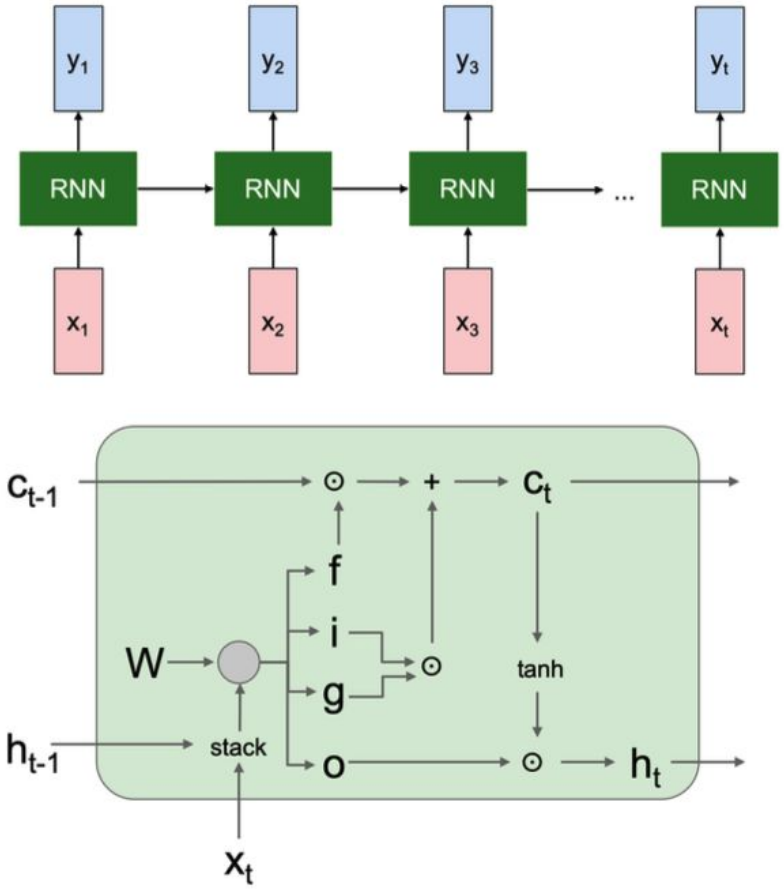
WALC 2024
Applied AI

Large Language Models (LLMs) at the Edge

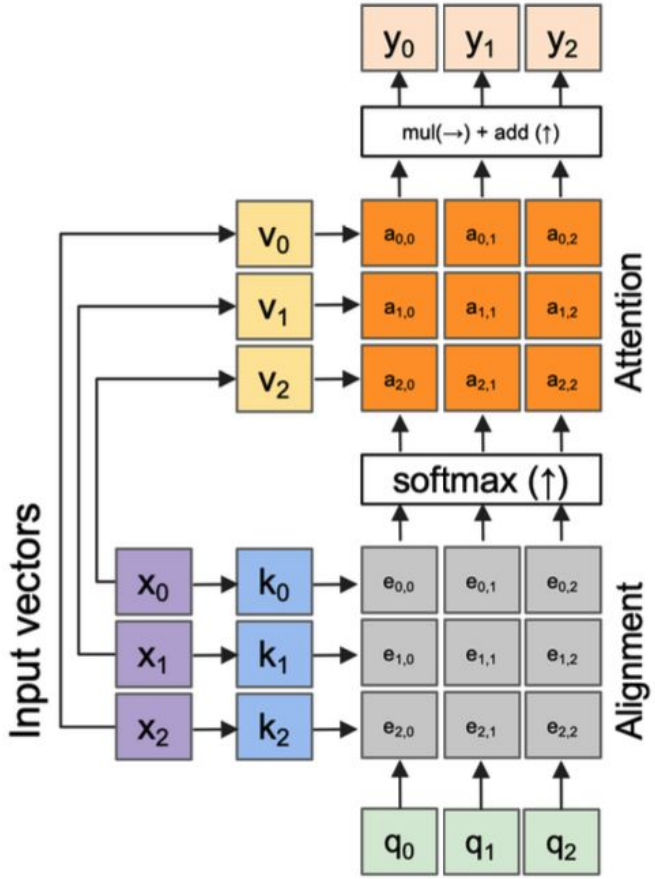
Prof. Marcelo J. Rovai
rovai@unifei.edu.br

UNIFEI - Federal University of Itajuba, Brazil
TinyML4D Academic Network Co-Chair

Recap: Models Beyond DNN and CNN



Recurrent neural network



Attention mechanism / Transformers

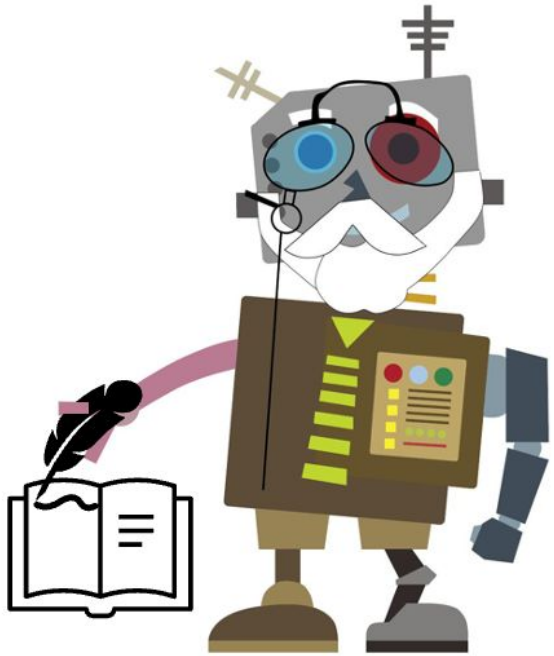
Machado de Assis Bot with RNN - GRU

The *robot writer* model is a **Recurrent Neural network (RNN/GRU)**. The model, with 4M parameters, was trained with a **150-characters sequence** from seven of his books: *Memorias Posthumas de Braz Cubas, Dom Casmurro, Quincas Borba, Papeis Avulsos, A Mão e a Luva, Esaú e Jacob*, and *Memorial de Ayres*.

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(1, 150, 256)	29,952
gru (GRU)	(1, 150, 1024)	3,938,304
dense (Dense)	(1, 150, 117)	119,925

Total params: 4,088,181 (15.60 MB)
Trainable params: 4,088,181 (15.60 MB)
Non-trainable params: 0 (0.00 B)



A LUVA DE CASMURRO II

A missa do coupé e um presente e o governo devia cazar logo no papel, a morte do autor, e todos os seus considerados de alegria. Era um espirito de vinte e cinco annos, e eu não estou alguns passos no cerebro, como de outra cousa. Deus me disse:

--Não digo que não. Se eu tivesse a intenção de um probosito. Palha acudiu a mulher, não havia nada. A noite vinha tambem para o seminario, tinha o aspecto do partido recto e de restaurar a minha mãe e do pae, pela primeira vez, a menor destinada a dispensar o chapéo, esperou que não vinhas com as suas mãos de creanças. A manhã della chegasse a baroneza e a maneira desta divida. Parece que é casada.

--Está bom, perdoa-lhe de todos os lados, a vida de que o comprar para o meu quarto de hora, e contavam com o fim de a anterior, e, a parede pouco tempo a alma de pessoas que definitivamente lhe interessam a menos para mim. De quando em quando, esses dous annos de conversação para o fim de deixar nenhuma pessoa que se dispersasse; mas não falo de uma cousa nem lhe pedia com a mão tremula, como se ella quizesse. Eu, apertando-lhe a mão, aliás o principio do governo, a proposito disso, com a desattenção de Estevão, e eu começou a aborrecel-o, e a solidão podia ser melhor, e a sympathia colloca da mãe, e não se sabe calar o enterro no meio do lagem, o que iam-se apanhados no chão, e para a mulher, não tendo visto, nem a mesma cousa.

Generative AI (GenAI)

Generative AI is an artificial intelligence system capable of creating new, original content across various mediums such as text, images, audio, and video. These systems learn patterns from existing data and use that knowledge to generate novel outputs that didn't previously exist.

Large Language Models (LLMs), Small Language Models (SLMs), and multimodal models can all be considered types of GenAI when used for generative tasks.

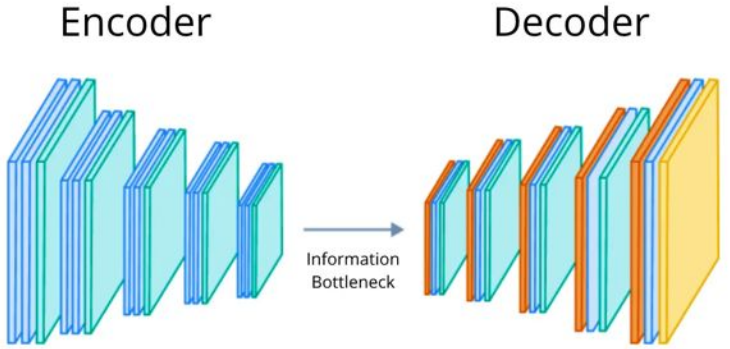
LLM / SLM

Large Language Model / Small Language Models

LLMs are **specialized deep learning models designed to understand and generate human language**, used for tasks like translation, summarization, and generating human-like text responses. SLMs are the same, but use a simpler, less resource-intensive approach (smaller in size).

Deep Learning models (or artificial neural networks)

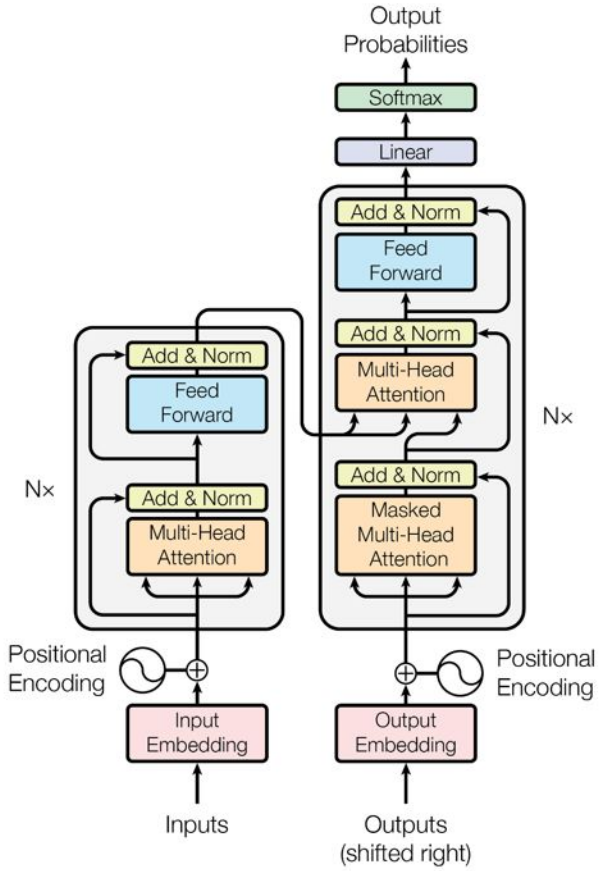
- **Autoencoders**: Used primarily for unsupervised learning tasks such as dimensionality reduction and feature extraction, autoencoders learn to compress data from the input layer into a shorter code and then reconstruct the output from this representation.



- **Transformer Models**: Highly effective in handling sequences, transformers use mechanisms like self-attention to weigh the importance of different words in a sentence, regardless of their position. The Transformer architecture, while innovative, can be seen as a derivative of earlier deep learning models, particularly those based on the concept of sequence modeling. However, the most direct lineage can be traced to the sequence-to-sequence (seq2seq) models that utilize **encoder-decoder** architectures. These earlier seq2seq models were often built using **recurrent neural networks (RNNs)** or their more advanced variants like **LSTMs (Long Short-Term Memory Networks) or GRUs (Gated Recurrent Units).**

