models quickly. It is free for business use under its current terms, but read them yourself.

Open WebUI. Open WebUI is a free chat screen that looks and feels much like ChatGPT, and it runs in your web browser. It connects to Ollama in the background. It adds things a team wants: separate logins for each person, saved conversations, and the ability to upload documents and ask questions about them. Other people in the office can use it from their own computers over your office network, while the model itself runs only on the one AI computer.

A simple way to think about it: Ollama is the engine, Open WebUI is the dashboard and steering wheel, and LM Studio is an all-in-one alternative for one person at one desk.

Your first afternoon: a step-by-step walkthrough on Windows 11

This assumes a Windows 11 computer with at least 16 GB of RAM. Set aside two to three hours. Menus and version numbers change, so if a screen looks a little different from what I describe, look for the closest match.

Step 1. Check your memory and graphics card. Right-click the Start button, choose Task Manager, and click the Performance tab. Click “Memory” to see your RAM. Click “GPU” to see your graphics card and its “Dedicated GPU memory.” That dedicated number is your VRAM. Write both down.

Step 2. Update your graphics driver (if you have an NVIDIA card). Go to nvidia.com, find the driver download page, and install the latest driver for your card. Restart if asked.

Step 3. Install Ollama. Go to ollama.com and click Download, then choose Windows. Run the installer file and click through it. When it finishes, Ollama runs quietly in the background. You will see a small llama icon near the clock.

Step 4. Download and run your first model. Open the Start menu, type “Terminal” (or “PowerShell”), and open it. Type the following and press Enter:

ollama run qwen3.5:9b

On an entry level laptop with no graphics card, start smaller instead:

ollama run llama3.2:3b

The first time, it will download the model, which can take several minutes depending on your internet speed. When you see a prompt that says “Send a message,” type a question and press Enter. To leave, type /bye.

Model names on Ollama change as new versions come out. If a name does not work, look it up in the model library on ollama.com and copy the exact name shown there.

Step 5. Give it a real test. Try three things from your actual work, using nothing sensitive yet: - “Rewrite this email to sound friendlier but still firm:” and paste an email. - “Summarize this in five bullet points:” and paste a page of a document. - “Write a first draft checklist for opening the office each morning.” Judge the results the way you would judge a new hire’s first draft.

Step 6. Add a chat screen (optional but recommended). Open WebUI is the friendliest screen, but installing it takes a few more steps. The Open WebUI website (docs.openwebui.com) lists the current install options for Windows. The common routes are installing Docker Desktop (a free program that runs packaged software) and then running the single command Open WebUI gives you, or installing Python and using its built-in installer. Once it is running, you open it in your web browser, create the first account (which becomes the administrator), and it finds Ollama on its own.

If that sounds like too much for day one, install LM Studio instead (lmstudio.ai). It gives you a chat window in a few clicks.

Step 7. Decide what is next. At the end of the afternoon, you will know whether the answers are good enough to be useful, and whether your current computer is fast enough. That tells you whether to stop, keep using what you have, or buy better equipment.

Part 6: Which Models to Try

New open models come out constantly. The names below are what hardware review and benchmark sites were recommending in October 2026. By the time you read this, there may be newer versions. The general idea will still hold: pick a well known family, and pick the largest size that fits comfortably in your memory.

Entry tier (16 GB RAM, no graphics card or a small one) - **Llama 3.2 3B** (Meta). Small and quick, fine for short rewrites and simple questions. - **Qwen 3.5 9B** (Alibaba). Strong for its size, good general assistant. May be slow without a graphics card. - **Gemma 4 E4B** (Google). Compact, fast, and can also read images.

Mid tier (12 to 16 GB VRAM) - **Gemma 4 12B** (Google). Several October 2026 reviews named it the best all-round pick for a 12 GB card. Good writing, can read images and scanned pages. - **Qwen 3.5 9B** (Alibaba). Fast, good with instructions, good for working with other tools. - **Qwen3 14B** or **Qwen 2.5 Coder 14B** (Alibaba). A step up in size; tight on 12 GB, comfortable on 16 GB. - **gpt-oss 20B** (OpenAI). A common pick for 16 GB cards. Reviews say it does not fit well on a 12 GB card.

