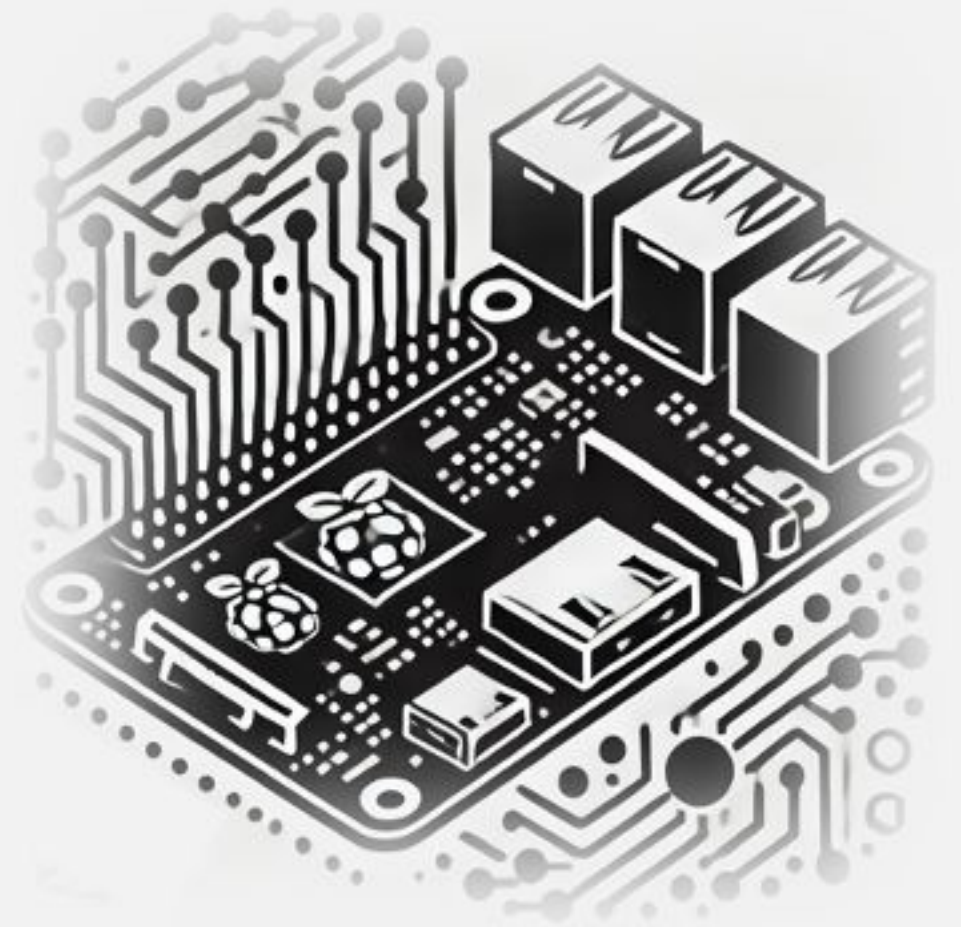


IESTI05 – Edge AI

Machine Learning System Engineering

18. Ollama Python Library



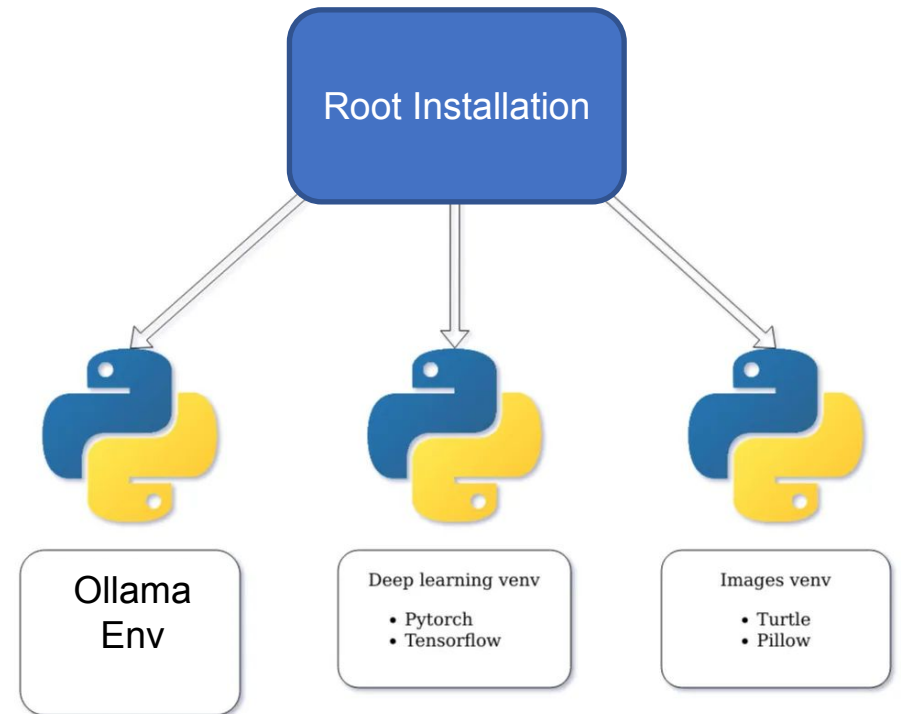
Activating the Virtual Environment

Activate the environment:

```
source ~/ollama/bin/activate
```

To **exit** the virtual environment, use:

```
deactivate
```

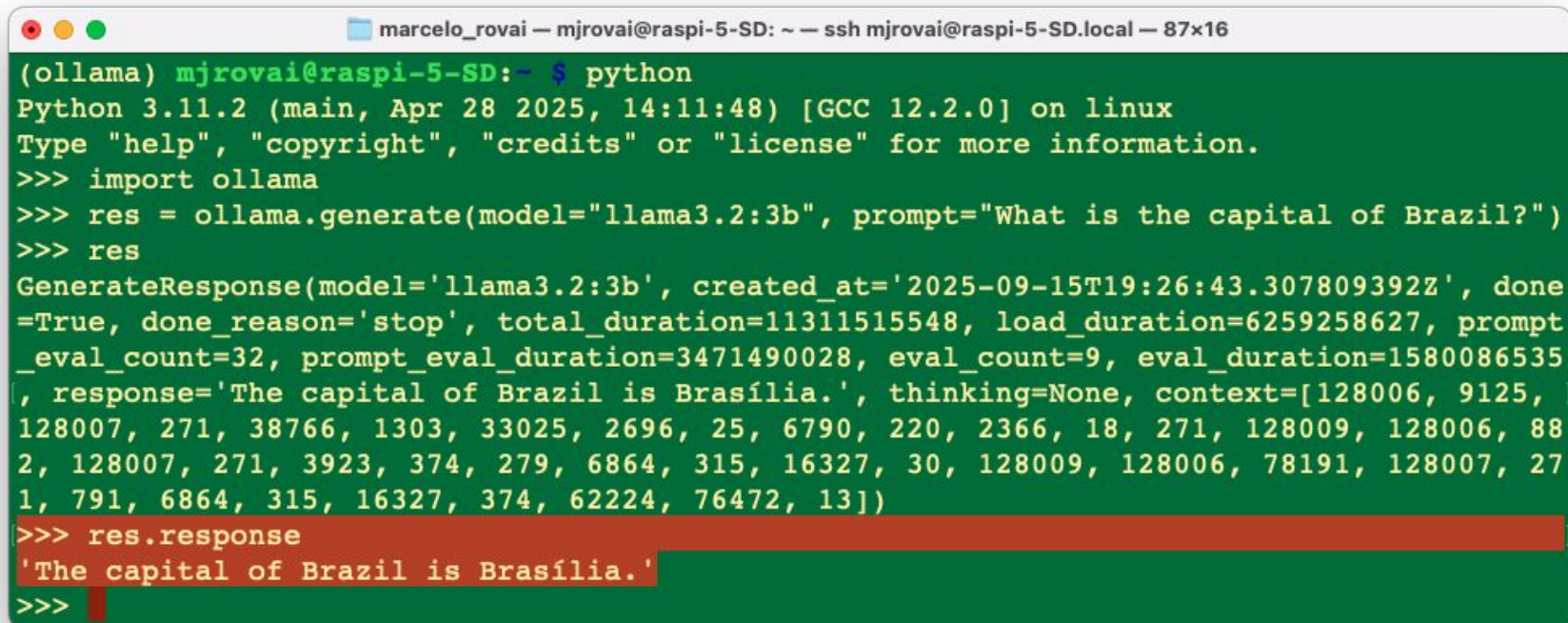


Ollama Python Library

The [Ollama Python library](#) simplifies interaction with advanced LLM models, enabling more sophisticated responses and capabilities, besides providing the easiest way to integrate Python 3.8+ projects with [Ollama](#).

==> Inside the Python env, run: `pip install ollama`

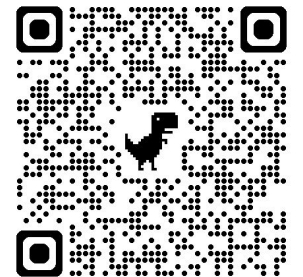
==> Test it using the Python interpreter:

A terminal window with a dark green background and yellow text. The window title bar shows 'marcelo_rovai — mjrovai@raspi-5-SD: ~ — ssh mjrovai@raspi-5-SD.local — 87x16'. The terminal shows a Python 3.11.2 prompt where the user imports 'ollama' and calls 'ollama.generate' with model 'llama3.2:3b' and prompt 'What is the capital of Brazil?'. The output is a 'GenerateResponse' object. The user then prints 'res.response', which outputs 'The capital of Brazil is Brasília.'.