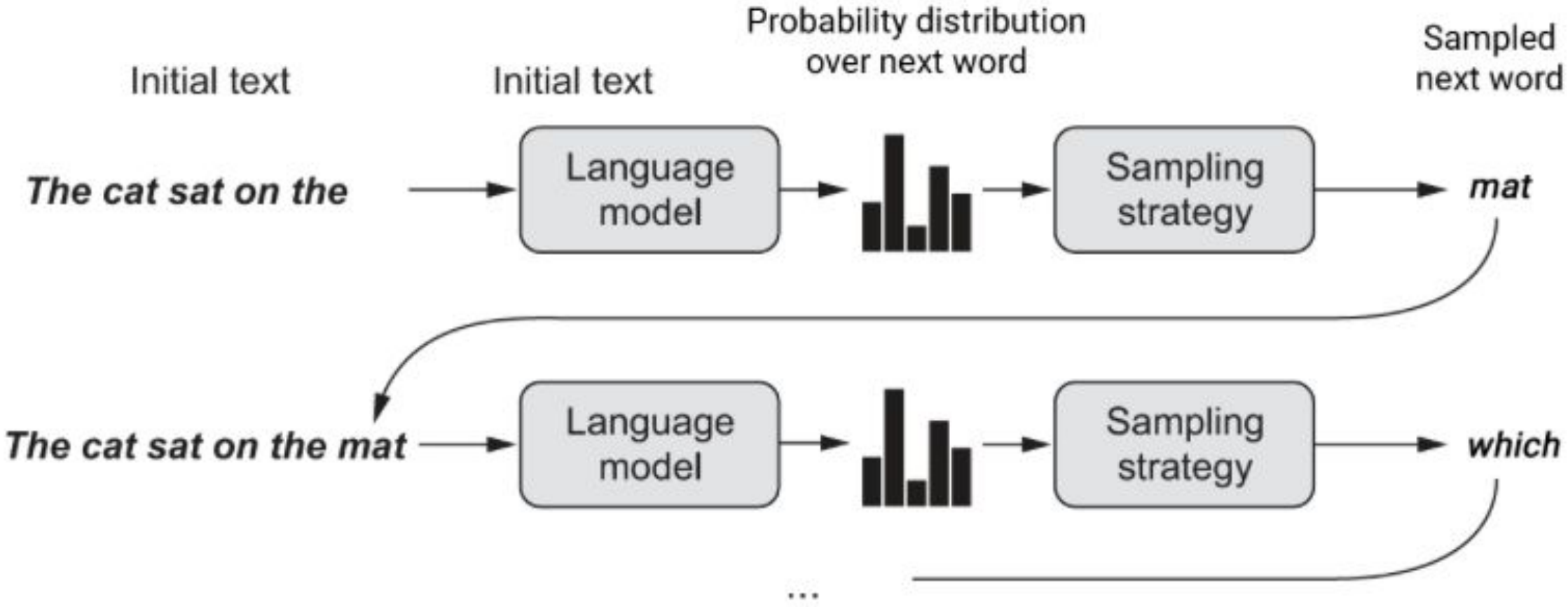
LLM/SLM – Large /Small Language Model

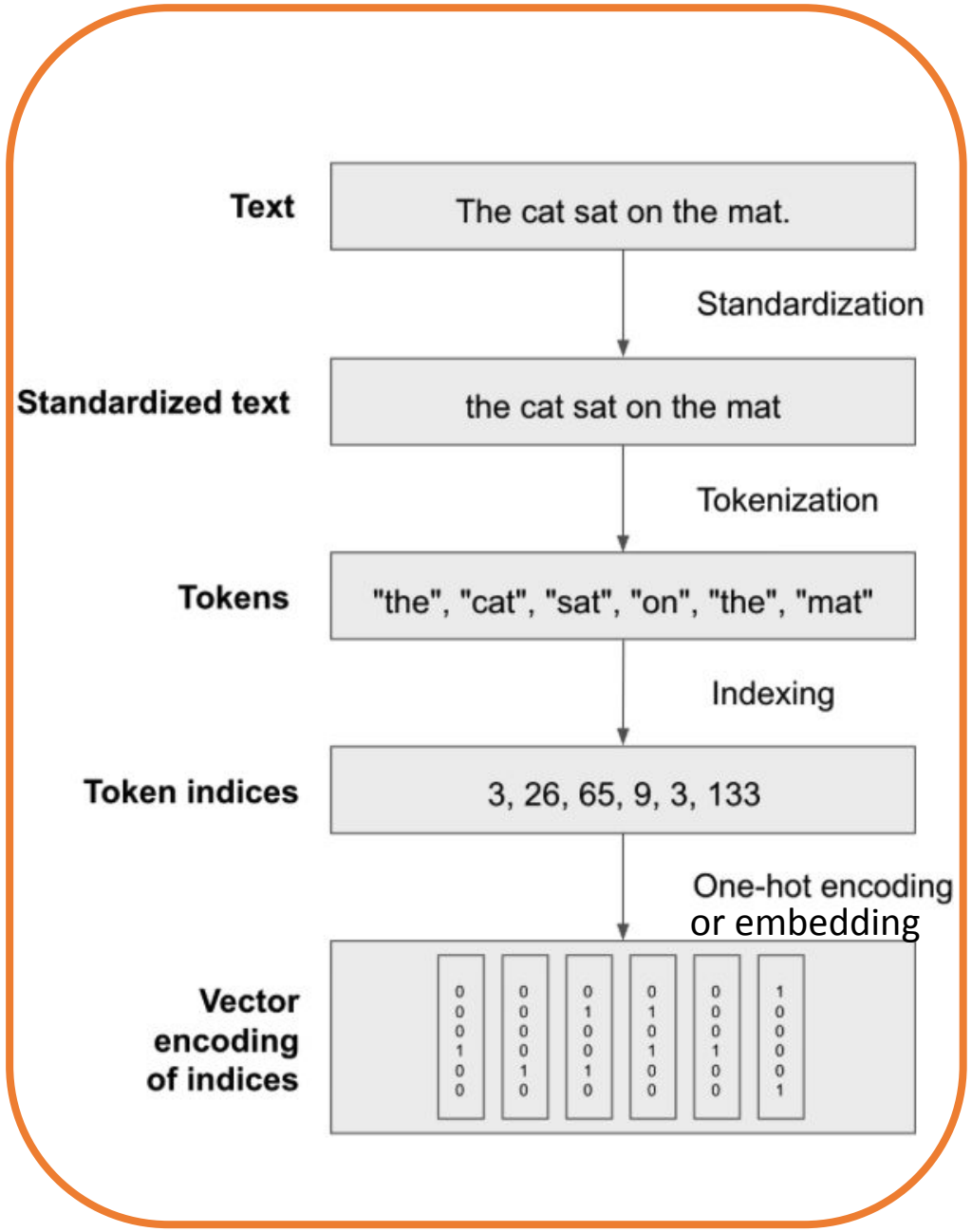
Large Language Models (LLMs) and SLMs are advanced neural networks based on the Transformer architecture that excel in understanding and generating human language. They represent a significant evolution from earlier sequence-based mode like RNNs, which surpass them in handling long-range dependencies and parallel processing efficiency.

