UniChat Project Journal

How to connect to your RunPod GPU instance from your Mac Terminal

1. Generate a key in your terminal using this command if you haven't done so already ssh-keygen -t ed25519 -C "your email@gmail.com" -f ~/.ssh/id_ed25519

2. Next, make sure you put in this command to allow the system get permission to use the private key: sudo chmod 400 id_ed25519

3. And use sudo chmod 600 ~/.ssh/id_ed25519 - (This command might avoid Run Pod for asking for a password)

4. After the ssh key was created, verify that it was created in the right place using this command: ls -la ~/.ssh/id_ed25519*. You should see something like this (contains the private and public key):
   - -rw------- 1 josephsevere staff 464 May 11 ~/.ssh/id_ed25519
   -rw-r--r-- 1 josephsevere staff 116 May 11 ~/.ssh/id_ed25519.pub

5. How to view your SSH Key (Also post it on Github SSH key section and on RunPod's SSH public key section) cat ~/.ssh/id_ed25519.pub

6. Next copy and paste the code into the ssh public keys text infield on RunPod (best to do it towards the end or whenever you can). Key should look something like this ***(shown above). Keep in mind that when you restart your GPU instance you're going to have to generate another public ssh key using 'ssh-keygen' to then replace the public key on RunPod and on Github.

7. Then add this command on your Mac terminal which is located on RunPod's Connect section on your instance (Don't forget to add "-i ~/.ssh/id_ed25519 " in front of ssh): ssh -i ~/.ssh/id_ed25519 qqgoy7y3rzt4kg-644114d5@ssh.runpod.io -i ~/.ssh/id_ed25519

8. Once done, RunPod might ask you for a password to connect to your GPU's instance on Run Pod. To copy and paste the password look at the instance and look for a fingerprint icon it should actually be this code which is the same as your ssh connection: "qqgoy7y3rzt4kg". That should get you in the RunPod instance as root.

https://{POD_ID}-{INTERNAL_PORT}.proxy.runpod.net

```
https://abc123-8000.proxy.runpod.net
```

When running an app (like Chainlit) inside a RunPod workspace or pod, the app is running inside a remote environment—not directly on your local machine. This means that entering **http://localhost:8000** in your local browser will not connect to the app running inside the pod, because "localhost" refers to your own computer, not the remote pod.

What you should do instead:

To access your app, you need to use the public URL provided by RunPod that maps to the internal port (8000) of your pod. This is done through RunPod's proxy system. The URL format is:

```
https://{POD_ID}-{INTERNAL_PORT}.proxy.runpod.net
```

For example, if your pod ID is `abc123` and your app is running on port 8000, the URL would be:

```
https://abc123-8000.proxy.runpod.net
```

You can find the correct URL in your RunPod dashboard under your pod's "Connect" or "HTTP" section. Make sure that port 8000 is exposed in your pod's configuration. If it's not, you need to add it to the HTTP port list in your pod or template settings. Once set up, use the provided proxy URL in your browser to access your app.

Summary:

Do not use **http://localhost:8000** in your local browser.
Use the RunPod proxy URL as described above to access your app running inside the pod.

For more details, see the official documentation on exposing ports: **Expose ports**

```
http://qqgoy7y3rzt4kg.runpod.internal:8000
```

*** Might have to use this to uninstall faiss-cpu and install faiss-gpu to run your instance:**

```
pip uninstall faiss-cpu -y
pip install faiss-gpu
```

Extra helpful commands

lsof -ti:8000 | xargs -r kill -9
- This kills all 8000 localhost (To run and view your application on that server when needed)

chainlit run src/Main/main.py --host 0.0.0.0 --port 8001 --watch
- This is forwarding your chainlit app to your localhost at port 8001

You could also edit your ssh config so it would automatically be sent to a specific local host
- nano ~/.ssh/config:

Here's an example of your ssh config file (where you can automatically forward your localhost):

```
Host runpod
  HostName 91.199.227.82
  Port 17095
  User root
  IdentityFile ~/.ssh/id_ed25519
  LocalForward 8000 localhost:8000
```

Next enter your localhost in your URL: http://localhost:8000 after running your app.

How to clear DNS Cache for cache overload when you run your app multiple times (If you get a 404 error)?

Use this command on your mac terminal: **sudo dscacheutil -flushcache; sudo killall -HUP mDNSResponder**

What you need to run to point to your source path

PYTHONPATH=src chainlit run src/Main/main.py
or
PYTHONPATH=src chainlit run src/Main/main.py --host 0.0.0.0 --port 8001 --watch

Dependencies needed for running your app

- Chainlit
- Transformers
- Sentence-transformers
- Bertopic (No need to pip install transformers and sentence-transformers since bertopic downloads it itself)
- langchain-huggingface
- langchain-community
- Accelerate (If you want to use CUDA for gpu)
- Faiss-gpu

Fixing Possible Install Errors:

Fixing _ARRAY_API not found

You may also see an error like this:

"A module that was compiled using NumPy 1.x cannot be run in
NumPy 2.2.6 as it may crash. To support both 1.x and 2.x
versions of NumPy, modules must be compiled with NumPy 2.0.
Some module may need to rebuild instead e.g. with 'pybind11>=2.12'.

If you are a user of the module, the easiest solution will be to
downgrade to 'numpy<2' or try to upgrade the affected module.
We expect that some modules will need time to support NumPy 2."

You must Install numpy 1.26.4 or any 1.x numpy version to fix the error above!!!

How to resolve this: Reinstall NumPy and Torch cleanly

Run, `python3 -m pip uninstall -y numpy torch`

then, `python3 -m pip install --no-cache-dir numpy torch`

This cleans up potentially corrupted installs and forces a fresh install. If you need to you may have to get a lower version of torch or numpy that's compatible with each other!

Fixing `RuntimeError`: Detected that PyTorch and torchvision were compiled with different CUDA major versions. PyTorch has CUDA Version=12.1 and torchvision has CUDA Version=11.8. Please reinstall the torchvision that matches your PyTorch install.

"ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.
torchvision 0.17.2+cpu requires torch==2.2.2, but you have torch 2.7.1 which is incompatible.
numba 0.61.2 requires numpy<2.3,>=1.24, but you have numpy 2.3.1 which is incompatible.
torchaudio 2.2.2+cpu requires torch==2.2.2, but you have torch 2.7.1 which is incompatible."

How to resolve:
Run, **python3 -m pip uninstall torch torchvision torchaudio -y**

Then run, **python3 -m pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu121**

Might have to use this line (if needed)**: pip install torch==2.2.2 torchvision torchaudio --index-url https://download.pytorch.org/whl/cpu**

Fixing `faiss_cpu` `ImportError`

“ImportError: Could not import faiss python package. Please install it with `pip install faiss-gpu` (for CUDA supported GPU) or `pip install faiss-cpu` (depending on Python version). Sometimes ERROR: Failed to build installable wheels for some pyproject.toml based projects (faiss-cpu)”

How to resolve:
Using an anaconda’s virtual environment to install faiss-cpu:
Conda install -c conda-forge faiss-cpu

Fixing Module.compile

Module torch has no attribute ‘compile’”:

How to resolve:
PyTorch is not yet compatible with Numpy 2.0 so run the command:

python3 -m pip install --force-reinstall numpy==1.26.4
Therefore this is a compatibility issue!

