About

Local LLM stands for well running LLM's on your own hardware at home or at least that you control. In other words your own ChatGPT or Claude LLM that you can talk to without you sharing any piece of data with these companies. This means your data stays local, completely private and you can safely create directories, notes and save all your conversations without the need of Big Tech. On top of that as we will see as well, you can use these local LLM's as plugins for your open source applications and on top of that as an engine for your local AI agent that we will explore after this.

Now, Local LLM's can run on any type of hardware that has enough resources such as a phone, tablet, laptop and PC or on your HomeLab server. We will mainly focus on getting it deployed on your home server, but will discuss the phone and desktop options as well. The reason

Phones have tighter memory, processing, and thermal limits, so they’re generally better suited to smaller models and brief tasks. A server can accommodate more capable models, but accessing it remotely adds network latency and means you’ll need to keep the server running and secure its connection. On-device processing can work offline and avoid sending prompts elsewhere; a self-hosted server can also keep data under your control, provided its remote access is configured securely.

Rule of thumb: Choose the phone for convenience and offline use; choose the server for capability and sustained workloads. A hybrid setup is often practical: use the phone for quick or sensitive tasks, and send heavier tasks to your server when you have a connection

You don’t necessarily need a separate harness like Ollama. Hermes needs access to a model, and that can happen in either of two ways:

- Hermes manages the local runtime itself: Hermes can download and run a local model using its managed `llama.cpp` runtime.
- Hermes connects to a separate model provider/server: You can run a model with Ollama, then configure Hermes to connect to Ollama’s local API endpoint.
- Use a hosted model: Hermes can connect to a cloud provider instead; in that setup, the provider runs the model.

So, Ollama is one possible model-serving runtime, not a mandatory part of every Hermes setup. Hermes itself handles the agent’s orchestration and tools; the model can be local or hosted. The model and runtime also need to support tool calling for the agent to use tools reliably.

Revision #1
Created 2026-11-20 17:46:10 UTC by dpasn
Updated 2026-11-20 18:08:44 UTC by dpasn