Jupyter-AI & Open LLMs in large science classes

ESPM 157

Carl Boettiger
UC Berkeley

CC BY

- 122 students
- ~ 2/3 bio sciences
- 1/3 data science
- Active learning classroom

Photo credits Mathew Burciaga / UC Berkeley

ESPM-157: Data Science for Global Change Ecology

Motivations

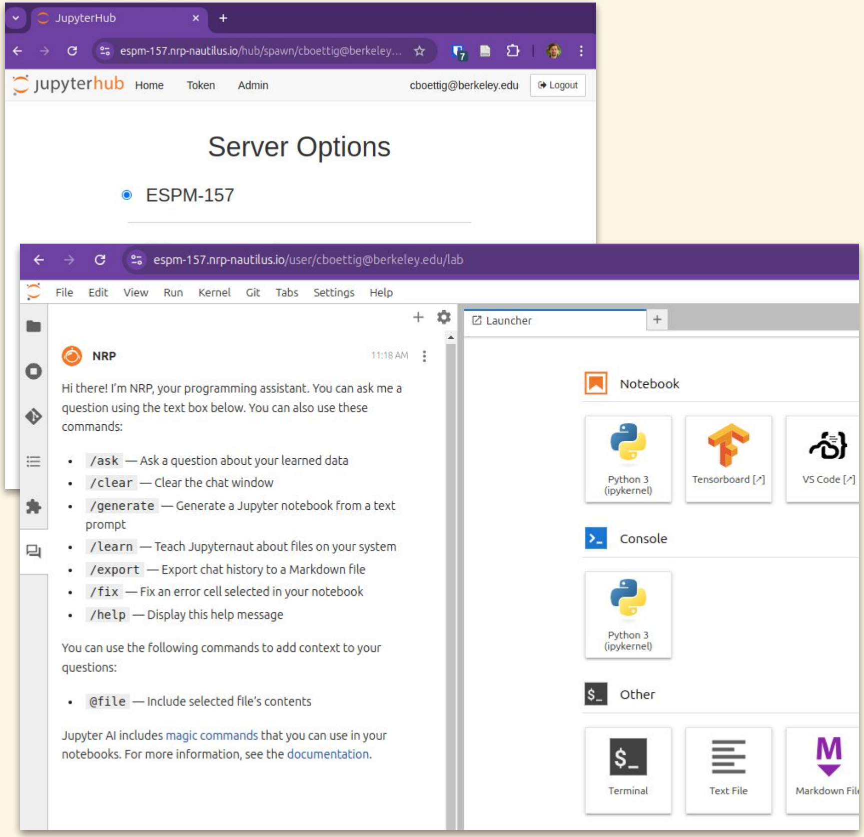
LLMs are here. Am I preparing my students for this future?

- Can we go further, faster by **coding** with LLM assistants?
- What are **environmental** implications of LLMs?
- What are the **ethical** implications?
- Can we make LLM use **safer** & more **reliable**?

Computing Resources

NRP NATIONAL RESEARCH PLATFORM

nvidia.com/gpu: “1”   x 122 students!



Server Options

● ESPM-157

NRP 11:18 AM

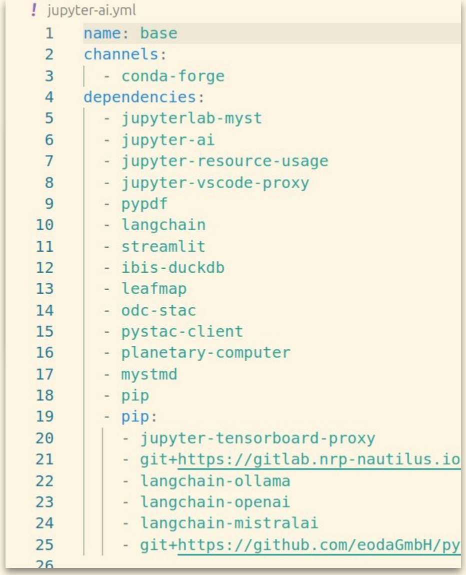
Hi there! I'm NRP, your programming assistant. You can ask me a question using the text box below. You can also use these commands:

- `/ask` — Ask a question about your learned data
- `/clear` — Clear the chat window
- `/generate` — Generate a Jupyter notebook from a text prompt
- `/learn` — Teach Jupyternaut about files on your system
- `/export` — Export chat history to a Markdown file
- `/fix` — Fix an error cell selected in your notebook
- `/help` — Display this help message

You can use the following commands to add context to your questions:

- `@file` — Include selected file's contents

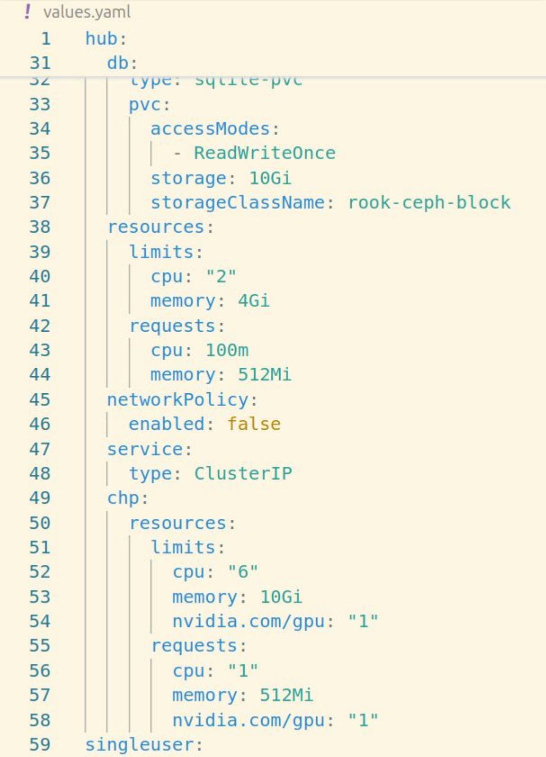
Jupyter AI includes magic commands that you can use in your notebooks. For more information, see the documentation.



! jupyter-ai.yml

```yaml
name: base
channels:
  - conda-forge
dependencies:
  - jupyterlab-myst
  - jupyter-ai
  - jupyter-resource-usage
  - jupyter-vscode-proxy
  - pypdf
  - langchain
  - streamlit
  - ibis-duckdb
  - leafmap
  - odc-stac
  - pystac-client
  - planetary-computer
  - mystmd
  - pip
  - pip:
    - jupyter-tensorboard-proxy
    - git+https://gitlab.nrp-nautilus.io
    - langchain-ollama
    - langchain-openai
    - langchain-mistralai
    - git+https://github.com/eodaGmbH/py
```

Environment.yml
Docker image



! values.yaml

```yaml
hub:
  db:
    type: sqlite-pvc
    pvc:
      accessModes:
        - ReadWriteOnce
      storage: 10Gi
      storageClassName: rook-ceph-block
  resources:
    limits:
      cpu: "2"
      memory: 4Gi
    requests:
      cpu: 100m
      memory: 512Mi
  networkPolicy:
    enabled: false
  service:
    type: ClusterIP
  chp:
    resources:
      limits:
        cpu: "6"
        memory: 10Gi
        nvidia.com/gpu: "1"
      requests:
        cpu: "1"
        memory: 512Mi
        nvidia.com/gpu: "1"
singleuser:
```