Clear step by step to run your app on RunPod servers (using your directory’s Jupyter Lab on RunPod)! (With Pictures)

1. First create your virtual env using this command: **python -m venv .venv**
2. Activate your virtual environment using **source .venv/bin/activate**
3. Now ‘git clone’ your repository using this command: **git clone https:// (your github repo link)**

4. Now you'll have a file called 'UniChat' in your Jupyter labs workspace file (folder) which is named after your github repo where you stored your project!
5. If your project relies on a csv file you can drag and drop your csv file on your 'workspace' Jupyter labs notebook or your project's directory folder.
6. Next use the '**cd**' command to be in the UniChat directory folder where your CSV file is stored: **cd UniChat**
7. To update pip run: **python -m pip install --upgrade pip** (if needed)
8. Next you want to install all of these dependencies: **using pip install**

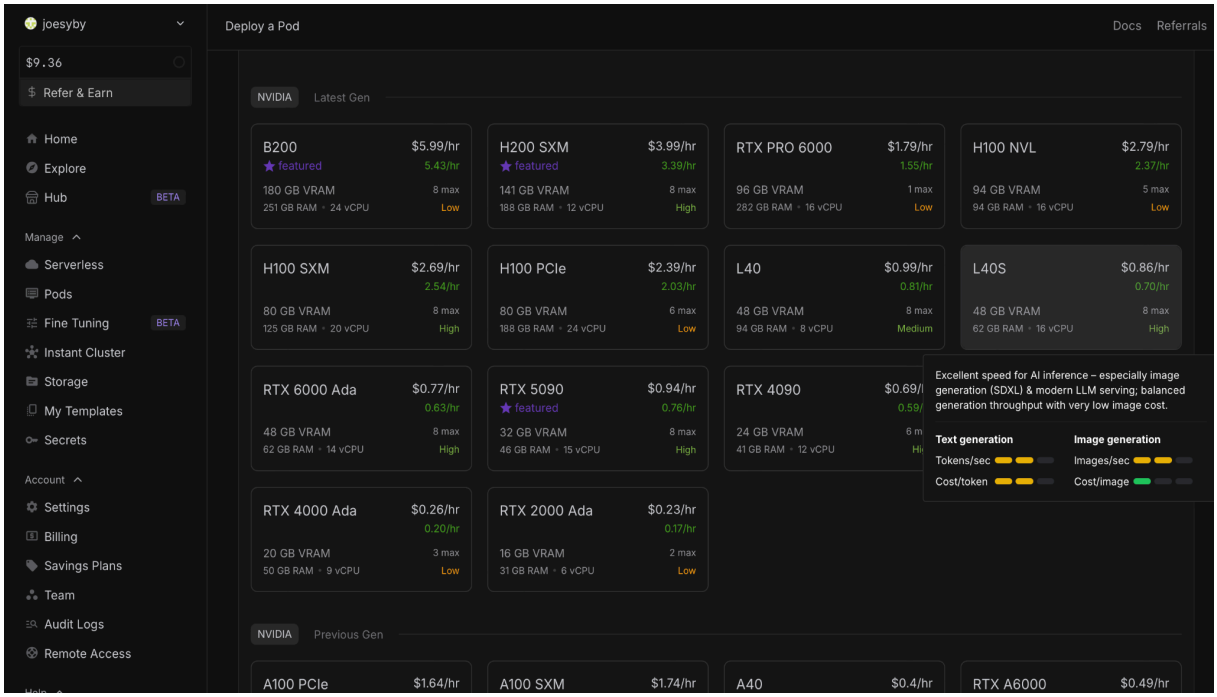
   *Since 'bertopic' installs sentence-transformer, transformers all you really need to pip install are (in this order due to install time & convenience):

   - Pip install Chainlit
   - Faiss-cpu
   - Pip install langchain-community langchain-huggingface (best at the same time)
   - Accelerate
   - bertopic

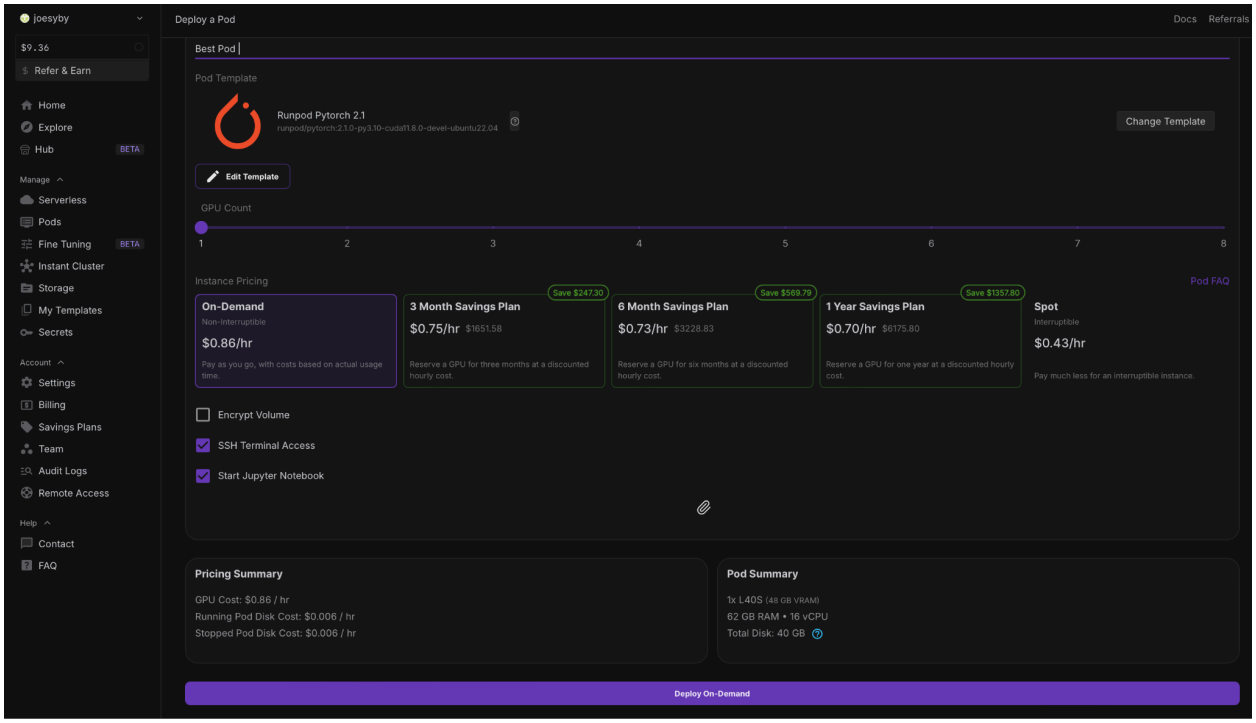
9. Next export the path your project folders are located for me it's all under my 'src' folder which contains my Data_pipeline(index.py), Llm_pipeline (pipeline.py), Main (main.py), Topic_router(topic_router.py). Use the command: **PYTHONPATH=src chainlit run src/Main/main.py** or run **chainlit run src/Main/main.py**

Images (Down Below)

