HIGH PERFORMANCE RESEARCH COMPUTING

Programming Using the Jupyter AI Assistant

March 27th, 2026

Keegan Smith

High Performance
Research Computing
DIVISION OF RESEARCH

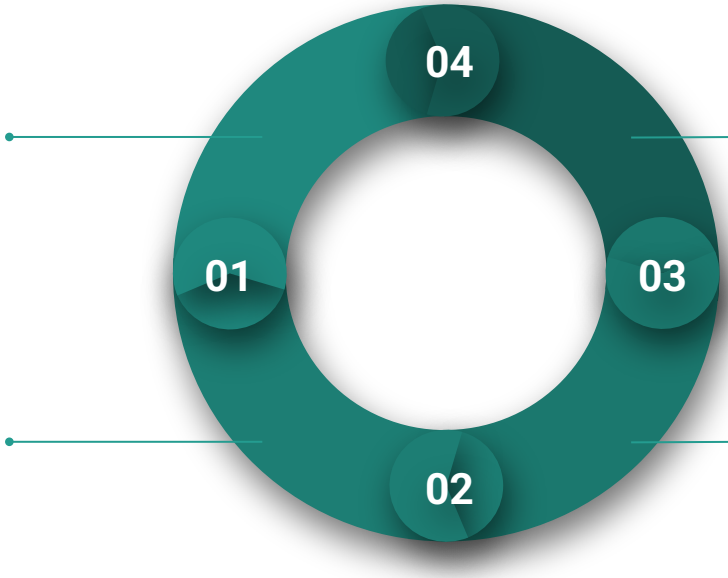
Schedule

Lab 1: Creating a session (15 minutes)

We will create a jupyter AI Assistant session from the ACES portal.

Lab 2: Debugging with Jupyter AI (15 mins)

We will go through some examples with two popular Python libraries: Pandas and Matplotlib for data exploration.



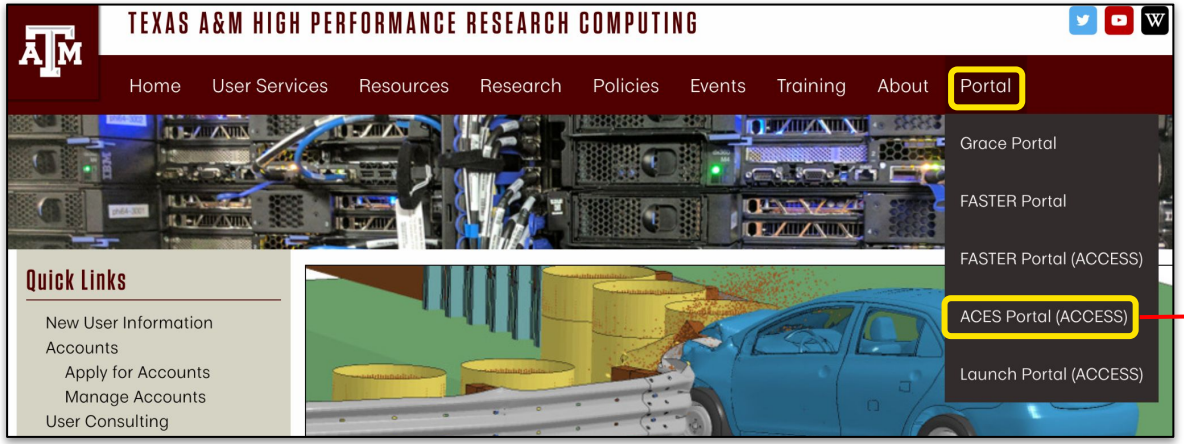
Lab 3: Creating a PyTorch Application (15 minutes)

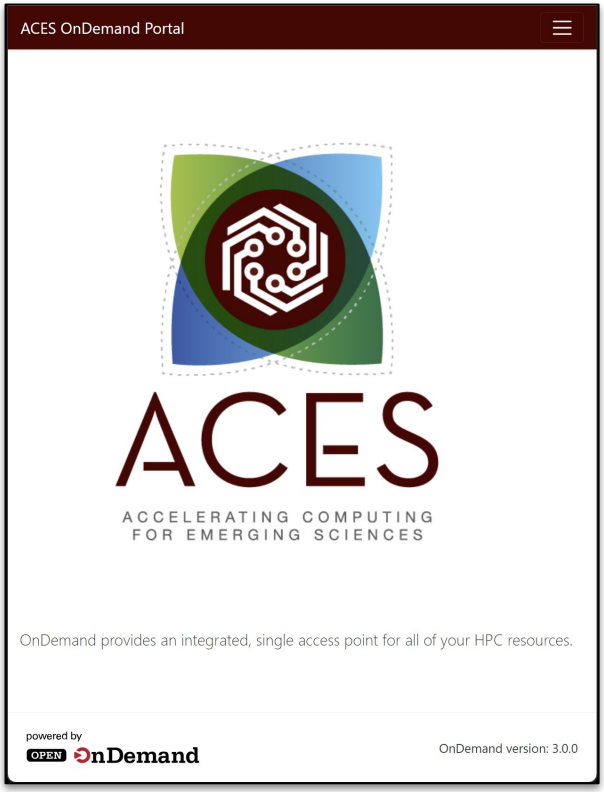
We will create and debug a Pytorch application from scratch with the assistance of Jupyter AI.

Lab 4: Creating a Tensorflow Application (15 minutes)

We will create and debug a Tensorflow application from scratch with the assistance of Jupyter AI.