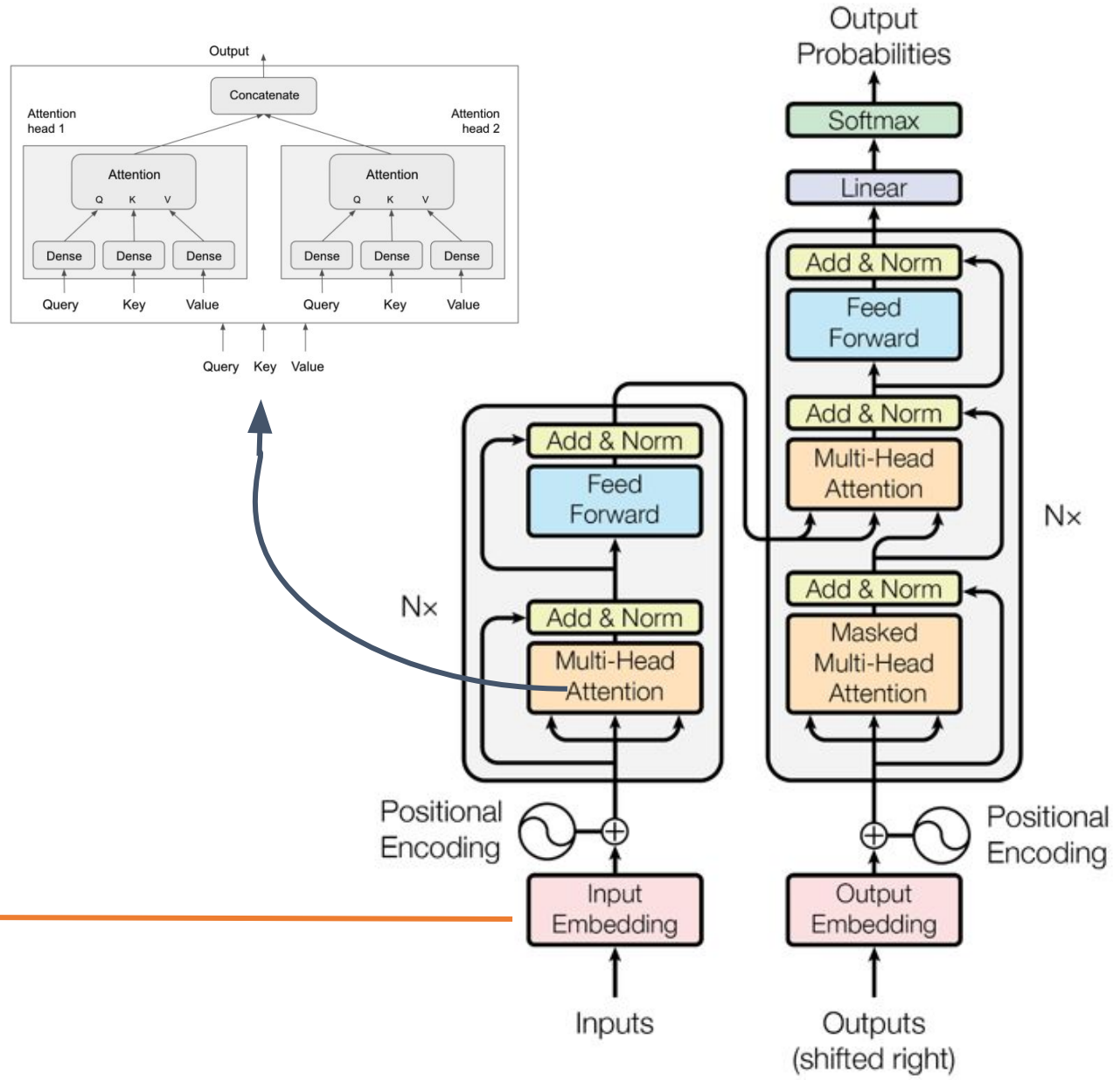


The Illustrated Transformer

LLM/SLM – Large /Small Language Model



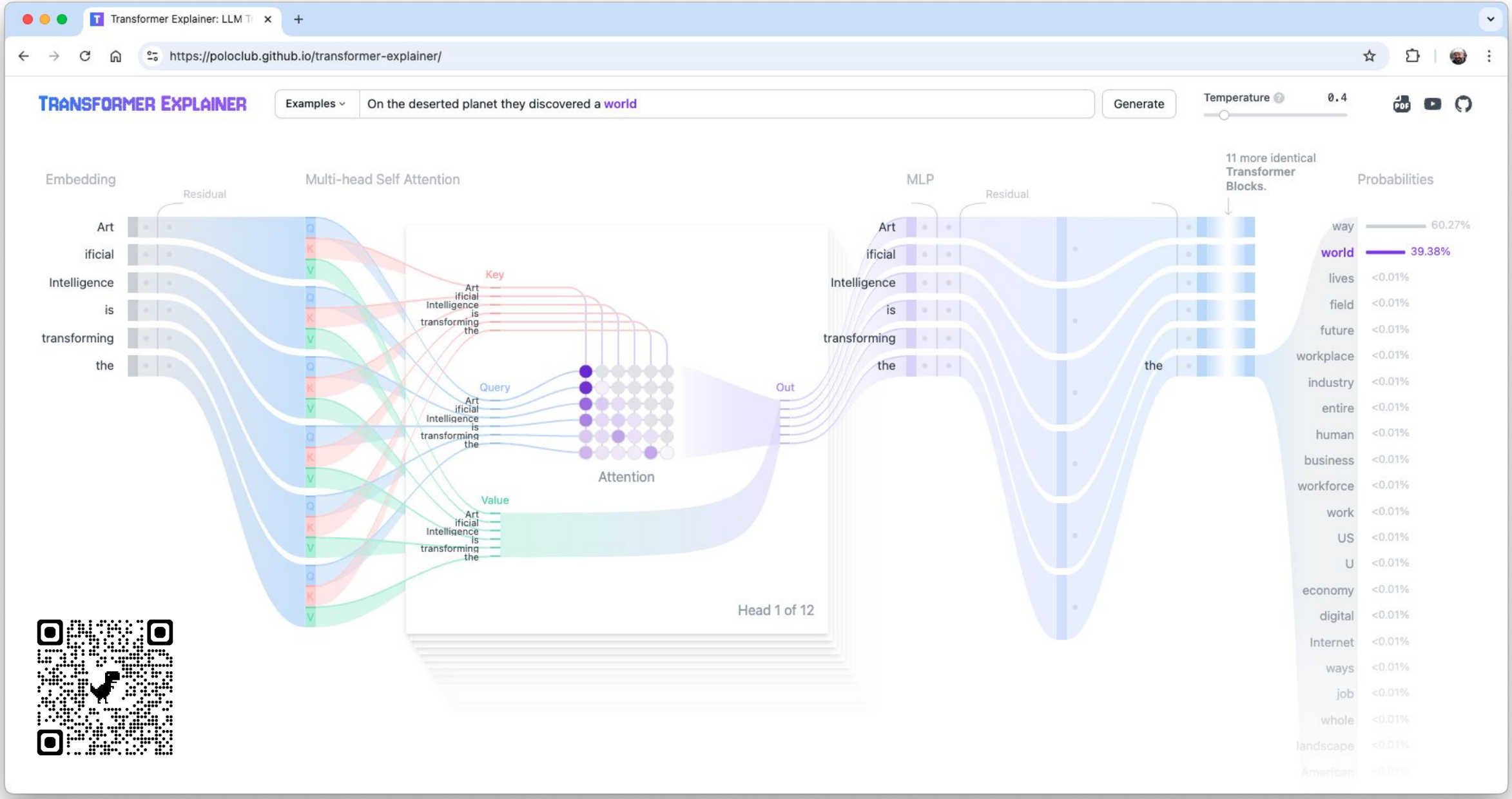




Encoder Decoder

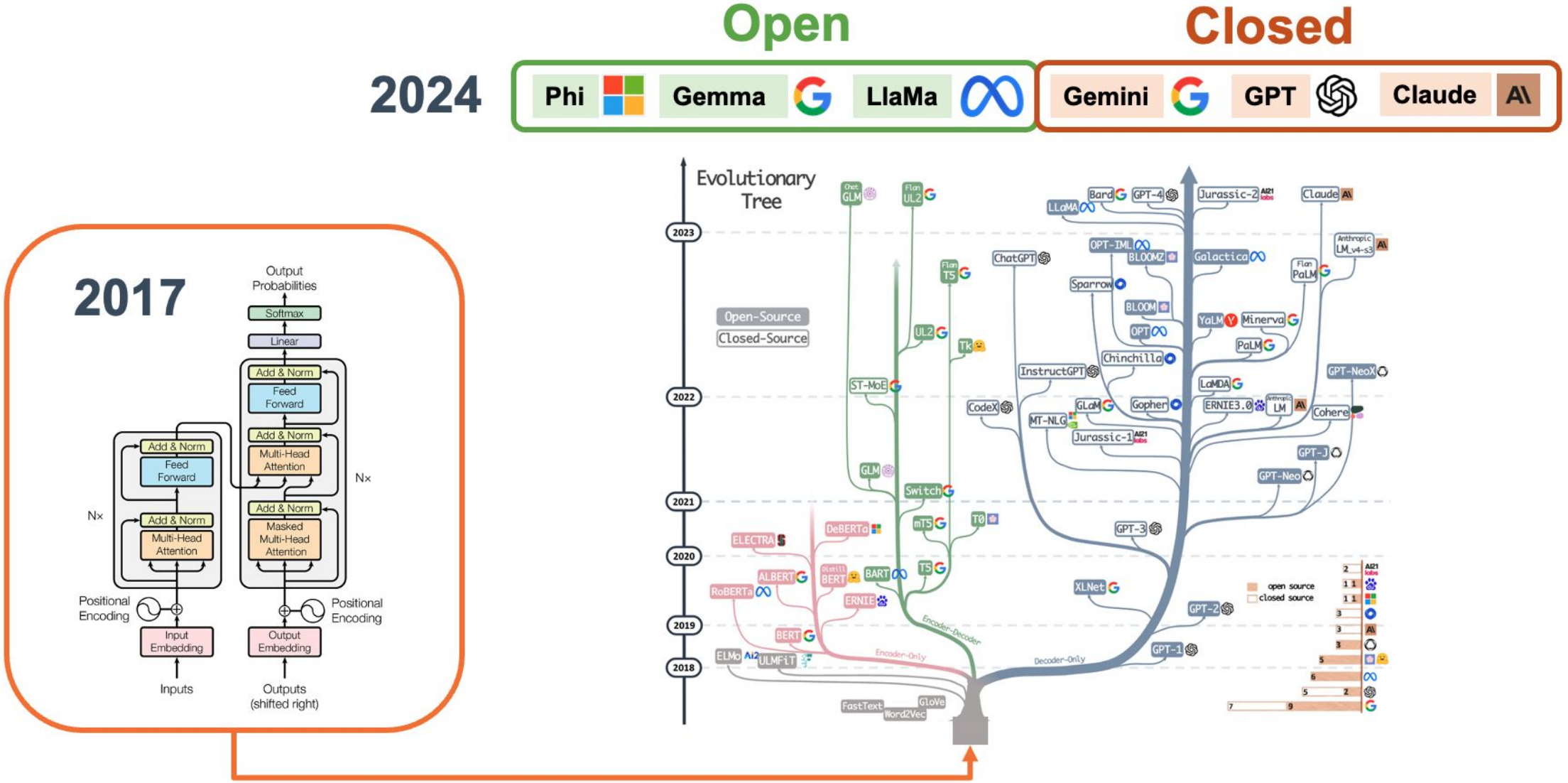
LLM/SLM – Large /Small Language Model

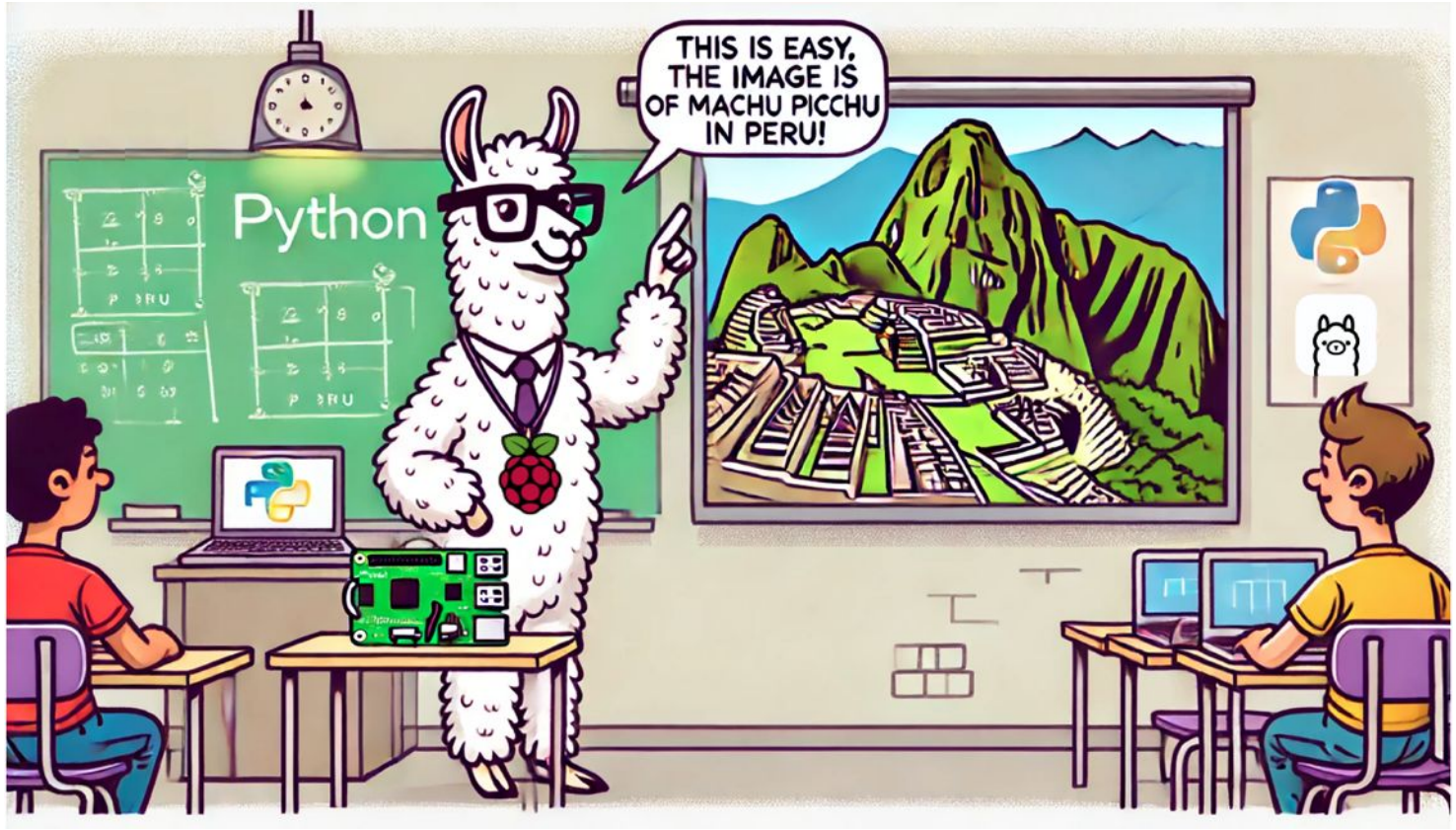
How large language models work



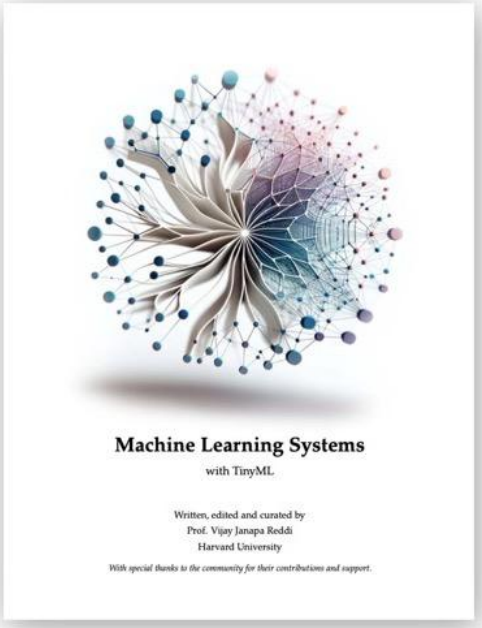
https://poloclub.github.io/transformer-explainer/