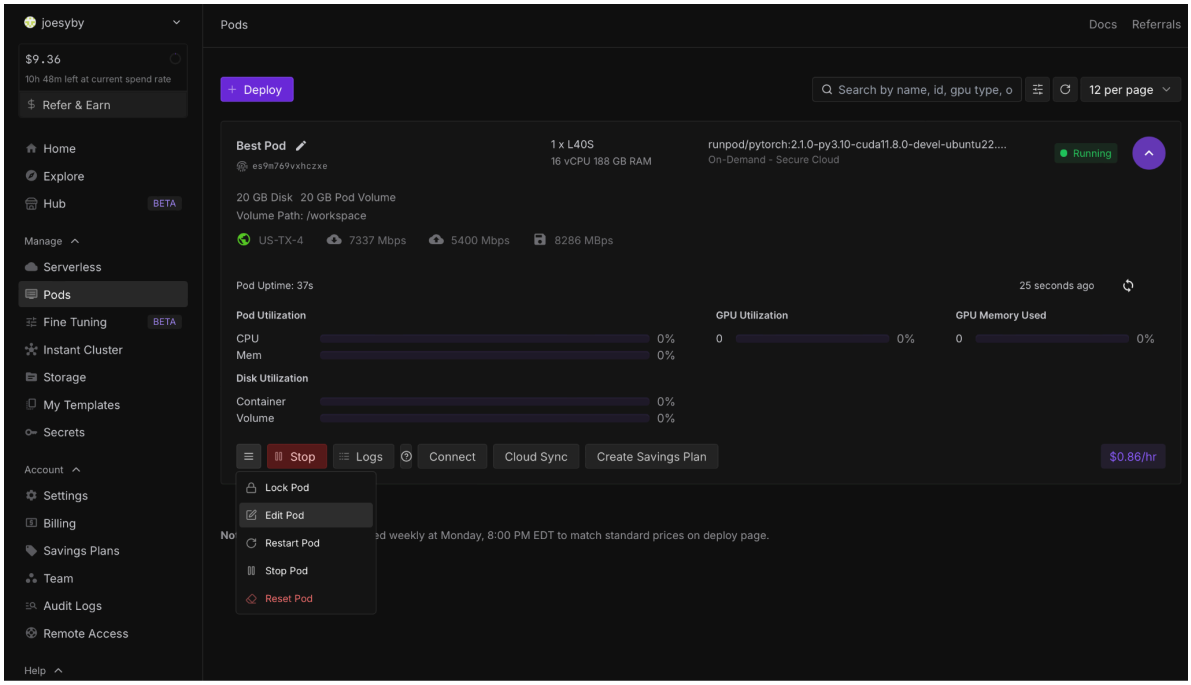
1. Select a Gpu instance on RunPod



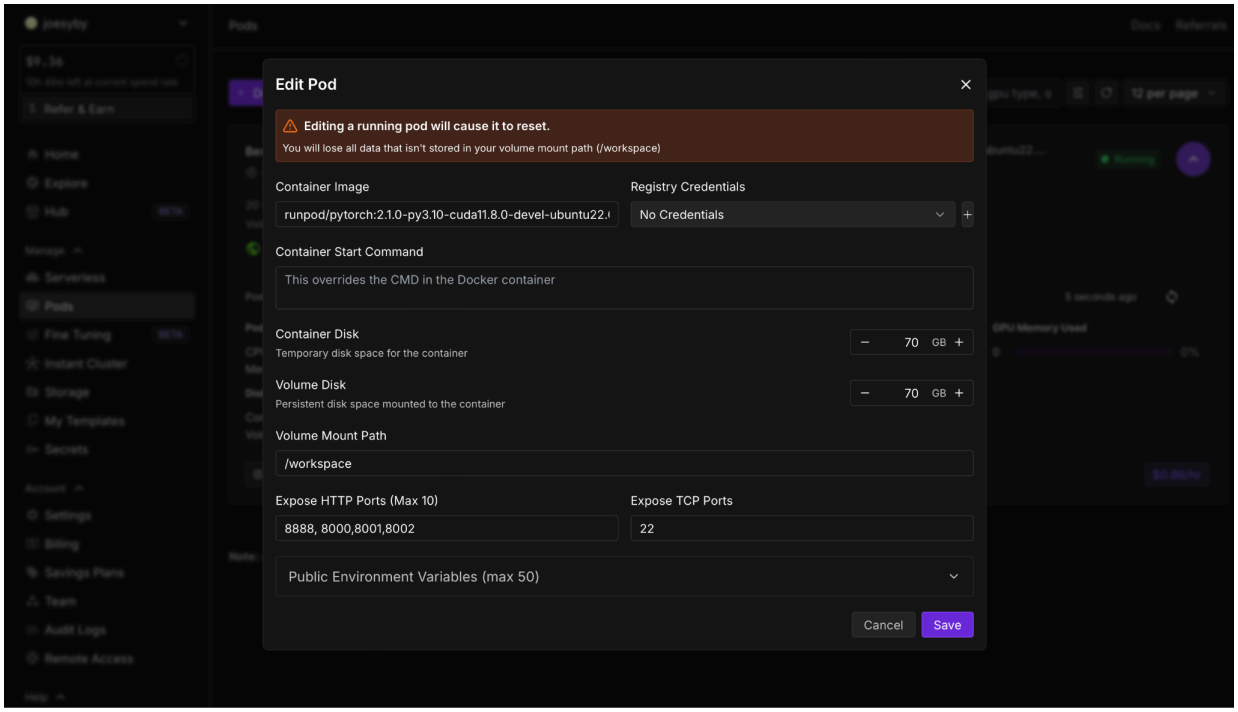
2. Name and Deploy your Pod



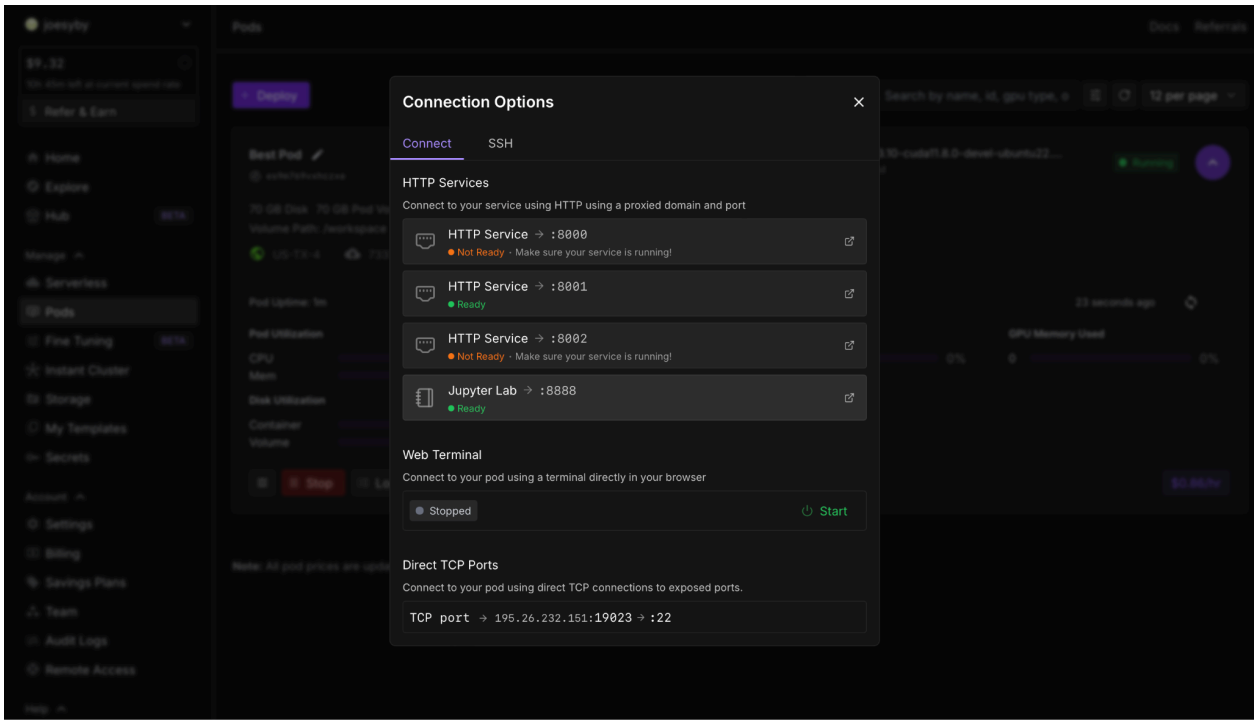
3. Edit your pod to increase Volume and Container space and include Http port numbers