High tier (24 GB+ VRAM or 64 GB+ Mac) - **Qwen 3.8 27B** (Alibaba). Among the most capable models that fit in this tier. - **Gemma 4 26B** (Google). - **Qwen 3.6 35B-A3B** (Alibaba). A “mixture of experts” model that runs faster than its size suggests. - **Llama 3.3 70B** (Meta). Needs over 40 GB of memory at 4-bit. Workable on a 64 GB Mac, but slow.

Mistral and DeepSeek. Mistral (France) and DeepSeek (China) both publish open models, and both appear in the Ollama library. DeepSeek’s largest new models are far too big for a business computer, but smaller “distilled” versions (shrunk copies trained to imitate the big one) are available. Some owners have concerns about using Chinese models. When a model runs locally with no internet connection, your data does not go back to the company that made it. That is the point of running it locally. Still, you are free to stick with models from U.S. and European companies if that makes you more comfortable.

Licenses. Most of these models are free for business use, but each has its own license, and some have conditions. Read the license on the model’s page before you build your business around one.

Part 7: Best Business Uses, and What to Leave to the Cloud

Where local AI earns its keep

First drafts of routine writing. Emails, customer letters, job postings, policy updates, meeting notes. A mid tier model handles these well. You edit, you send.

Summarizing documents. Long contracts, vendor proposals, inspection reports, meeting transcripts. “Give me the five things I need to know” is a job these models do well. Always check anything important against the original.

Questions about your own files. With Open WebUI or a similar tool, you can load your employee handbook, procedures, price sheets, and past proposals, and then ask questions like “What is our policy on overtime for part-time staff?” The tool finds the relevant pages and the model answers from them. The industry calls this **RAG** (retrieval augmented generation), which just means “look it up in your documents first, then answer.” This is one of the most useful setups for a small business, and one of the easiest to get slightly wrong, so test it with questions you already know the answers to.

Sensitive information. This is where local AI makes the most sense: - Financials: reviewing statements, categorizing expenses, drafting notes for your accountant. - HR: drafting performance reviews, writing up incidents, summarizing employee feedback. - Customer information: drafting replies that include account details, cleaning up customer lists. - Anything covered by a confidentiality agreement.

First drafts of SOPs. SOP stands for standard operating procedure, the written “how we do it here” steps for a task. Talk through the task (or type rough notes), and have the model turn it into a clean, numbered procedure. The person who actually does the job reviews and corrects it.

What to leave to cloud AI

- **The hardest thinking.** Complex strategy questions, tricky financial modeling, legal or tax reasoning where the details matter. Use the strongest model available, and still verify.
- **Anything needing current information.** Market prices, recent regulations, news, competitor research. Cloud tools can search the web.
- **Large or unusual projects.** Very long documents, detailed analysis across many files, or tasks where the local model keeps getting it wrong.
- **Polished final pieces.** Marketing copy or important client communication where quality matters most.

The hybrid approach

My recommendation for most businesses is not “local or cloud.” It is both, with a simple rule:

Private or routine goes local. Hard or public goes to the cloud.

If it contains employee, customer, or financial information, it stays on the local machine. If it is a hard problem that does not involve private details, or it needs current information from the web, use a cloud service. When a hard problem does involve private details, remove the names and numbers first, ask the cloud tool, then apply the answer yourself.

Write that rule down and share it with your team. It does more for your data safety than any piece of software.

Part 8: What I Am Setting Up and Why

I am practicing what I am describing here. A new desktop arrives at my office the week of October 26, 2026, bought for one job: running a couple of local models for part of my business. I have not used it yet, so I cannot share results. What I can share is what I bought, why, and what I expect it can and cannot do, based on the research in this guide.

The machine: - Processor: Intel Core i7-14700 - Memory: 64 GB RAM - Graphics card: NVIDIA GeForce RTX 5070 with 12 GB VRAM - Storage: 1 TB NVMe SSD (a fast type of solid state drive) - Motherboard: Gigabyte B760 - Power supply: 600 watts - Operating system: Windows 11 - Software: Ollama, with a chat front end on top

Why this setup. This is a mid tier machine. The 64 GB of RAM is more than the graphics card alone needs, and that is on purpose. When a model is too big for the 12 GB card, the extra RAM lets part of it run from regular memory. That is slower, but it works, and it gives me room to try larger models. The RTX 5070 is fast for models that fit inside its 12 GB.

What it should run well. Based on published October 2026 tests on 12 GB cards: - Models in the 8B to 12B range, such as Gemma 4 12B and Qwen 3.5 9B, should fit fully on the card and run fast, well faster than reading speed. - 14B models should fit with careful settings, but with less room for long conversations or long documents.