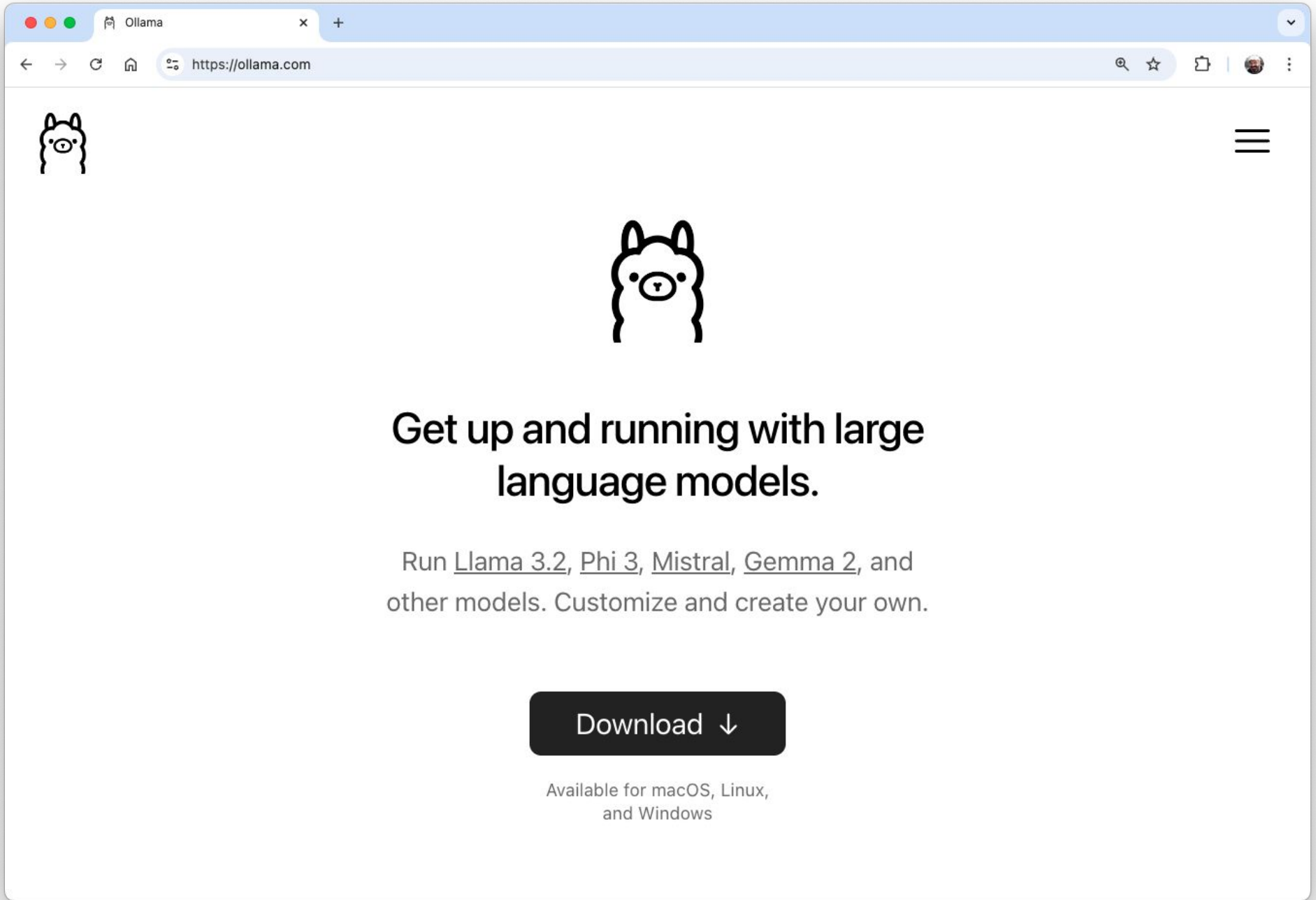
Transformers to LLMs and SLMs

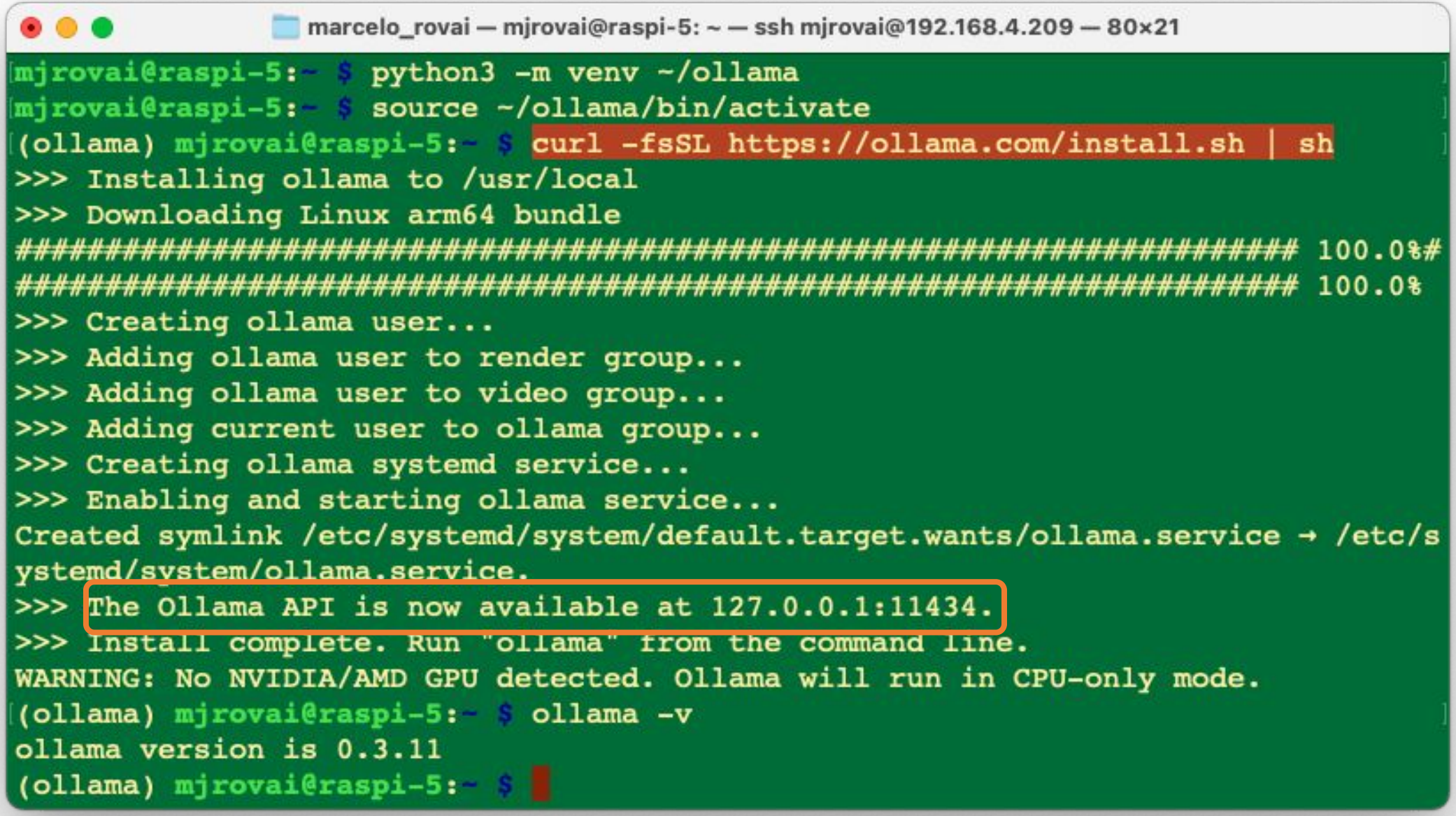


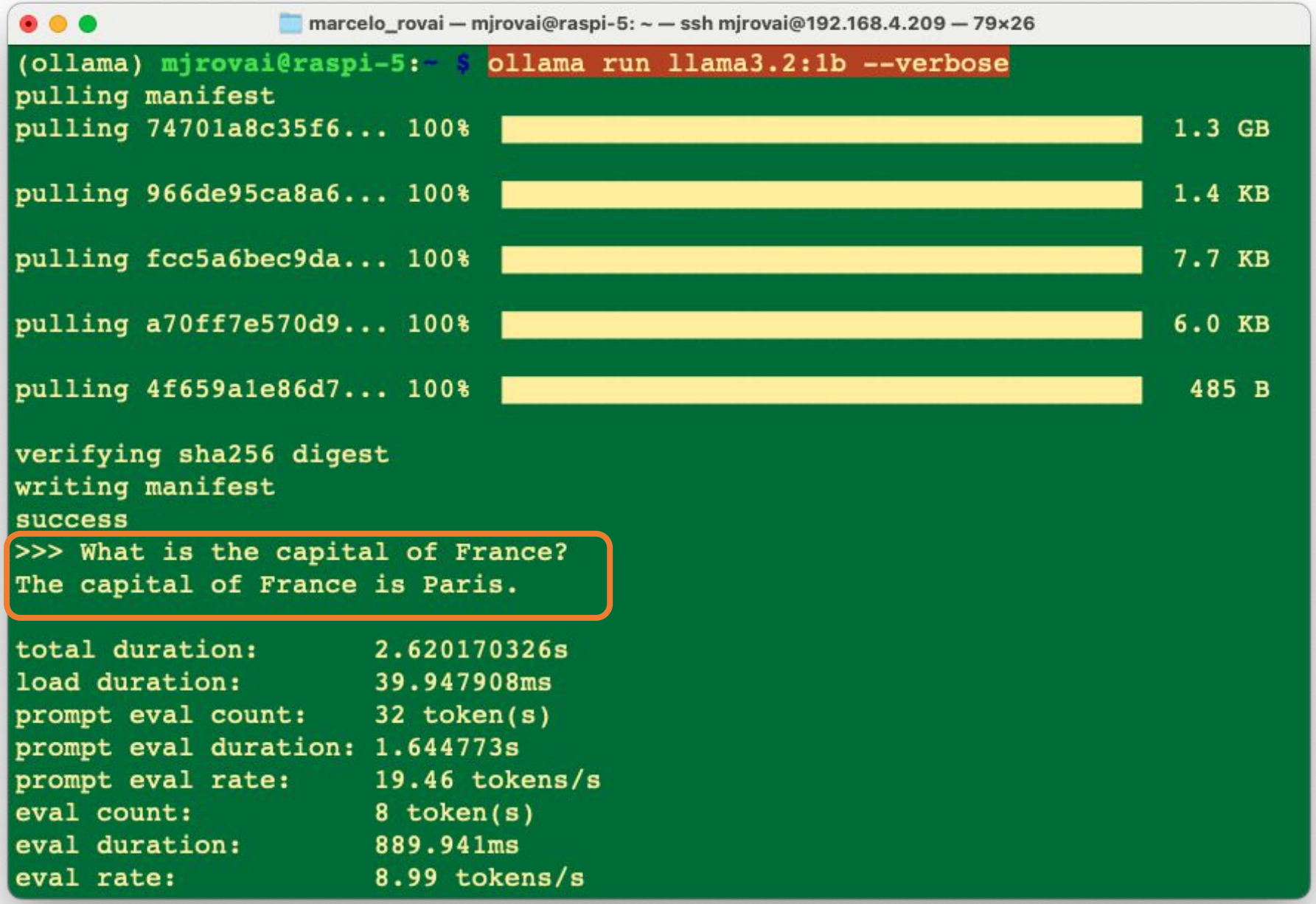


Running Large Language Models on Raspberry Pi at the Edge




Ollama
https://ollama.com
Get up and running with large language models.
Run Llama 3.2, Phi 3, Mistral, Gemma 2, and other models. Customize and create your own.
Download ↓
Available for macOS, Linux, and Windows


marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 80×21
mjrovai@raspi-5:~ $ python3 -m venv ~/ollama
mjrovai@raspi-5:~ $ source ~/ollama/bin/activate
(ollama) mjrovai@raspi-5:~ $ curl -fsSL https://ollama.com/install.sh | sh
>>> Installing ollama to /usr/local
>>> Downloading Linux arm64 bundle
100.0%#
100.0%
>>> Creating ollama user...
>>> Adding ollama user to render group...
>>> Adding ollama user to video group...
>>> Adding current user to ollama group...
>>> Creating ollama systemd service...
>>> Enabling and starting ollama service...
Created symlink /etc/systemd/system/default.target.wants/ollama.service → /etc/s
ystemd/system/ollama.service.
>>> The Ollama API is now available at 127.0.0.1:11434.
>>> Install complete. Run "ollama" from the command line.
WARNING: No NVIDIA/AMD GPU detected. Ollama will run in CPU-only mode.
(ollama) mjrovai@raspi-5:~ $ ollama -v
ollama version is 0.3.11
(ollama) mjrovai@raspi-5:~ $

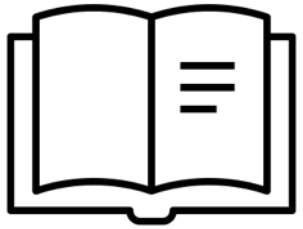

marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 79×26
(ollama) mjrovai@raspi-5:~ $ ollama run llama3.2:1b --verbose
pulling manifest
pulling 74701a8c35f6... 100% 1.3 GB
pulling 966de95ca8a6... 100% 1.4 KB
pulling fcc5a6bec9da... 100% 7.7 KB
pulling a70ff7e570d9... 100% 6.0 KB
pulling 4f659a1e86d7... 100% 485 B
verifying sha256 digest
writing manifest
success
>>> What is the capital of France?
The capital of France is Paris.
total duration: 2.620170326s
load duration: 39.947908ms
prompt eval count: 32 token(s)
prompt eval duration: 1.644773s
prompt eval rate: 19.46 tokens/s
eval count: 8 token(s)
eval duration: 889.941ms
eval rate: 8.99 tokens/s

Multimodal Models

LLaVa-PHI3

```
marcelo_rovai — mjrovai@raspi-5: ~/Documents/OLLAMA — ssh mjrovai@192.168.4.209 — 84×36
(ollama) mjrovai@raspi-5:~/Documents/OLLAMA $ pwd
/home/mjrovai/Documents/OLLAMA
(ollama) mjrovai@raspi-5:~/Documents/OLLAMA $ ollama run llava-phi3:3.8b --verbose
>>> Describe the image /home/mjrovai/Documents/OLLAMA/image_test_1.jpg
Added image '/home/mjrovai/Documents/OLLAMA/image_test_1.jpg'
The image captures a breathtaking view of Paris, France. The cityscape is
dotted with buildings in various shades of white and gray, interspersed with
lush green trees that add a touch of nature to the urban setting.

In the heart of the scene stands the Eiffel Tower, an iconic symbol of Paris,
its iron lattice structure reaching up into the clear blue sky. The tower's
distinctive silhouette is unmistakable against the backdrop of the sky, which
is a vibrant shade of blue with just a few clouds scattered across it.

The Seine River gracefully winds its way through the city, bordered by an
array of buildings on both sides. The river is lined with several bridges that
connect different parts of the city and facilitate movement for pedestrians
and vehicles alike.

Above all these elements, a few birds can be seen soaring freely in the sky,
their presence adding life to the scene. Their flight paths crisscross over
the river and the buildings, creating dynamic patterns that draw the eye.

Overall, this image presents a beautiful daytime snapshot of Paris - its
architectural marvels, natural beauty, and bustling city life coexisting in
harmony.

total duration:       3m55.972199346s
load duration:        16.198011ms
prompt eval count:    1 token(s)
prompt eval duration: 2m19.561783s
prompt eval rate:     0.01 tokens/s
eval count:           276 token(s)
eval duration:        1m36.330959s
eval rate:            2.87 tokens/s
>>> Send a message (/? for help)
```

llava-phi-3 is a LLaVA model (**L**arge **La**nguage and **V**ision **A**ssistant) fine-tuned from Microsoft Phi-3 mini



= 147K tokens

~ 350 pages

~ 300 words/page

1 word = ~ 1.4 token

A **4-bit** quantized 3.8 billion parameter * language model trained on 3.3 trillion tokens**, whose overall performance, as measured by both academic benchmarks and internal testing, rivals that of models such as Mixtral 8x7B and GPT-3.5

* 2.4 GB

** 22.5 Million books - 17% of all books written in the world

llava-phi-3 (2.9 GB)

Ollama

```
mjrovai@rpi-5: ~/Documents/OLLAMA
Help
ute.

/Documents/OLLAMA $
/Documents/OLLAMA $ python calc_distance_image.py /
home/mjrovai/Documents/OLLAMA/image_test_1.jpg

 The image shows Paris, with lat:48.86 and long: 2.35, located in
France and about 11,630 kilometers away from Santiago, Chile.

 [INFO] ==> The code (running llava-phi3), took 232.60845186299412
 seconds to execute.

mjrovai@rpi-5:~/Documents/OLLAMA $
```

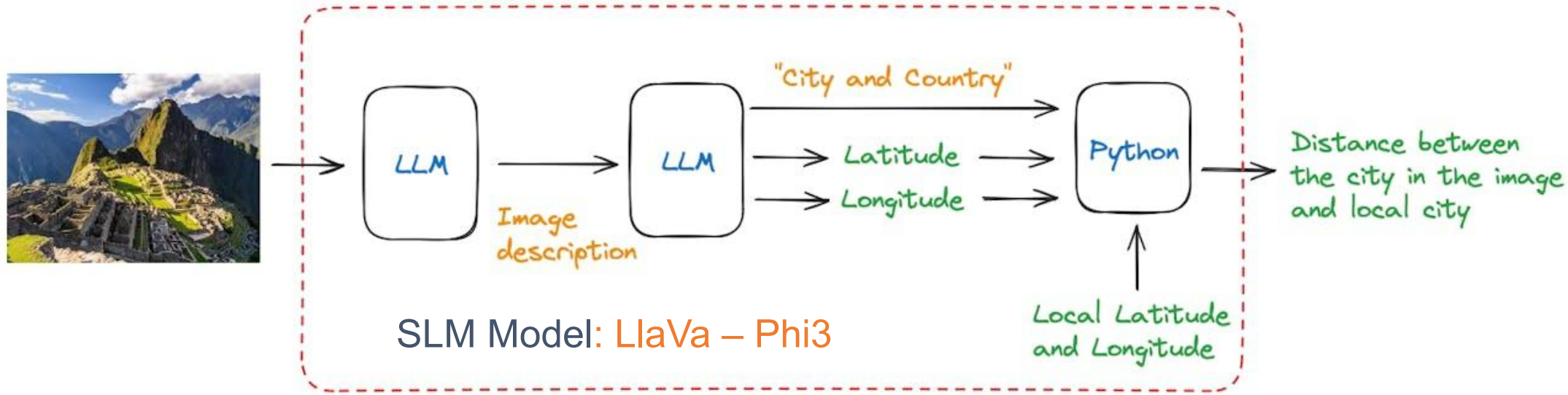
```
mjrovai@rpi-5: ~
File  Edit  Tabs  Help
>>> Awswer with one short sentence, what is the capital of France and its distanc
... e in Km from Santiago, Chile
 The capital of France is Paris and it is around 12,674 kilometers away
from Santiago, Chile.

total duration:       13.860074968s
load duration:        1.537039ms
prompt eval count:    27 token(s)
prompt eval duration: 5.925386s
prompt eval rate:     4.56 tokens/s
eval count:           26 token(s)
eval duration:        7.539223s
eval rate:            3.45 tokens/s
>>> Send a message (/? for help)
```

(13 seconds)

```
mjrovai@rpi-5: ~/Documents/OLLAMA
Help

/Documents/OLLAMA $
/Documents/OLLAMA $ python calc_distance_image.py
/home/mjrovai/Documents/OLLAMA/image_test_3.jpg

 The image shows Machu Picchu, with lat:-13.16 and long: -72.54,
 located in Peru and about 2,250 kilometers away from Santiago,
Chile.

 [INFO] ==> The code (running llava-phi3), took 267.579568572007
7 seconds to execute.

mjrovai@rpi-5:~/Documents/OLLAMA $
```

(4 minutes)

