

NDP LLM-as-a-Service on NRP Tutorial



**ILKAY
ALTINTAS**



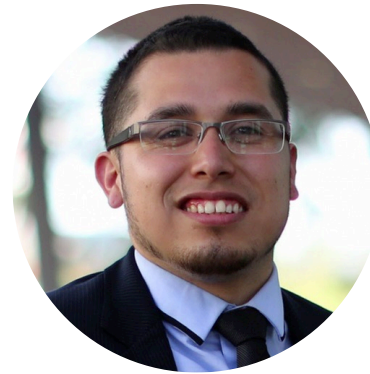
**SUBHASIS
DASGUPTA**



**SERGEY
GURVICH**



**PEDRO
RAMONETTI**



**ISMAEL
PEREZ**



**AMARNATH
GUPTA**

ClimateGPT: a GPT for Climate Research

ēcī Home⁰¹ ClimateGPT⁰² ClimateGPT+⁰³ Responsible AI⁰⁴ Governance⁰⁵ About⁰⁶



Developed by a team of researchers at RWTH Aachen University, in collaboration with Erasmus AI and others.

- A GPT model
 - Trained on over 10 billion web pages and millions of open-access academic articles for interdisciplinary research on climate change
 - 7B parameter models built with 300 billion tokens
- The models are:
 - Available through the HuggingFace interface
 - Well-maintained 7B, 13B, and 70B versions
 - Built on the Llama 2 architecture

Demo Tutorial: How do we add NAIRR context to ClimateGPT?

LLM: eci-io/climategpt-7b

Text: Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource

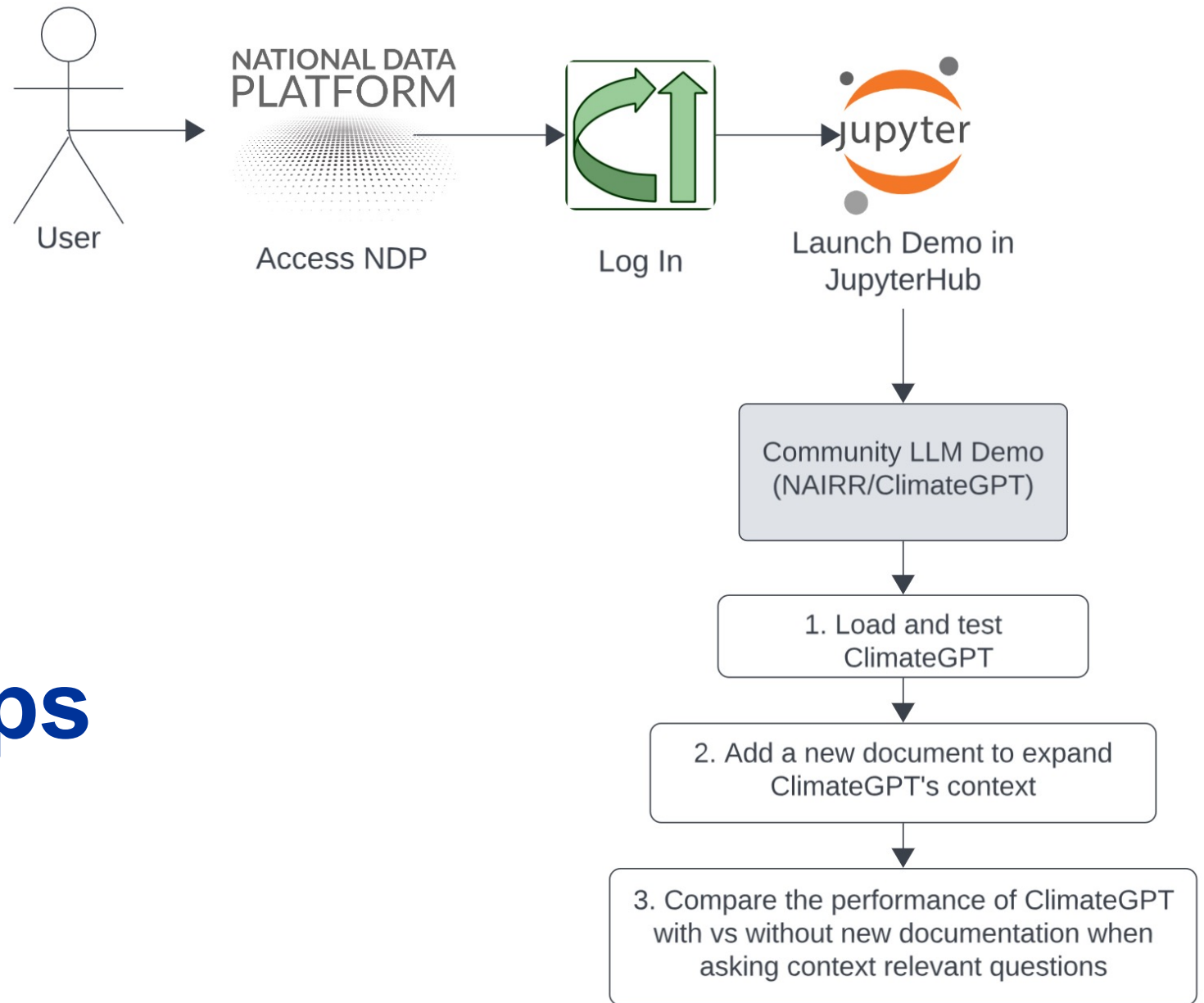
Question/Query: What task forces are recommended in the report?

