

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/wheel/cu121`

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

The screenshot shows the RunPod web interface. On the left is a sidebar with navigation links like Home, Explore, Hub, Manage, Serverless, Pods, Fine Tuning, Instant Cluster, Storage, My Templates, Secrets, Account, Settings, Billing, Savings Plans, Team, Audit Logs, and Remote Access. A balance of \$9.36 is displayed, along with a 'Refer & Earn' button. The main area is titled 'Deploy a Pod' and shows a grid of GPU instances under the 'NVIDIA Latest Gen' tab. The instances listed are:

GPU Model	Cost	VRAM	CPU	Performance
B200	\$5.99/hr	180 GB VRAM 251 GB RAM	8 max 24 vCPU	Low
H200 SXM	\$3.99/hr	141 GB VRAM 188 GB RAM	8 max 12 vCPU	High
RTX PRO 6000	\$1.79/hr	96 GB VRAM 282 GB RAM	1 max 16 vCPU	Low
H100 NVL	\$2.79/hr	94 GB VRAM 94 GB RAM	5 max 16 vCPU	Low
H100 SXM	\$2.69/hr	80 GB VRAM 125 GB RAM	8 max 20 vCPU	High
H100 PCIe	\$2.39/hr	80 GB VRAM 188 GB RAM	6 max 24 vCPU	Low
L40	\$0.99/hr	48 GB VRAM 94 GB RAM	8 max 8 vCPU	Medium
L40S	\$0.86/hr	48 GB VRAM 62 GB RAM	8 max 16 vCPU	High
RTX 6000 Ada	\$0.77/hr	48 GB VRAM 62 GB RAM	8 max 14 vCPU	High
RTX 5090	\$0.94/hr	32 GB VRAM 46 GB RAM	8 max 15 vCPU	High
RTX 4090	\$0.69/hr	24 GB VRAM 41 GB RAM	6 max 12 vCPU	High
RTX 4000 Ada	\$0.26/hr	20 GB VRAM 50 GB RAM	3 max 9 vCPU	Low
RTX 2000 Ada	\$0.23/hr	16 GB VRAM 31 GB RAM	2 max 6 vCPU	Low

A note below the grid states: "Excellent speed for AI inference – especially image generation (SDXL) & modern LLM serving; balanced generation throughput with very low image cost." Below the grid are sections for Text generation and Image generation, each with a legend for Tokens/sec, Images/sec, Cost/token, and Cost/image.

At the bottom, there are tabs for 'NVIDIA Previous Gen' and a row of GPU models with their costs: A100 PCIe (\$1.64/hr), A100 SXM (\$1.74/hr), A40 (\$0.4/hr), RTX A6000 (\$0.49/hr).

2. Name and Deploy your Pod

The screenshot shows the 'Deploy a Pod' page for a 'Best Pod' template named 'Runpod Pytorch 2.1'. The sidebar on the left is identical to the first screenshot. The main area has a 'Pod Template' section with a preview of the Runpod logo and the template name. It includes a 'GPU Count' slider set to 1, and sections for 'Instance Pricing' and 'Pod FAQ'.

Instance Pricing:

- On-Demand:** \$0.86/hr
- 3 Month Savings Plan:** \$0.75/hr (\$1651.58) - Save \$247.30
- 6 Month Savings Plan:** \$0.73/hr (\$3228.83) - Save \$569.79
- 1 Year Savings Plan:** \$0.70/hr (\$6175.80) - Save \$1357.80

Pod FAQ: Pay much less for an interruptible instance.