Function Calling

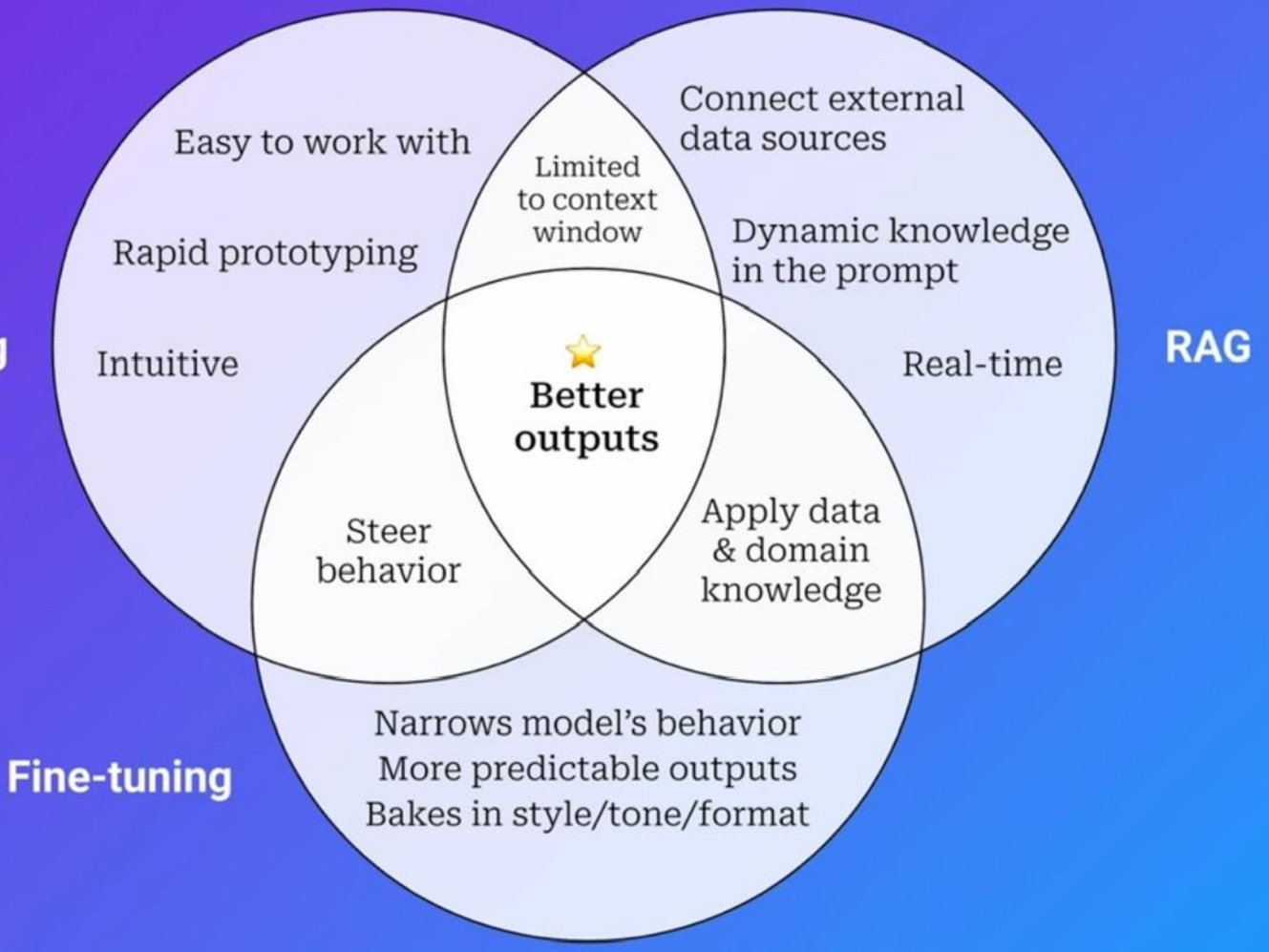


LLMs: Optimization Techniques

LLMs: Common Optimization Techniques

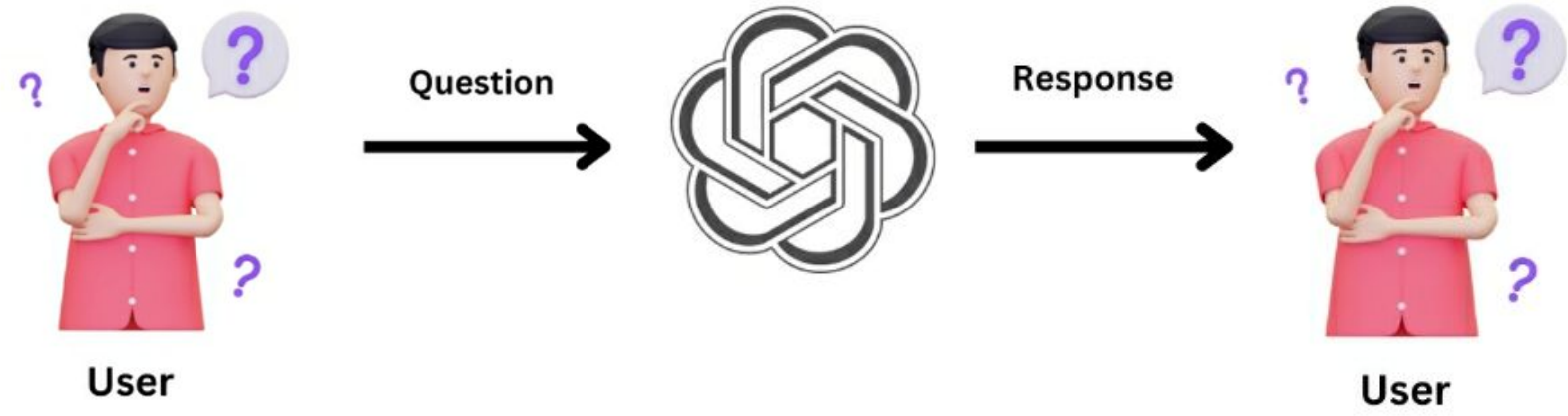
1. **Prompt Engineering**: Tailor your interactions.
2. **Fine-tuning**: Perfect the model's tasks.
3. **RAG**: Enhance with relevant data.

Comparison of Techniques

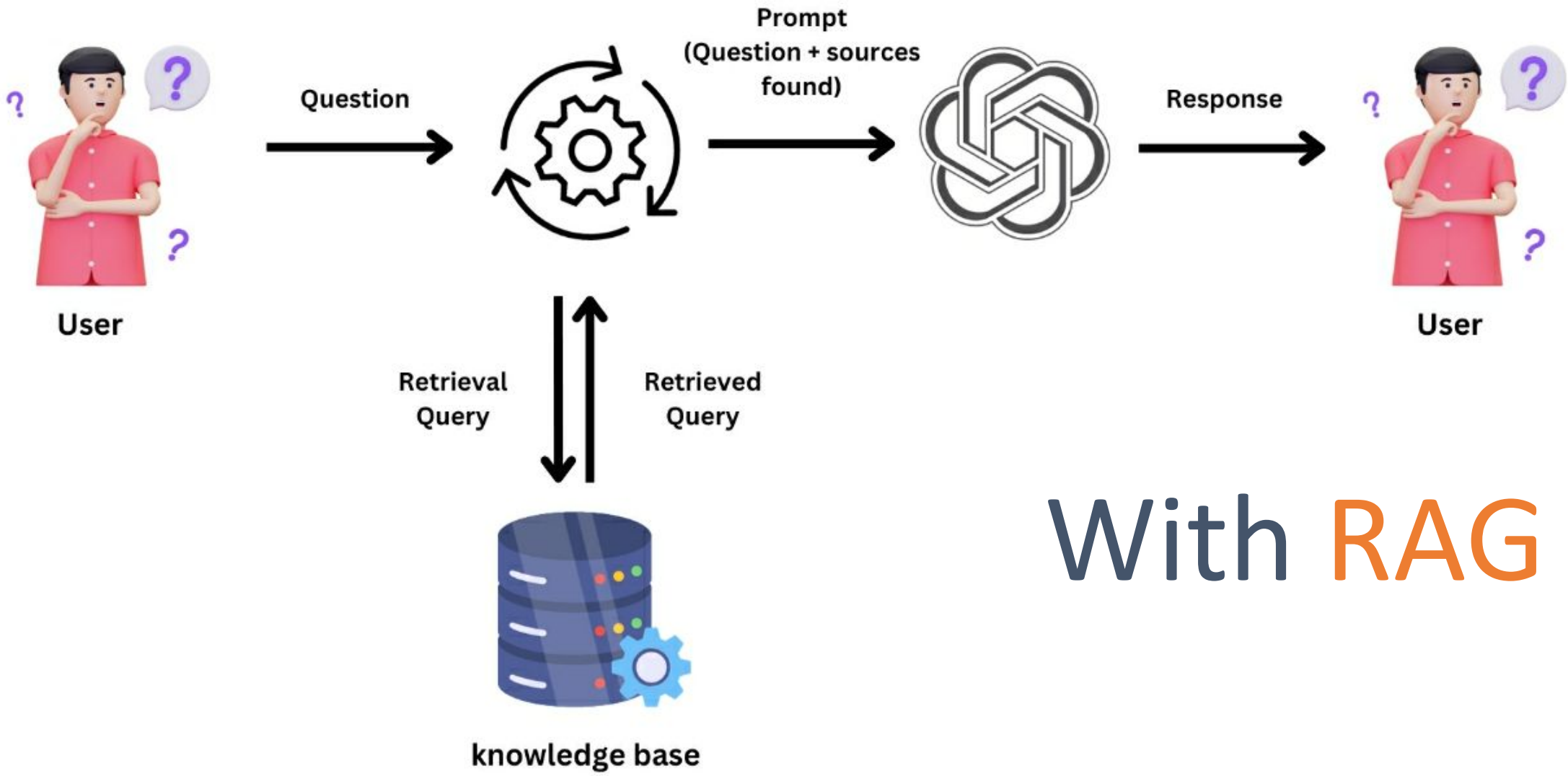


Retrieval-Augmented Generation (RAG)

“A method created by the FAIR team at Meta to enhance the accuracy of Large Language Models (LLMs) and reduce false information or “hallucinations.”

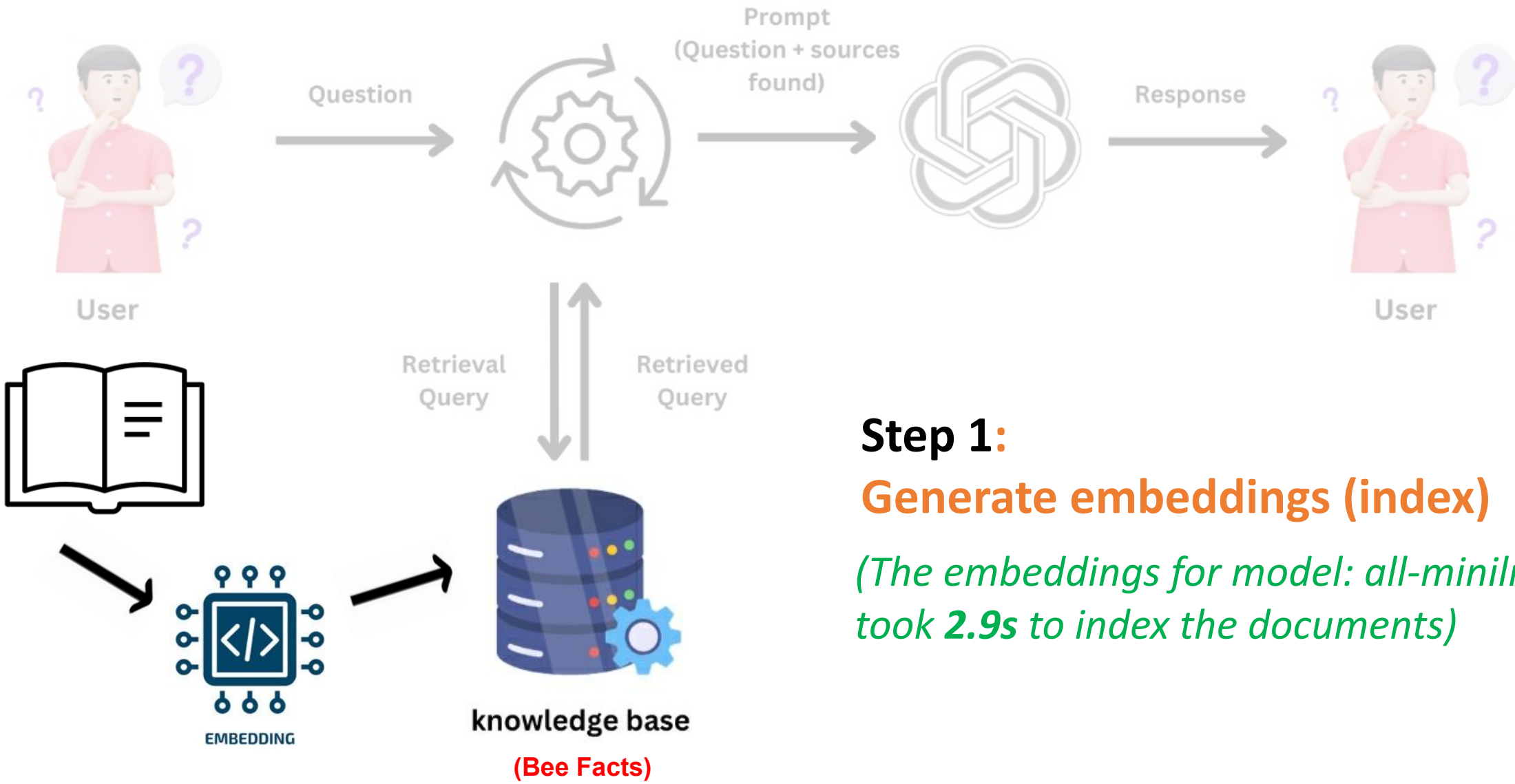


Usual Prompt



With RAG

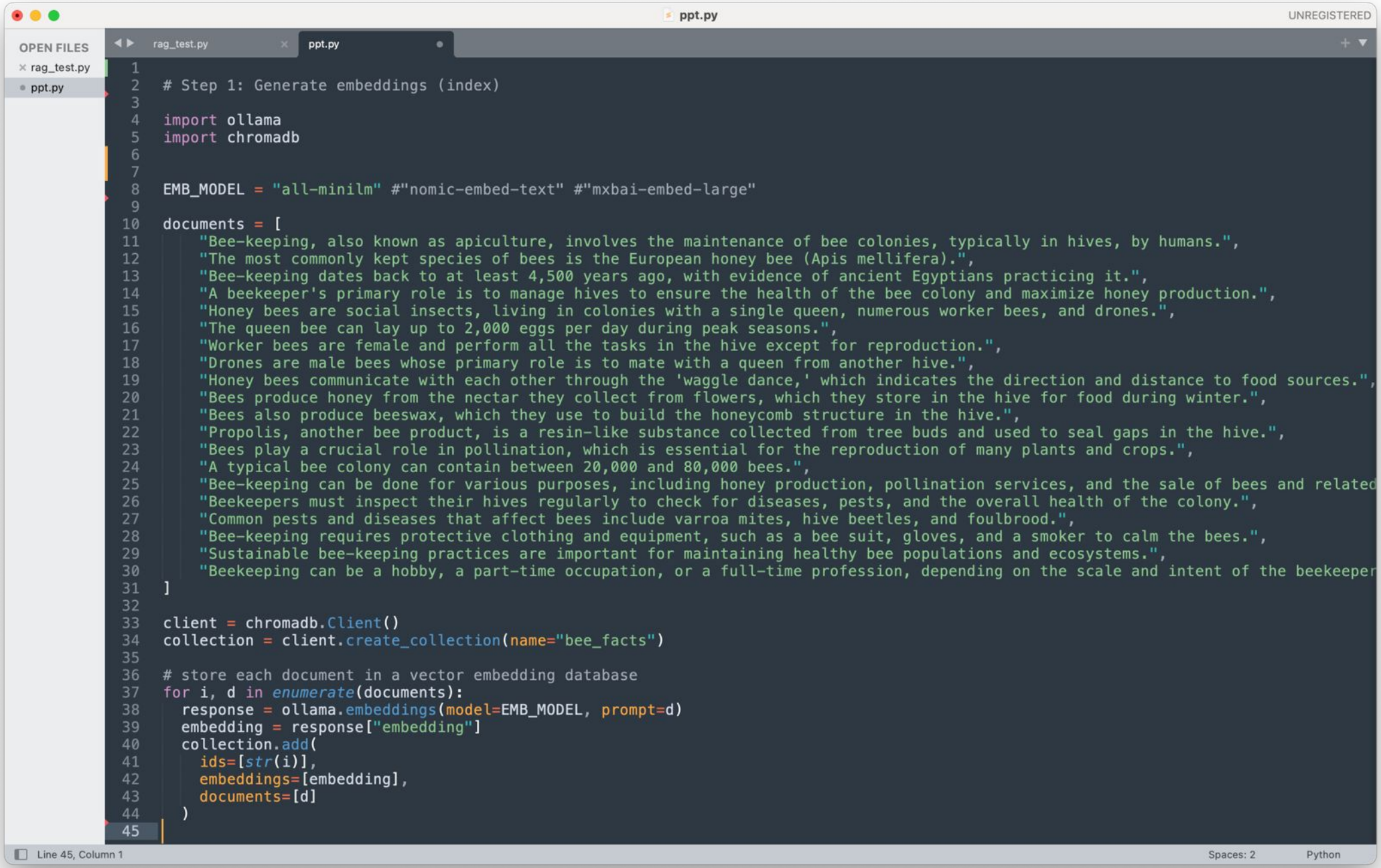
Bouchard, Louis-François; Peters, Louie . Building LLMs for Production: Enhancing LLM Abilities and Reliability with Prompting, Fine-Tuning, and RAG (p. 282). Kindle Edition.