Response: The report recommends the formation of several task forces to address specific issues related to the use of the NAIRR. These task forces include:

1. A task force to develop a process for identifying and documenting the NAIRR's capabilities and requirements.
2. A task force to develop a process for evaluating and selecting AI tools for inclusion in the NAIRR.
3. A task force to develop guidelines for the use of AI tools in the NAIRR.
4. A task force to develop training and support materials for the use of the NAIRR.
5. A task force to periodically survey the evolving AI tool landscape and provide advice on additions or deletions from the NAIRR standard virtual.
6. A task force to address any other issues related to the use of the NAIRR that may arise.

It's important to note that the report does not provide specific details about the membership or duration of these task forces. It's likely that the Operating Entity and resource providers would work together to form these task forces and determine their membership and goals.

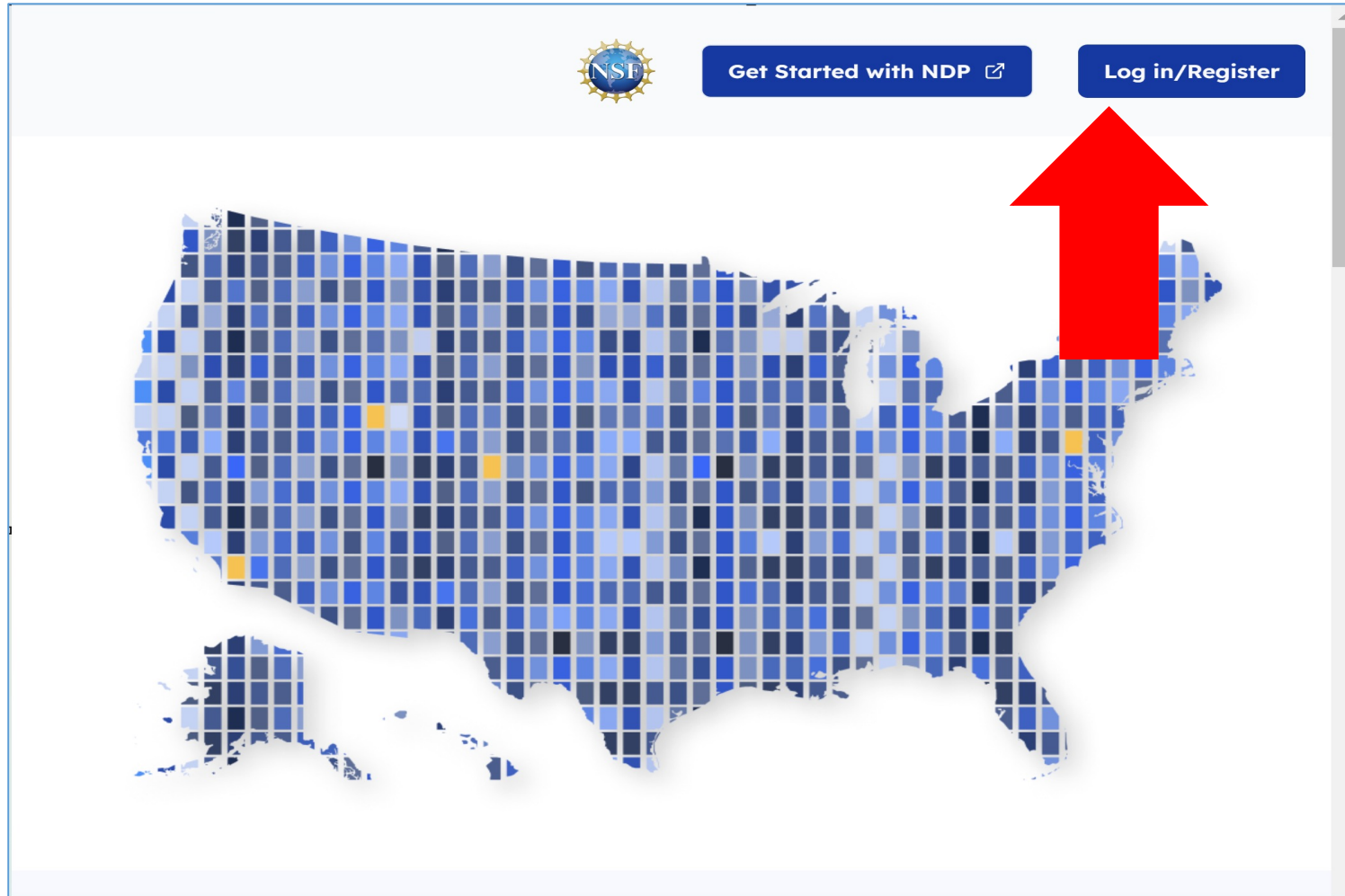
Tutorial Steps



Step 1: Go to nationaldatapatform.org

The screenshot shows the homepage of the National Data Platform. At the top left, it says 'NATIONAL DATA PLATFORM v0.1 alpha version' with a 'Release Notes' link. The navigation menu includes 'Catalog', 'Services', 'Education Hub', and 'About'. On the right, there are buttons for 'Get Started with NDP' and 'Log in/Register'. The main heading is 'Open Data, Equitable Access and AI Services for All', followed by the subtext 'Building the nation's federated data ecosystem. Explore data. Run analyses. Transform AI education.' Below this is a button 'Explore our catalog of datasets'. A large map of the United States is displayed, composed of a grid of blue and white squares. At the bottom, there are four statistics: 4615 data collections and livestreams, 5 data and AI services, 56 registered users, and 0 educational data challenges. Logos for partner institutions are shown on the left: SDSC (San Diego Supercomputer Center), SCI (The University of Utah), EarthScope Consortium, UC San Diego, The University of Utah, and the University of Colorado Boulder.

Step 2: Click on Log in/Register



Step 3: Select CI Logon

You are using v0.1 alpha version of the NDP platform. Please report any issues [here](#).

LOG IN


USERNAME or EMAIL


PASSWORD

Remember me [Forgot Password?](#)

SIGN IN

Or sign in with

 **CI Logon**



Step 4: Click on the Select an identity Provider dropdown and search your institution. Click on Log On.


Select an Identity Provider

ORCID ▾ ?

Remember this selection ?

Log On

By selecting "Log On", you agree to the [privacy_policy](#).



Step 5: Log in with your institutional credentials

SINGLE SIGN-ON UC San Diego

Passwords and Access Enroll in Two-Step Login Get Help

Signing on using: Active Directory

User name (or email address) Or sign on with:

 Active Directory ▾

Password:

[Reset password](#)

LOGIN

i Sign out and close your browser when you're finished.

Step 6: In your dashboard, click on JupyterHub

The screenshot shows the National Data Platform (NDP) dashboard for user Pedro Antonio Ramonetti Vega. The dashboard includes a left sidebar with navigation options: My Dashboard, My Uploads, Analysis Hub, Catalogs, and Education Hub. The main content area is titled 'My Dashboard' and features a welcome message, a list of recent updates, and a 'Quick Explore' section. The 'Quick Explore' section contains four buttons: 'Explore Data Catalog', 'Upload Data to Catalog', 'JupyterHub', and 'Education Gateway'. A red arrow points to the 'JupyterHub' button. To the right of the 'Quick Explore' section is a 'Your Profile' card showing the user's name, email, and last login time.

NATIONAL DATA PLATFORM
v0.1 alpha version
Release Notes

My Dashboard

Welcome back, Pedro Antonio Ramonetti Vega!

- January 19, 2024: More details ironed out. Alpha version of NDP is live!
- January 12, 2024: User dashboard drafted. JupyterHub integrated.
- January 5, 2024: String search integrated.
- December 22, 2023: CKAN data catalog is integrated into NDP. Large data and streaming uses cases are underway. Education gateway flow is being laid out.
- December 8, 2023: Landing page complete with Keycloak login. The foundation of the catalog of datasets, search, and education gateway is set.
- November 2023: Outlined user personas, assessed the needs of our users, and iron out the details of the initial features of NDP.

Quick Explore

- Explore Data Catalog
- Upload Data to Catalog
- JupyterHub**
- Education Gateway

Your Profile

Pedro Antonio Ramonetti Vega
pramonettivega@ucsd.edu
Last Logged In: January 18, 2024 8:45 PM

Step 7: Go to Image and select LLM Service Client. Click on Start.

Region
Any

GPUs
0

Cores
1

RAM, GB
16

GPU type
NVIDIA GeForce GTX 1080 Ti

/dev/shm for pytorch

Image
Minimal NDP Starter Jupyter Lab

- Minimal NDP Starter Jupyter Lab
- Physics Guided Machine Learning Starter Code
- SAGE Pilot Streaming Data Starter Code
- EarthScope Consortium Streaming Data Starter Code
- NAIRR Pilot - NASA Harmonized Landsat Sentinel-2 (HLS) Starter Code
- LLM Training (CUDA 12.3, tested with 1 GPU, 12 cores, 64GB RAM, NVIDIA A100-80GB)
- LLM Service Client (Minimal, No CUDA)**

will be deleted

Start

Today's Demo Tutorial running on NRP/Nautilus

- **LLM Service Client:** Built for question and answer using pre-trained LLM model for ClimateGPT
 - Load any model and add your documents
- **LLM Training:** Built for customizing models for your domain
 - Update and fine-tune an existing model permanently for your domain using a large corpus of documents
 - Contact ndp@sdsc.edu for more information on training services

Step 8: Wait for the server to launch

NATIONAL DATA PLATFORM Home Token pramonettivega@ucsd.edu Logout

Your server is starting up.