ACES Portal

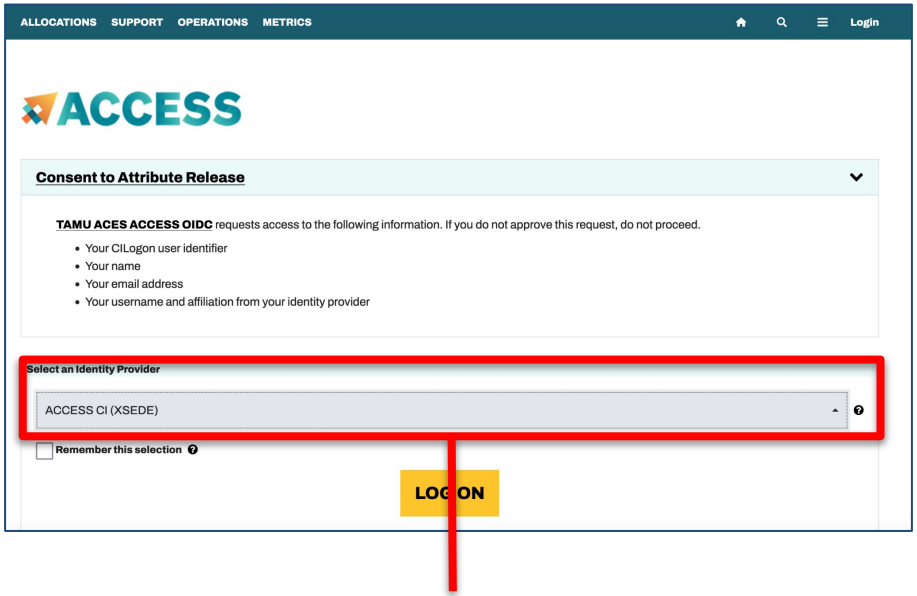




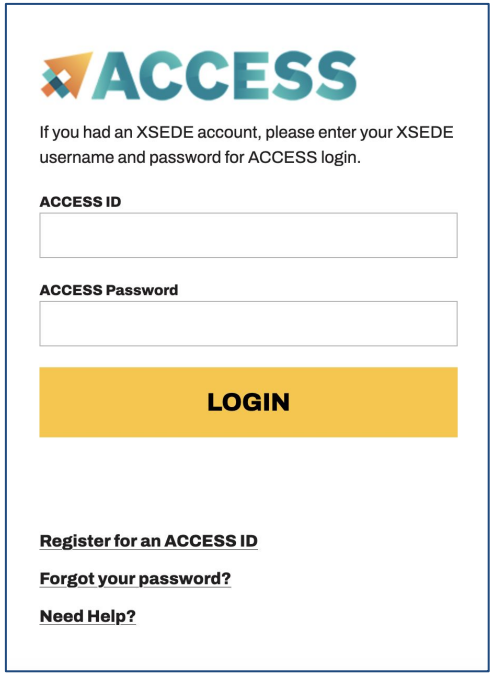
ACES Portal portal-aces.hprc.tamu.edu
is the web-based user interface for the ACES cluster

Open OnDemand (OOD) is an advanced web-based graphical interface framework for HPC users

Accessing ACES via the Portal (ACCESS)

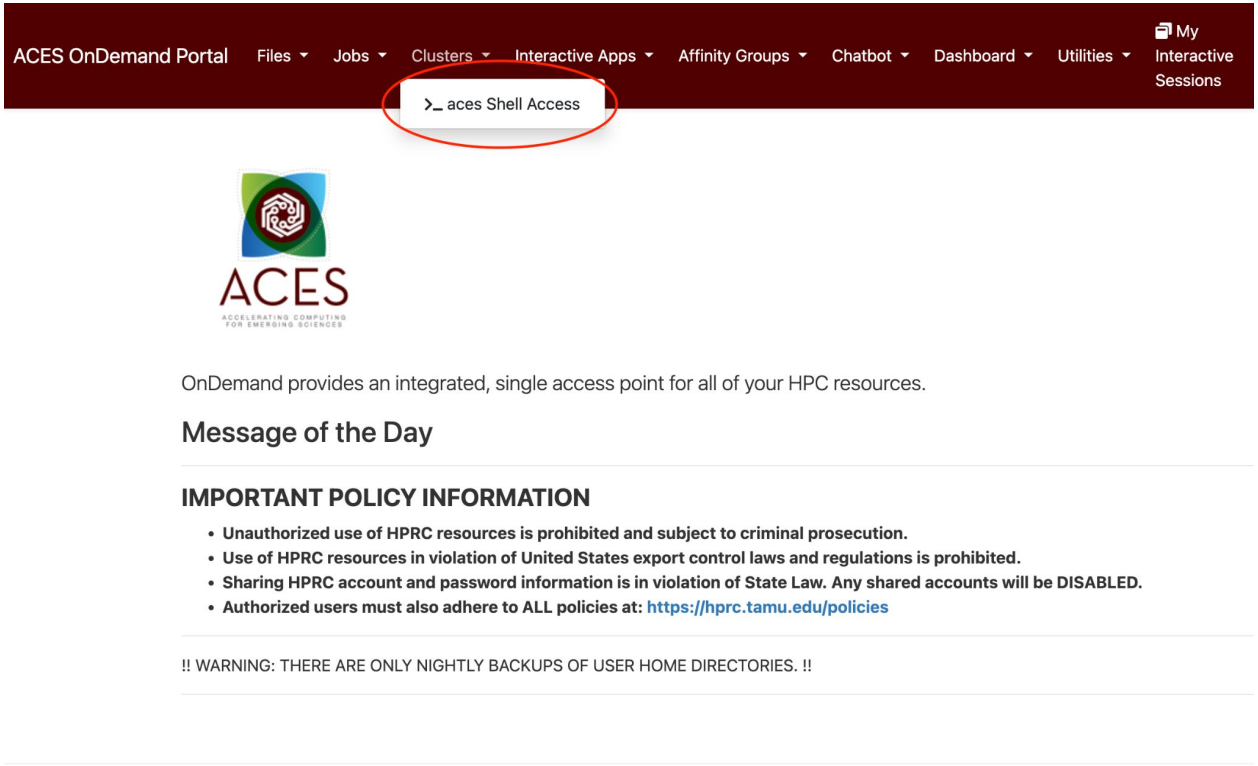


Select the Identity Provider appropriate for your account.