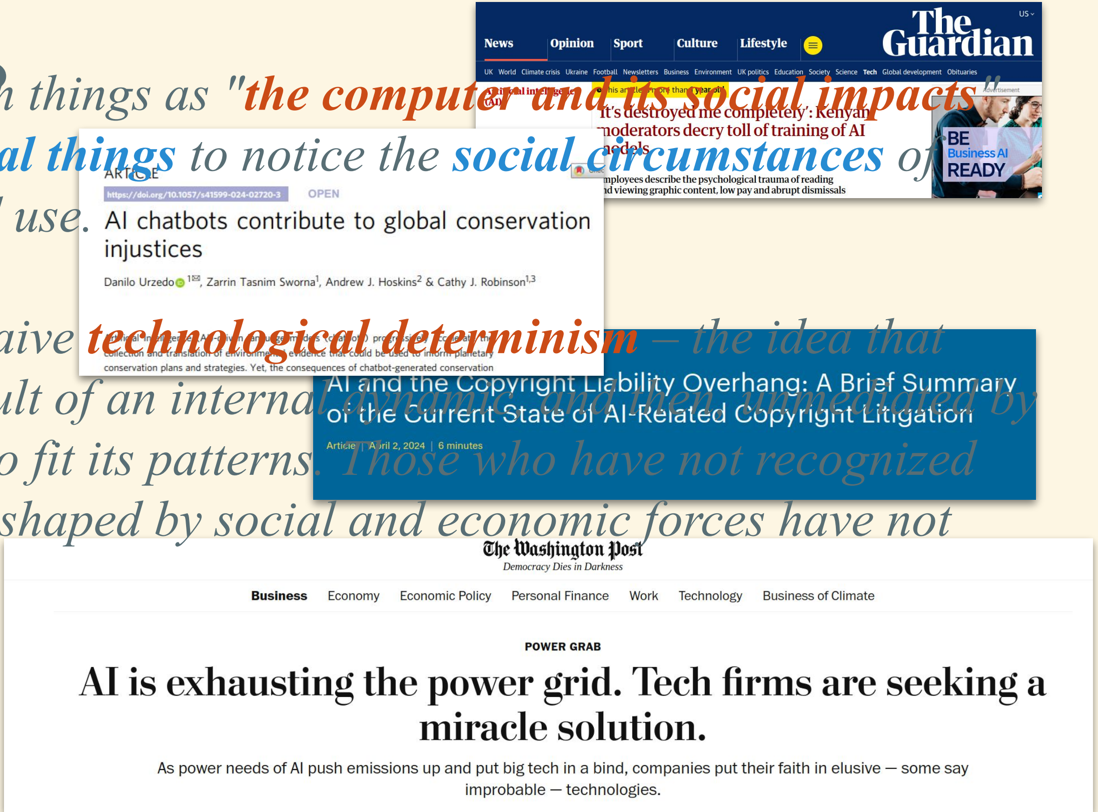
k8s Helm chart

Challenges with respect to AI

- Student data privacy?
- Energy footprint?
- Ethical concerns:
  - copyright,
  - labor practices,
  - biases

*those who focus uncritically on such things as "**the computer and its social impacts**" but who fail to **look behind technical things** to notice the **social circumstances** of their development, deployment, and use. This view provides an **antidote** to naive **technological determinism** – the idea that technology develops as the sole result of an internal dynamic and then, unmediated by any other influence, molds society to fit its patterns. Those who have not recognized the ways in which technologies are shaped by social and economic forces have not gotten very far.*

- *Langdon Winner, 1980*



News
Opinion
Sport
Culture
Lifestyle
The Guardian
US
UK World Climate crisis Ukraine Football Newsletters Business Environment UK politics Education Society Science Tech Global development Obituaries
Advertisement
BE Business AI READY

Artificial intelligence (AI)

This article is more than 1 year old

'It's destroyed me completely': Kenyan moderators decry toll of training of AI models

mployees describe the psychological trauma of reading nd viewing graphic content, low pay and abrupt dismissals

ARTICLE

https://doi.org/10.1057/s41599-024-02720-3 OPEN

AI chatbots contribute to global conservation injustices

Danilo Urzedo[1✉], Zarrin Tasnim Sworna[1], Andrew J. Hoskins[2] & Cathy J. Robinson[1,3]

Artificial intelligence (AI)-driven language models (chatbots) progressively accelerate the collection and translation of environmental evidence that could be used to inform planetary conservation plans and strategies. Yet, the consequences of chatbot-generated conservation

AI and the Copyright Liability Overhang: A Brief Summary of the Current State of AI-Related Copyright Litigation