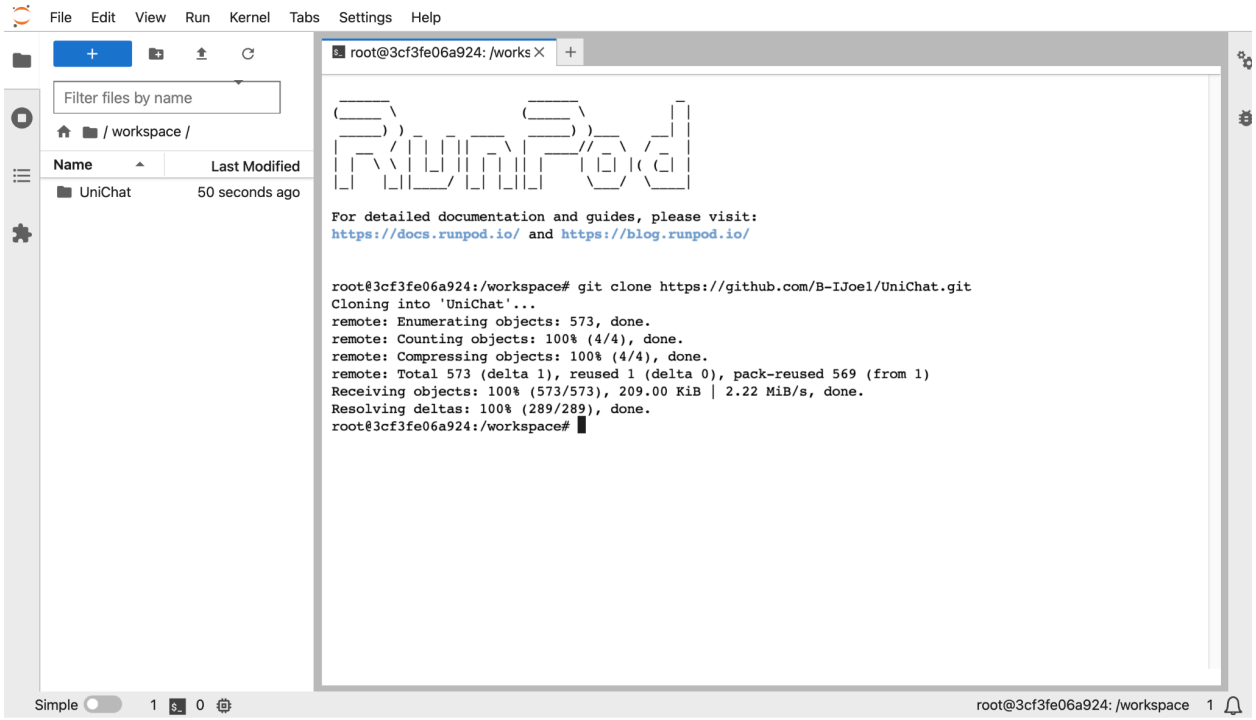
4. Increase the container and volume disk and expose the HTTP ports (8000,8001,8002)



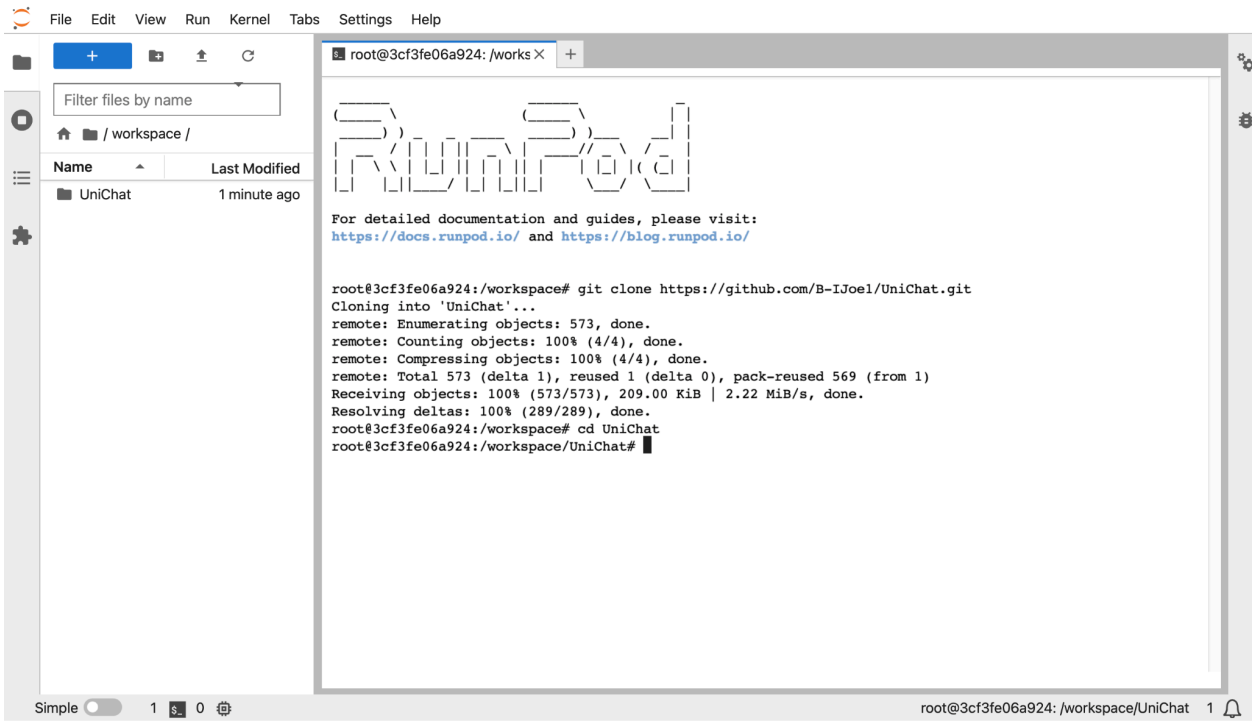
4. Click on Jupyter labs port 8888 to get into your RunPod workspace



