AMI Name: Gigabits-Local AI-Ubuntu 24.04

1. Lunch Instance with above AMI

   Default Username: ubuntu
   Storage: Higer Recommended (Based on model, that you want to run)
   Recommended RAM: Higer Recommended
   Recommended CPU: Higher as much as possible (GPU Recommended)

2. Verification:

i. systemctl status local-ai

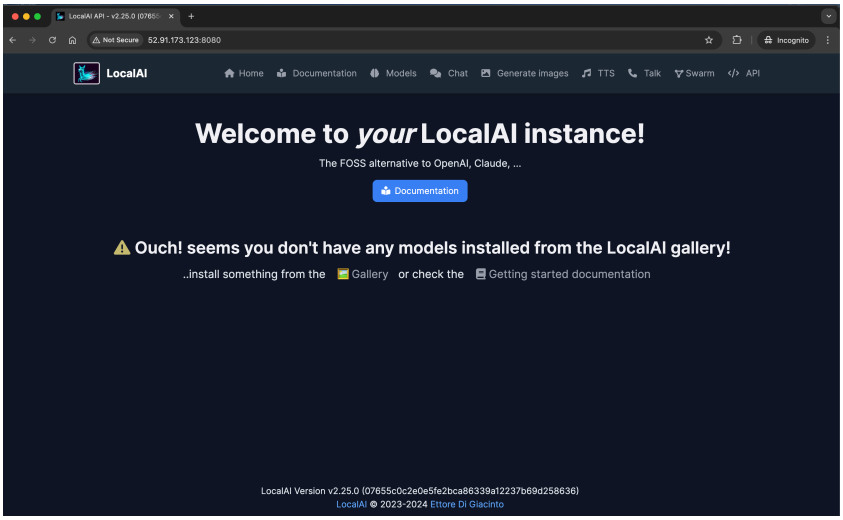
```
root@ip-172-31-22-142:~# systemctl status local-ai
● local-ai.service - LocalAI Service
     Loaded: loaded (/etc/systemd/system/local-ai.service; enabled; preset: enabled)
     Active: active (running) since Wed 2025-01-29 16:22:13 UTC; 4s ago
   Main PID: 14712 (local-ai)
      Tasks: 10 (limit: 19170)
     Memory: 1.2G (peak: 1.2G)
        CPU: 3.135s
     CGroup: /system.slice/local-ai.service
             └─14712 /usr/local/bin/local-ai run

Jan 29 16:22:13 ip-172-31-22-142 systemd[1]: Started local-ai.service - LocalAI Service.
Jan 29 16:22:13 ip-172-31-22-142 local-ai[14712]: 4:22PM INF env file found, loading environment variables from file envFile=/etc/localai.env
Jan 29 16:22:13 ip-172-31-22-142 local-ai[14712]: 4:22PM INF Setting logging to info
Jan 29 16:22:13 ip-172-31-22-142 local-ai[14712]: 4:22PM INF Starting LocalAI using 4 threads, with models path: /usr/share/local-ai/models
Jan 29 16:22:13 ip-172-31-22-142 local-ai[14712]: 4:22PM INF LocalAI version: v2.25.0 (07655c0c2e0e5fe2bca86339a12237b69d258636)
Jan 29 16:22:13 ip-172-31-22-142 local-ai[14712]: 4:22PM INF Preloading models from /usr/share/local-ai/models
root@ip-172-31-22-142:~#
```