Deployment Options:

- Encrypt Volume
- SSH Terminal Access
- Start Jupyter Notebook

Pricing Summary:

GPU Cost: \$0.86 / hr
Running Pod Disk Cost: \$0.006 / hr
Stopped Pod Disk Cost: \$0.006 / hr

Pod Summary:

1x L40S (48 GB VRAM)
62 GB RAM + 16 vCPU
Total Disk: 40 GB

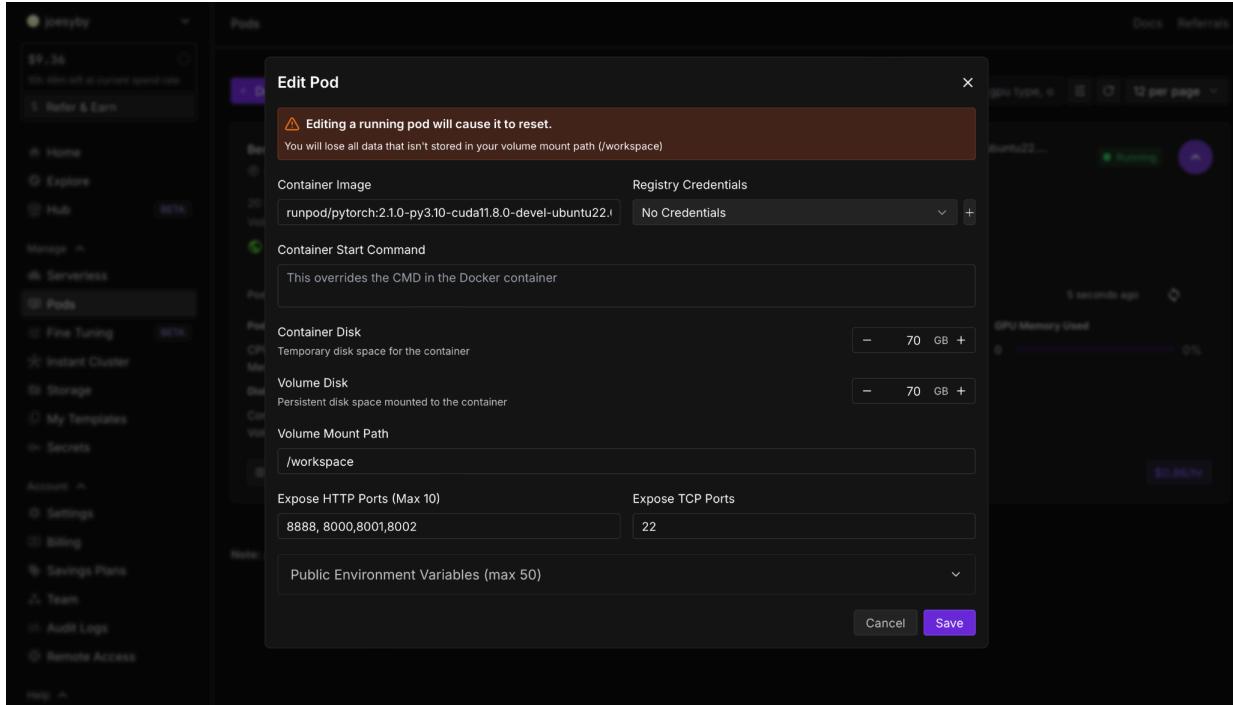
Buttons:

Deploy On-Demand

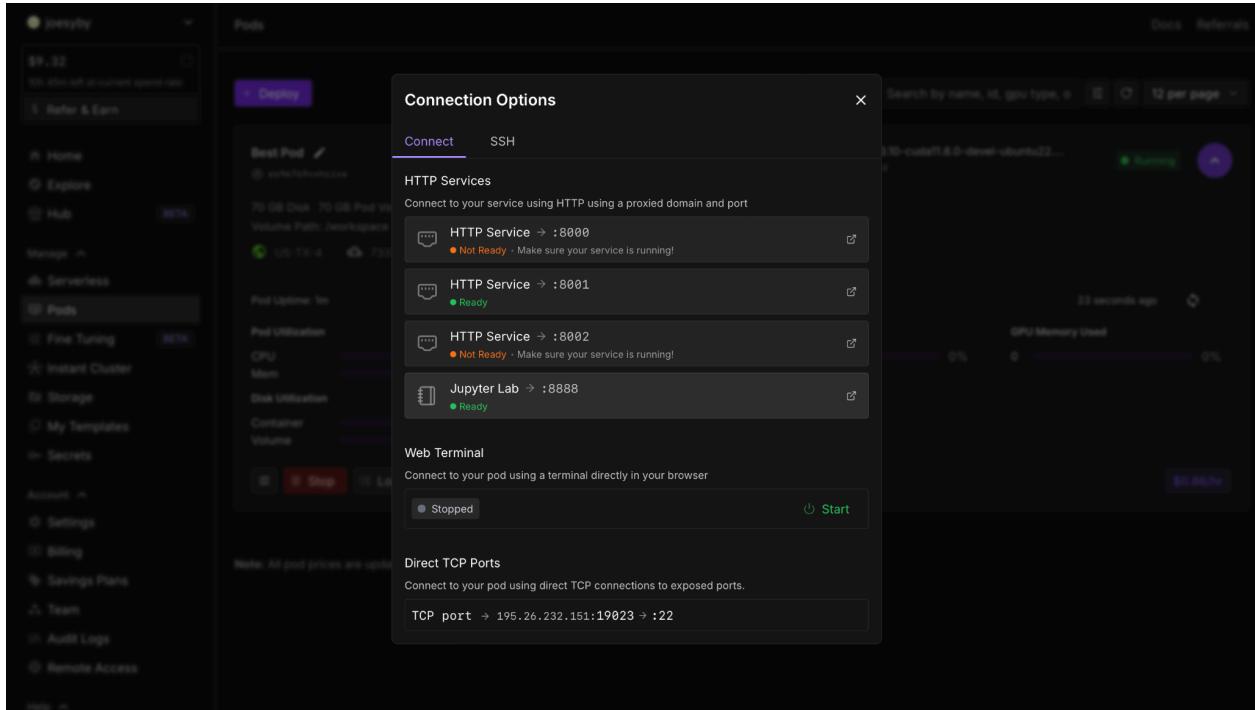
3. Edit your pod to increase Volume and Container space and include Http port numbers

The screenshot shows the RunPods web interface. On the left is a sidebar with user information (\$9.36, 10h 48m left at current spend rate), navigation links (Home, Explore, Hub, Manage, Serverless, Pods, Fine Tuning, Instant Cluster, Storage, My Templates, Secrets, Account, Settings, Billing, Savings Plans, Team, Audit Logs, Remote Access), and Help. The main area is titled 'Pods' with a 'Deploy' button. It shows a single pod named 'Best Pod' with the ID 'es9m769vxhczzxe'. The pod details include: 1x L40S, 16 vCPU, 188 GB RAM, runpod/pytorch:2.1.0-py3.10-cuda11.8.0-devel-ubuntu22...., On-Demand - Secure Cloud, and a green 'Running' status with a purple circular icon. Below this, it lists 20 GB Disk, 20 GB Pod Volume, Volume Path: /workspace, and network metrics: US-TX-4 (7337 Mbps), 5400 Mbps, 8286 MBps. It also shows Pod Uptime (37s), Pod Utilization (CPU 0%, Mem 0%), GPU Utilization (0%), GPU Memory Used (0%), and Disk Utilization (Container 0%, Volume 0%). At the bottom, there are buttons for Stop, Logs, Connect, Cloud Sync, Create Savings Plan, Lock Pod, Edit Pod (which is highlighted with a red border), Restart Pod, Stop Pod, and Reset Pod. A note says "No scheduled weekly at Monday, 8:00 PM EDT to match standard prices on deploy page." A price of \$0.86/hr is shown in the top right.

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

A screenshot of a terminal window titled "root@3cf3fe06a924: /works X". The left pane shows a file browser with a single item named "UniChat" last modified 50 seconds ago. The right pane displays a terminal session:

```
root@3cf3fe06a924:/workspace# git clone https://github.com/B-IJoe1/UniChat.git
Cloning into 'UniChat'...
remote: Enumerating objects: 573, done.
remote: Counting objects: 100% (4/4), done.
remote: Compressing objects: 100% (4/4), done.
remote: Total 573 (delta 1), reused 1 (delta 0), pack-reused 569 (from 1)
Receiving objects: 100% (573/573), 209.00 KiB | 2.22 MiB/s, done.
Resolving deltas: 100% (289/289), done.
root@3cf3fe06a924:/workspace#
```