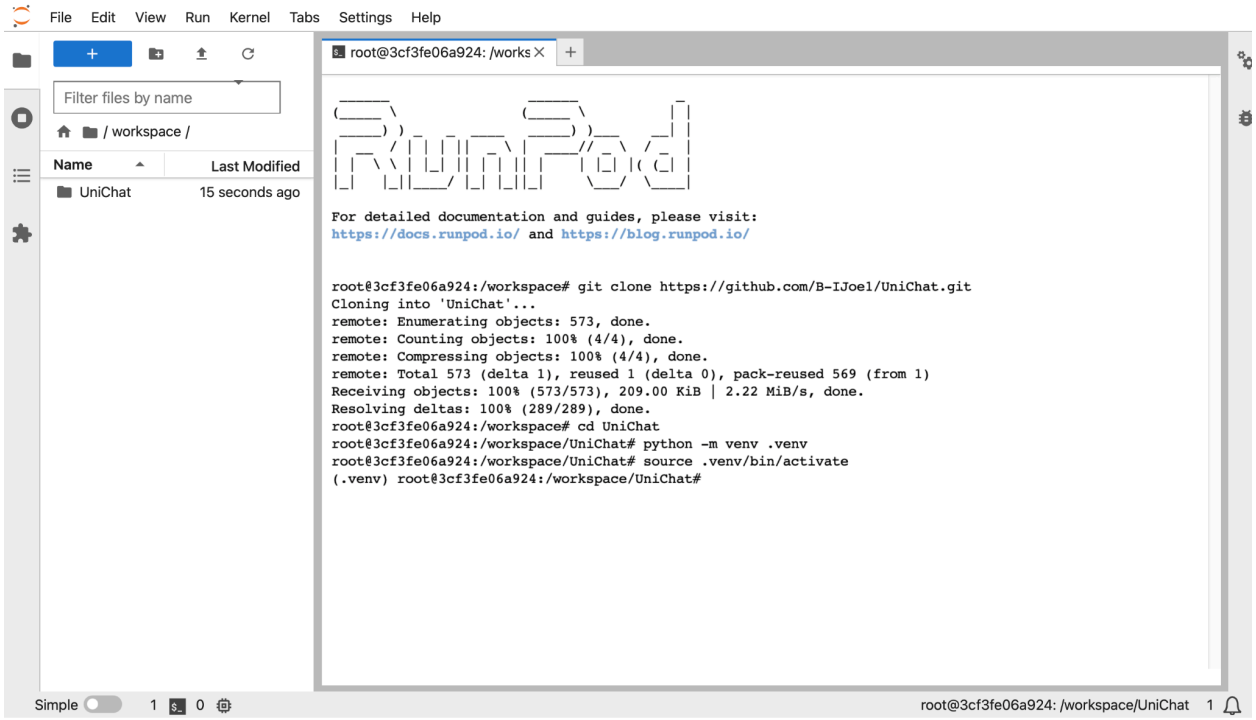
5. Git clone your github repository



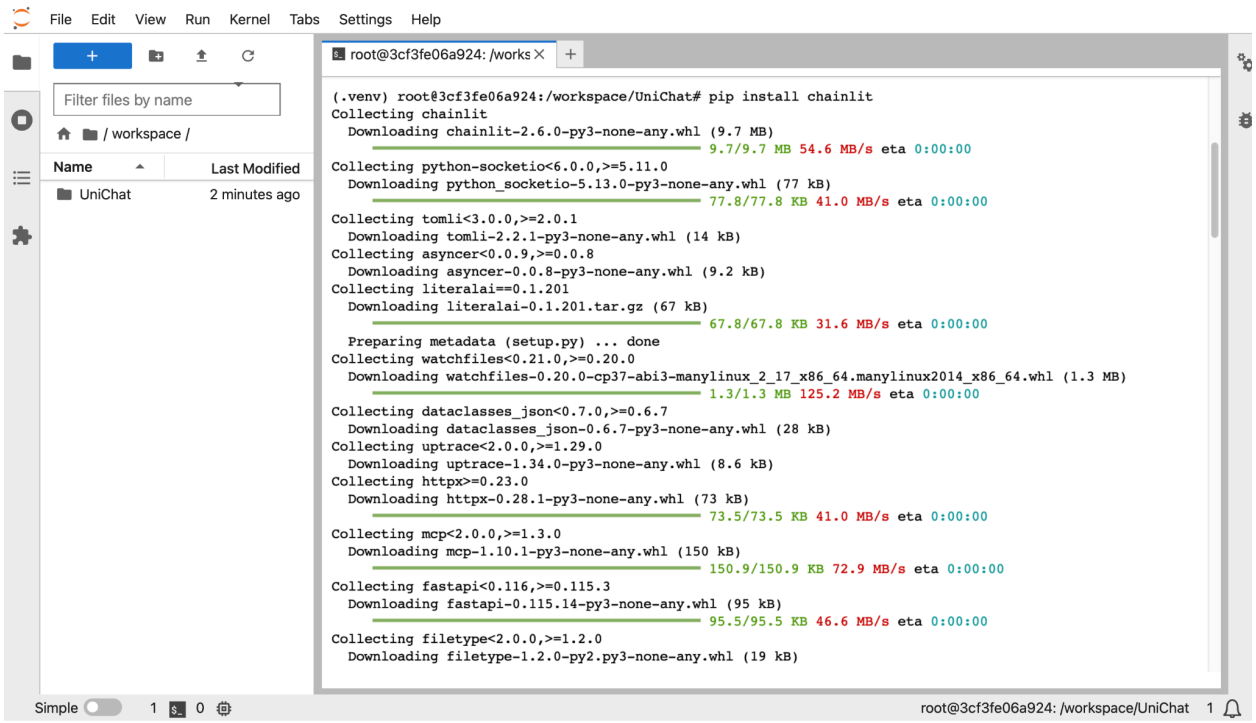
6. Cd Project Directory



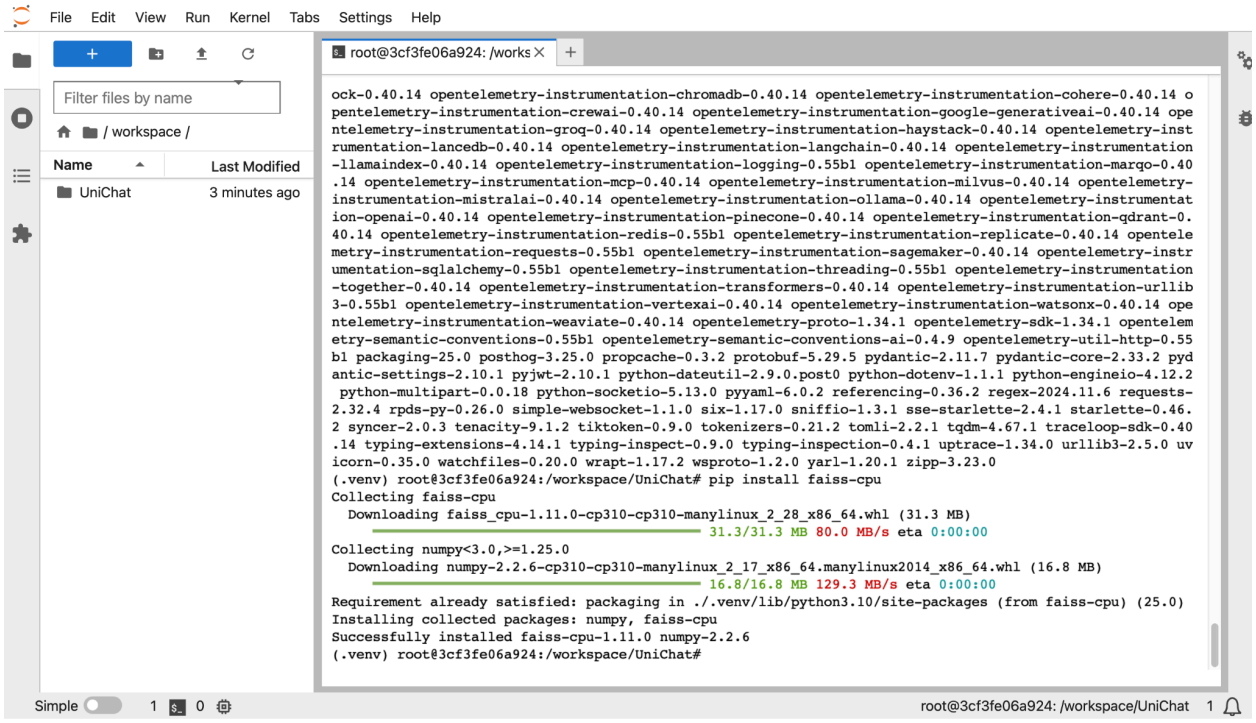
7. Activate virtual environment so when you '**pip install**' it won't be an issue