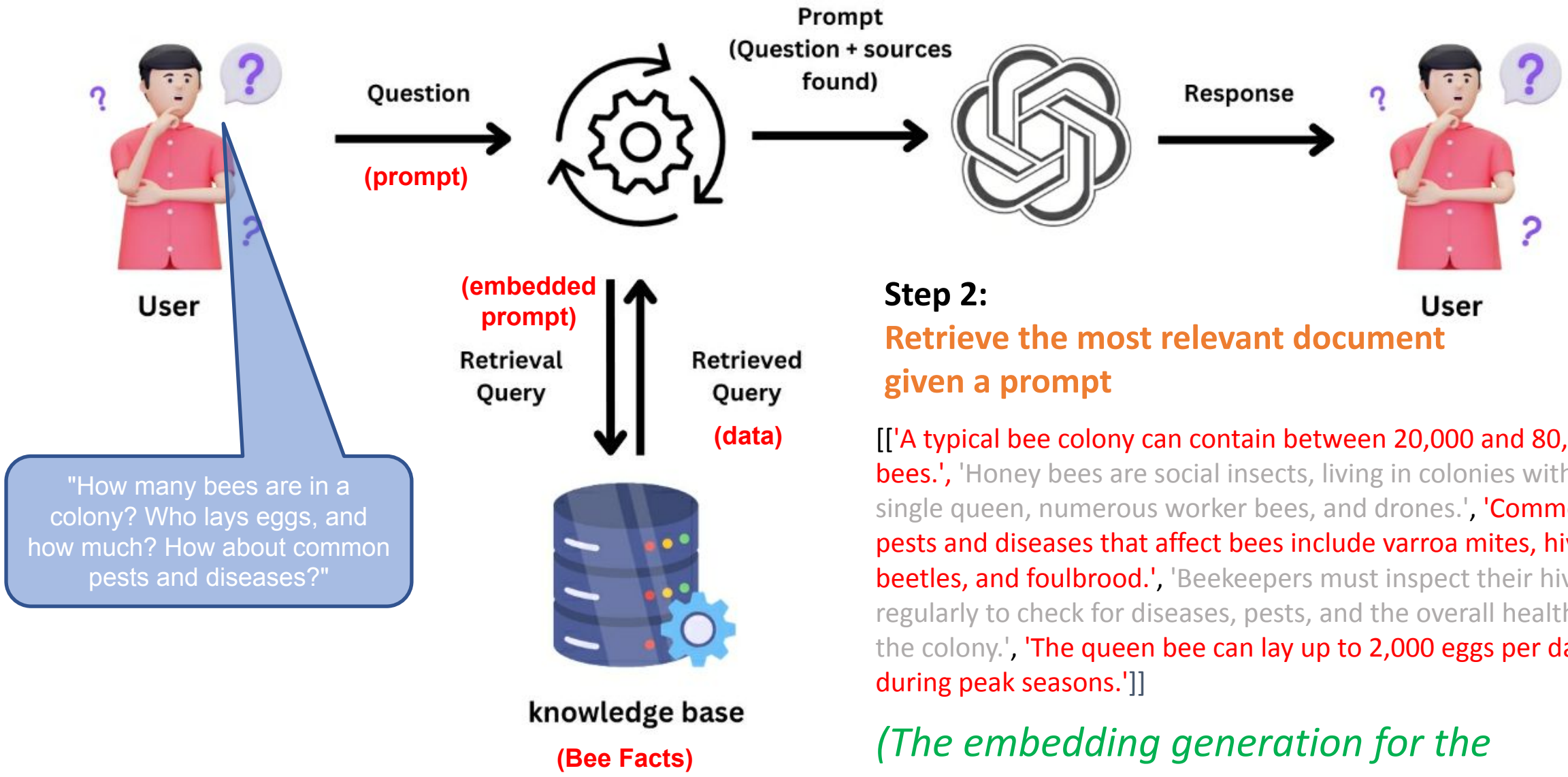


Step 1:

Generate embeddings (index)

*(The embeddings for model: all-minilm, took **2.9s** to index the documents)*


ppt.py
UNREGISTERED
OPEN FILES
rag_test.py
ppt.py
rag_test.py
ppt.py
# Step 1: Generate embeddings (index)
import ollama
import chromadb
EMB_MODEL = "all-minilm" #"nomic-embed-text" #"mxbai-embed-large"
documents = [
"Bee-keeping, also known as apiculture, involves the maintenance of bee colonies, typically in hives, by humans.",
"The most commonly kept species of bees is the European honey bee (Apis mellifera).",
"Bee-keeping dates back to at least 4,500 years ago, with evidence of ancient Egyptians practicing it.",
"A beekeeper's primary role is to manage hives to ensure the health of the bee colony and maximize honey production.",
"Honey bees are social insects, living in colonies with a single queen, numerous worker bees, and drones.",
"The queen bee can lay up to 2,000 eggs per day during peak seasons.",
"Worker bees are female and perform all the tasks in the hive except for reproduction.",
"Drones are male bees whose primary role is to mate with a queen from another hive.",
"Honey bees communicate with each other through the 'waggle dance,' which indicates the direction and distance to food sources.",
"Bees produce honey from the nectar they collect from flowers, which they store in the hive for food during winter.",
"Bees also produce beeswax, which they use to build the honeycomb structure in the hive.",
"Propolis, another bee product, is a resin-like substance collected from tree buds and used to seal gaps in the hive.",
"Bees play a crucial role in pollination, which is essential for the reproduction of many plants and crops.",
"A typical bee colony can contain between 20,000 and 80,000 bees.",
"Bee-keeping can be done for various purposes, including honey production, pollination services, and the sale of bees and related
"Beekeepers must inspect their hives regularly to check for diseases, pests, and the overall health of the colony.",
"Common pests and diseases that affect bees include varroa mites, hive beetles, and foulbrood.",
"Bee-keeping requires protective clothing and equipment, such as a bee suit, gloves, and a smoker to calm the bees.",
"Sustainable bee-keeping practices are important for maintaining healthy bee populations and ecosystems.",
"Beekeeping can be a hobby, a part-time occupation, or a full-time profession, depending on the scale and intent of the beekeeper
]
client = chromadb.Client()
collection = client.create_collection(name="bee_facts")
# store each document in a vector embedding database
for i, d in enumerate(documents):
response = ollama.embeddings(model=EMB_MODEL, prompt=d)
embedding = response["embedding"]
collection.add(
ids=[str(i)],
embeddings=[embedding],
documents=[d]
)
Line 45, Column 1
Spaces: 2
Python

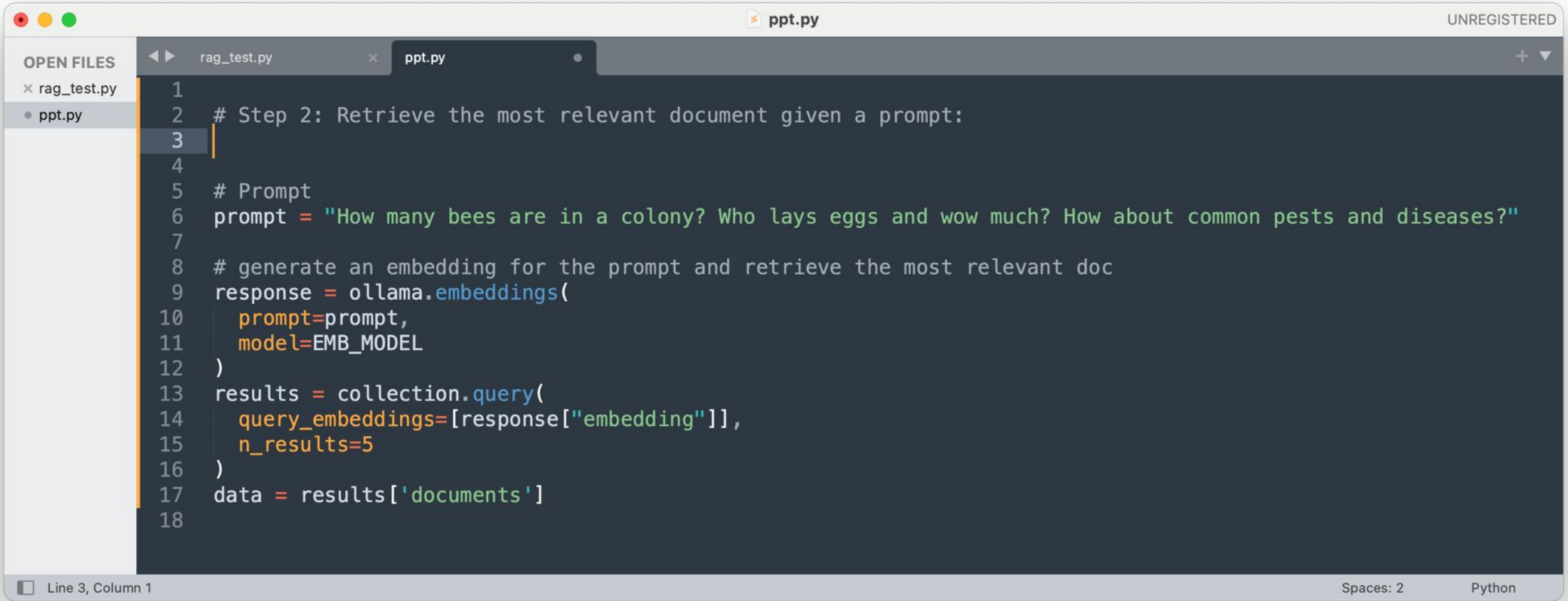


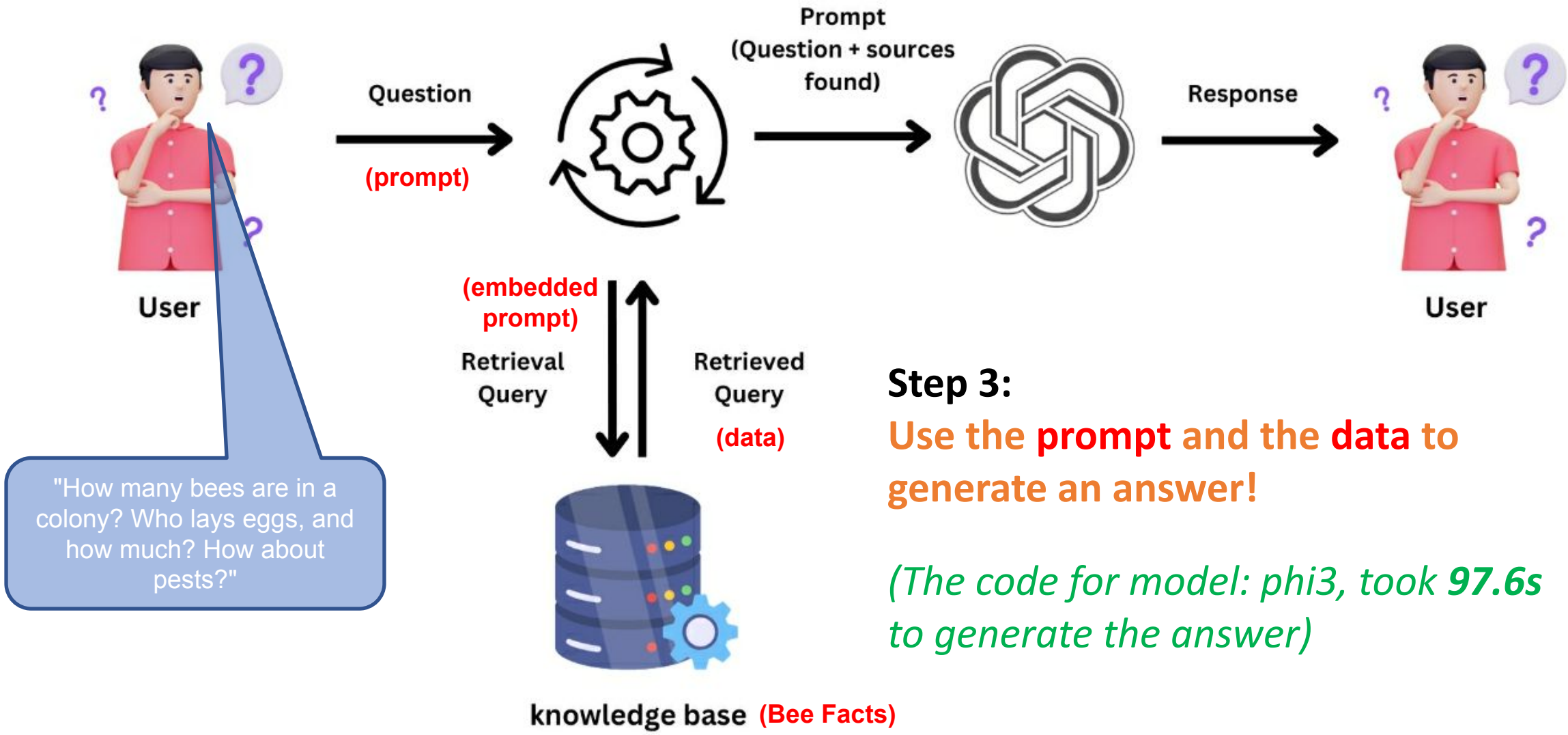
Step 2:

Retrieve the most relevant document given a prompt

[['A typical bee colony can contain between 20,000 and 80,000 bees.', 'Honey bees are social insects, living in colonies with a single queen, numerous worker bees, and drones.', 'Common pests and diseases that affect bees include varroa mites, hive beetles, and foulbrood.', 'Beekeepers must inspect their hives regularly to check for diseases, pests, and the overall health of the colony.', 'The queen bee can lay up to 2,000 eggs per day during peak seasons.']]