What it can stretch to. Some reviewers report running a 27B model (Qwen 3.8 27B) on a 12 GB card by shrinking it to about 3-bit, and running Qwen 3.6 35B-A3B by keeping part of it in system RAM. Both are possible, but they are tighter, slower, or lose a bit of quality. I plan to test them, not count on them.

The honest limits. - **12 GB of VRAM is the ceiling that matters.** It is the biggest limit on this machine. A 16 GB card would have given more room, and a 24 GB card much more. I chose this card as a balance of speed, price, and availability, knowing that. - **Long documents use memory too.** The longer the document or conversation, the more memory the model needs beyond its own size. On 12 GB, a big document plus a big model will not both fit on the card. - **70B models are not realistic on this machine** for daily use. They would mostly run from system RAM and be very slow. - **It will not match the top cloud models.** I expect it to handle routine drafting, summaries, and private document work. I will still use cloud AI for the hardest problems.

What I plan to do with it. Drafting and summarizing work that involves private client and business information, building a searchable knowledge base from my own documents, and testing which models do which jobs best. Once I have real experience with it, I will write a follow-up with actual results, good and bad.

Part 9: Is This Right for You?

Go through this list. The more you check, the more sense local AI makes for your business.

- ☐ We handle sensitive information (financial, HR, customer, medical, legal) that I would rather not paste into a cloud service.
- ☐ Several people on my team would use AI regularly, and per-person subscriptions are adding up.
- ☐ Most of what we would use AI for is routine: drafts, summaries, rewrites, looking things up in our own documents.

- ☐ Someone on my team (or I) is comfortable installing software and following step-by-step instructions.
- ☐ I am willing to spend an afternoon testing it on an existing computer before buying anything.
- ☐ I can budget roughly $1,800 to $2,500 for a dedicated machine if the test goes well.
- ☐ I am fine with answers that are good but not always as strong as the best cloud tools.
- ☐ I will still use cloud AI for the hardest problems and anything needing current information.

If you checked most of these, start with the free test in Part 5 on a computer you already own.

If you checked only a few, cloud AI with a clear rule about what not to paste in is probably enough for now. Revisit this in a year. The models keep getting better and smaller.

Keep Learning

If you want more plain-English guidance like this, sign up for the free **Main Street AI Brief** at agenticsolutions.biz. It covers practical AI for small business owners and managers: what is worth trying, what is not, and how to put it to work without the hype. When I have real results from my own local AI machine, that is where I will share them first.

Sources

Prices and model recommendations were checked in October 2026. Prices change often.

1. ComputingForGeeks, “Ollama Models List: Best Models by VRAM,” October 2026. https://computingforgeeks.com/ollama-models-cheat-sheet/
2. Atomic Chat, “Best Local LLMs for 12GB VRAM in 2026,” September 2026. https://atomic.chat/blog/guides/best-local-llm-12gb
3. SpecPicks, “Best Local LLM for 12GB VRAM in 2026.” https://specpicks.com/reviews/best-local-llm-12gb-vram-2026
4. OpenClawDC, “Best Local LLM for 12GB VRAM (2026),” October 1, 2026. https://openclawdc.com/blog/best-local-llm-12gb-vram/
5. Local LLM benchmark results (JConradoN), GitHub. https://github.com/JConradoN/local-llm-benchmark
6. BestValueGPU, “RTX 5070 Price History,” October 2026. https://bestvaluegpu.com/history/new-and-used-rtx-5070-price-history-and-specs/
7. videocardprices.com, “GPU and AI Hardware Prices,” updated October 24, 2026. https://videocardprices.com/
8. rtx.live, RTX 50 series prices and MSRPs. https://rtx.live/
9. BasketPrice, “GPU prices, October 2026.” https://www.basketprice.com/gpu
10. ReadyGPU, “Mac mini M4 Pro 64GB” and “Mac mini M4 Pro 64GB vs Mac Studio M4 Max 64GB,” October 2026. https://readygpu.com/us/machine/mac-mini-m4-pro-64gb
11. LLMCheck, “Best Local LLMs for Mac mini M4 Pro 64GB.” https://llmcheck.net/best-llm/mac-mini-m4-pro-64gb/
12. Ollama, downloads and model library. https://ollama.com
13. LM Studio. https://lmstudio.ai
14. Open WebUI documentation. https://docs.openwebui.com