You will be redirected automatically when it's ready for you.

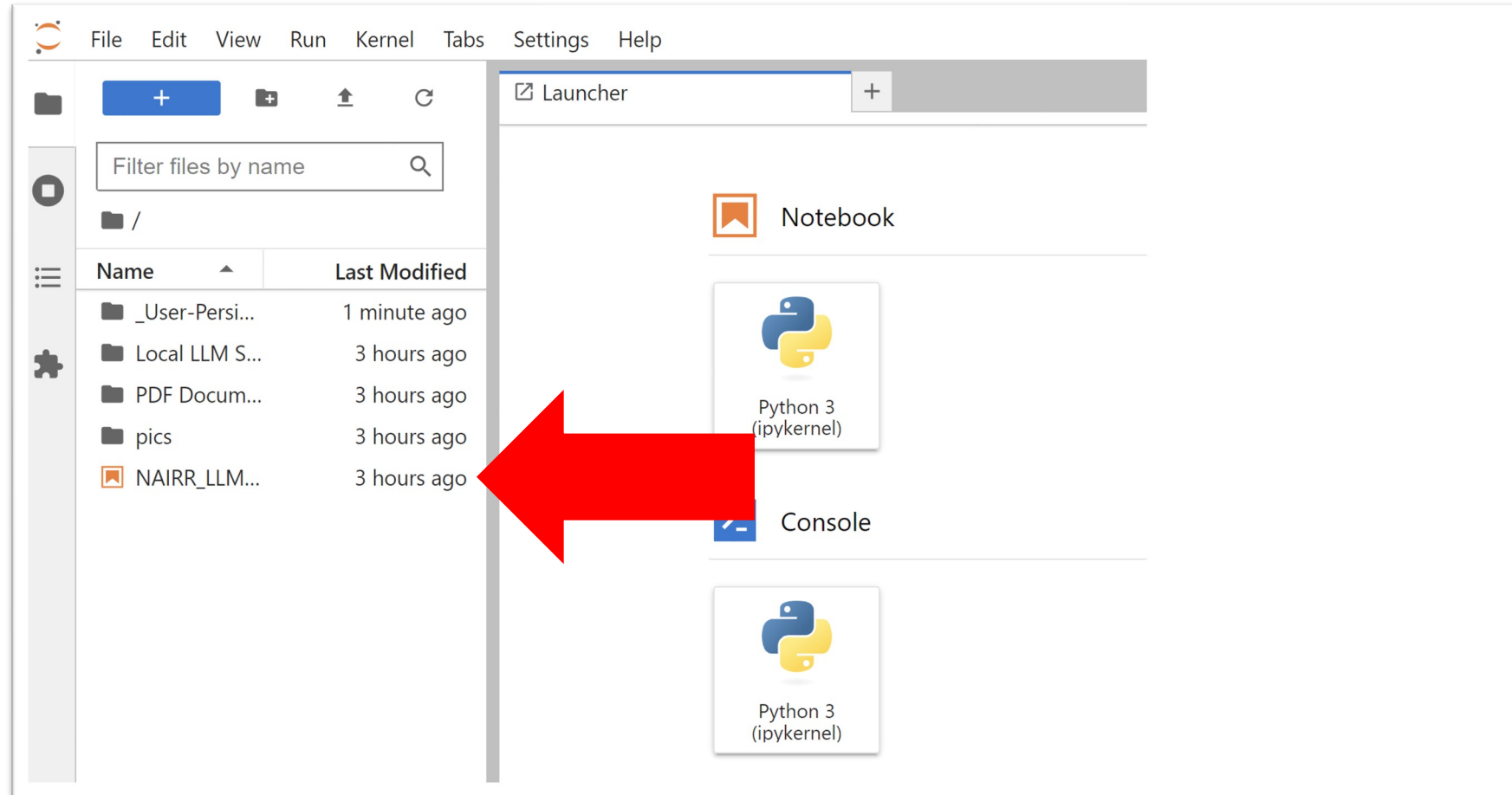


2024-03-18T19:18:20Z [Normal] AttachVolume.Attach succeeded for volume "pvc-dabaeab0-9c6c-4e2e-8f0c-d8dc3397bfcf"

Event log

It will take a couple of minutes.