Log-in using your ACCESS or institutional credentials.

Setting up


ACES OnDemand Portal
Files
Jobs
Clusters
Interactive Apps
Affinity Groups
Chatbot
Dashboard
Utilities
My Interactive Sessions
>_ aces Shell Access
ACES
ACCELERATING COMPUTING FOR EMERGING SCIENCES
OnDemand provides an integrated, single access point for all of your HPC resources.
Message of the Day
IMPORTANT POLICY INFORMATION
Unauthorized use of HPRC resources is prohibited and subject to criminal prosecution.
Use of HPRC resources in violation of United States export control laws and regulations is prohibited.
Sharing HPRC account and password information is in violation of State Law. Any shared accounts will be DISABLED.
Authorized users must also adhere to ALL policies at: https://hprc.tamu.edu/policies
!! WARNING: THERE ARE ONLY NIGHTLY BACKUPS OF USER HOME DIRECTORIES. !!

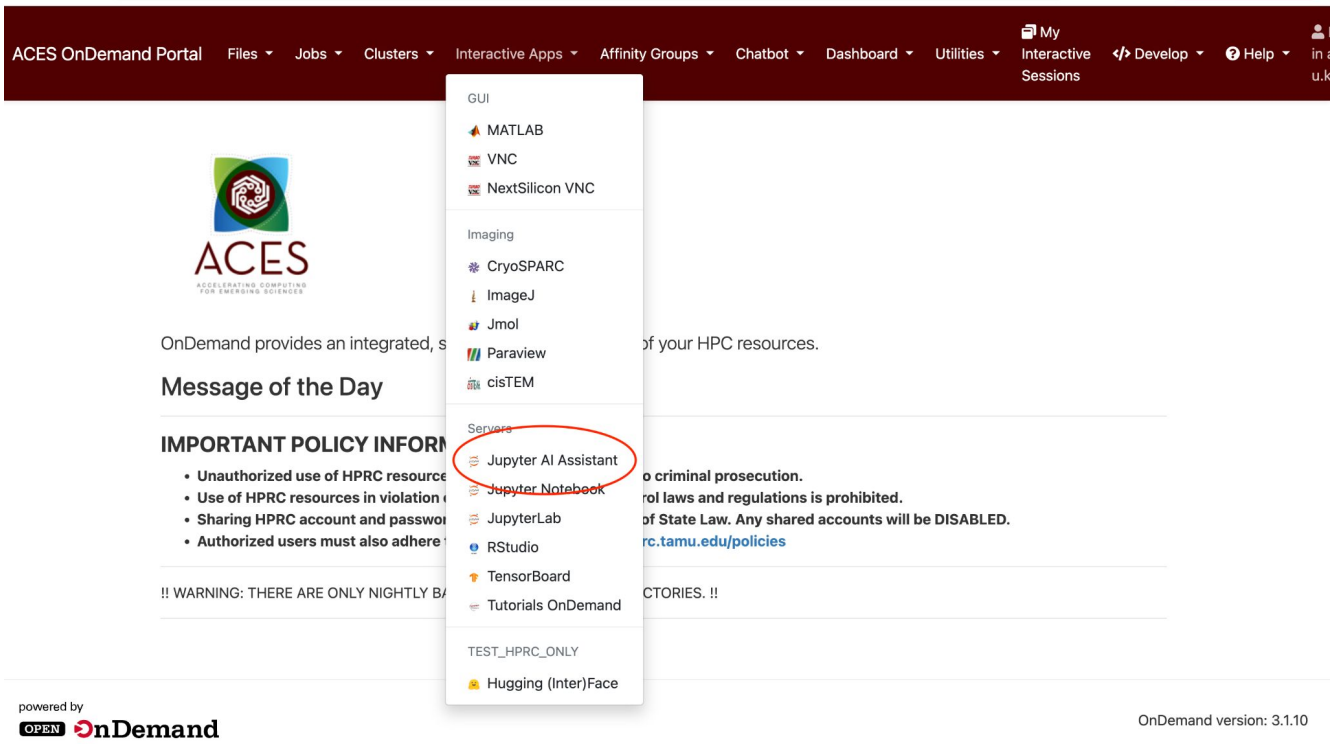
Setting up

In your terminal, type:

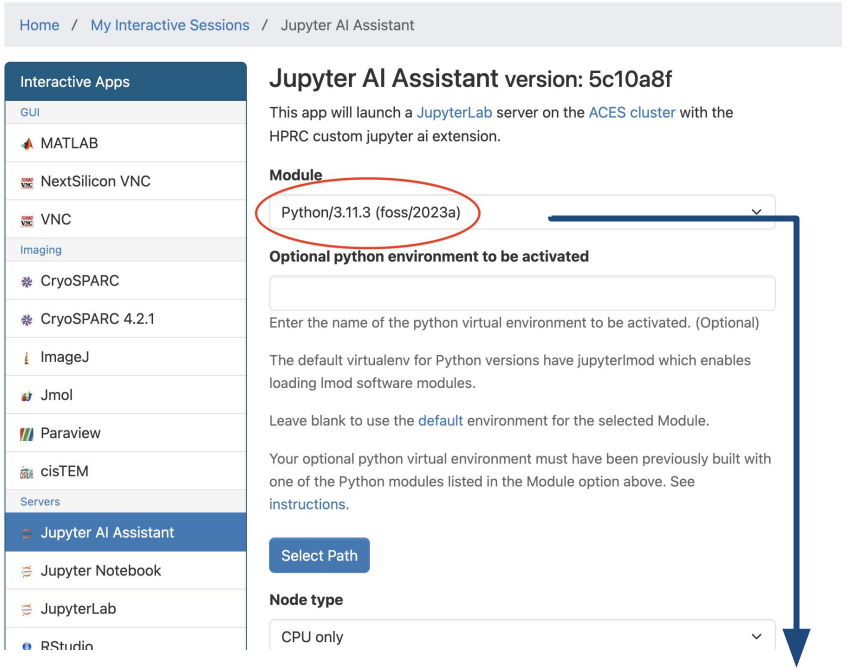
```
bash /scratch/training/jupyter_ai_intro/setup.sh
```

This will put the notebook `jupyter_ai_intro.ipynb` in your $SCRATCH directory.

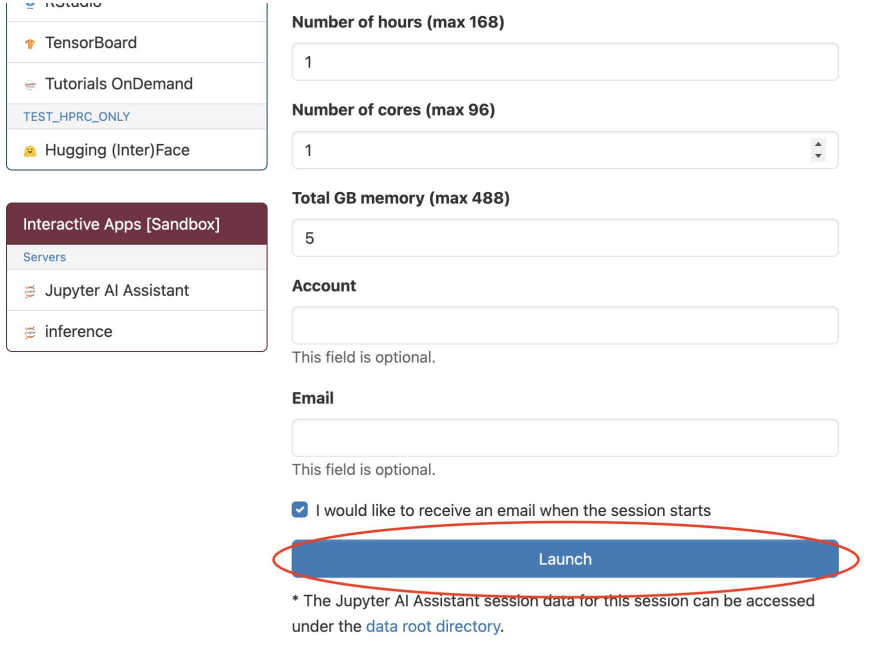
Starting Jupyter AI Assistant



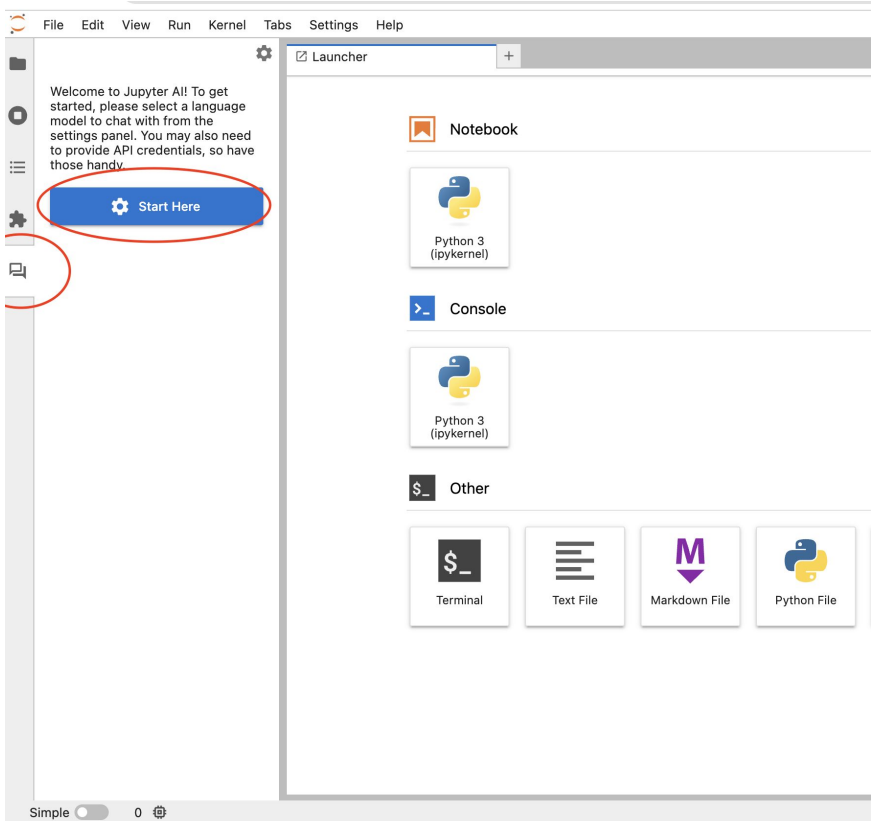
Starting Jupyter AI Assistant



Python/3.11.3 (foss/2023a)