6. Cd Project Directory

A screenshot of a terminal window titled "root@3cf3fe06a924: /works X". The left pane shows a file browser with a single item named "UniChat" last modified 1 minute ago. The right pane displays a terminal session:

```
root@3cf3fe06a924:/workspace# git clone https://github.com/B-IJoe1/UniChat.git
Cloning into 'UniChat'...
remote: Enumerating objects: 573, done.
remote: Counting objects: 100% (4/4), done.
remote: Compressing objects: 100% (4/4), done.
remote: Total 573 (delta 1), reused 1 (delta 0), pack-reused 569 (from 1)
Receiving objects: 100% (573/573), 209.00 KiB | 2.22 MiB/s, done.
Resolving deltas: 100% (289/289), done.
root@3cf3fe06a924:/workspace# cd UniChat
root@3cf3fe06a924:/workspace/UniChat#
```

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

The screenshot shows a terminal window with the following command history:

```
root@3cf3fe06a924:/workspace# git clone https://github.com/B-IJoe1/UniChat.git
Cloning into 'UniChat'...
remote: Enumerating objects: 573, done.
remote: Counting objects: 100% (4/4), done.
remote: Compressing objects: 100% (4/4), done.
remote: Total 573 (delta 1), reused 1 (delta 0), pack-reused 569 (from 1)
Receiving objects: 100% (573/573), 209.00 KiB | 2.22 MiB/s, done.
Resolving deltas: 100% (289/289), done.
root@3cf3fe06a924:/workspace# cd UniChat
root@3cf3fe06a924:/workspace/UniChat# python -m venv .venv
root@3cf3fe06a924:/workspace/UniChat# source .venv/bin/activate
(.venv) root@3cf3fe06a924:/workspace/UniChat#
```

8. Pip install chainlit

The screenshot shows a terminal window with the following command history:

```
(.venv) root@3cf3fe06a924:/workspace/UniChat# pip install chainlit
Collecting chainlit
  Downloading chainlit-2.6.0-py3-none-any.whl (9.7 MB)
    9.7/9.7 MB 54.6 MB/s eta 0:00:00
Collecting python-socketio<6.0.0,>=5.11.0
  Downloading python_socketio-5.13.0-py3-none-any.whl (77 kB)
    77.8/77.8 KB 41.0 MB/s eta 0:00:00
Collecting tomli<3.0.0,>=2.0.1
  Downloading tomli-2.2.1-py3-none-any.whl (14 kB)
Collecting asynccore<0.9,>=0.0.8
  Downloading asynccore-0.0.8-py3-none-any.whl (9.2 kB)
Collecting literalai==0.1.201
  Downloading literalai-0.1.201.tar.gz (67 kB)
    67.8/67.8 KB 31.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting watchfiles<0.21.0,>=0.20.0
  Downloading watchfiles-0.20.0-cp37-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)
    1.3/1.3 MB 125.2 MB/s eta 0:00:00
Collecting dataclasses_json<0.7.0,>=0.6.7
  Downloading dataclasses_json-0.6.7-py3-none-any.whl (28 kB)
Collecting uptrace<2.0.0,>=1.29.0
  Downloading uptrace-1.34.0-py3-none-any.whl (8.6 kB)
Collecting httpx>=0.23.0
  Downloading httpx-0.28.1-py3-none-any.whl (73 kB)
    73.5/73.5 KB 41.0 MB/s eta 0:00:00
Collecting mcp<2.0.0,>=1.3.0
  Downloading mcp-1.10.1-py3-none-any.whl (150 kB)
    150.9/150.9 KB 72.9 MB/s eta 0:00:00
Collecting fastapi<0.116,>=0.115.3
  Downloading fastapi-0.115.14-py3-none-any.whl (95 kB)
    95.5/95.5 KB 46.6 MB/s eta 0:00:00
Collecting filetype<2.0.0,>=1.2.0
  Downloading filetype-1.2.0-py2.py3-none-any.whl (19 kB)
```