Article | April 2, 2024 | 6 minutes

The Washington Post

Democracy Dies in Darkness

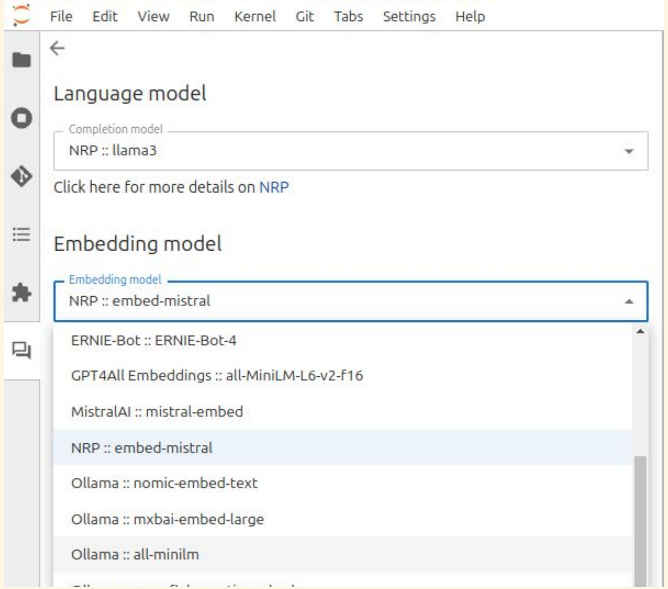
POWER GRAB

AI is exhausting the power grid. Tech firms are seeking a miracle solution.

As power needs of AI push emissions up and put big tech in a bind, companies put their faith in elusive — some say improbable — technologies.

Local, open models:

```
$ ollama serve
$ ollama pull nomic-embed-text
```


File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help

Language model

Completion model: NRP :: llama3

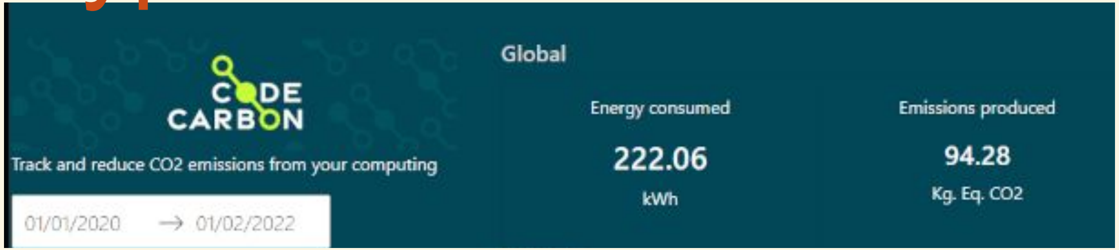
Click here for more details on NRP

Embedding model

Embedding model: NRP :: embed-mistral

ERNIE-Bot :: ERNIE-Bot-4
GPT4All Embeddings :: all-MiniLM-L6-v2-f16
MistralAI :: mistral-embed
NRP :: embed-mistral
Ollama :: nomic-embed-text
Ollama :: mxbai-embed-large
Ollama :: all-minilm

CO2 footprint data, not hype:



Track and reduce CO2 emissions from your computing

01/01/2020 → 01/02/2022

Global

Energy consumed

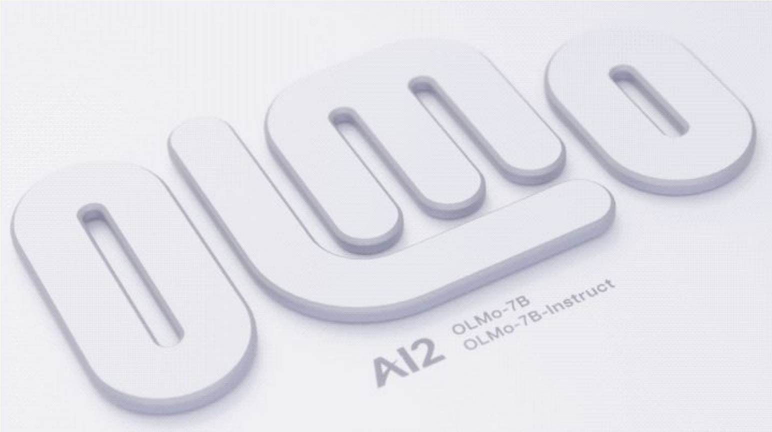
222.06

kWh

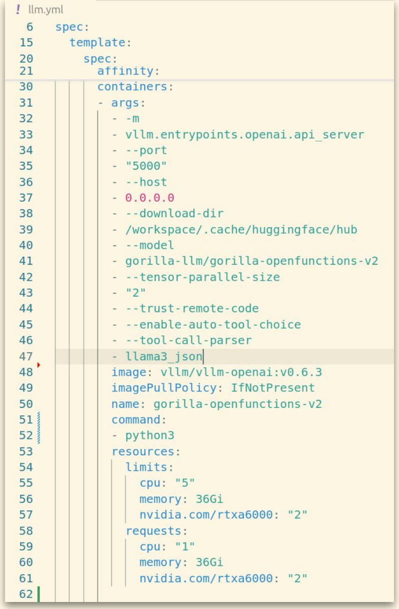
Emissions produced

94.28

Kg. Eq. CO2

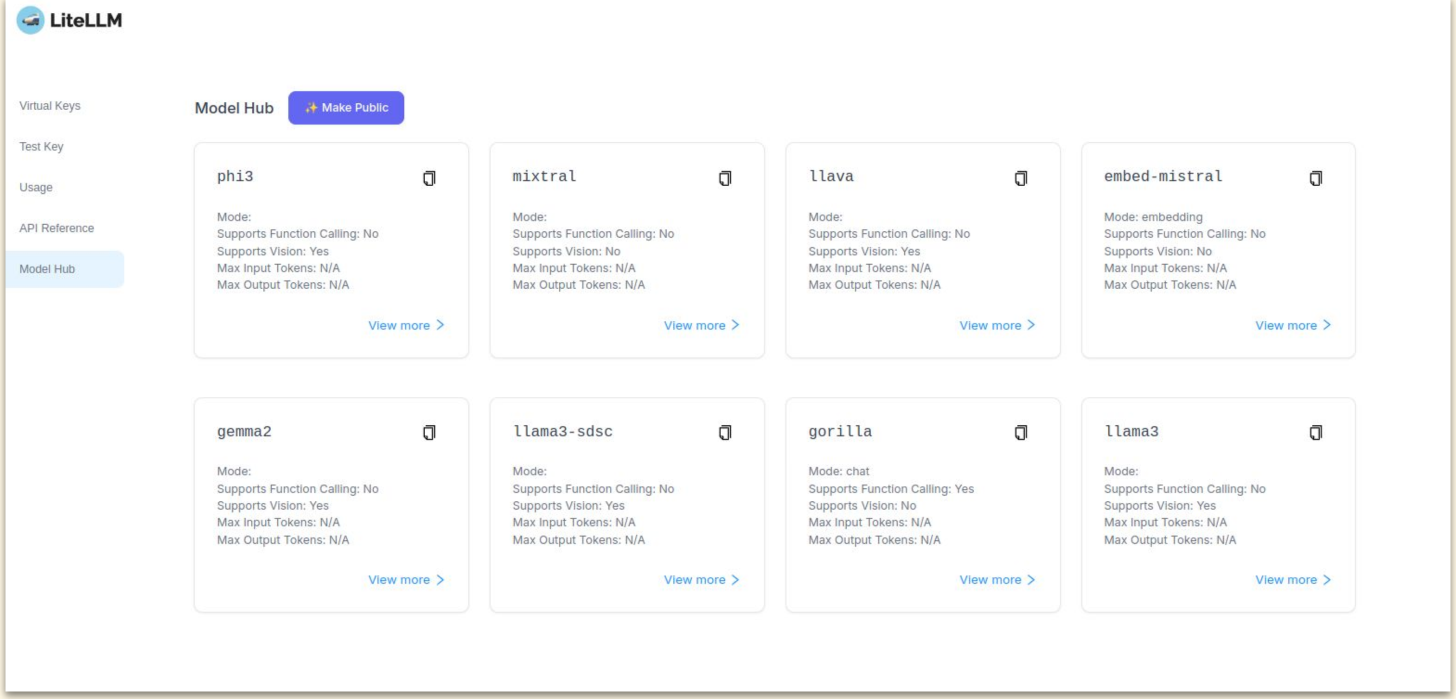


Researcher developed + hosted



llm.yml

```yaml
spec:
  template:
    spec:
      affinity:
      containers:
      - args:
        - -m
        - vllm.entrypoints.openai.api_server
        - --port
        - "5000"
        - --host
        - 0.0.0.0
        - --download-dir
        - /workspace/.cache/huggingface/hub
        - --model
        - gorilla-llm/gorilla-openfunctions-v2
        - --tensor-parallel-size
        - "2"
        - --trust-remote-code
        - --enable-auto-tool-choice
        - --tool-call-parser
        - llama3_json
        image: vllm/vllm-openai:v0.6.3
        imagePullPolicy: IfNotPresent
        name: gorilla-openfunctions-v2
        command:
        - python3
        resources:
          limits:
            cpu: "5"
            memory: 36Gi
            nvidia.com/rtxa6000: "2"
          requests:
            cpu: "1"