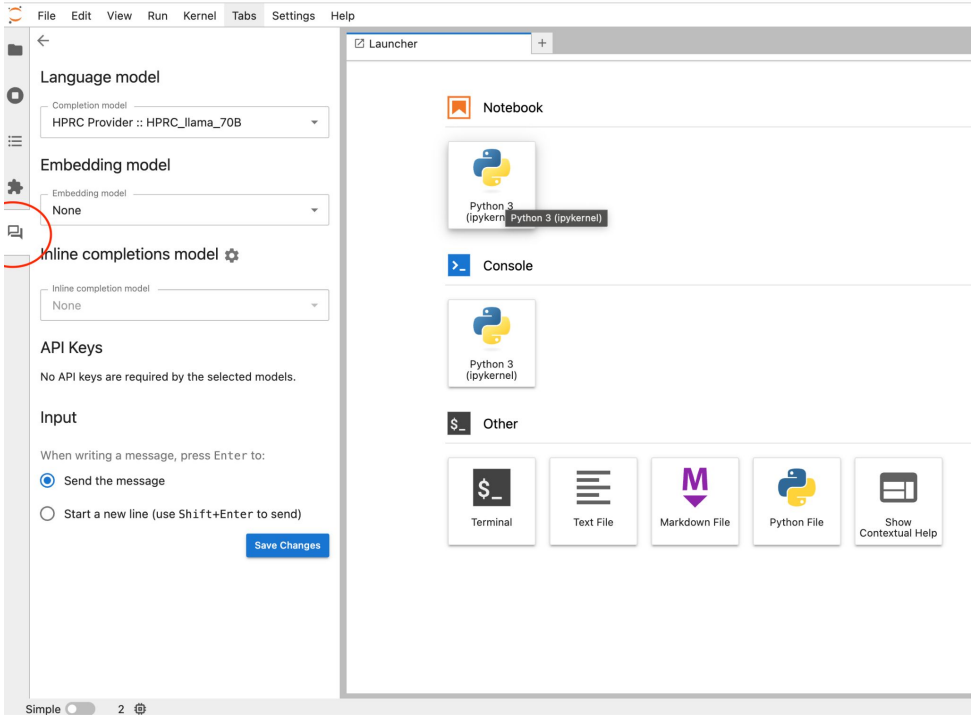


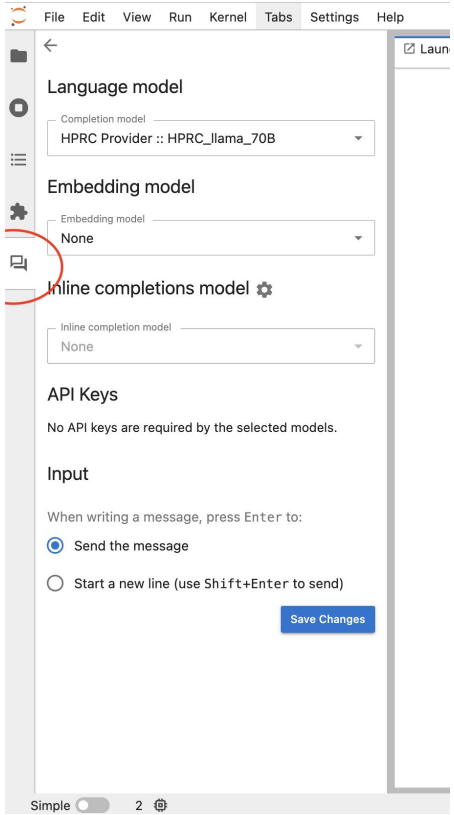
Using Jupyter AI Assistant



Using Jupyter AI Assistant

Click on the message icon on the left side of the screen



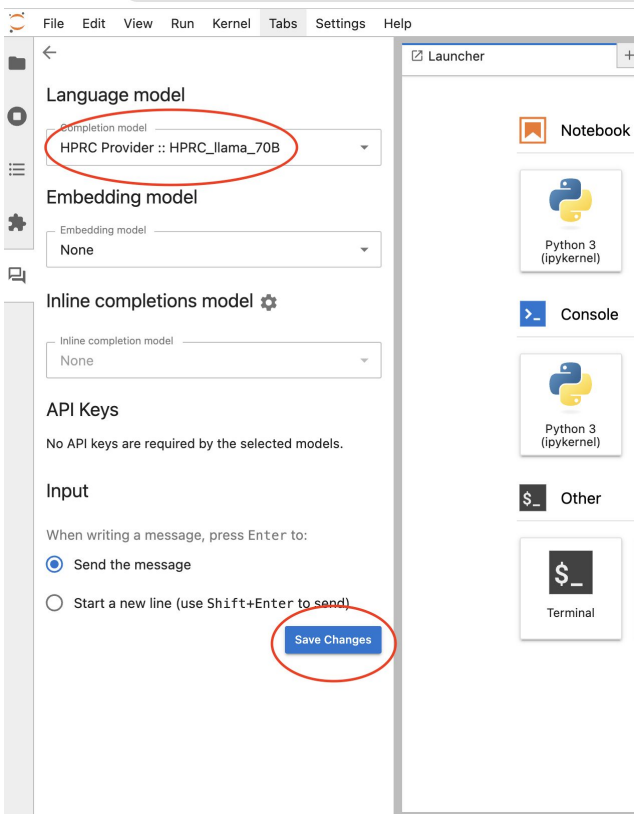


Using Jupyter AI Assistant

Select the “HPRC Provider :: HPRC_llama_70B” Completion model from the dropdown