```
(ollama) mjrovai@raspi-5-SD:~ $ python
Python 3.11.2 (main, Apr 28 2025, 14:11:48) [GCC 12.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import ollama
>>> res = ollama.generate(model="llama3.2:3b", prompt="What is the capital of Brazil?")
>>> res
GenerateResponse(model='llama3.2:3b', created_at='2025-09-15T19:26:43.307809392Z', done=True, done_reason='stop', total_duration=11311515548, load_duration=6259258627, prompt_eval_count=32, prompt_eval_duration=3471490028, eval_count=9, eval_duration=1580086535, response='The capital of Brazil is Brasília.', thinking=None, context=[128006, 9125, 128007, 271, 38766, 1303, 33025, 2696, 25, 6790, 220, 2366, 18, 271, 128009, 128006, 882, 128007, 271, 3923, 374, 279, 6864, 315, 16327, 30, 128009, 128006, 78191, 128007, 271, 791, 6864, 315, 16327, 374, 62224, 76472, 13])
>>> res.response
The capital of Brazil is Brasília.
>>>
```

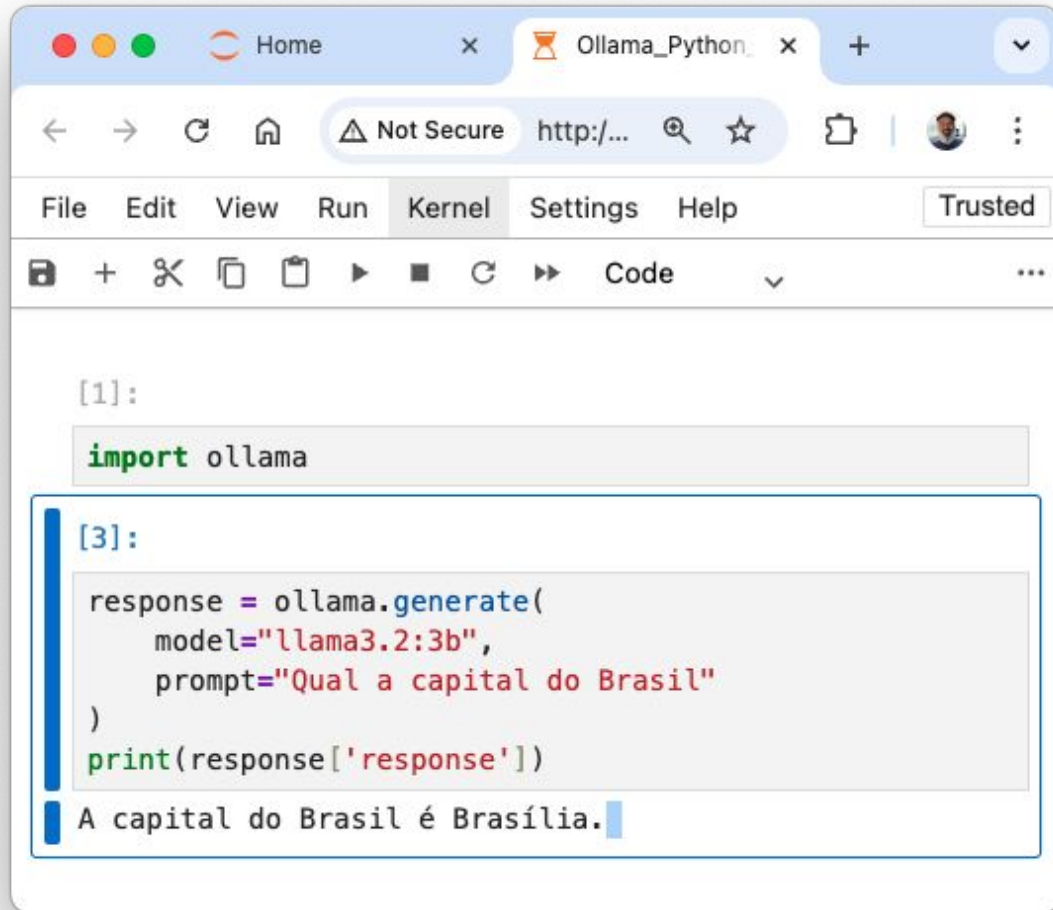
Testing SLMs with Python:

Ollama_Python_Intro.ipynb

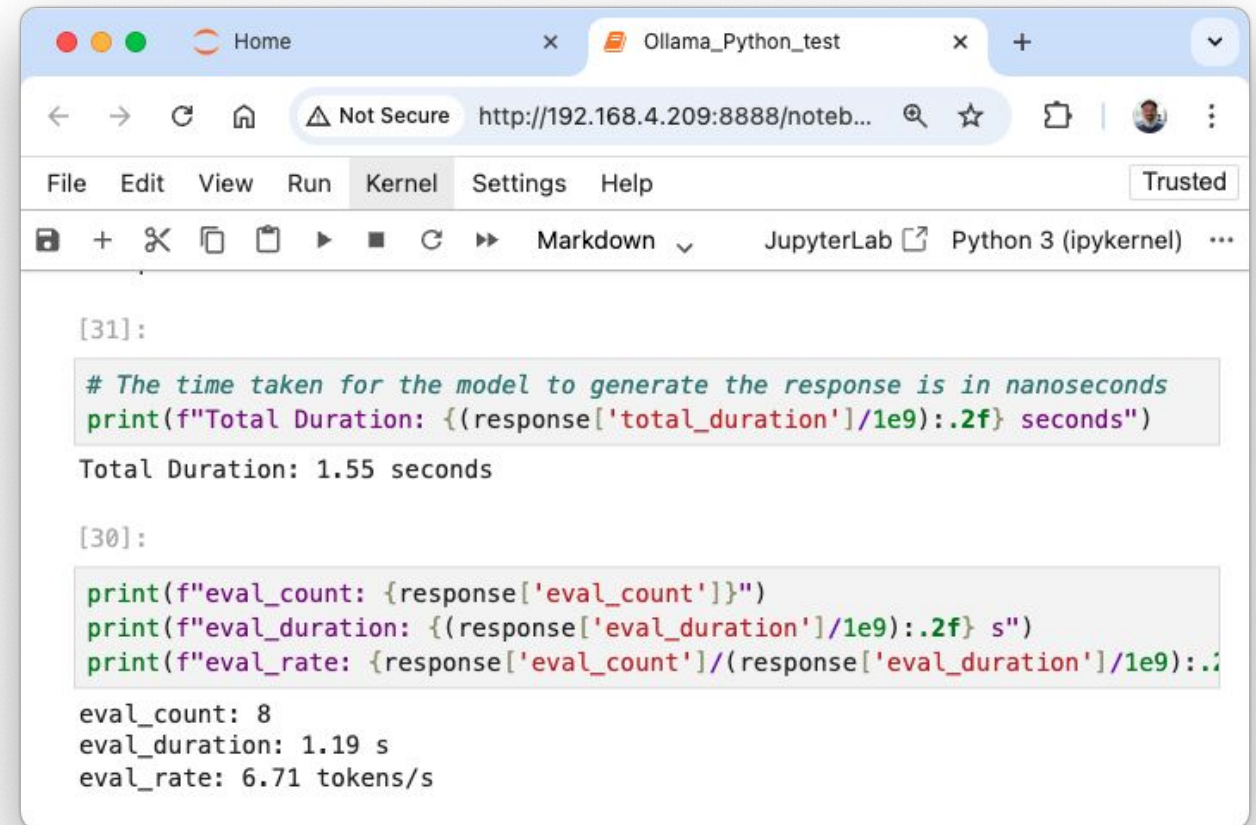


Ollama Python Library on Jupyter Notebook

```
jupyter notebook --ip=192.168.4.209 --no-browser
```



```
[1]:  
import ollama  
  
[3]:  
response = ollama.generate(  
    model="llama3.2:3b",  
    prompt="Qual a capital do Brasil"  
)  
print(response['response'])  
A capital do Brasil é Brasília.
```



```
[31]:  
# The time taken for the model to generate the response is in nanoseconds  
print(f"Total Duration: {(response['total_duration']/1e9):.2f} seconds")  
Total Duration: 1.55 seconds  
  
[30]:  
print(f"eval_count: {response['eval_count']}")  
print(f"eval_duration: {(response['eval_duration']/1e9):.2f} s")  
print(f"eval_rate: {response['eval_count']/(response['eval_duration']/1e9):.2f} tokens/s")  
eval_count: 8  
eval_duration: 1.19 s  
eval_rate: 6.71 tokens/s
```

Ollama Chat

Do not
Work!



```
[34]: response = simple_query ("Qual a capital do Brasil?")
      print(f"{response['response']}")
      A capital do Brasil é Brasília.

[35]: response = simple_query ("E a do Perú?")
      print(f"{response['response']}")
      Parece que você está falando em espanhol!

      Sim, "e la" é uma expressão comum em castelhano para dizer "e ela".
```