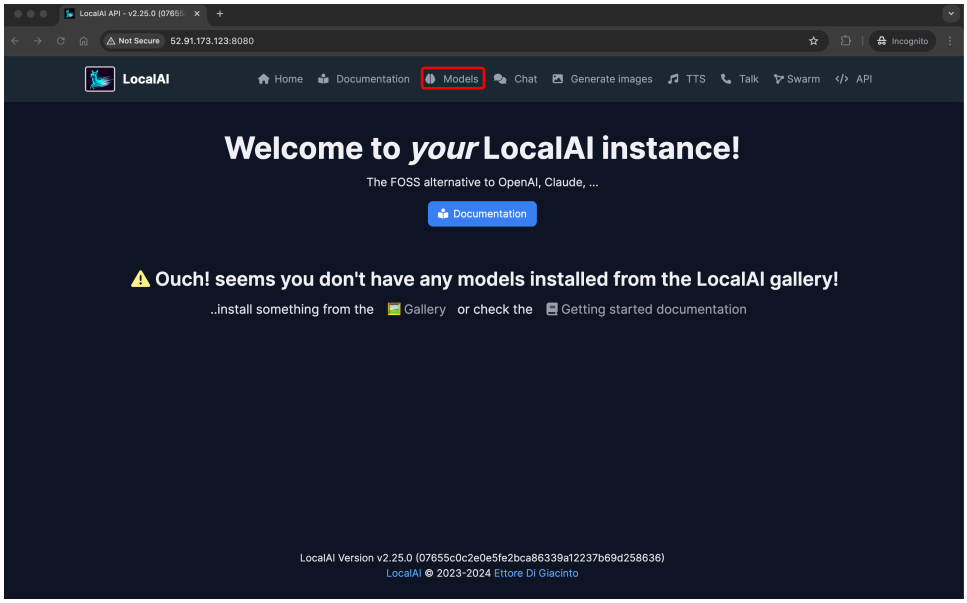
ii. ss -tulnp *(make sure that the port: 80 is listen)*

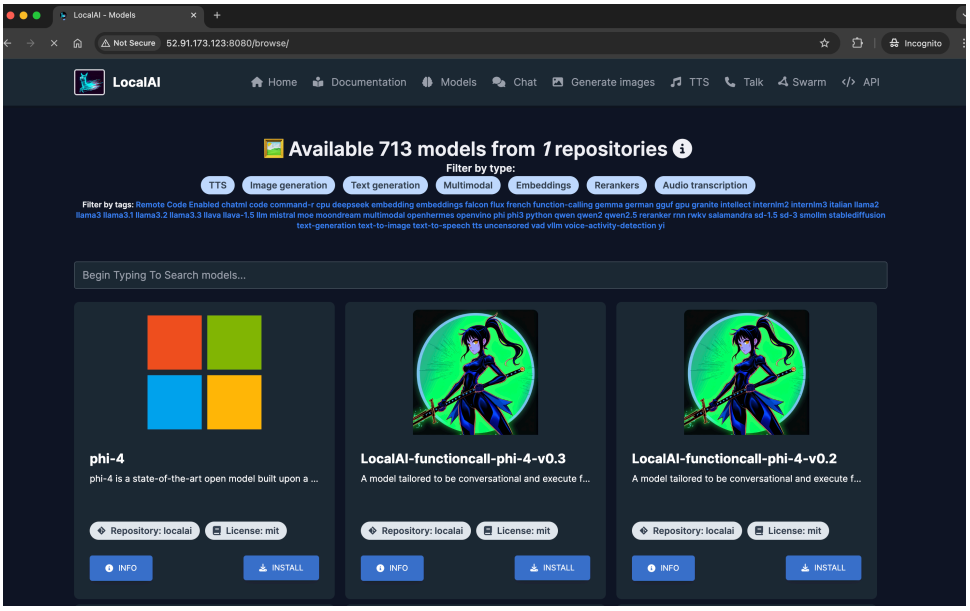
```
root@ip-172-31-22-142:~# ss -tulnp
Netid      State       Recv-Q      Send-Q            Local Address:Port            Peer Address:Port      Process
udp        UNCONN      0           0                    127.0.0.54:53                 0.0.0.0:*          users:(("systemd-resolve",pid=14228,fd=16))
udp        UNCONN      0           0                 127.0.0.53%lo:53                 0.0.0.0:*          users:(("systemd-resolve",pid=14228,fd=14))
udp        UNCONN      0           0              172.31.22.142%enX0:68               0.0.0.0:*          users:(("systemd-network",pid=14243,fd=22))
udp        UNCONN      0           0                    127.0.0.1:323                0.0.0.0:*          users:(("chronyd",pid=5232,fd=5))
udp        UNCONN      0           0                        [::1]:323                   [::]:*          users:(("chronyd",pid=5232,fd=6))
tcp        LISTEN      0           4096              127.0.0.53%lo:53                 0.0.0.0:*          users:(("systemd-resolve",pid=14228,fd=15))
tcp        LISTEN      0           4096                   0.0.0.0:8080               0.0.0.0:*          users:(("local-ai",pid=14712,fd=7))
tcp        LISTEN      0           4096                 127.0.0.54:53                 0.0.0.0:*          users:(("systemd-resolve",pid=14228,fd=17))
tcp        LISTEN      0           4096                       *:22                       *:*          users:(("sshd",pid=14223,fd=3),("systemd",pid=1,f
d=221))
root@ip-172-31-22-142:~#
```