            memory: 36Gi
            nvidia.com/rtxa6000: "2"
```



phi3
Mode:
Supports Function Calling: No
Supports Vision: Yes
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

mixtral
Mode:
Supports Function Calling: No
Supports Vision: No
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

llava
Mode:
Supports Function Calling: No
Supports Vision: Yes
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

embed-mistral
Mode: embedding
Supports Function Calling: No
Supports Vision: No
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

gemma2
Mode:
Supports Function Calling: No
Supports Vision: Yes
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

llama3-sdsc
Mode:
Supports Function Calling: No
Supports Vision: Yes
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

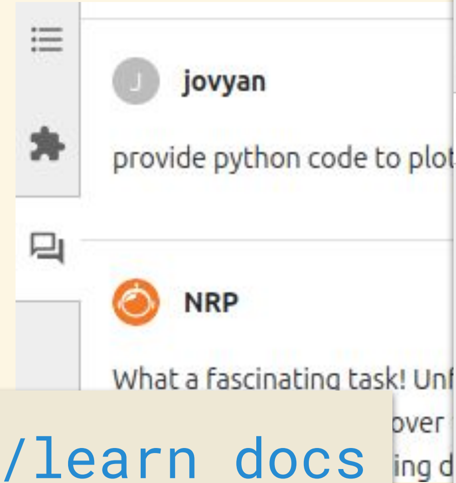
gorilla
Mode: chat
Supports Function Calling: Yes
Supports Vision: No
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

llama3
Mode:
Supports Function Calling: No
Supports Vision: Yes
Max Input Tokens: N/A
Max Output Tokens: N/A
View more >

```
> base_url = “https://llm.nrp-nautilus.io/”
```