*(The embedding generation for the prompt and data retrieval took **0.1s**)*

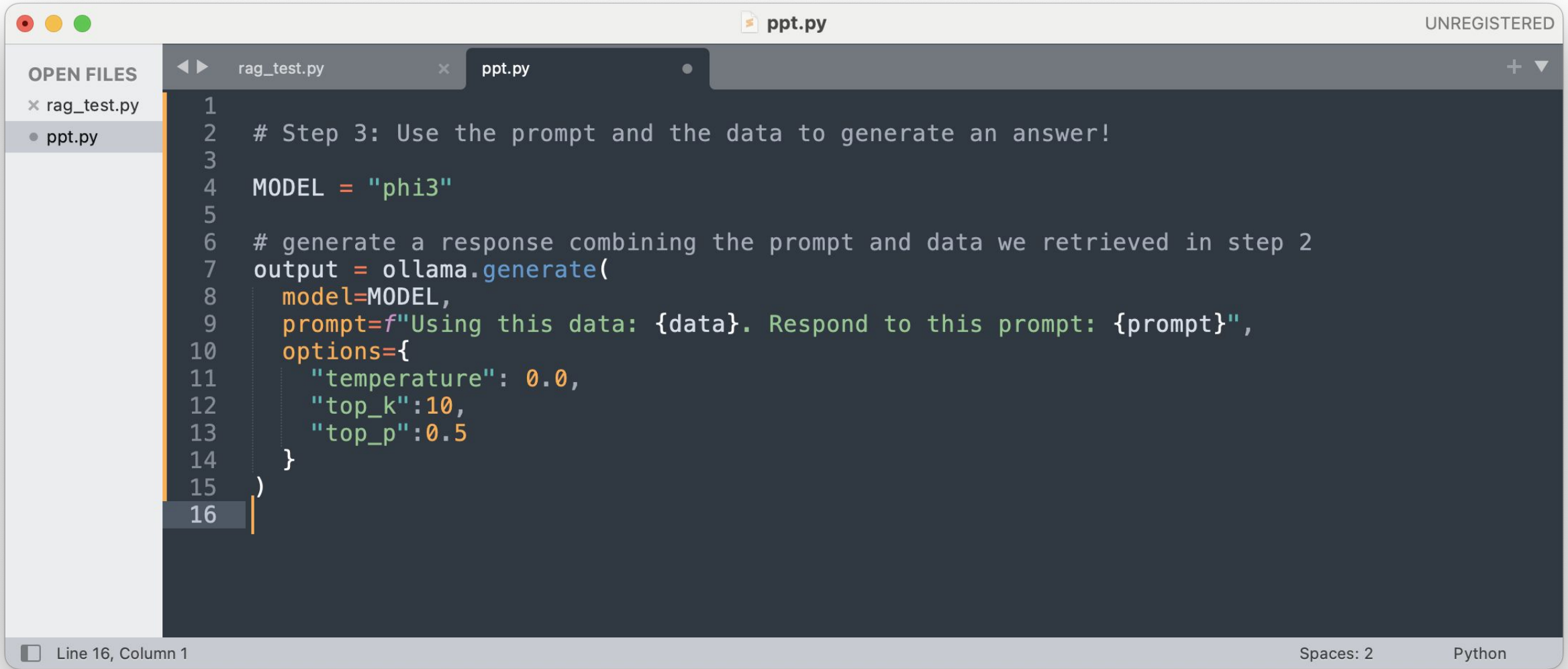

ppt.py
UNREGISTERED
OPEN FILES
rag_test.py
ppt.py
rag_test.py
ppt.py
# Step 2: Retrieve the most relevant document given a prompt:
# Prompt
prompt = "How many bees are in a colony? Who lays eggs and wow much? How about common pests and diseases?"
# generate an embedding for the prompt and retrieve the most relevant doc
response = ollama.embeddings(
prompt=prompt,
model=EMB_MODEL
)
results = collection.query(
query_embeddings=[response["embedding"]],
n_results=5
)
data = results['documents']
Line 3, Column 1
Spaces: 2
Python



Step 3:

Use the prompt and the data to generate an answer!

*(The code for model: phi3, took **97.6s** to generate the answer)*

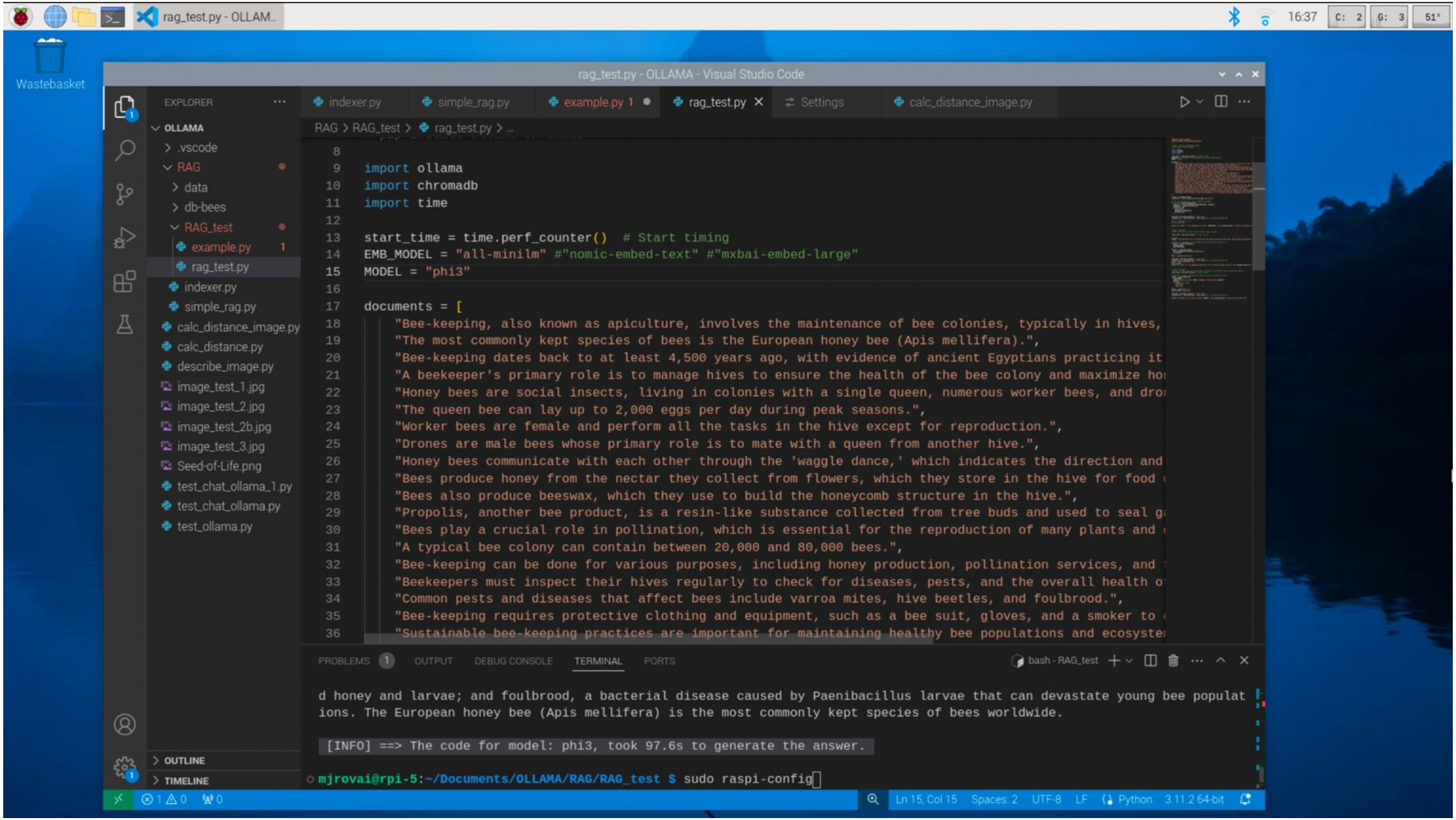

ppt.py
UNREGISTERED
OPEN FILES
rag_test.py
ppt.py
rag_test.py
ppt.py
# Step 3: Use the prompt and the data to generate an answer!
MODEL = "phi3"
# generate a response combining the prompt and data we retrieved in step 2
output = ollama.generate(
model=MODEL,
prompt=f"Using this data: {data}. Respond to this prompt: {prompt}",
options={
"temperature": 0.0,
"top_k":10,
"top_p":0.5
}
)
Line 16, Column 1
Spaces: 2
Python

Question:

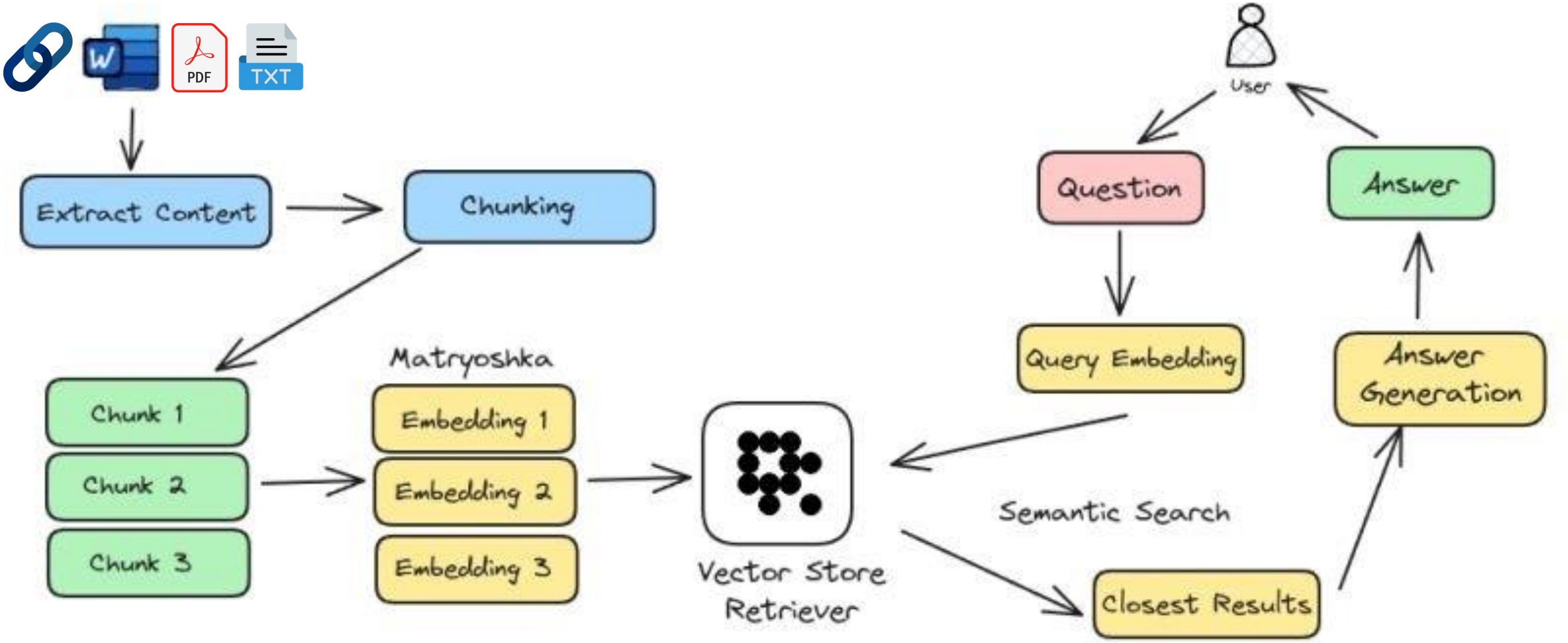
"How many bees are in a colony? Who lays eggs, and how much? How about common pests and diseases?"