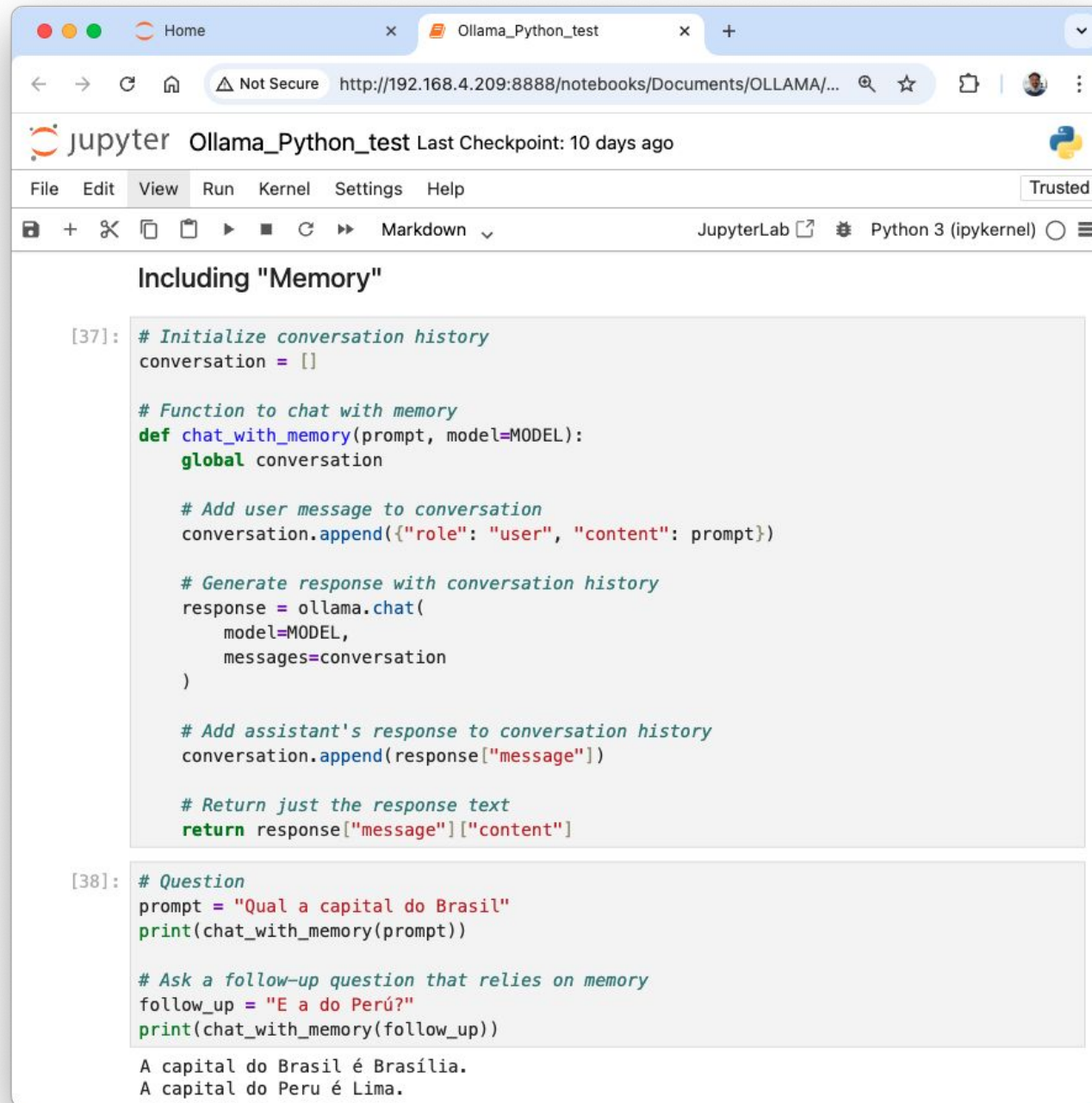
Work!



```
(ollama) mjrovai@raspi-5-SD:~ $ ollama run llama3.2:3b
>>> Qual a capital do Brasil?
A capital do Brasil é Brasília.

>>> E a do Perú?
A capital do Peru é Lima.

>>> Send a message (/? for help)
```



The screenshot shows a JupyterLab interface in a web browser. The browser's address bar shows a local URL: `http://192.168.4.209:8888/notebooks/Documents/OLLAMA/...`. The JupyterLab header includes the title 'Ollama_Python_test' and a 'Last Checkpoint: 10 days ago' message. The main menu bar contains 'File', 'Edit', 'View', 'Run', 'Kernel', 'Settings', and 'Help'. Below the menu is a toolbar with icons for file operations and a 'Markdown' dropdown. On the right side of the toolbar, it says 'JupyterLab', 'Python 3 (ipykernel)', and a 'Trusted' badge.