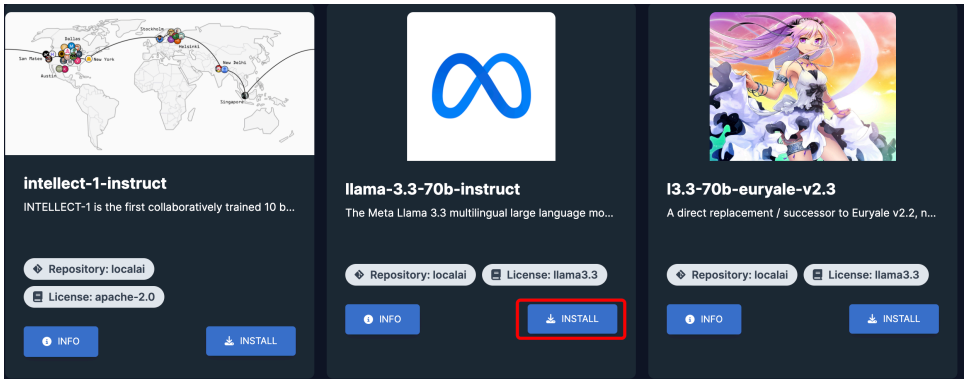
iii. On Browser, http://your_public_ip
E.g. http://52.91.173.123

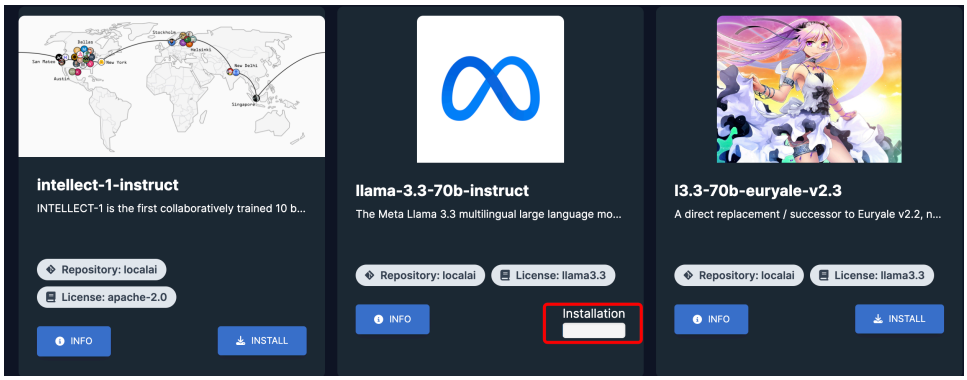


Install the model that you want to run from the Models section.









iv. **local-ai --help** *(to run from terminal)*

```
root@ip-172-31-22-142:~# local-ai --help
4:45PM INF env file found, loading environment variables from file envFile=/etc/localai.env
Usage: local-ai <command> [flags]

LocalAI is a drop-in replacement OpenAI API for running LLM, GPT and genAI models locally on CPU, GPUs with consumer grade hardware.

Some of the models compatible are:
  - Vicuna
  - Koala
  - GPT4ALL
  - GPT4ALL-J
  - Cerebras
  - Alpaca
  - StableLM (ggml quantized)

For a list of compatible models, check out: https://localai.io/model-compatibility/index.html

Copyright: Ettore Di Giacinto

Version: v2.25.0 (07655c0c2e0e5fe2bca86339a12237b69d258636)

Flags:
  -h, --help                  Show context-sensitive help.
      --log-level=LOG-LEVEL   Set the level of logs to output [error,warn,info,debug,trace] ($LOCALAI_LOG_LEVEL)

Commands:
  run [<models> ...] [flags]
    Run LocalAI, this the default command if no other command is specified. Run 'local-ai run --help' for more information

  federated [flags]
    Run LocalAI in federated mode

  models list [flags]
    List the models available in your galleries

  models install [<models> ...] [flags]
    Install a model from the gallery

  tts --model=STRING <text> ... [flags]
    Convert text to speech

  sound-generation --backend=STRING --model=STRING <text> ... [flags]
    Generates audio files from text or audio
```

Thank You