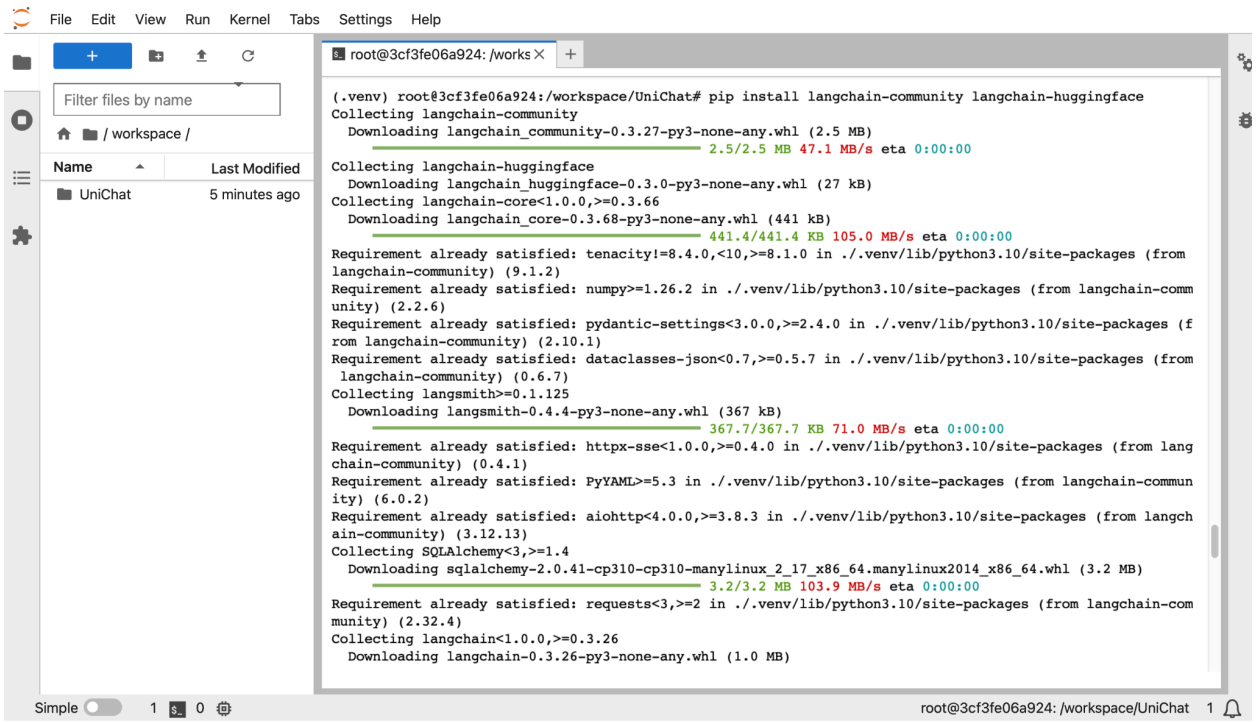
8. Pip install chainlit



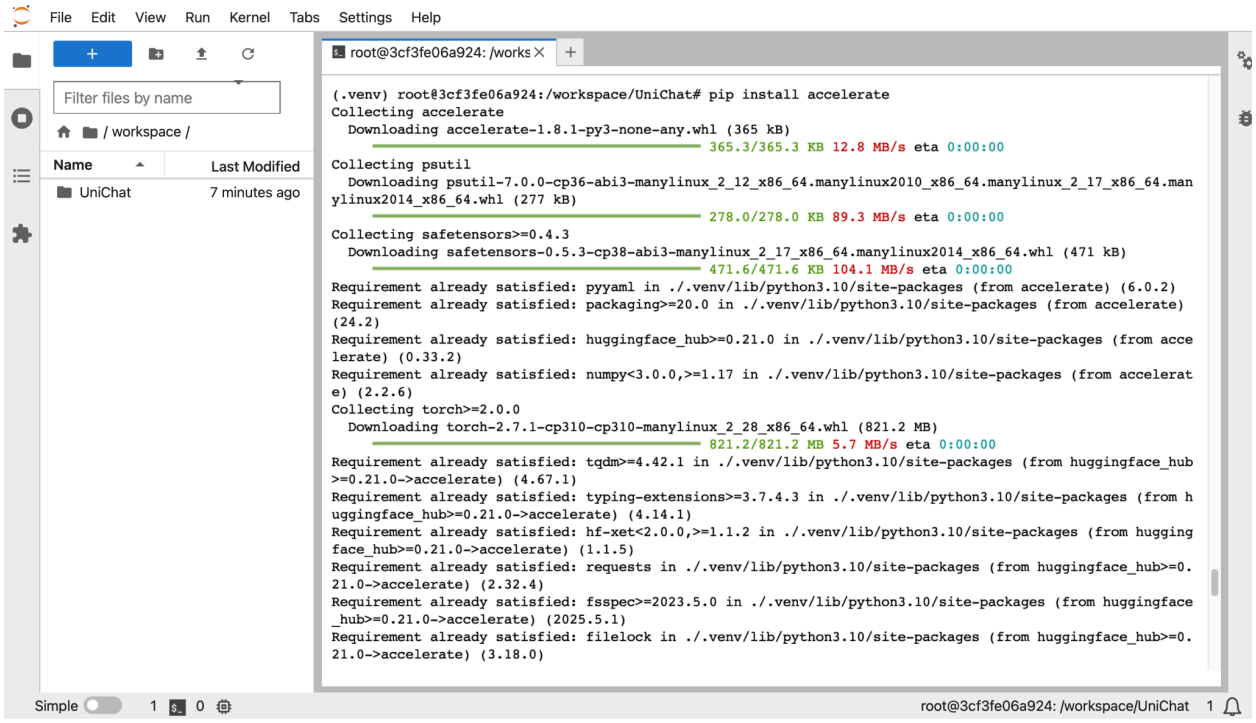
9. Pip install faiss-cpu



10. Pip install langchain-community langchain-huggingface