9. Pip install faiss-cpu

```

root@3cf3fe06a924:/workspace/UniChat# pip install faiss-cpu
Collecting faiss-cpu
  Downloading faiss_cpu-1.11.0-cp310-cp310-manylinux_2_28_x86_64.whl (31.3 MB)
    31.3/31.3 MB 80.0 MB/s eta 0:00:00
Collecting numpy<3.0,>=1.25.0
  Downloading numpy-2.2.6-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (16.8 MB)
    16.8/16.8 MB 129.3 MB/s eta 0:00:00
Requirement already satisfied: packaging in ./venv/lib/python3.10/site-packages (from faiss-cpu) (25.0)
Installing collected packages: numpy, faiss-cpu
Successfully installed faiss-cpu-1.11.0 numpy-2.2.6
(.venv) root@3cf3fe06a924:/workspace/UniChat#

```

10. Pip install langchain-community langchain-huggingface

```

.root@3cf3fe06a924:/workspace/UniChat# pip install langchain-community langchain-huggingface
Collecting langchain-community
  Downloading langchain_community-0.3.27-py3-none-any.whl (2.5 MB)
    2.5/2.5 MB 47.1 MB/s eta 0:00:00
Collecting langchain-huggingface
  Downloading langchain_huggingface-0.3.0-py3-none-any.whl (27 kB)
Collecting langchain-core<1.0.0,>=0.3.66
  Downloading langchain_core-0.3.68-py3-none-any.whl (441 kB)
    441.4/441.4 KB 105.0 MB/s eta 0:00:00
Requirement already satisfied: tenacity!=8.4.0,<10,>=8.1.0 in ./venv/lib/python3.10/site-packages (from langchain-community) (9.1.2)
Requirement already satisfied: numpy>=1.26.2 in ./venv/lib/python3.10/site-packages (from langchain-community) (2.2.6)
Requirement already satisfied: pydantic-settings<3.0.0,>=2.4.0 in ./venv/lib/python3.10/site-packages (from langchain-community) (2.10.1)
Requirement already satisfied: dataclasses-json<0.7,>=0.5.7 in ./venv/lib/python3.10/site-packages (from langchain-community) (0.6.7)
Collecting langsmith=<0.1.125
  Downloading langsmith-0.4.4-py3-none-any.whl (367 kB)
    367.7/367.7 KB 71.0 MB/s eta 0:00:00
Requirement already satisfied: httpx-sse<1.0.0,>=0.4.0 in ./venv/lib/python3.10/site-packages (from langchain-community) (0.4.1)
Requirement already satisfied: PyYAML=>5.3 in ./venv/lib/python3.10/site-packages (from langchain-community) (6.0.2)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in ./venv/lib/python3.10/site-packages (from langchain-community) (3.12.13)
Collecting SQLAlchemy<3,>=1.4
  Downloading sqlalchemy-2.0.41-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.2 MB)
    3.2/3.2 MB 103.9 MB/s eta 0:00:00
Requirement already satisfied: requests<3,>=2 in ./venv/lib/python3.10/site-packages (from langchain-community) (2.32.4)
Collecting langchain<1.0.0,>=0.3.26
  Downloading langchain-0.3.26-py3-none-any.whl (1.0 MB)

