How to set up an instance of Open WebUI and Ollama for use in the setting up of local instances of LLM chatbots similar to ChatGPT and DeepSeek R1 through VMware.

1. Make a VM of whatever OS is your choice. As long as it is compatible with Docker, then it should work fine.
2. Ensure your VM has a lot of hardware behind it, both in memory and in storage. LLM’s are very intensive. I made a test VM I called James\_DeepSeek which I gave 32 GB of Ram and 128 GB of storage. This may have been unnecessary, but I wanted to aim bigger so I could in theory use more powerful model. Also the more CPU’s the better.
3. Once your VM is set up, install docker.

Run the following command: “docker run -d -p 3000:8080 -v ollama:/root/.ollama -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:ollama”

1. Once the installation is complete, go to your VM’s IP with a port of 3000 (for example if your VM’s IP is 192.168.6.19, then go to 192.168.6.19:3000 in your browser)
2. Set up an administrator account and then you are in!
3. There are no models to start, so I recommend looking at <https://ollama.com/search> and looking for a model you think you can handle. Do some shopping.
4. Run the ollama pull command in the search bar in the web ui and it should download and install the model for you!
5. In theory you can cut off internet connection from these LLM’s and they will work the same, but navigating the setting of environment variables to run in offline mode by default in these containers is difficult and I have yet to be able to do it successfully.