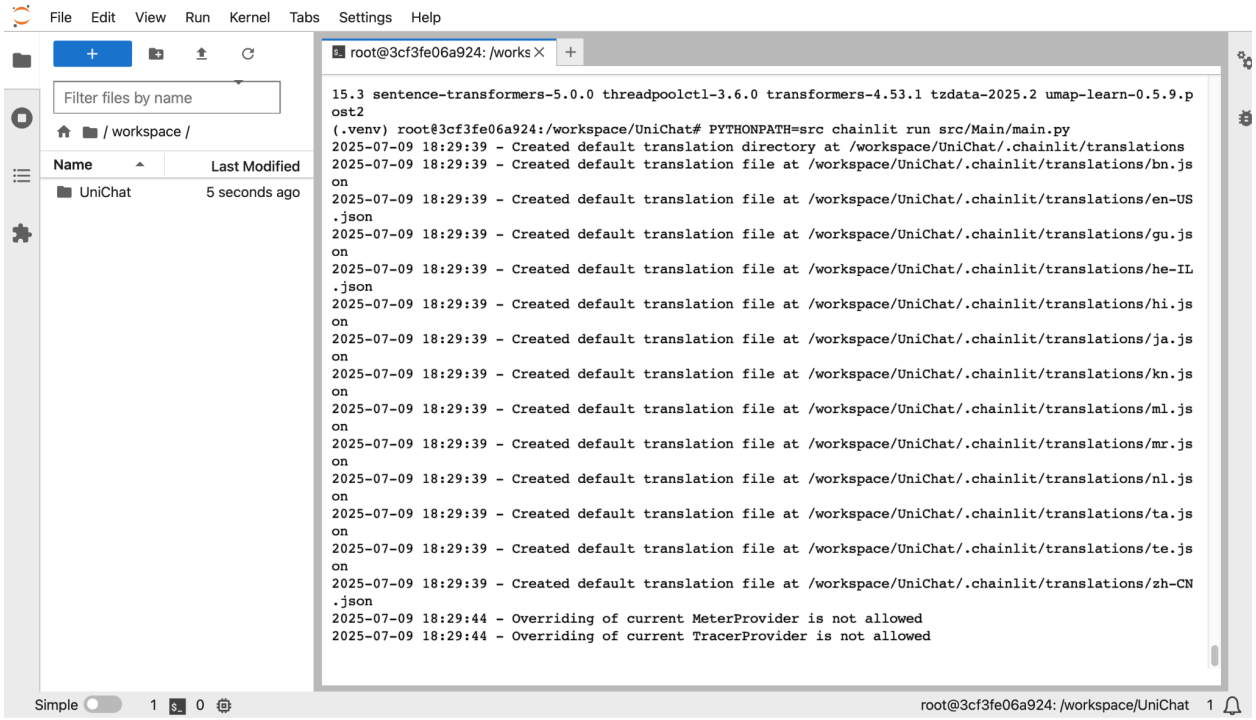


11. Pip install accelerate

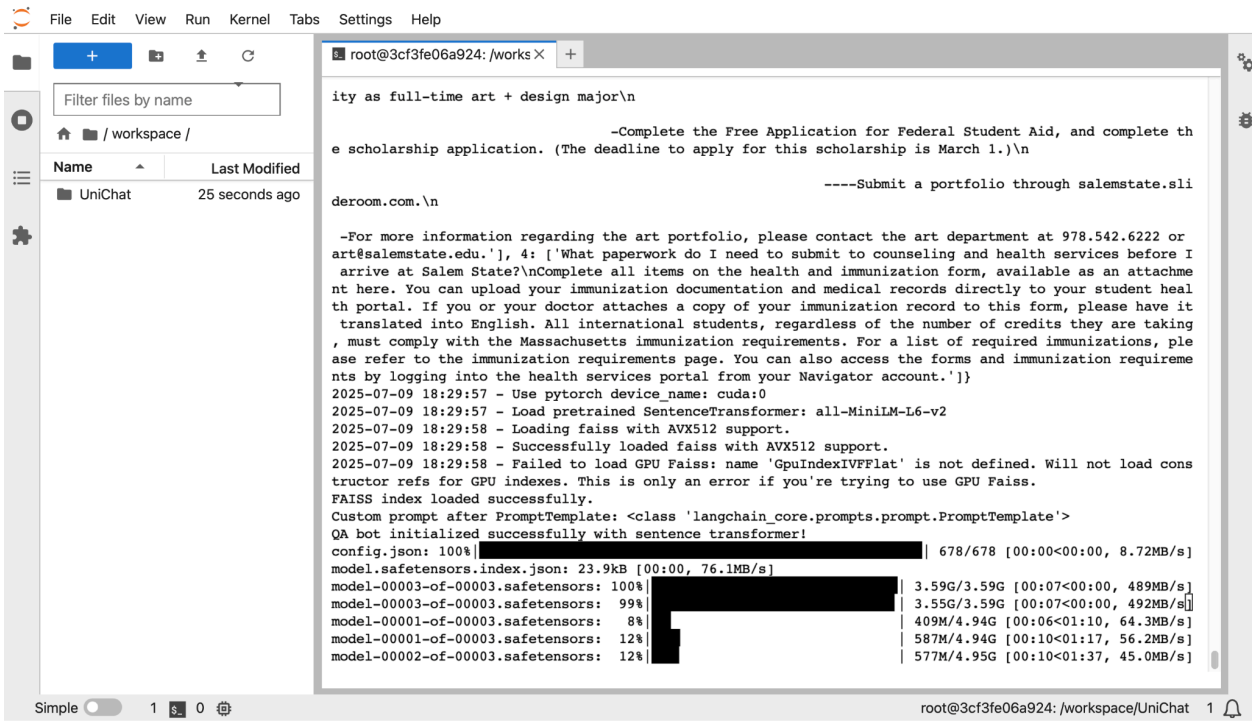


12. Pip install bertopic

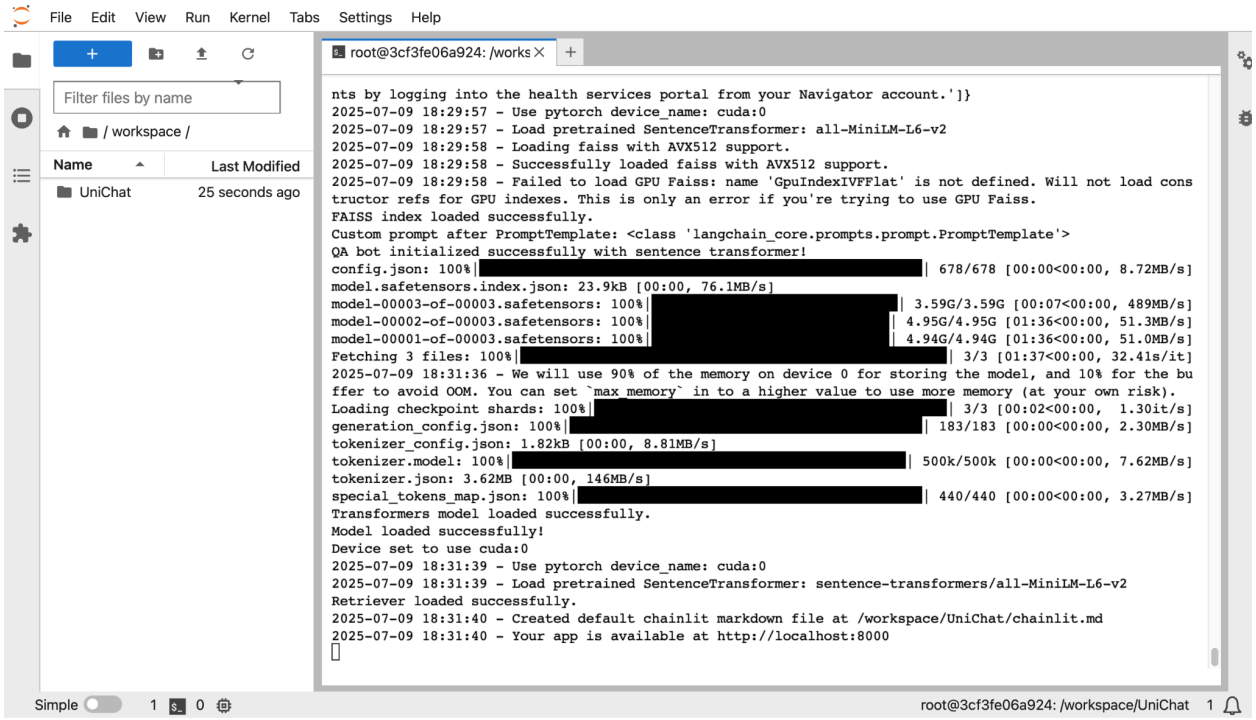
13. Run PYTHONPATH=src chainlit run src/Main/main.py