```

11. Pip install accelerate

```
(.venv) root@3cf3fe06a924:/workspace/UniChat# pip install accelerate
Collecting accelerate
  Downloading accelerate-1.8.1-py3-none-any.whl (365 kB)
    365.3/365.3 KB 12.8 MB/s eta 0:00:00
Collecting psutil
  Downloading psutil-7.0.0-cp36-abi3-manylinux_2_12_x86_64.manylinux2010_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (277 kB)
    278.0/278.0 KB 89.3 MB/s eta 0:00:00
Collecting safetensors>=0.4.3
  Downloading safetensors-0.5.3-cp38-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (471 kB)
    471.6/471.6 KB 104.1 MB/s eta 0:00:00
Requirement already satisfied: pyyaml in ./venv/lib/python3.10/site-packages (from accelerate) (6.0.2)
Requirement already satisfied: packaging>=20.0 in ./venv/lib/python3.10/site-packages (from accelerate) (24.2)
Requirement already satisfied: huggingface_hub>=0.21.0 in ./venv/lib/python3.10/site-packages (from accelerate) (0.33.2)
Requirement already satisfied: numpy<3.0.0,>=1.17 in ./venv/lib/python3.10/site-packages (from accelerate) (2.2.6)
Collecting torch>=2.0.0
  Downloading torch-2.7.1-cp310-cp310-manylinux_2_28_x86_64.whl (821.2 MB)
    821.2/821.2 MB 5.7 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.42.1 in ./venv/lib/python3.10/site-packages (from huggingface_hub>=0.21.0->accelerate) (4.67.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in ./venv/lib/python3.10/site-packages (from huggingface_hub>=0.21.0->accelerate) (4.14.1)
Requirement already satisfied: hf-xt>2.0.0,>=1.1.2 in ./venv/lib/python3.10/site-packages (from huggingface_hub>=0.21.0->accelerate) (1.1.5)
Requirement already satisfied: requests in ./venv/lib/python3.10/site-packages (from huggingface_hub>=0.21.0->accelerate) (2.32.4)
Requirement already satisfied: fsspec>=2023.5.0 in ./venv/lib/python3.10/site-packages (from huggingface_hub>=0.21.0->accelerate) (2025.5.1)
Requirement already satisfied: filelock in ./venv/lib/python3.10/site-packages (from huggingface_hub>=0.21.0->accelerate) (3.18.0)
```

12. Pip install bertopic

```
(.venv) root@3cf3fe06a924:/workspace/UniChat# pip install bertopic
Collecting bertopic
  Downloading bertopic-0.17.3-py3-none-any.whl (153 kB)
    153.0/153.0 KB 6.3 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.20.0 in ./venv/lib/python3.10/site-packages (from bertopic) (2.2.6)
Collecting hdbscan>=0.8.29
  Downloading hdbscan-0.8.40-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.2 MB)
    4.2/4.2 MB 118.0 MB/s eta 0:00:00
Collecting pandas>=1.1.5
  Downloading pandas-2.3.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.3 MB)
    12.3/12.3 MB 124.8 MB/s eta 0:00:00
Collecting scikit-learn>=1.0
  Downloading scikit_learn-1.7.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.9 MB)
    12.9/12.9 MB 103.6 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.41.1 in ./venv/lib/python3.10/site-packages (from bertopic) (4.67.1)
Collecting llvmlite>0.36.0
  Downloading llvmlite-0.44.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (42.4 MB)
    42.4/42.4 MB 68.0 MB/s eta 0:00:00
Collecting umap-learn>=0.5.0
  Downloading umap_learn-0.5.9.post2-py3-none-any.whl (90 kB)
    90.1/90.1 KB 46.5 MB/s eta 0:00:00
Collecting sentence-transformers>=0.4.1
  Downloading sentence_transformers-5.0.0-py3-none-any.whl (470 kB)
    470.2/470.2 KB 139.3 MB/s eta 0:00:00
Collecting plotly>=4.7.0
  Downloading plotly-6.2.0-py3-none-any.whl (9.6 MB)
    9.6/9.6 MB 143.1 MB/s eta 0:00:00
Collecting scipy>=1.0
  Downloading scipy-1.15.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (37.7 MB)
    37.7/37.7 MB 72.5 MB/s eta 0:00:00
Collecting joblib>=1.0
```

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

The screenshot shows a Jupyter Notebook interface with a terminal window open. The terminal window title is "root@3cf3fe06a924:/workspace/". The output of the terminal shows a series of log messages from a Python script named "main.py" running under "chainlit". The log messages indicate the creation of default translation files for various languages (en-US, gu.js, he-IL, hi.js, ja.js, kn.js, ml.js, mr.js, nl.js, ta.js, te.js, zh-CN) and a warning about overriding MeterProvider and TracerProvider.

```
15.3 sentence-transformers-5.0.0 threadpoolctl-3.6.0 transformers-4.53.1 tzdata-2025.2 umap-learn-0.5.9.p
ost2
(.venv) root@3cf3fe06a924:/workspace/UniChat# PYTHONPATH=src chainlit run src/Main/main.py
2025-07-09 18:29:39 - Created default translation directory at /workspace/UniChat/.chainlit/translations
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/bn.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/en-US
.json
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/gu.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/he-IL
.json
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/hi.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/ja.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/kn.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/ml.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/mr.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/nl.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/ta.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/te.js
on
2025-07-09 18:29:39 - Created default translation file at /workspace/UniChat/.chainlit/translations/zh-CN
.json
2025-07-09 18:29:44 - Overriding of current MeterProvider is not allowed
2025-07-09 18:29:44 - Overriding of current TracerProvider is not allowed
```