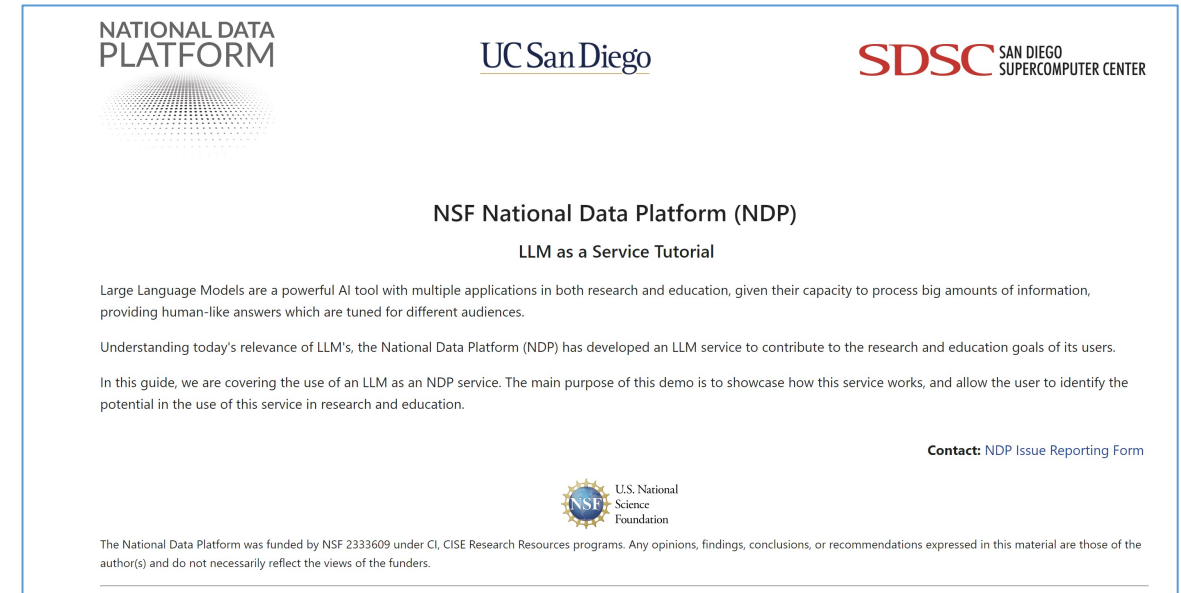
Step 9: Select NAIRR_LLM_chat.ipynb



Step 10: Hands On

Part 1: Explore Q&A using ClimateGPT

- Run cells under "A. Set Up"
 - Run using Shift+Enter on the cell
 - Functions built
 - `make_query`, `run_conversation` and `save_conversation_to_file`
- Run cells under "B. Running the conversation"
 - Use example questions or ask your own
 - **Type ENTER after each question. DO NOT TYPE SHIFT+ENTER**
 - Answer might take some time



The screenshot shows the NSF National Data Platform (NDP) LLM as a Service Tutorial page. At the top, there are logos for the National Data Platform, UC San Diego, and SDSC (San Diego Supercomputer Center). The main heading is "NSF National Data Platform (NDP) LLM as a Service Tutorial". Below this, there is introductory text about Large Language Models and the NDP's LLM service. A "Contact: NDP Issue Reporting Form" link is provided. At the bottom, there is a logo for the U.S. National Science Foundation and a disclaimer stating that the NDP was funded by NSF 2333609 under CI, CISE Research Resources programs, and that the opinions expressed are those of the author(s).

Part 2: Add New Document and Compare Q&A Results

- Run cells under "C. Adding new context to ClimateGPT"
 - Loading the document can take up to a couple of minutes
 - Run queries and compare results

3. Go to the next cell *POST*, and repeat the process. Now the model will provide an answer relying on the document.

PRE-Context

```
[6]: # PRE - In this cell, we will ask the question without giving any context to ClimateGPT
conversation()
```

Question (q=quit, s=save previous answer): What are NAIRR's goals with respect to human capital?

Answer: The North American Interfraternity Conference's (NAIRR) goals with respect to human capital are focused on promoting the growth and development of fraternity and sorority communities. Specifically, NAIRR aims to:

1. Enhance leadership development: NAIRR provides educational programs, resources, and guidance to help fraternity and sorority members develop leadership skills and become effective leaders in their communities.
2. Foster brotherhood and sisterhood: NAIRR supports the development of meaningful relationships among fraternity and sorority members through various programs and initiatives, aiming to cultivate a strong sense of brotherhood and sisterhood within these communities.
3. Promote diversity, equity, and inclusion: NAIRR recognizes the importance of creating a welcoming and inclusive environment for all members of fraternity and sorority communities. They work to support diversity, equity, and inclusion efforts in these communities.
4. Improve risk management: NAIRR provides resources and guidance to help fraternity and sorority communities manage risks effectively, ensuring a safe and positive experience for members.
5. Support the overall success of fraternity and sorority communities: NAIRR aims to help fraternity and sorority communities thrive by providing resources, advocacy, and support to member organizations.