The notebook content is titled 'Including "Memory"'. It contains two code cells. The first cell, labeled '[37]:', defines a function `chat_with_memory` that maintains a conversation history. The second cell, labeled '[38]:', uses this function to answer two questions: 'Qual a capital do Brasil' and 'E a do Perú?'. The output of the second cell shows the answers: 'A capital do Brasil é Brasília.' and 'A capital do Peru é Lima.'

```
[37]: # Initialize conversation history
conversation = []

# Function to chat with memory
def chat_with_memory(prompt, model=MODEL):
    global conversation

    # Add user message to conversation
    conversation.append({"role": "user", "content": prompt})

    # Generate response with conversation history
    response = ollama.chat(
        model=MODEL,
        messages=conversation
    )

    # Add assistant's response to conversation history
    conversation.append(response["message"])

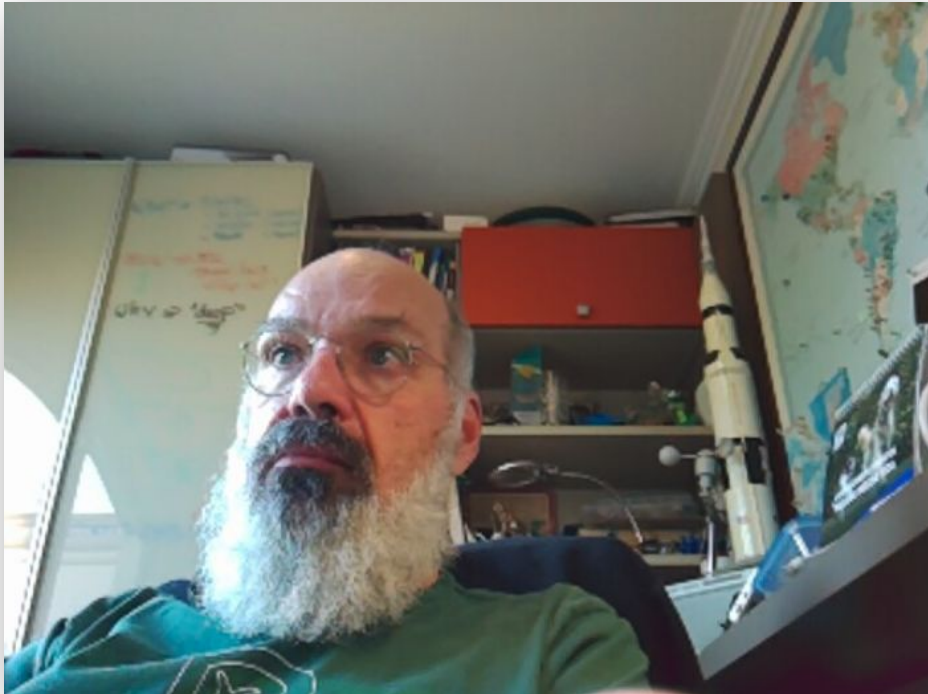
    # Return just the response text
    return response["message"]["content"]

[38]: # Question
prompt = "Qual a capital do Brasil"
print(chat_with_memory(prompt))

# Ask a follow-up question that relies on memory
follow_up = "E a do Perú?"
print(chat_with_memory(follow_up))

A capital do Brasil é Brasília.
A capital do Peru é Lima.
```