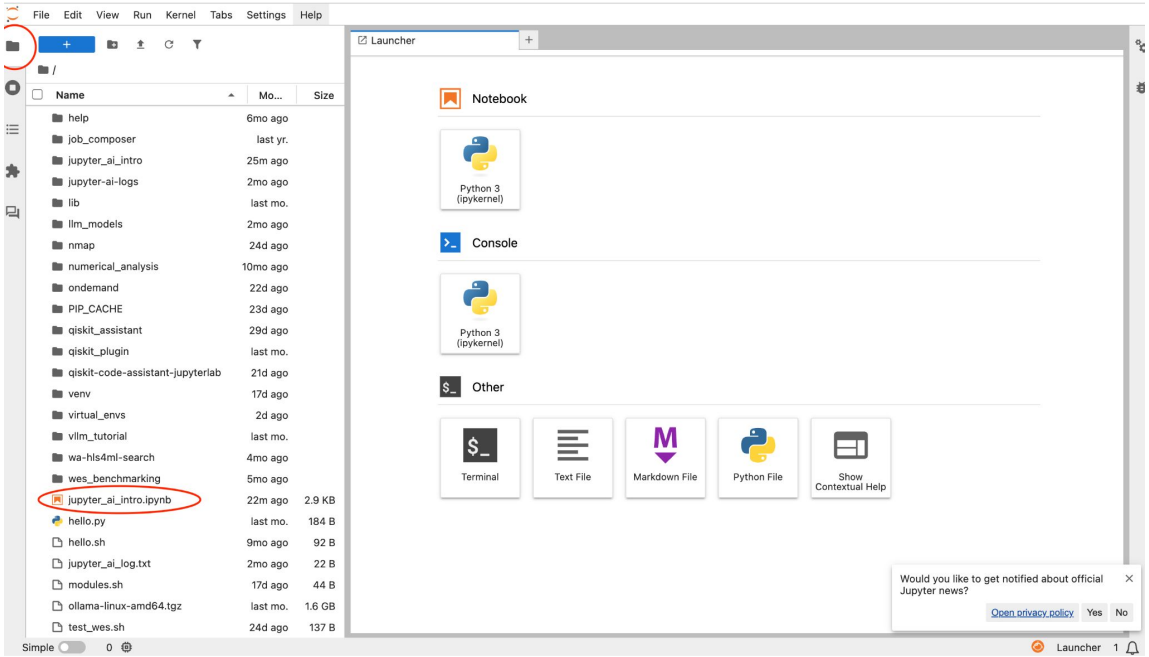
Click, “Save Changes”

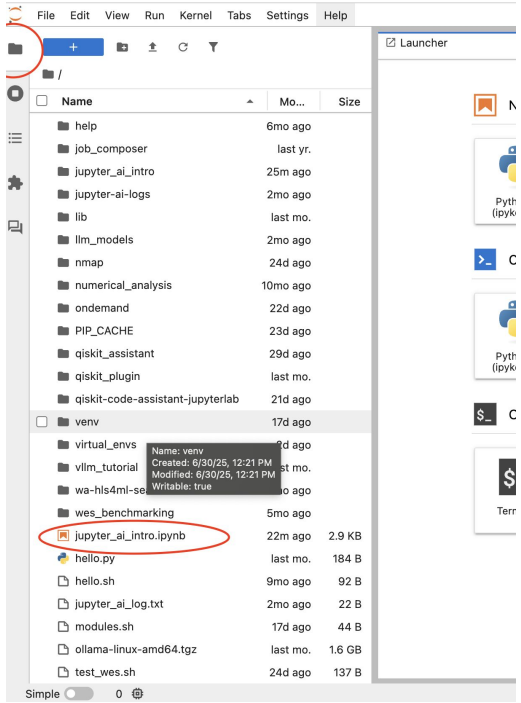
Click the back arrow on the top left to open the chat window



Main Content

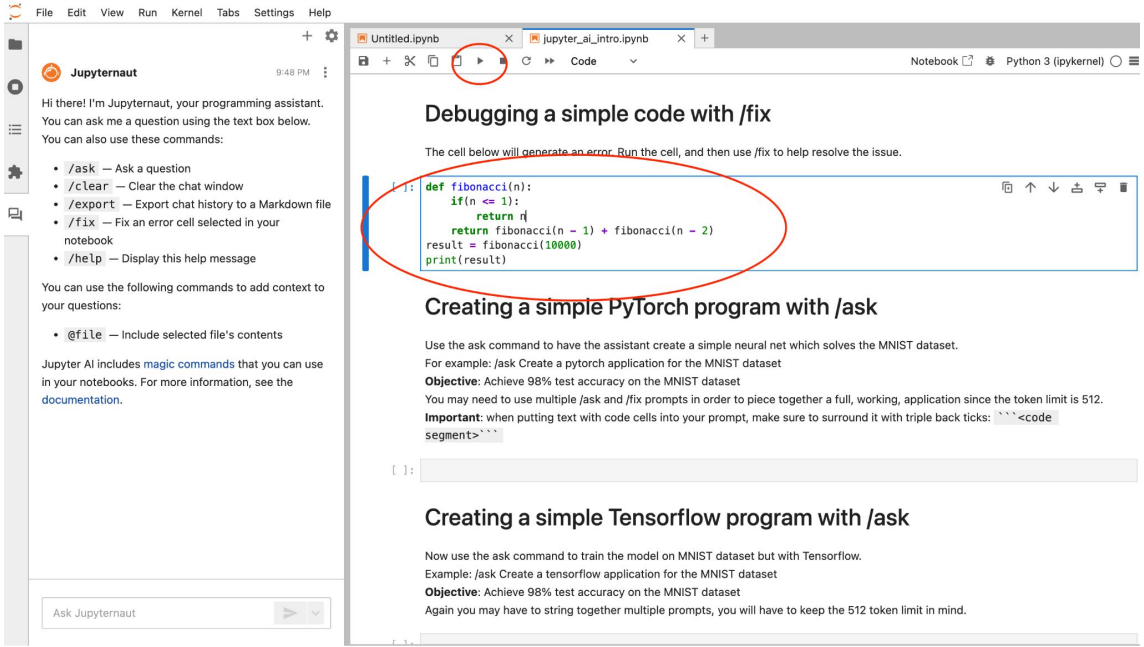
Open the “jupyter_ai_intro.ipynb” notebook in your jupyter ai assistant session:





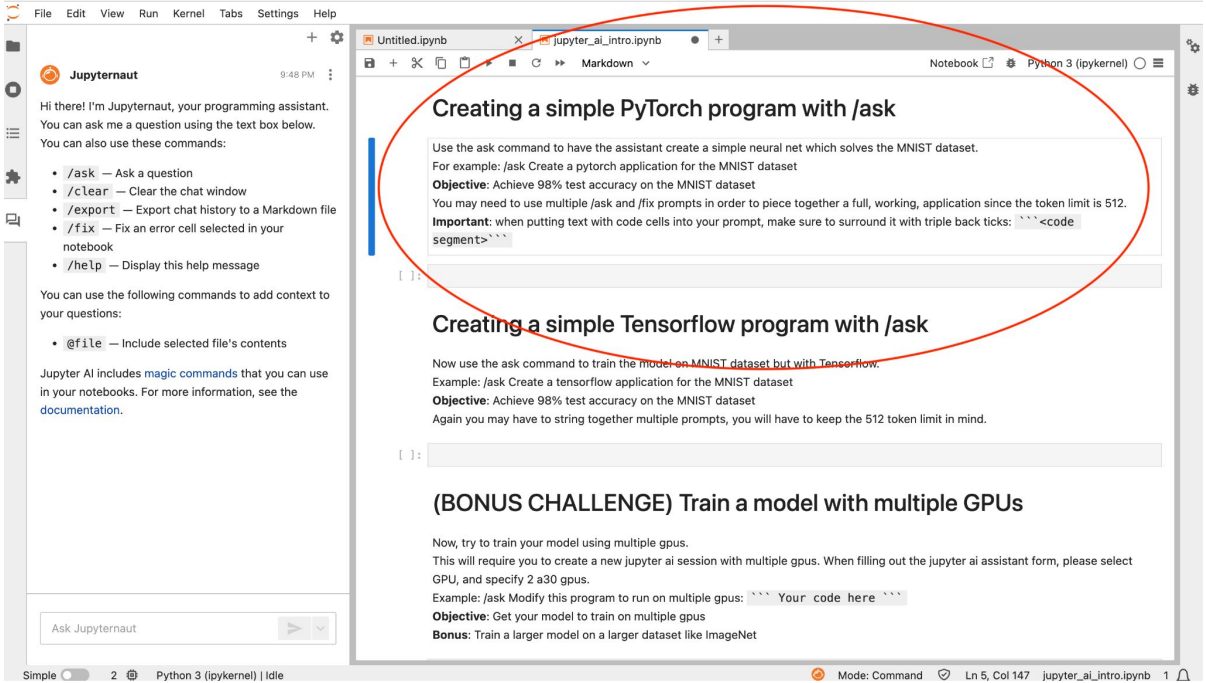
Debugging a simple program

- Click on the highlighted code cell
- Run the code cell with the highlighted button at the top of the screen
- Use the /fix command in the chat window to help debug the issue



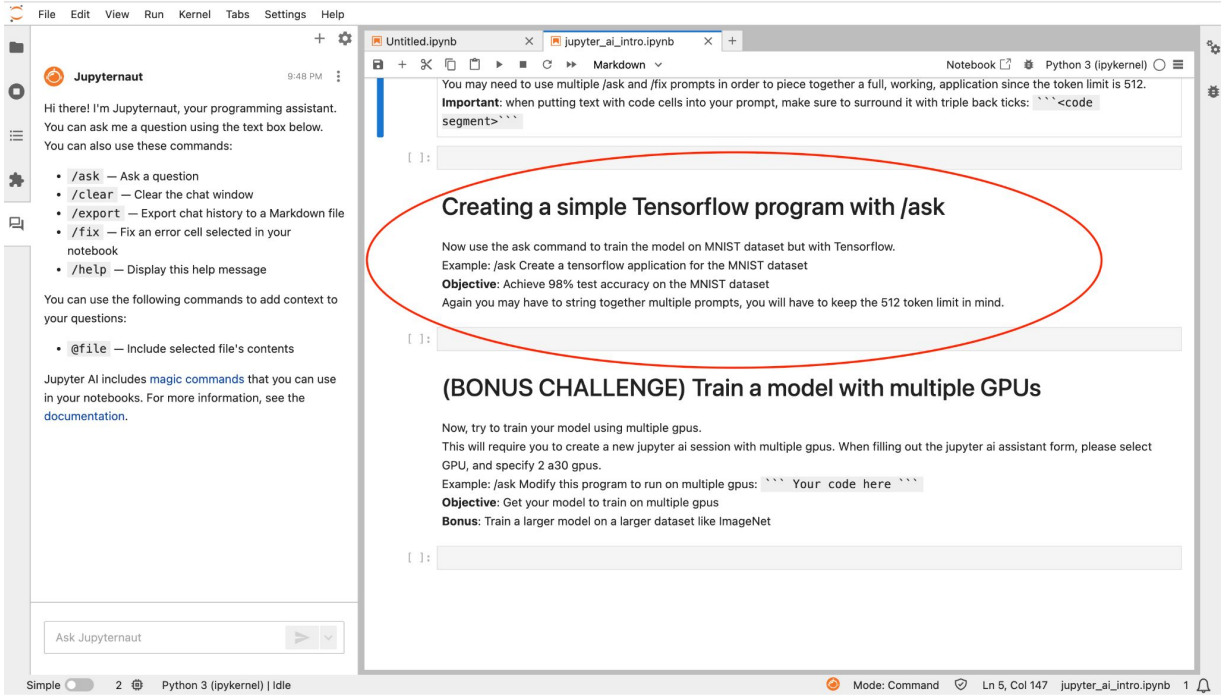
Creating a simple PyTorch program

- Use the /ask command to help develop code for solving the MNIST dataset with pytorch



