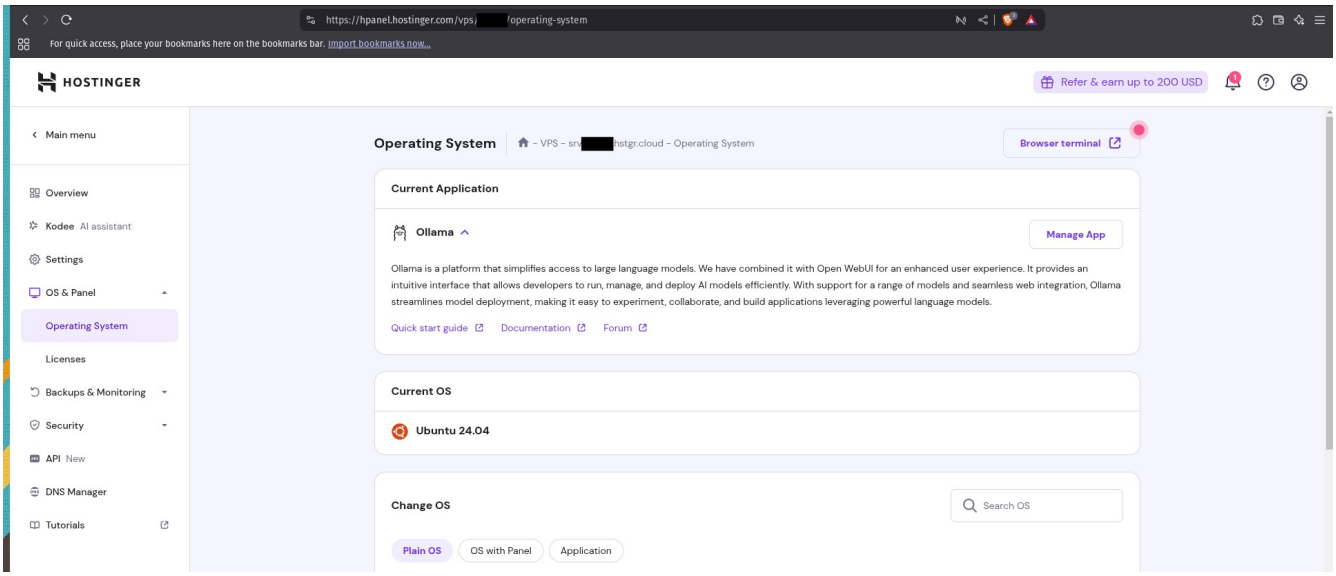
A.I. Rolodex

The option this guide uses is a VPS (Virtual Private Server) which is a cloud compute instance on the internet. Think of it as someone's computer that you rent for a monthly fee. You can also choose to install and run this entire setup on your own local computer and network.

1. Visit Hostinger.com and choose your plan (I used KVM2 at $6.99/mo for 24 months). Create your account login with Hostinger and enter payment details. Document your password in a password manager.

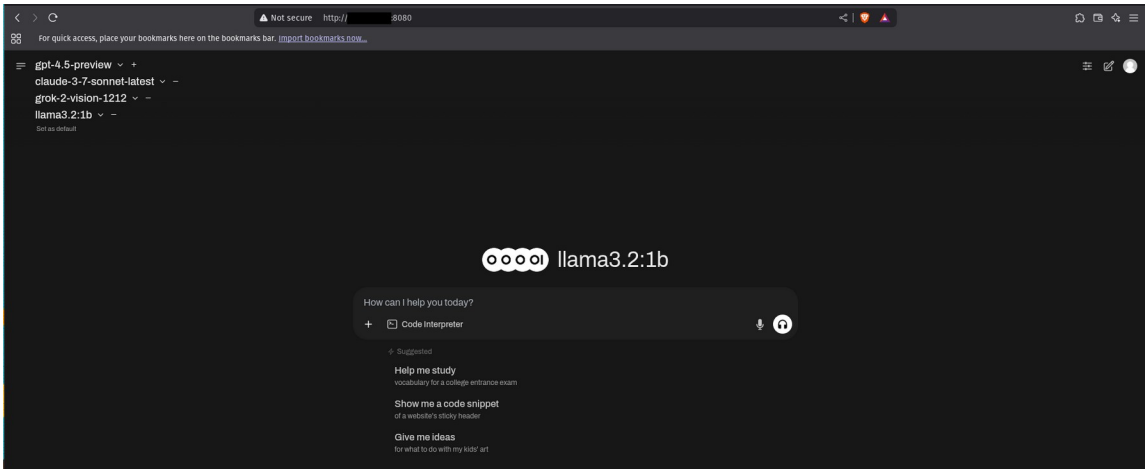
2. Once logged back into the dashboard, click on Application, and choose Ollama / Ubuntu 24.04. This will take a few minutes to deploy, so let it finish and come back in about 10 minutes.