Image Description



```
Home x Ollama_Python_test x +
Not Secure http://192.168.4.209:8888/notebooks/Documents/OLLAMA/OI...
jupyter Ollama_Python_test Last Checkpoint: 10 days ago
File Edit View Run Kernel Settings Help Trusted
JupyterLab Python 3 (ipykernel)

Image Description

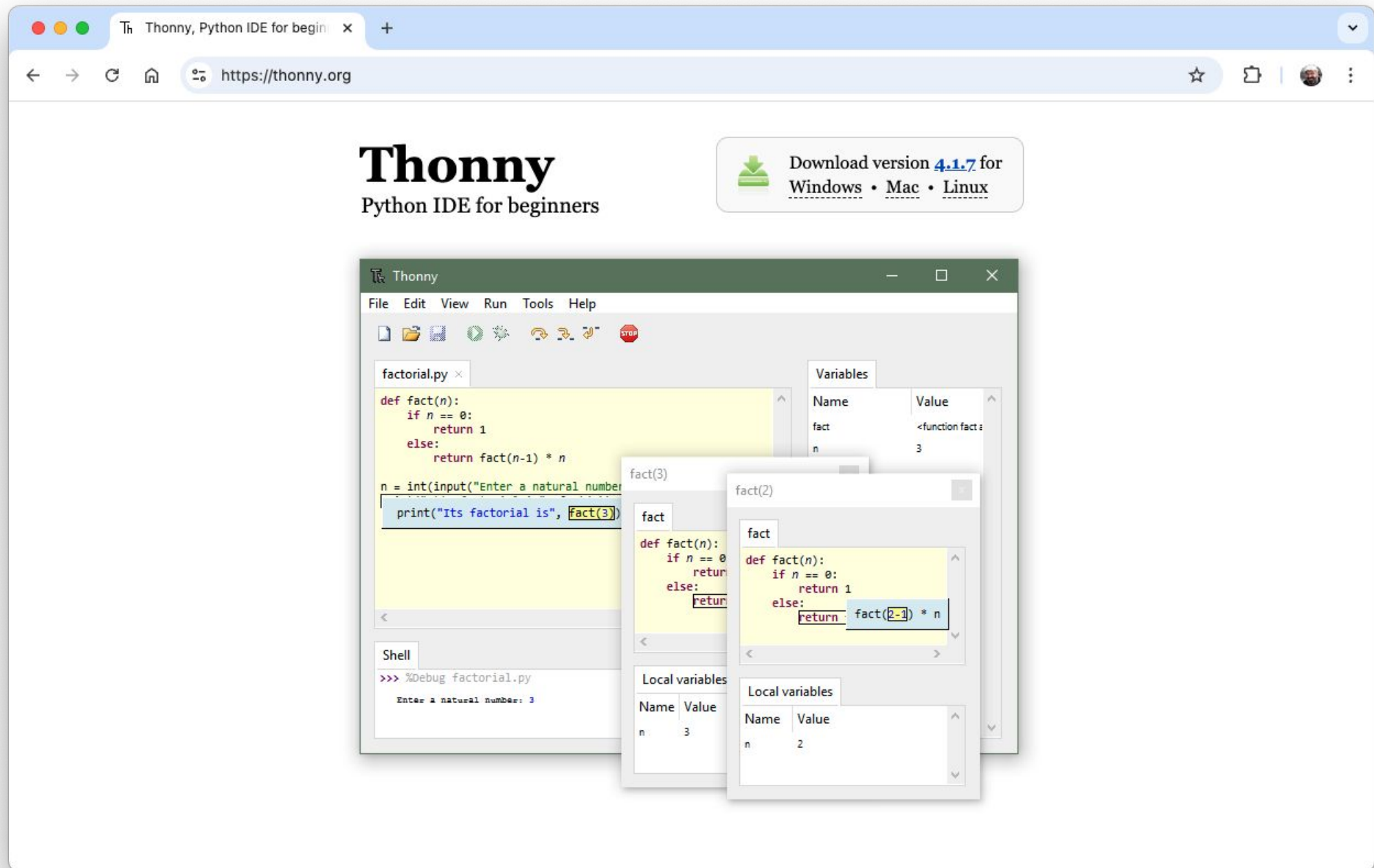
[4]: img_path = "/home/mjrovai/Documents/OLLAMA/capt_image.jpg"
MODEL = "moondream:latest"

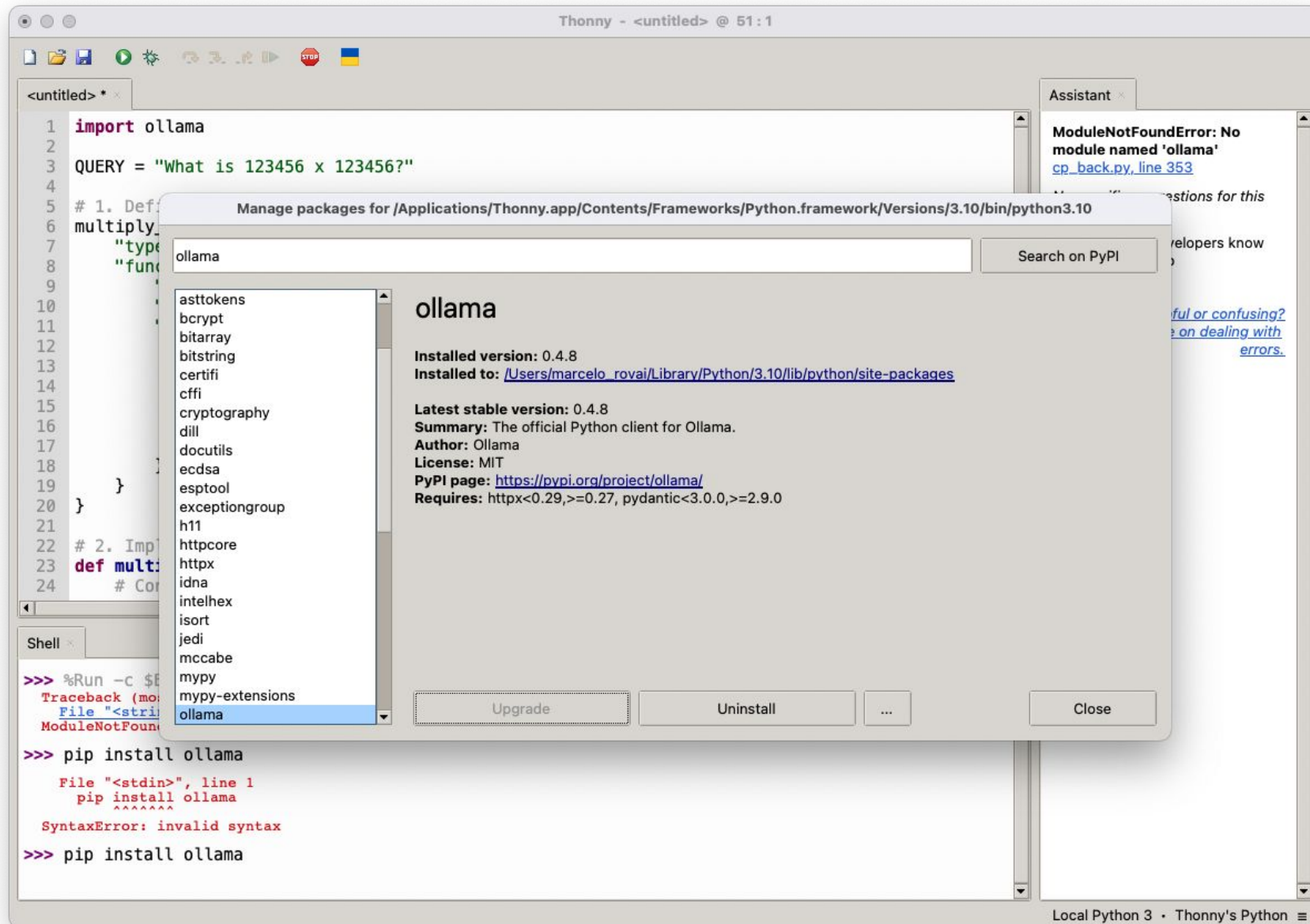
[6]: def image_description(img_path):
    with open(img_path, 'rb') as file:
        response = ollama.chat(
            model=MODEL,
            messages=[
                {
                    'role': 'user',
                    'content': "'return the description of the image'",
                    'images': [file.read()],
                },
            ],
            options = {
                'temperature': 0,
            }
        )
    #print(response)
    return response['message']['content']

[7]: response = image_description(img_path)
print(response)

The image features a bearded man wearing glasses sitting in an office chair. He is looking up at something, possibly a map or a picture on the wall behind him. The room has various objects scattered around, including several books and a clock on the wall near the center-left side of the scene.
```