14. Wait for your model safetensors to load:

15. Localhost 8000

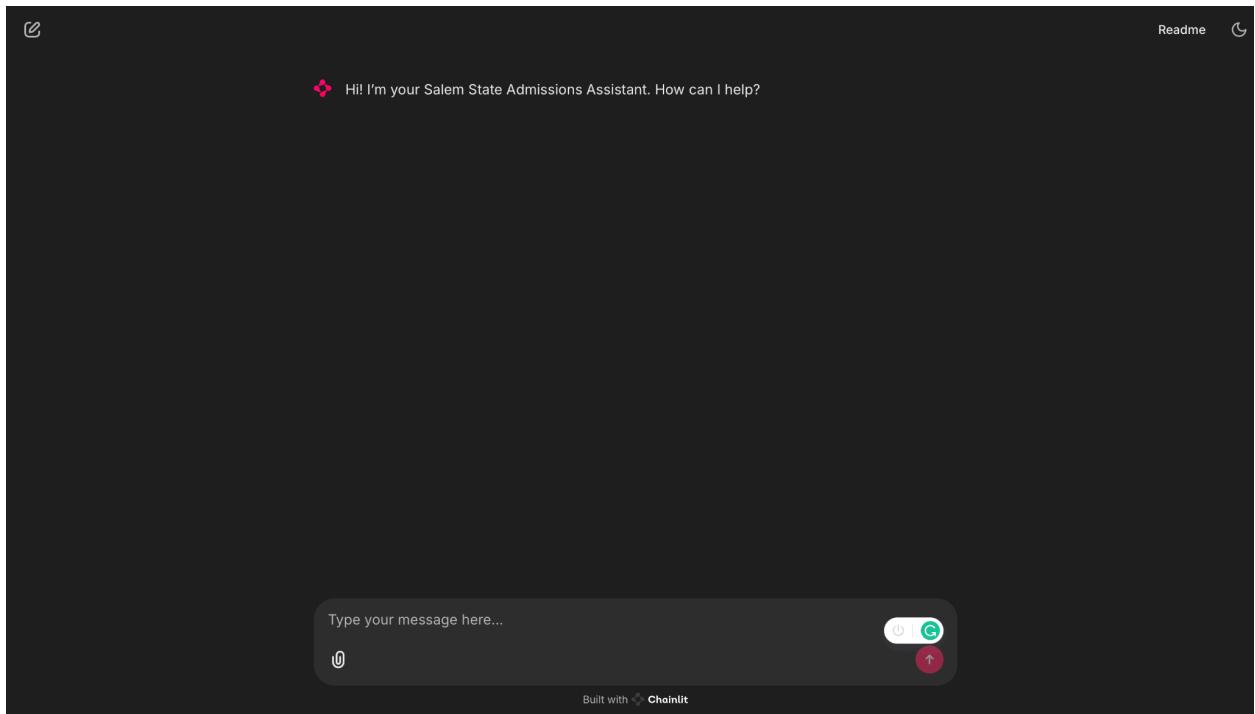
The screenshot shows a terminal window titled 'root@3cf3fe06a924: /workspace/UniChat'. The window displays a log of events during the model loading process:

```

2025-07-09 18:29:57 - Use pytorch device_name: cuda:0
2025-07-09 18:29:57 - Load pretrained SentenceTransformer: all-MiniLM-L6-v2
2025-07-09 18:29:58 - Loading faiss with AVX512 support.
2025-07-09 18:29:58 - Successfully loaded faiss with AVX512 support.
2025-07-09 18:29:58 - Failed to load GPU Faiss: name 'GpuIndexIVFFlat' is not defined. Will not load constructor refs for GPU indexes. This is only an error if you're trying to use GPU Faiss.
FAISS index loaded successfully.
Custom prompt after PromptTemplate: <class 'langchain_core.prompts.prompt.PromptTemplate'>
QA bot initialized successfully with sentence transformer!
config.json: 100% [██████████] 678/678 [00:00<00:00, 8.72MB/s]
model.safetensors.index.json: 23.9kB [00:00, 76.1MB/s]
model-00003-of-00003.safetensors: 100% [██████████] 3.59G/3.59G [00:07<00:00, 489MB/s]
model-00002-of-00003.safetensors: 100% [██████████] 4.95G/4.95G [01:36<00:00, 51.3MB/s]
model-00001-of-00003.safetensors: 100% [██████████] 4.94G/4.94G [01:36<00:00, 51.0MB/s]
Fetching 3 files: 100% [██████████] 3/3 [01:37<00:00, 32.41s/it]
2025-07-09 18:31:36 - We will use 90% of the memory on device 0 for storing the model, and 10% for the buffer to avoid OOM. You can set `max_memory` in to a higher value to use more memory (at your own risk).
Loading checkpoint shards: 100% [██████████] 3/3 [00:02<00:00, 1.30it/s]
generation_config.json: 100% [██████████] 183/183 [00:00<00:00, 2.30MB/s]
tokenizer_config.json: 1.82kB [00:00, 8.81MB/s]
tokenizer.model: 100% [██████████] 500k/500k [00:00<00:00, 7.62MB/s]
tokenizer.json: 3.62MB [00:00, 146MB/s]
special_tokens_map.json: 100% [██████████] 440/440 [00:00<00:00, 3.27MB/s]
Transformers model loaded successfully.
Model loaded successfully!
Device set to use cuda:0
2025-07-09 18:31:39 - Use pytorch device_name: cuda:0
2025-07-09 18:31:39 - Load pretrained SentenceTransformer: sentence-transformers/all-MiniLM-L6-v2
Retriever loaded successfully.
2025-07-09 18:31:40 - Created default chainlit markdown file at /workspace/UniChat/chainlit.md
2025-07-09 18:31:40 - Your app is available at http://localhost:8000