3. Once the OS has finished installation, click on 'VPS Management page'. Click on 'Manage' to the right of your Ollama instance.



4. Click 'Manage App' button near top right which will open your OpenWebUI on your local network at the IP address you were given by Hostinger VPS, on port 8080.

5. Once the OpenWebUI tab launches and loads, create your admin account login and document it in your password manager. By this step, you can now start using llama3.2:1b or later version AI model.



6. Open a new tab and go to openai.com/api/, create an account and document your credentials in your password manager. Add payment details and top up your account with at least $5. Create an API key.

7. Go back to your Open WebUI tab and click account button on top right and select Admin Panel. Visit Settings, then click on Connections tab. You should see a section under OpenAI API that says ‘Manage OpenAI API Connections with a link like this: https://api.openai.com/v1.

Paste your OpenAI API key to the right in ‘API’ field and click save. You can now access all of OpenAI’s models including the latest ones, in addition to your llama 3.2 model (click the top left menu icon and select ‘New Chat’ to begin using your AI models. You can select as many as three models to run concurrently (this is the true power of this setup). You can continue the chat with all two or three models concurrently, or at the bottom right of the response, you can select ‘Merge’ and it will use all two or three models and distill it together into one. Very fascinating to watch this in action.

8. Optional: Setup additional OpenWebUI User accounts if you wish to give your family or friends access to your AI models. You may wish to add additional models first though before doing this.

9. We are going to add LiteLLM Proxy Server that installs along side OpenWebUI that allows us to run nearly any other AI model such as Grok, DeepSeek, Claude, etc.

10. Back in your Hostinger VPS account dashboard, click on ‘Browser Terminal’ button to launch a terminal instance of your VPS, we need to install and configure LiteLLM here.

11. In the terminal prompt in your new tab, type the command:

```
git clone https://github.com/BerriAI/litellm
```

12. Type ‘ls’ to ensure we downloaded litellm folder. Jump inside that folder by typing ‘cd litellm’, once inside that folder, type ‘nano .env’ and hit enter to enter the hidden config file. Inside the file, you will enter:

```
LITELLM_MASTER_KEY="sk-xxxxxxxxxxxx"
```

(generate any random alphanumeric characters here including uppercase, lowercase and numbers.)

On the next line, enter:

```
LITELLM_SALT_KEY="sk-yyyyyyyyyy"
```

So your .env file should look like this with the ‘xxx’ and ‘yyy’ strings being any length random characters, make sure to copy this into your password manager. Your file should look like this:

```
LITELLM_MASTER_KEY="sk-xxxxxxxxxxxx"
LITELLM_SALT_KEY="sk-yyyyyyyyyy"
```

Hit Ctl+X, Y and Enter to save and exit this file.

This ‘sk-xxxxxxxxxx” string will be your password to login to the LiteLLM dashboard, so store this in your password manager. After you save/exit the .env file, type and enter this into your terminal to build our LiteLLM config:

```
docker-compose up -d
```

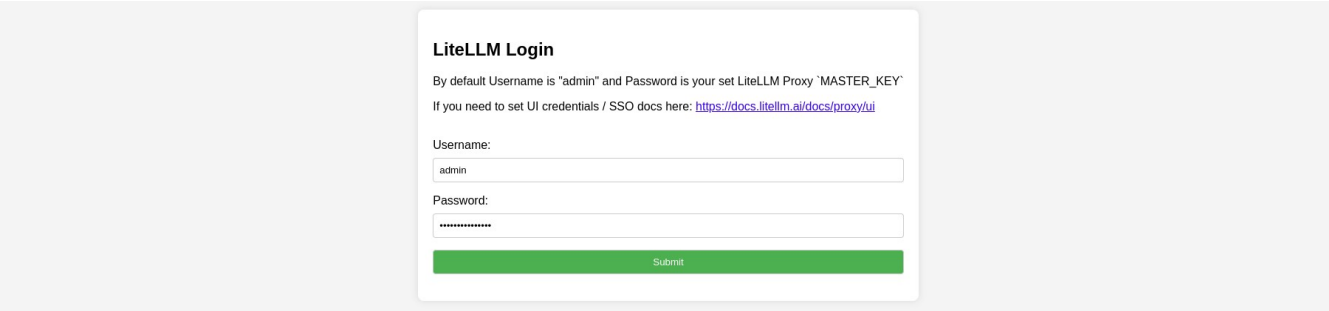
13. While that is brewing, go to your OpenAI tab/account and click on API Keys tab on left side. Create a new key and name it something like ‘LiteLLM API Key’, then copy the full API key string into your password manager.