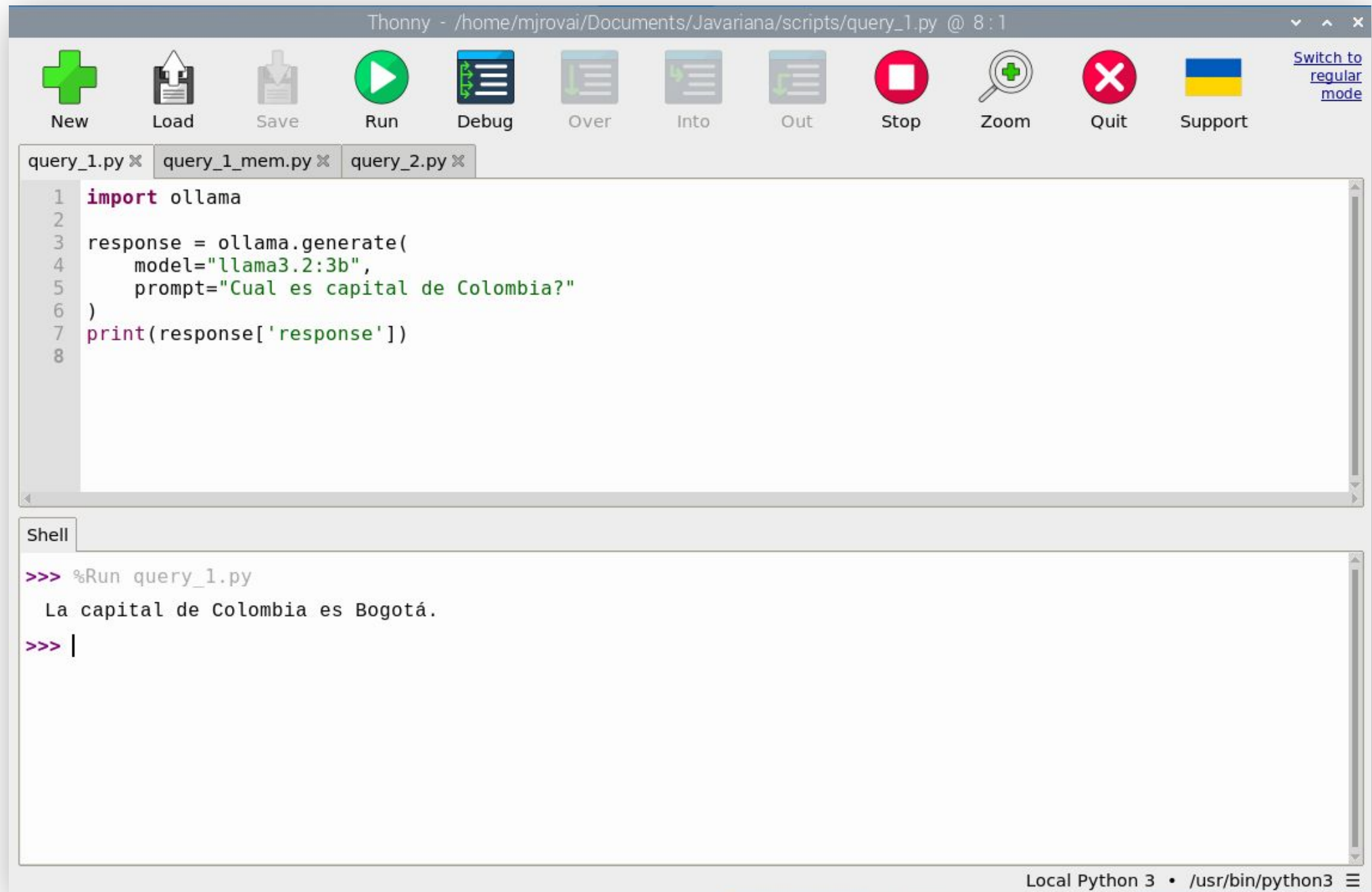
Optional Ollama Python Library Using Thonny IDE



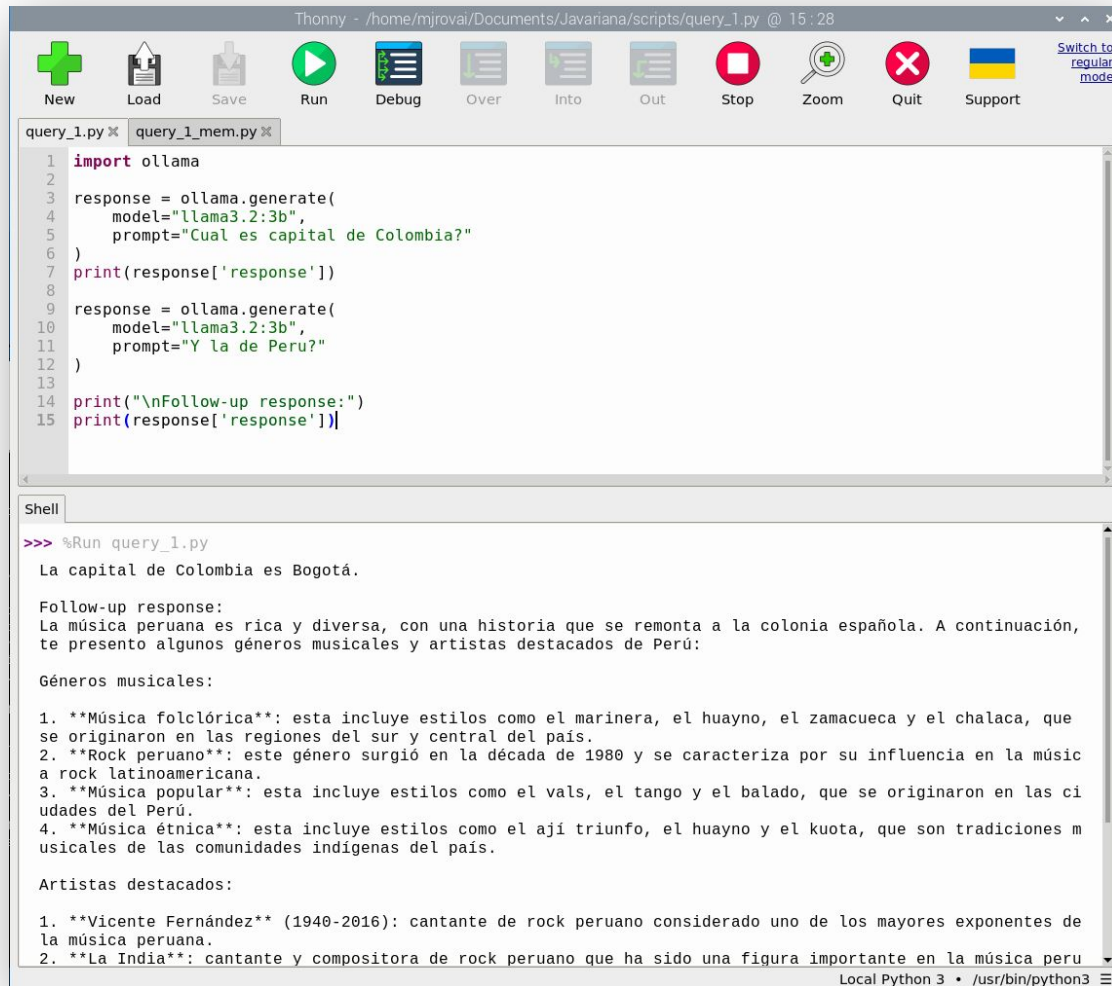


Install Ollama or any Python package on Thonny

Step	Action
1	Open Thonny IDE
2	Tools → Manage Packages
3	Search for ollama
4	Ensure Ollama server is installed & running
5	Pull a model with
6	Use



ollama.generate()



```
1 import ollama
2
3 response = ollama.generate(
4     model="llama3.2:3b",
5     prompt="Cual es capital de Colombia?"
6 )
7 print(response['response'])
8
9 response = ollama.generate(
10     model="llama3.2:3b",
11     prompt="Y la de Peru?"
12 )
13
14 print("\nFollow-up response:")
15 print(response['response'])
```