Challenges with respect to AI

- LLMs **don't**



jovyan

provide python code to plot

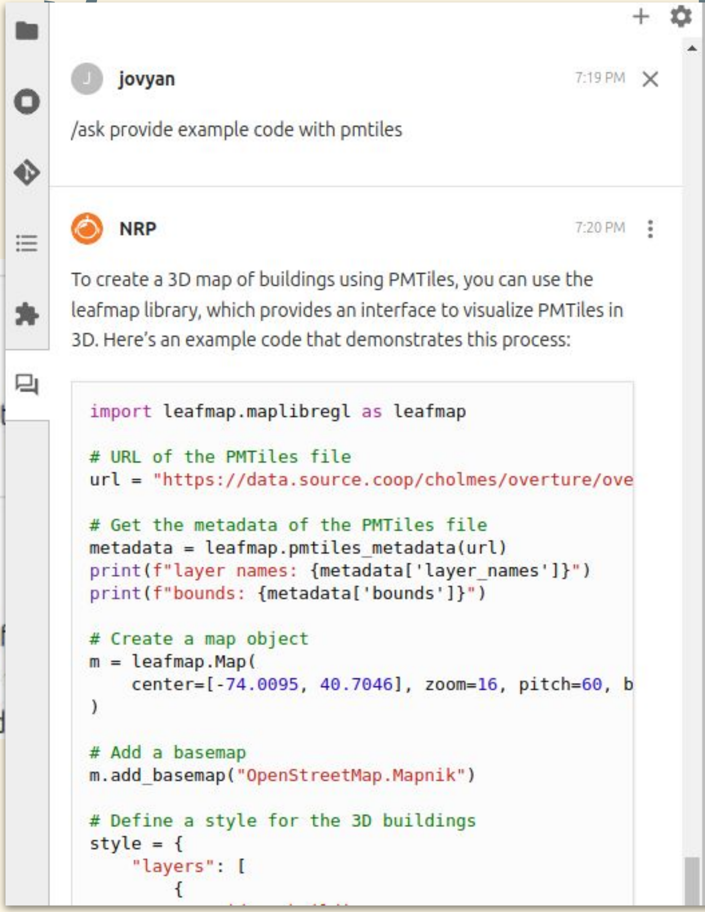
NRP

What a fascinating task! Unf

over

ing d

```
/learn docs
/ask …
```



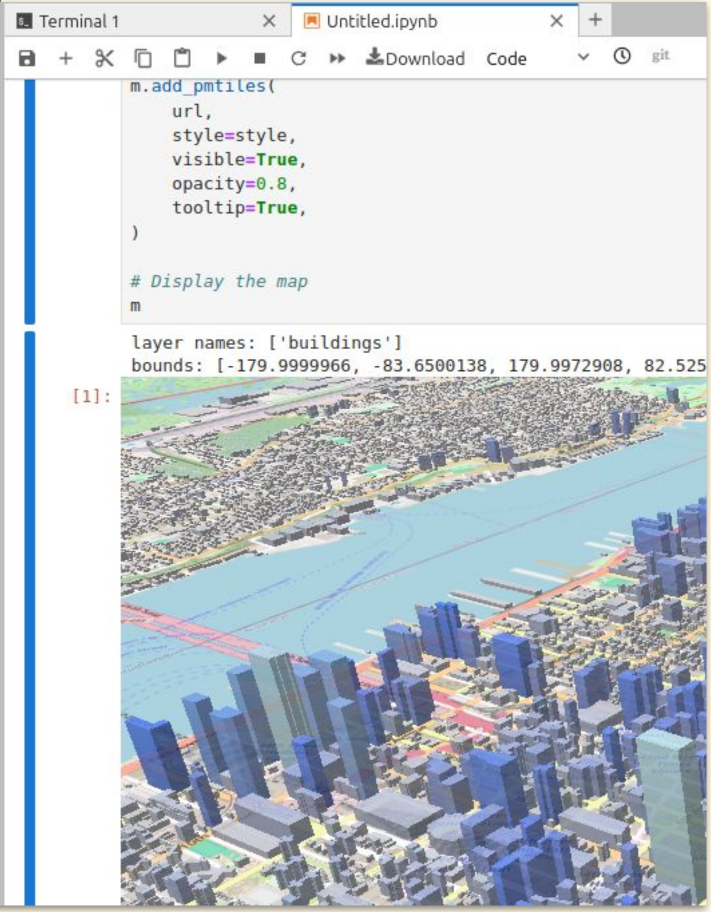
jovyan 7:19 PM

/ask provide example code with pmtiles

NRP 7:20 PM

To create a 3D map of buildings using PMTiles, you can use the leafmap library, which provides an interface to visualize PMTiles in 3D. Here's an example code that demonstrates this process:

```
import leafmap.maplibregl as leafmap

# URL of the PMTiles file
url = "https://data.source.coop/cholmes/overture/ove

# Get the metadata of the PMTiles file
metadata = leafmap.pmtiles_metadata(url)
print(f"layer names: {metadata['layer_names']}")
print(f"bounds: {metadata['bounds']}")

# Create a map object
m = leafmap.Map(
    center=[-74.0095, 40.7046], zoom=16, pitch=60, b
)

# Add a basemap
m.add_basemap("OpenStreetMap.Mapnik")

# Define a style for the 3D buildings
style = {
    "layers": [
        {
```