By focusing on these goals, NAIRR works to create well-rounded, engaged, and successful fraternity and sorority communities that benefit members and contribute positively to higher education and society as a whole.

Question (q=quit, s=save previous answer): q

Looking at the responses, we can confirm the model is given proper answers which are constructed taking the added document as context.

POST-Context

```
[*]: # POST - We add True to indicate the model to make use of the new document
conversation(True)
```

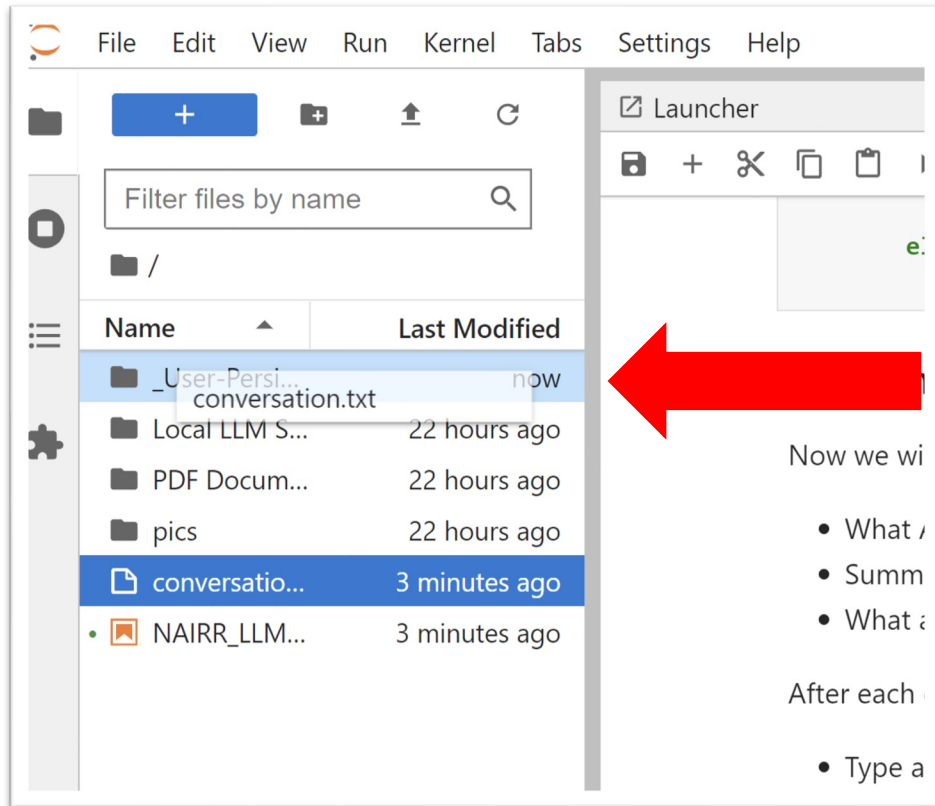
Question (q=quit, s=save previous answer): What are NAIRR's goals with respect to human capital?

Answer: The NAIRR aims to provide a comprehensive set of AI tools and resources for federal government use, and it recognizes the importance of human capital in the successful implementation and use of these tools. The Operating Entity is responsible for ensuring that the NAIRR includes appropriate training and support materials for users, and for sharing responsibility with resource providers for providing this training and support. The NAIRR also aims to incorporate lessons learned from its implementation and use to improve its governance, management, and operations over time.

Question (q=quit, s=save previous answer):