Shell

```
>>> %Run query_1.py
La capital de Colombia es Bogotá.

Follow-up response:
La música peruana es rica y diversa, con una historia que se remonta a la colonia española. A continuación,
te presento algunos géneros musicales y artistas destacados de Perú:

Géneros musicales:

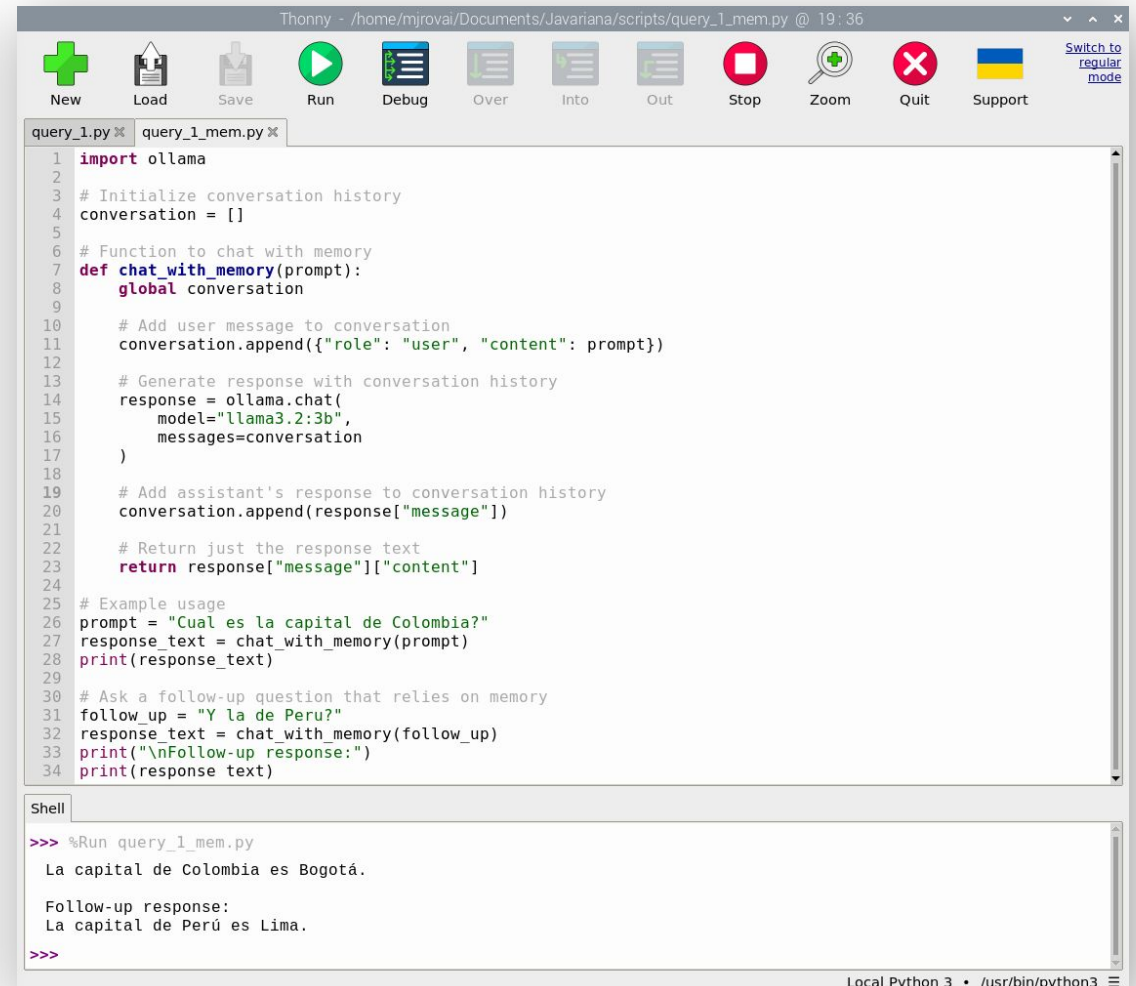
1. **Música folclórica**: esta incluye estilos como el marinera, el huayno, el zamacueca y el chalaca, que
se originaron en las regiones del sur y central del país.
2. **Rock peruano**: este género surgió en la década de 1980 y se caracteriza por su influencia en la música
a rock latinoamericana.
3. **Música popular**: esta incluye estilos como el vals, el tango y el balado, que se originaron en las ci
udades del Perú.
4. **Música étnica**: esta incluye estilos como el ají triunfo, el huayno y el kuota, que son tradiciones m
usicales de las comunidades indígenas del país.

Artistas destacados:

1. **Vicente Fernández** (1940-2016): cantante de rock peruano considerado uno de los mayores exponentes de
la música peruana.
2. **La India**: cantante y compositora de rock peruano que ha sido una figura importante en la música peru
```

Local Python 3 • /usr/bin/python3

ollama.chat()



```
1 import ollama
2
3 # Initialize conversation history
4 conversation = []
5
6 # Function to chat with memory
7 def chat_with_memory(prompt):
8     global conversation
9
10     # Add user message to conversation
11     conversation.append({"role": "user", "content": prompt})
12
13     # Generate response with conversation history
14     response = ollama.chat(
15         model="llama3.2:3b",
16         messages=conversation
17     )
18
19     # Add assistant's response to conversation history
20     conversation.append(response["message"])
21
22     # Return just the response text
23     return response["message"]["content"]
24
25 # Example usage
26 prompt = "Cual es la capital de Colombia?"
27 response_text = chat_with_memory(prompt)
28 print(response_text)
29
30 # Ask a follow-up question that relies on memory
31 follow_up = "Y la de Peru?"
32 response_text = chat_with_memory(follow_up)
33 print("\nFollow-up response:")
34 print(response_text)
```

Shell

```
>>> %Run query_1_mem.py
La capital de Colombia es Bogotá.

Follow-up response:
La capital de Perú es Lima.

>>>
```

Local Python 3 • /usr/bin/python3

Questions?



Prof. Marcelo J. Rovai

rovai@unifei.edu.br



UNIFEI