

# Self-Hosted AI

Running LLMs on local hardware, or potentially self-managed cloud services.

- [Ollama](#)
- [Open WebUI](#)
  - [Error Troubleshooting](#)
  - [Making Docker see NVIDIA GPUs on Linux](#)
- [Stable Diffusion](#)
  - [AUTOMATIC1111](#)

# Ollama

A FOSS tool that helps with interacting directly with LLMs locally.

# Open WebUI

A useful web utility for easier, but also more powerful, interaction with local LLMs.

Open WebUI

# Error Troubleshooting

If you have an error such as the following:

```
Ollama:500, message='Internal Server Error', url=URL('http://127.0.0.1:11434/api/chat')
```

Likely you need to update Ollama, as Open WebUI has been updated too far to be able to interact with your current Ollama version (or vice versa). This is true even if you have Open WebUI installed via a package manager such as Pinokio.

The current way to update Ollama on your local machine is `curl https://ollama.ai/install.sh | sh`

.

Further, you can check your current Ollama version with, who would have guessed it, `ollama --version`.

[Open WebUI](#)

# Making Docker see NVIDIA GPUs on Linux

## Fedora Linux:

Run `sudo dnf install nvidia-container-toolkit nvidia-driver` to install NVIDIA's toolkit for interacting with containers, as well as the proprietary drivers for your NVIDIA GPU.

`sudo nvidia-ctk runtime configure --runtime=docker` configures Docker to be able to use the GPU.

`sudo systemctl restart docker` restarts the Docker service.

`docker info | grep Runtimes` checks if Docker can see the "nvidia" runtime.

# Stable Diffusion

Stable Diffusion

# AUTOMATIC1111