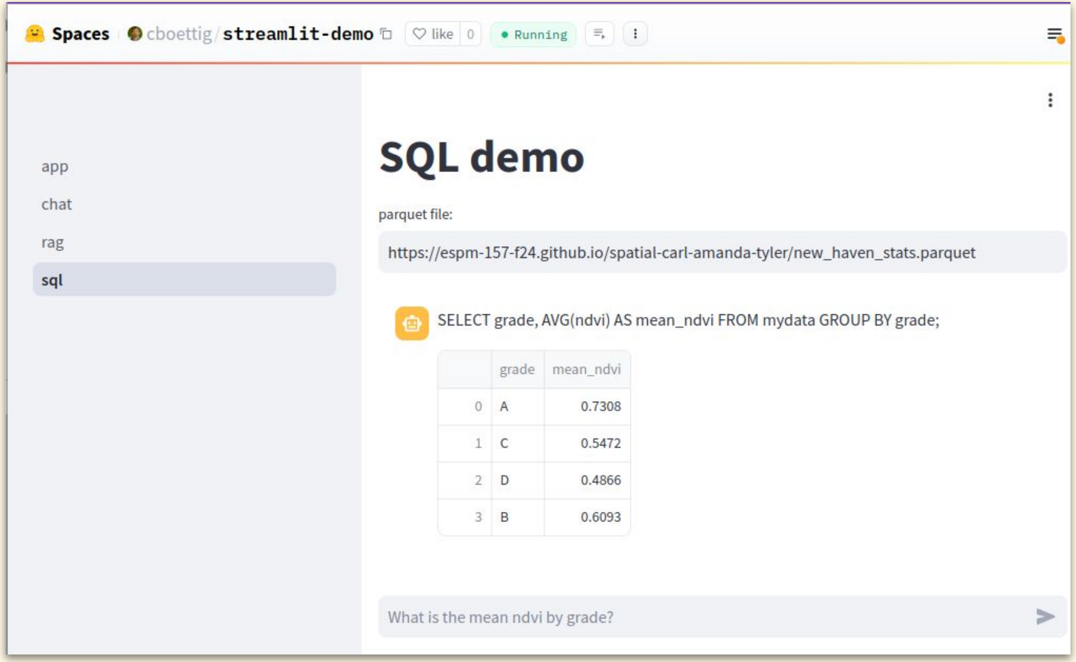


```
m.add_pmtiles(
    url,
    style=style,
    visible=True,
    opacity=0.8,
    tooltip=True,
)

# Display the map
m
```

```
layer names: ['buildings']
bounds: [-179.9999966, -83.6500138, 179.9972908, 82.525
```

RAG & TAG – Building LLM Agents

```
st.title("SQL demo")

## dockerized streamlit app wants to read from os.
api_key = os.getenv("LITELLM_KEY")
if api_key is None:
    api_key = st.secrets["LITELLM_KEY"]

parquet = st.text_input("parquet file", "https:/

eng = sqlalchemy.create_engine("duckdb:///:memory:
con = ibis.duckdb.from_connection(eng.raw_connecti
tbl = con.read_parquet(parquet, "mydata")

# langchain can also talk to this connection and s
db = SQLDatabase(eng, view_support=True)

# Build the template for system prompt
template = '''
You are a {dialect} expert. Given an input questio
Always return all columns from a query (select *)
Wrap each column name in double quotes (") to deno
Pay attention to use only the column names you can
Be careful to not query for columns that do not ex
Also, pay attention to which column is in which ta
Pay attention to use today() function to get the c
Respond with only the SQL query to run.  Do not re
Only use the following tables:
{table_info}
Question: {input}
'''

from langchain_core.prompts import PromptTemplate
```



SQL demo

parquet file:

https://espm-157-f24.github.io/spatial-carl-amanda-tyler/new_haven_stats.parquet

SELECT grade, AVG(ndvi) AS mean_ndvi FROM mydata GROUP BY grade;

	grade	mean_ndvi
0	A	0.7308
1	C	0.5472
2	D	0.4866
3	B	0.6093

What is the mean ndvi by grade?

Project Contributions & Future engagement

- Teach **safe LLM** use, not **AI abstinence**
- **Open models** mitigate some risks (energy, privacy)
- NRP as platform for collaborative innovation!
  - **Shared** hardware, software, models