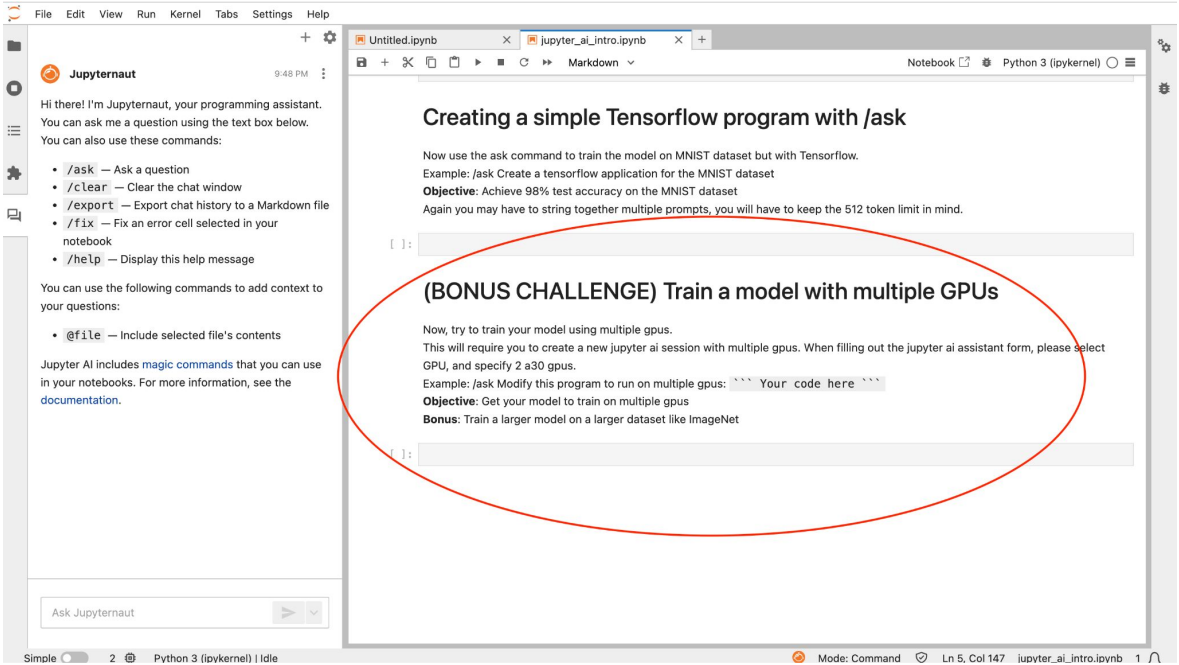
Creating a simple Tensorflow program

- Use the /ask command to help develop code for solving the MNIST dataset with Tensorflow



Bonus Challenge

- Create a new Jupyter AI Assistant session with 2 A30 GPUs
- Use /ask and /fix to help develop code which uses multiple gpus to train a large model on the ImageNet dataset



Technical Challenges

- Running LLM inference server on Intel PVC requires a full exclusive node
  - OneCCL does not play well with slurm, see https://community.intel.com/t5/oneAPI-Registration-Download/ccl-issue-with-Multi-GPU-AI-Training-Data-Parallel-with-Intel/td-p/1609982
- vLLM on xpu device would sometimes hang indefinitely when queuing prompts
  - Workaround: design an in-take web server to distribute prompts onto the backend LLM inference servers

Acknowledgements

This work was supported by

- the National Science Foundation (NSF), award numbers:
  - 2112356 - ACES - Accelerating Computing for Emerging Sciences
  - 1925764 - SWEETER - SouthWest Expertise in Expanding, Training, Education and Research
  - 2019129 - FASTER - Fostering Accelerated Scientific Transformations, Education, and Research
  - 2411377 - GOODLUCK - Growing Open OnDemand: Leveraging Unified Community Knowledge
- Staff and students at Texas A&M High-Performance Research Computing.
- ACCESS CCEP pilot program, Tier-II

Need Help?

First check the FAQ

- Knowledge Base
- Send us a ticket using the dashboard tab on our web portal
- Email further questions to help@hprc.tamu.edu

Help us help you -- when you contact us, tell us:

- Which cluster you're using
- Your username
- Job id(s) if any
- Location of your jobfile, input/output files
- Application used, if any
- Module(s) loaded, if any
- Error messages
- Steps you have taken, so we can reproduce the problem

High Performance Research Computing

DIVISION OF RESEARCH

https://hprc.tamu.edu

HPRC Helpdesk:

help@hprc.tamu.edu
Phone: 979-845-0219

Take our short course survey!

HPRC Survey

https://u.tamu.edu/hprc_shortcourse_survey

Thank you! Any Questions?

High Performance Research Computing

DIVISION OF RESEARCH

https://hprc.tamu.edu

HPRC Helpdesk:

help@hprc.tamu.edu
Phone: 979-845-0219

Help us help you. Please include details in your request for support, such as, Cluster (ACES, Faster, Grace, Launch), NetID ACCESS ID or Username, Job information (Job ID(s), Location of your jobfile, input/output files, Application, Module(s) loaded, Error messages, etc), and Steps you have taken, so we can reproduce the problem.