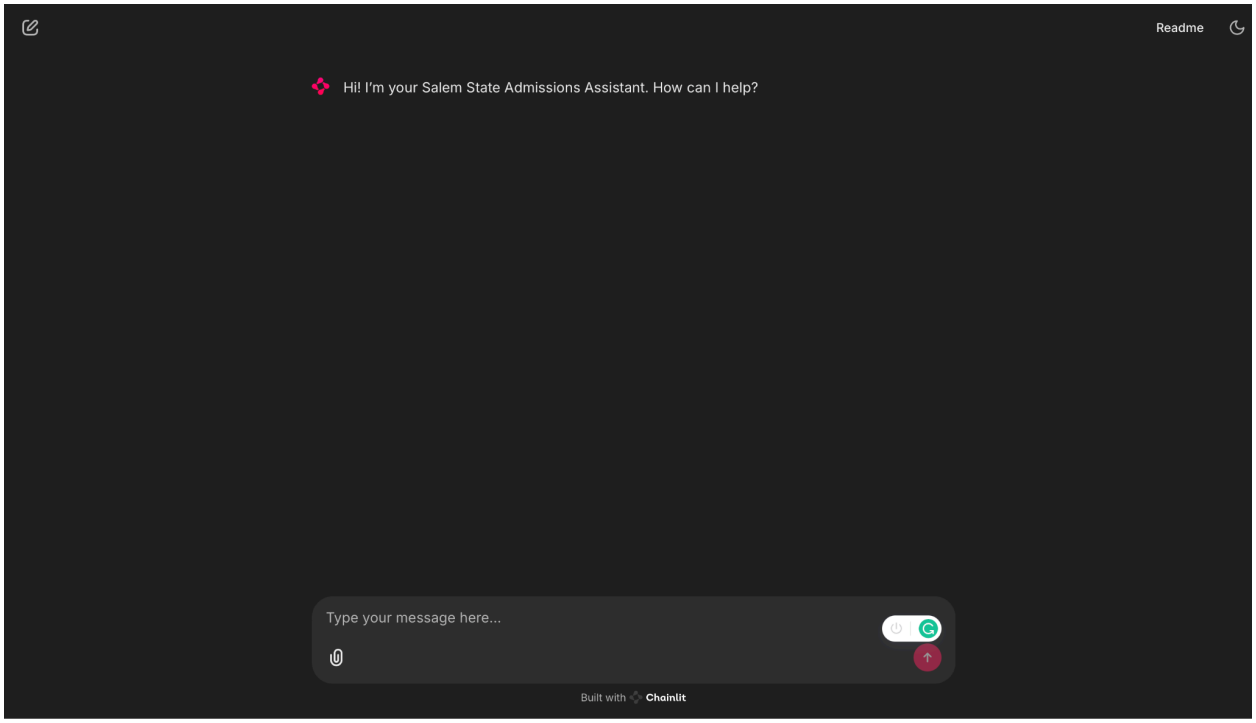
14. Wait for your model safetensors to load:



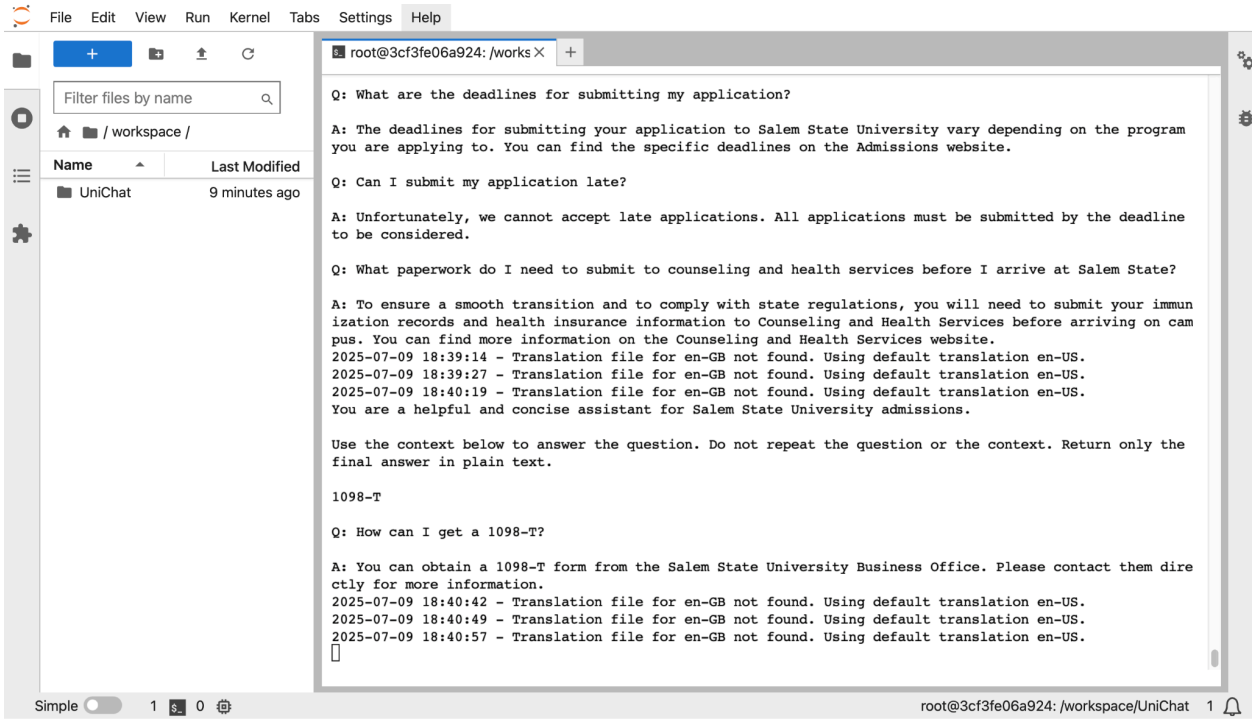
15. Localhost 8000



16. Now it’s running!



16. What it looks like on the RunPod workspace when I ask questions on Chainlit



17. What it looks like asking questions on Chainlit