**Steps to get GPU access from RCIC.**

1. Send an email to [hpc-support@uci.edu](mailto:hpc-support@uci.edu), requesting to get access on HPC clusters.

Key Points to mention

1. Your UCI ID ([your\_uci\_id@uci.edu](mailto:your_uci_id@uci.edu))
2. Tier – We would like to have the free GPU tier of 1000 hours.
3. Task we will be running
   1. Inference Task – Model(Llama 3.1B)
   2. Model is expected to use around 16GB VRAM.
   3. Tell them that you would use the A30 GPUs that they have. Those are good enough for us.
   4. We will be storing all our files in the DFS file system.

**You (as a user) have access to /pub/<ucinetid>, which 1TB of DFS storage. DFS storage on /pub cannot be "pooled" among users, it's meant for individual use.**

**You (as a user) can submit to the free-gpu queue**

**You (as a user) cannot purchase storage or cycles from RCIC.**

**How to access the HPC cluster?**

1. Open your preferred terminal, make sure you are connected to the UCI VPN.
2. Run this command “ssh <your\_uci\_id>@hpc.rcic.edu”

Here's what I did (just as a regular user).

 to run the gemma LLM locally, I did the following to get an interactive session on a gpu node via srun:

srun --pty -p free-gpu --gres=gpu:A30:1 /bin/bash -i

 (Note: if the cluster is busy and all A30s are busy, you will have to wait. This very moment nearly all GPUs are in use)

Here's what I did to get the gemma model and then run it

# NOTE: Change the following to reflect your pub directory (/pub/stayde)

# Do these once

mkdir /pub/stayde/ollama

mkdir /pub/stayde/ollama/models

# Running OR downloading ollama always change into the ollama directory

cd /pub/stayde/ollama

# Download the latest ollama binary and unpack it. Only need to do this when you want a new version of #ollama

wget [https://github.com/ollama/ollama/releases/download/v0.5.11/ollama-linux-amd64.tgz](https://urldefense.com/v3/__https:/github.com/ollama/ollama/releases/download/v0.5.11/ollama-linux-amd64.tgz__;!!CzAuKJ42GuquVTTmVmPViYEvSg!J2CCXRayVwgvqnT4h0x04T0XIgprt2nbmkH-t7QifS8JPKxqWXQxOuGlaKoUqLj2frFSQXqLkrvtLdFvUMNcnQ$)  
tar xzvf  ollama-linux-amd64.tgz

# Next 3 lines should always be performed when running . It loads the Cuda module, tells ollama where to store models, and

# starts the ollama server

module load cuda/12.2.0

export OLLAMA\_MODELS=/pub/stayde/ollama/models  
bin/ollama serve &> ollama-serve.log &

# Download a new model (or get updates to an existing, already-downloaded model). Change the model name for

# what you want  
bin/ollama pull gemma    # (took 4-5 minutes)

# Run the model and interact with it. Ollama has extensive documentation  about different ways to interact with the ollama server  
bin/ollama run gemma

Don't forget to exit your srun session when you are done. That will kill all processes (e.g. the ollama server) and release the gpu for the next person.