Below are the instructions needed to setup a working environment to run our RAG pipeline in Palmetto as of 10/23/24

1. SSH into Palmetto and go to the directory you want to setup your environment in
    a. ssh username@slogin.palmetto.clemson.edu
2. Create a Palmetto task:
    a. salloc --nodes 1 --ntasks-per-node 1 --cpus-per-task 8 --mem 8G --time 3:00:00
3. Run wget https://www.python.org/ftp/python/3.10.15/Python-3.10.15.tar.xz to install Python 3.10.15
    a. Newer versions are incompatible with LangChain
4. Run tar -xvf Python-3.10.15.tar.xz to extract the file
5. cd into Python-3.10.15 and run ./configure --with-pydebug
6. Run make -s -j8
    a. Now python is setup
7. Use cd .. or navigate to whatever directory you want the environment in to
8. Run Python-3.10.15/python -m venv .venv
9. Use vim .bashrc and add source ~/.venv/bin/activate
    a. Adjust the path to where python is

This process installs the correct version of Python and sets the virtual environment to start up connection to Palmetto

Next we setup Ollama

It would not hurt to clone our github into Palmetto first

1. Create a job
    a. salloc --nodes 1 --ntasks-per-node 1 --cpus-per-task 8 --mem 8G --time 3:00:00 --gpus a100:1
2. Copy Ollama from github
    a. Wget https://github.com/ollama/ollama/releases/download/v0.3.14/ollama-linux-amd64.tgz
3. Extract Ollama
    a. tar -xzvf ollama-linux-amd64.tgz
4. Run these commands to start an Ollama server and install LLaMA 3
    a. $ cd /bin (from home dir)
    b. $ tmux
    c. > Ctrl + B (wait like 0.5s), C
    d. $ ./ollama serve
    e. > Ctrl + B, 0
    f. $ ./ollama run llama3.2:1b
5. Run ./ollama serve whenever you want to run the model
    a. Pip install:

i. Sqlite3-binary
ii. Langchain
iii. Langchain_community
iv. Langchain_ollama
v. Langchain_core
vi. Chromadb
vii. pypdf