14. In your Hostinger VPS dashboard, grab your VPS IP address and type it into a fresh tab on your search bar, with port 4000 like so:

<xx.xx.xx.xx:4000> Use your own given IP address here followed by ‘:4000’

Click on ‘LiteLLM Admin Panel on /ui’

Enter ‘admin’ as username and enter the password you created in step 12 (the random string you made in this line: LITELLM_MASTER_KEY=”sk-xxxxxxxxxxxx” where sk-xxxxxxxxxx is your password to use on the LiteLLM admin panel.



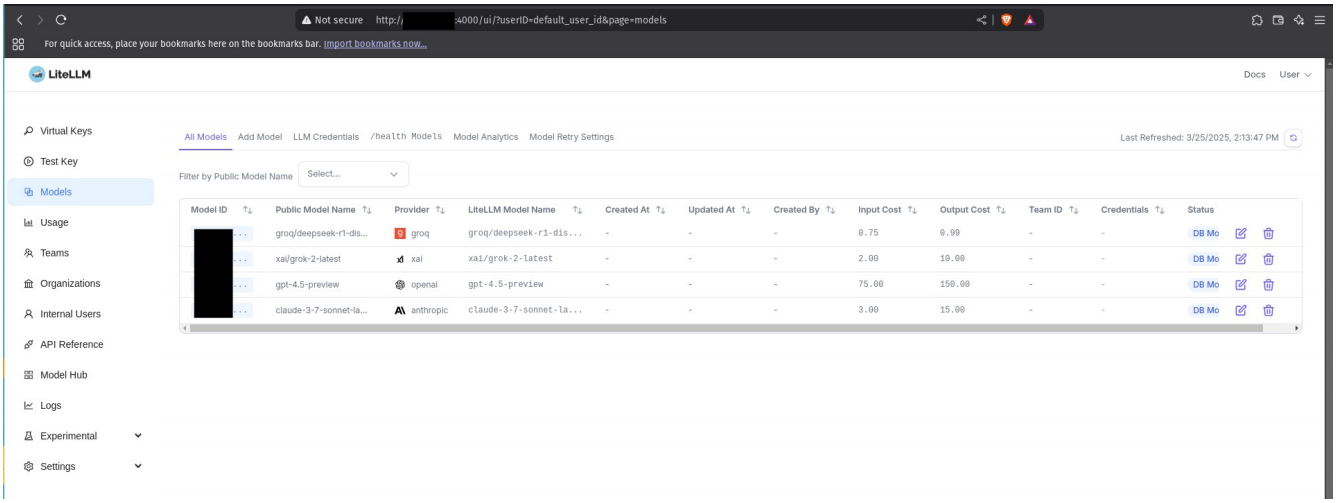
LiteLLM login screen using your VPS IP address:4000 (port 4000)

15. In the LiteLLM dashboard, click on Models tab on left, then Add Model and you can begin adding any additional AI models you want here. Select the Provider (such as Anthropic, Groq, Google, etc) and then select the model you wish.

(If you wish to add additional models in step 15, you must first register an account at that respective provider such as x.ai, anthropic.com, etc. Once you have an account, top up the account with at least $5 or so and then create an API key. You will use this key in step 15 when adding AI models within LiteLLM dashboard.

Your Open WebUI will have an OpenAI API key, and then a LiteLLM API key, which invokes all of your other models that you add within LiteLLM. So your respective API keys for whichever models you wish to add stay within LiteLLM settings, don’t try and plug those API keys directly into Open WebUI, just the one LiteLLM key. This was confusing for me at first, and is important to understand how LiteLLM plugin works alongside OpenWebUI using your various API keys.

You can quickly see that a password manager documenting all of the logins and API keys is a critical task in this VPS deployment!



LiteLLM Dashboard pictured above

16. Optional: In LiteLLM dashboard, select Virtual Keys in top left, we are going to add our own key for any additional users you wish to allow access to this VPS on your network. Selecting the user and then Settings, then Edit will allow you to put restrictions such as dollar limits, usage limits, and even custom restrictions on what the user is or is not allowed to do.

You can further fine tune which users get access to what within your Open WebUI interface under Admin panel. This whole setup allows for complete customization of your deployed instance of what I call an 'AI Rolodex.' You can even turn on or off logging to be able to have oversight of who is searching for what (great for your kids.)

There's plenty more tuning and additional AI models that you can play with here, with any amount of users you wish. You can access this locally by default, or remotely by logging into your VPS. You further have the option to point this VPS to your own custom domain so that you can access this easier by not having to remember an IP address, but your own domain name instead. Have fun with this. Keep an eye on the dollar amount you use until you get an idea of how much this costs, of course this will vary widely depending on your usage, and which/how many models you run.

Stay safe out there, and, stay free!

GrapheneGoat.com

Documentation

Ollama VPS: https://support.hostinger.com/en/articles/9310983-how-to-use-the-ollama-vps-template

Open WebUI: https://docs.openwebui.com/

LiteLLM: https://docs.litellm.ai/