Screenshot of a pre-context and post-context response

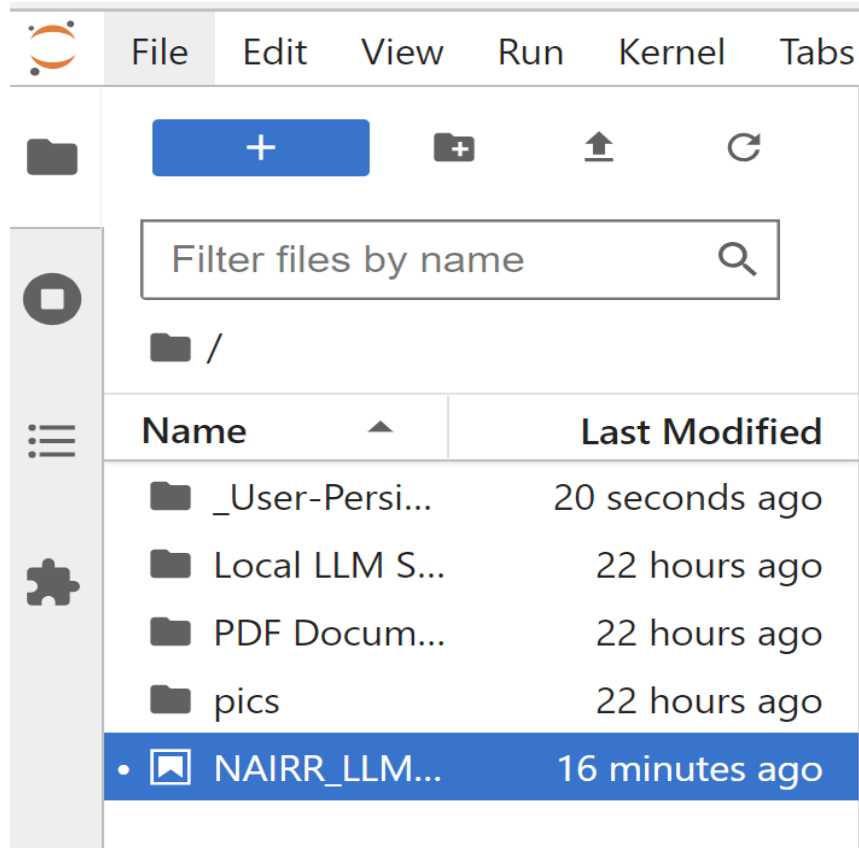
Step 11: Save your notebook and outputs



To save your notebook and or output, you can:

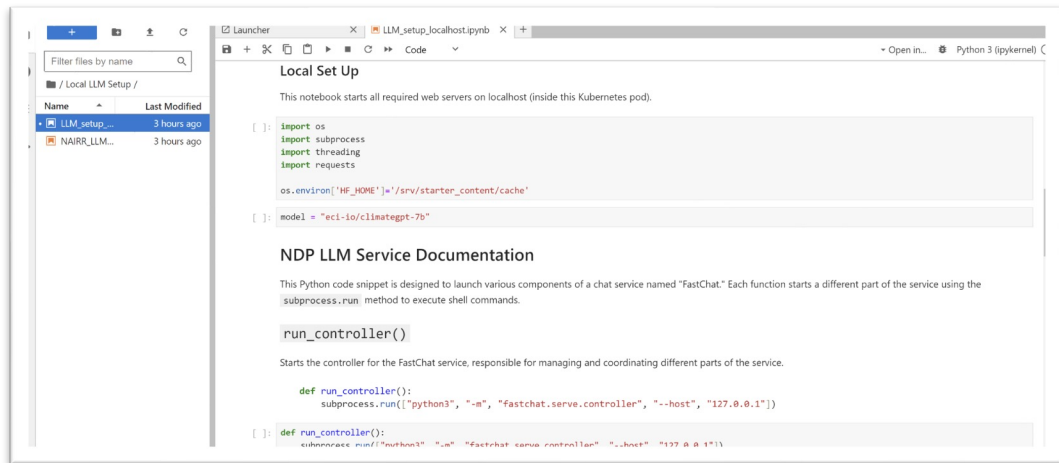
- Select the files and drag them into `_User-Persistent-Storage_`
- Copy-Paste the files into `_User-Persistent-Storage_`
- Download them locally

Local LLM: We provide the set-up code to host your own LLM within your server



To make this service work, it is necessary to reserve a server with at least one GPU instance.

Other NDP LLM Notebooks



```
Local Set Up

This notebook starts all required web servers on localhost (inside this Kubernetes pod).

[ ]: import os
import subprocess
import threading
import requests

os.environ['HF_HOME'] += '/srv/starter_content/cache'

[ ]: model = "ecl-io/climategpt-7b"

NDP LLM Service Documentation

This Python code snippet is designed to launch various components of a chat service named "FastChat." Each function starts a different part of the service using the subprocess.run method to execute shell commands.

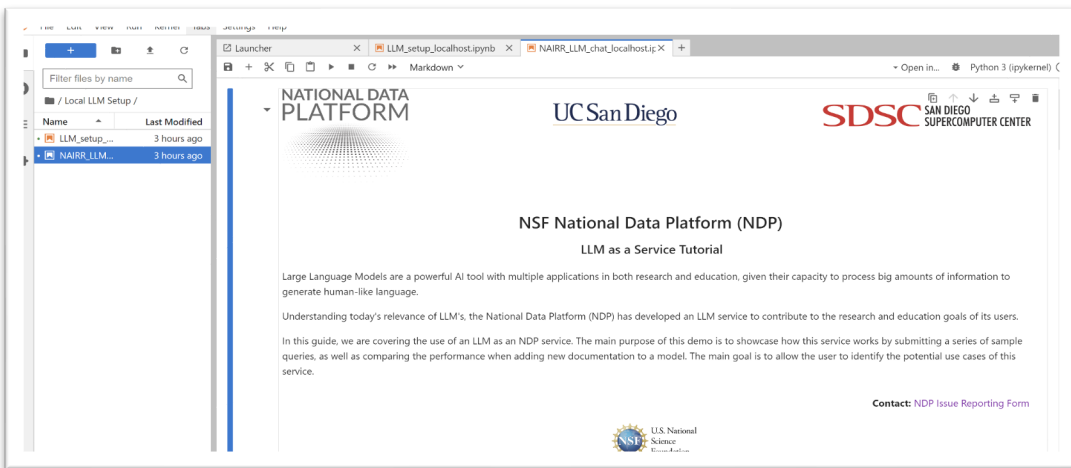
run_controller()

Starts the controller for the FastChat service, responsible for managing and coordinating different parts of the service.

def run_controller():
    subprocess.run(["python3", "-m", "fastchat.serve.controller", "--host", "127.0.0.1"])

[ ]: def run_controller():
    subprocess.run(["python3", "-m", "fastchat.serve.controller", "--host", "127.0.0.1"])
```

