

To use LLaMA with Ollama on your local PC, you can follow this streamlined guide. This assumes you're using LLaMA 2 via Ollama, which is the most straightforward way to run Meta's LLaMA models locally.

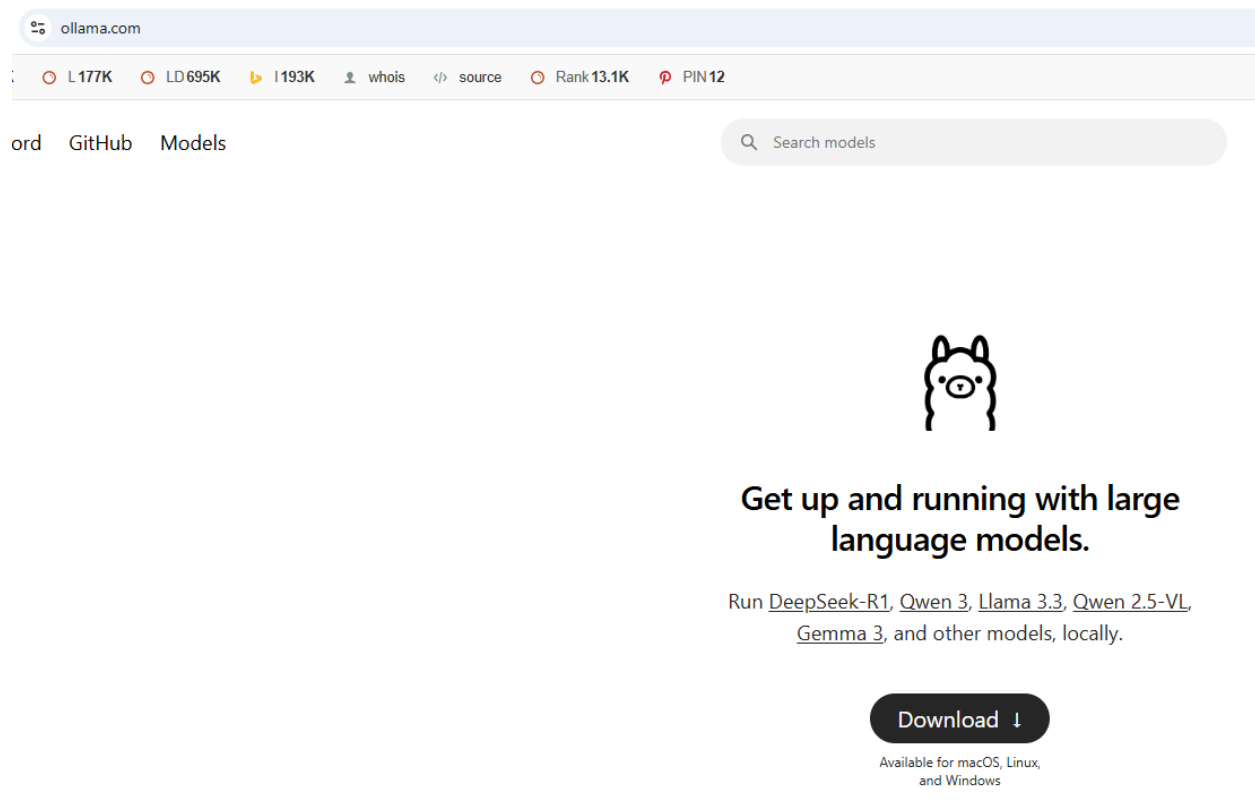
🔗 Step-by-Step: Run LLaMA Locally Using Ollama

📄 1. System Requirements

- OS: macOS, Linux, or Windows (with WSL2)
- RAM: 8 GB minimum (more for larger models)
- Optional: NVIDIA GPU for better performance

🔗 2. Install Ollama

Download from <https://ollama.com/> & install it.



The image shows a screenshot of the Ollama website homepage. At the top, the browser address bar shows 'ollama.com'. Below the address bar, there are several navigation links: 'ord', 'GitHub', and 'Models'. A search bar with the placeholder text 'Search models' is also visible. The main content area features a large, stylized llama logo. Below the logo, the text reads 'Get up and running with large language models.' Underneath this, it says 'Run [DeepSeek-R1](#), [Qwen 3](#), [Llama 3.3](#), [Qwen 2.5-VL](#), [Gemma 3](#), and other models, locally.' At the bottom, there is a prominent 'Download ↓' button, and below it, the text 'Available for macOS, Linux, and Windows'.

3. Run LLaMA Model

Open your command prompt:-

```
ollama run llama2
```

This will:

- Automatically download the LLaMA 2 model (quantized .gguf version)
- Start an interactive shell

You can now type prompts directly into the terminal.

4. Example Interaction

```
> ollama run llama2
```

```
> What is the capital of Italy?  
Rome.
```