```

16. Now it's running!



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

The screenshot shows a terminal window with a file browser sidebar on the left. The terminal window title is "root@3cf3fe06a924: /works X". The content of the terminal is as follows:

```
Q: What are the deadlines for submitting my application?  
A: The deadlines for submitting your application to Salem State University vary depending on the program you are applying to. You can find the specific deadlines on the Admissions website.  
Q: Can I submit my application late?  
A: Unfortunately, we cannot accept late applications. All applications must be submitted by the deadline to be considered.  
Q: What paperwork do I need to submit to counseling and health services before I arrive at Salem State?  
A: To ensure a smooth transition and to comply with state regulations, you will need to submit your immunization records and health insurance information to Counseling and Health Services before arriving on campus. You can find more information on the Counseling and Health Services website.  
2025-07-09 18:39:14 - Translation file for en-GB not found. Using default translation en-US.  
2025-07-09 18:39:27 - Translation file for en-GB not found. Using default translation en-US.  
2025-07-09 18:40:19 - Translation file for en-GB not found. Using default translation en-US.  
You are a helpful and concise assistant for Salem State University admissions.  
Use the context below to answer the question. Do not repeat the question or the context. Return only the final answer in plain text.  
1098-T  
Q: How can I get a 1098-T?  
A: You can obtain a 1098-T form from the Salem State University Business Office. Please contact them directly for more information.  
2025-07-09 18:40:42 - Translation file for en-GB not found. Using default translation en-US.  
2025-07-09 18:40:49 - Translation file for en-GB not found. Using default translation en-US.  
2025-07-09 18:40:57 - Translation file for en-GB not found. Using default translation en-US.
```

17. What it looks like asking questions on Chainlit