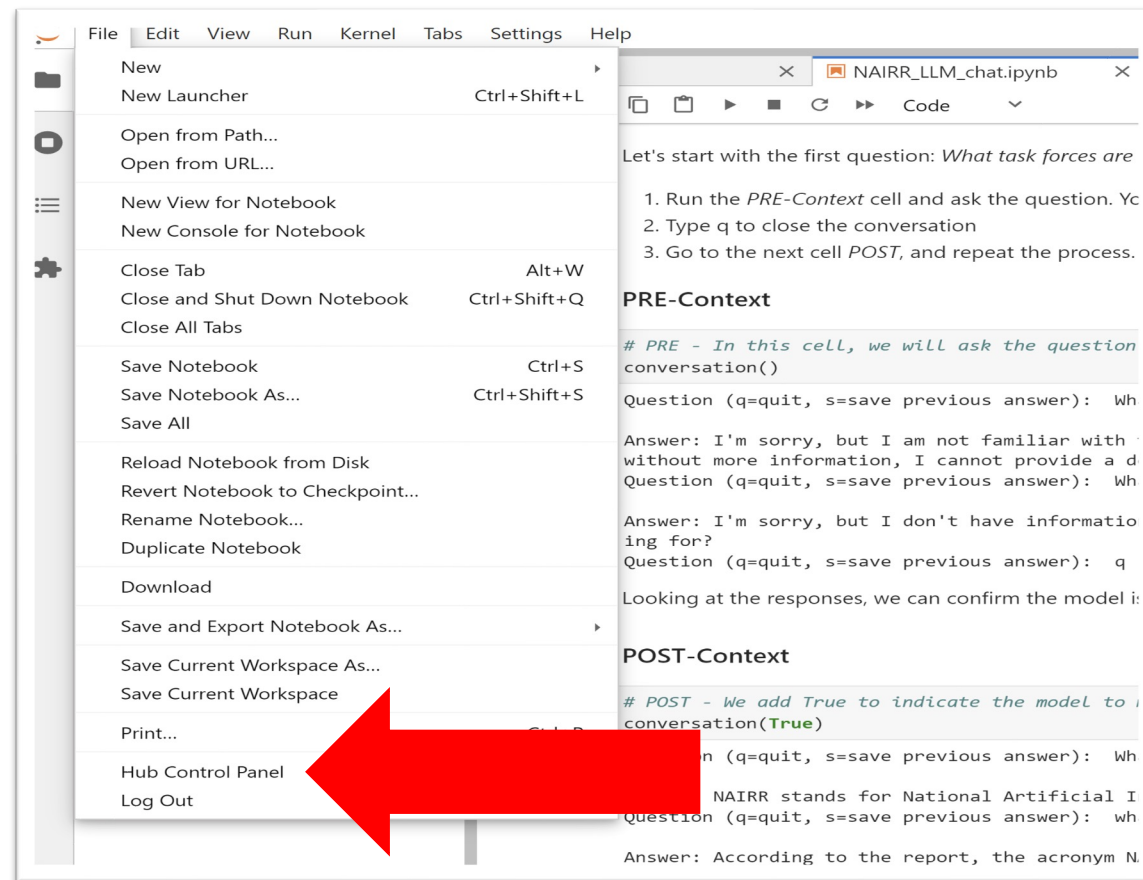
LLM_setup_localhost.ipynb: This notebook allows users to start their own host API server, and to load their model from HuggingFace into their server.



NAIRR_LLM_chat_localhost.ipynb: The code explored today, which connects to the localhost server instead of Community's LLM server. Users can connect and start interacting with their model.

Contact ndp@sdsc.edu for support

Step 12: Click on *File* and select *Hub Control Panel*



Step 13. Stop your Server

[Stop My Server](#) [My Server](#)

Named Servers

In addition to your default server, you may have additional server(s) with names. This allows you to have more than one server running at the same time.

Server name	URL	Last activity	Actions
<input type="text" value="Name your server"/>			Add New Server

How can we work with you?



Contact: Ilkay Altintas, Ph.D.

Email: ialtintas@ucsd.edu