Response

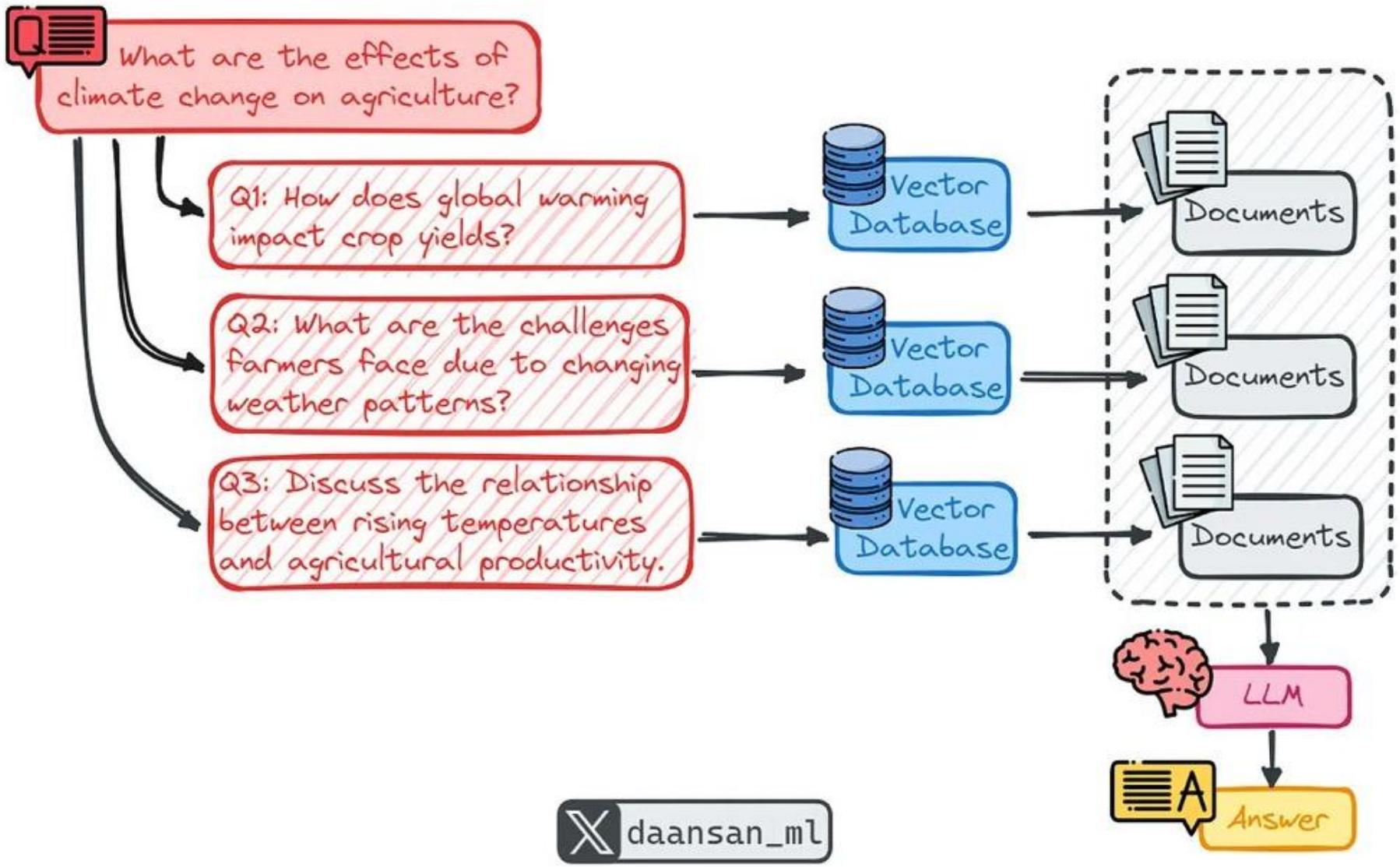
A typical bee colony contains between 20,000 and 80,000 bees. The queen bee is responsible for laying the majority of these eggs; she can produce up to 2,000 eggs per day during peak seasons. Beekeepers must regularly inspect their hives not only to monitor egg-laying but also to check for common pests and diseases that affect bees such as varroa mites, hive beetles, and foulbrood disease.

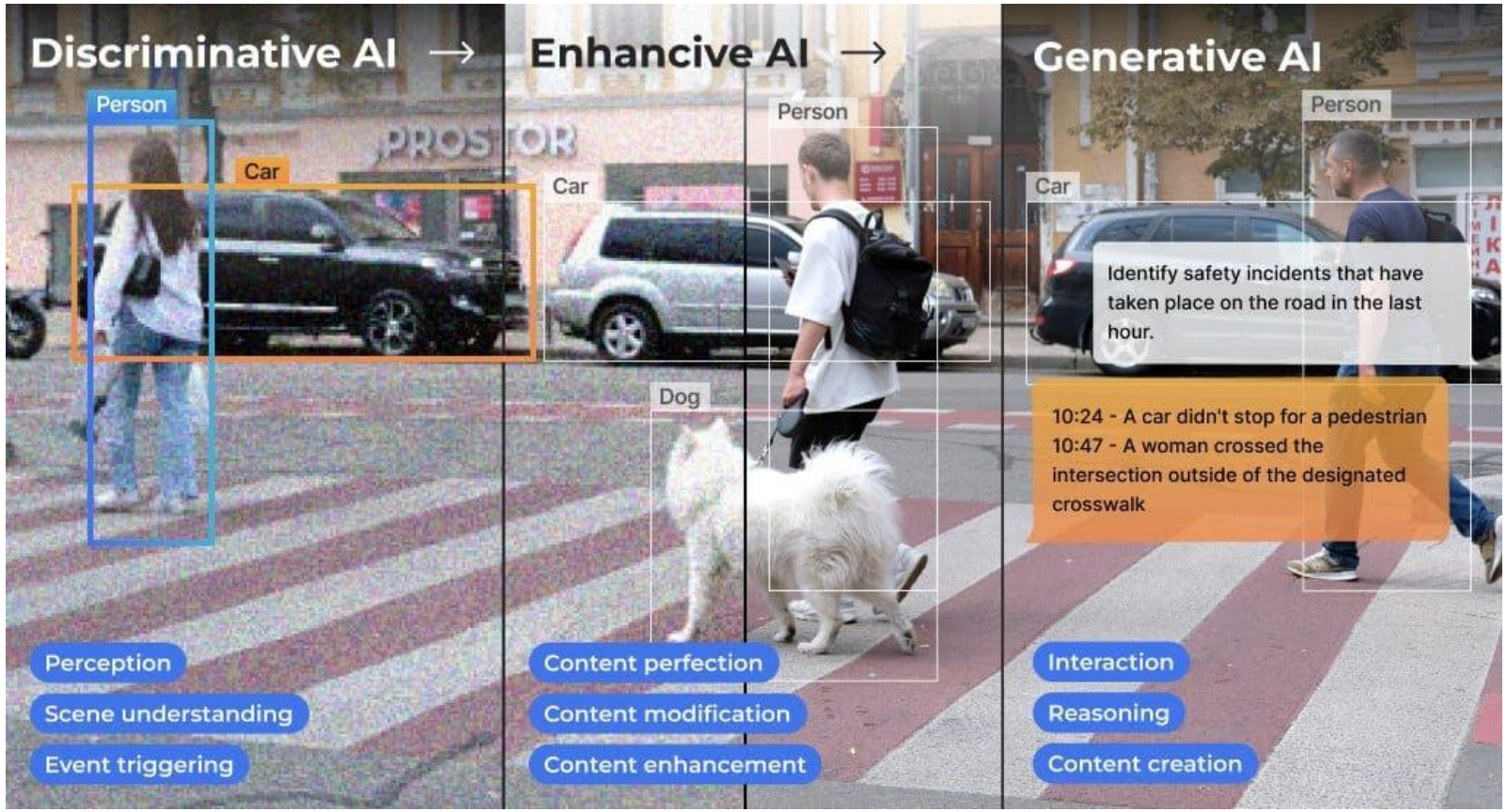

rag_test.py - OLLAM..
16:37
Wastebasket
rag_test.py - OLLAMA - Visual Studio Code
EXPLORER
OLLAMA
.vscode
RAG
data
db-bees
RAG_test
example.py
rag_test.py
indexer.py
simple_rag.py
calc_distance_image.py
calc_distance.py
describe_image.py
image_test_1.jpg
image_test_2.jpg
image_test_2b.jpg
image_test_3.jpg
Seed-of-Life.png
test_chat_ollama_1.py
test_chat_ollama.py
test_ollama.py
OUTLINE
TIMELINE
indexer.py
simple_rag.py
example.py 1
rag_test.py
Settings
calc_distance_image.py
RAG > RAG_test > rag_test.py > ...
import ollama
import chromadb
import time
start_time = time.perf_counter() # Start timing
EMB_MODEL = "all-minilm" #"nomic-embed-text" #"mxbai-embed-large"
MODEL = "phi3"
documents = [
"Bee-keeping, also known as apiculture, involves the maintenance of bee colonies, typically in hives,
"The most commonly kept species of bees is the European honey bee (Apis mellifera).",
"Bee-keeping dates back to at least 4,500 years ago, with evidence of ancient Egyptians practicing it
"A beekeeper's primary role is to manage hives to ensure the health of the bee colony and maximize ho
"Honey bees are social insects, living in colonies with a single queen, numerous worker bees, and dro
"The queen bee can lay up to 2,000 eggs per day during peak seasons.",
"Worker bees are female and perform all the tasks in the hive except for reproduction.",
"Drones are male bees whose primary role is to mate with a queen from another hive.",
"Honey bees communicate with each other through the 'waggle dance,' which indicates the direction and
"Bees produce honey from the nectar they collect from flowers, which they store in the hive for food
"Bees also produce beeswax, which they use to build the honeycomb structure in the hive.",
"Propolis, another bee product, is a resin-like substance collected from tree buds and used to seal g
"Bees play a crucial role in pollination, which is essential for the reproduction of many plants and
"A typical bee colony can contain between 20,000 and 80,000 bees.",
"Bee-keeping can be done for various purposes, including honey production, pollination services, and
"Beekeepers must inspect their hives regularly to check for diseases, pests, and the overall health o
"Common pests and diseases that affect bees include varroa mites, hive beetles, and foulbrood.",
"Bee-keeping requires protective clothing and equipment, such as a bee suit, gloves, and a smoker to
"Sustainable bee-keeping practices are important for maintaining healthy bee populations and ecosyste
PROBLEMS 1
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
bash - RAG_test
d honey and larvae; and foulbrood, a bacterial disease caused by Paenibacillus larvae that can devastate young bee populat
ions. The European honey bee (Apis mellifera) is the most commonly kept species of bees worldwide.
[INFO] ==> The code for model: phi3, took 97.6s to generate the answer.
mjrovai@rpi-5:~/Documents/OLLAMA/RAG/RAG_test $ sudo raspi-config
Ln 15, Col 15
Spaces: 2
UTF-8
LF
Python
3.11.2 64-bit

RAG: Simple Query



Advanced RAG: Multi Query



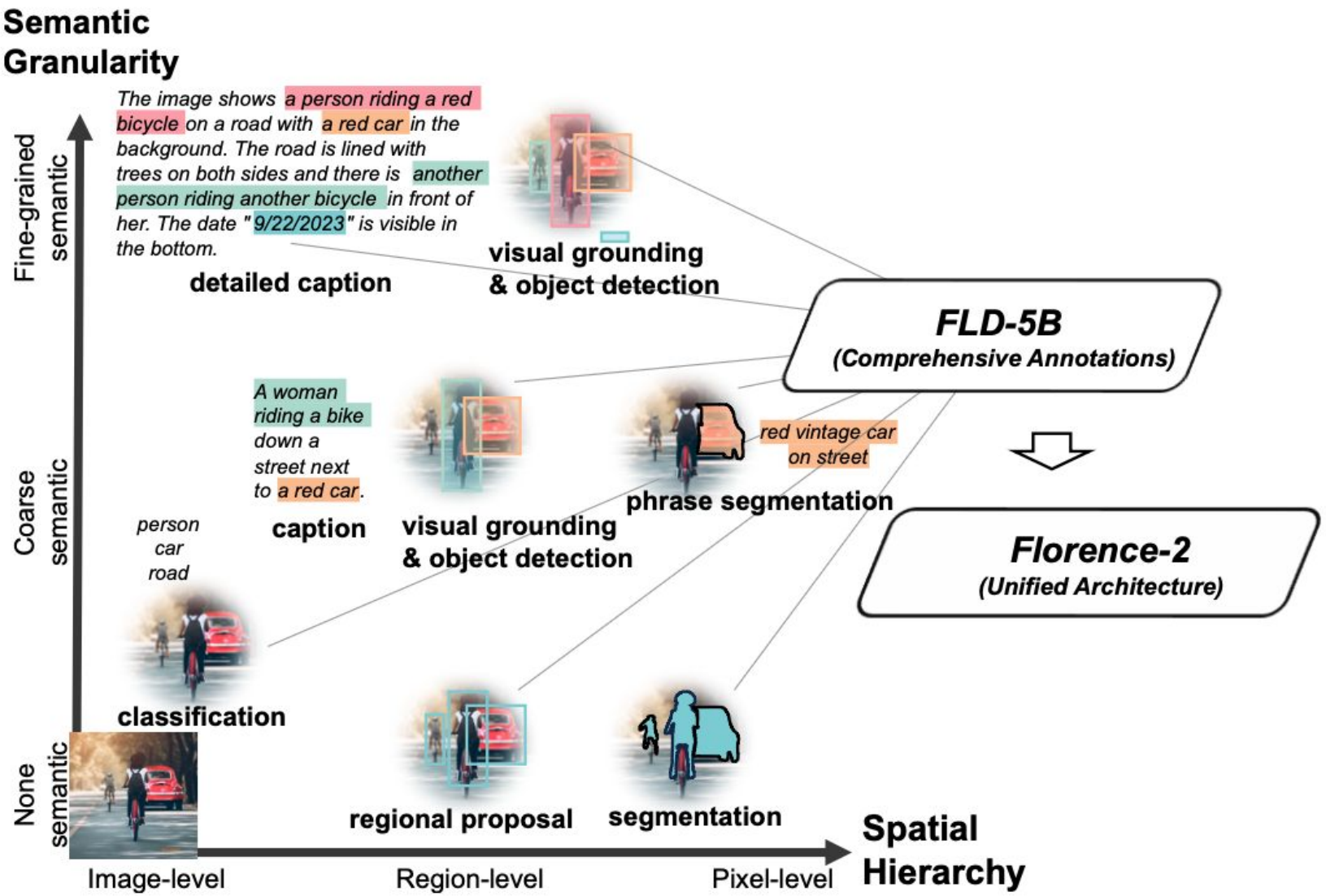


“In the vast landscape of artificial intelligence (AI), one of the most intriguing journeys has been the evolution of AI on the edge. This journey has taken us from classic machine vision to the realms of discriminative AI, enhancive AI, and now, the groundbreaking frontier of generative AI. Each step has brought us closer to a future where intelligent systems seamlessly integrate with our daily lives, offering an immersive experience of not just perception but also creation at the palm of our hand.”

Avi Baum, CTO at Hailo

Florence-2

Advancing a Unified Representation for a Variety of Vision Tasks



Paper: https://arxiv.org/abs/2311.06242

Image Encoder

What does the image describe?

Locate the objects in the image.

Locate the phrases in the caption: **A woman riding a bike**.

What does the **region (0.41, 0.15, 0.63, 0.73)** describe?

What is the polygon mask of region **(0.41, 0.15, 0.63, 0.73)**?

Multi-task prompts

visual embeddings

text + location embeddings

Transformer Encoders

Transformer Decoders

text + location tokens

The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.

person (0.41, 0.15, 0.63, 0.73) … car (0.58, 0.26, 0.89, 0.61)

A women riding a bike (0.41, 0.15, 0.63, 0.73

person riding red bicycle on road

(0.48, 0.19, 0.48, 0.18, 0.49, 0.17, ...)

microsoft/BitNet

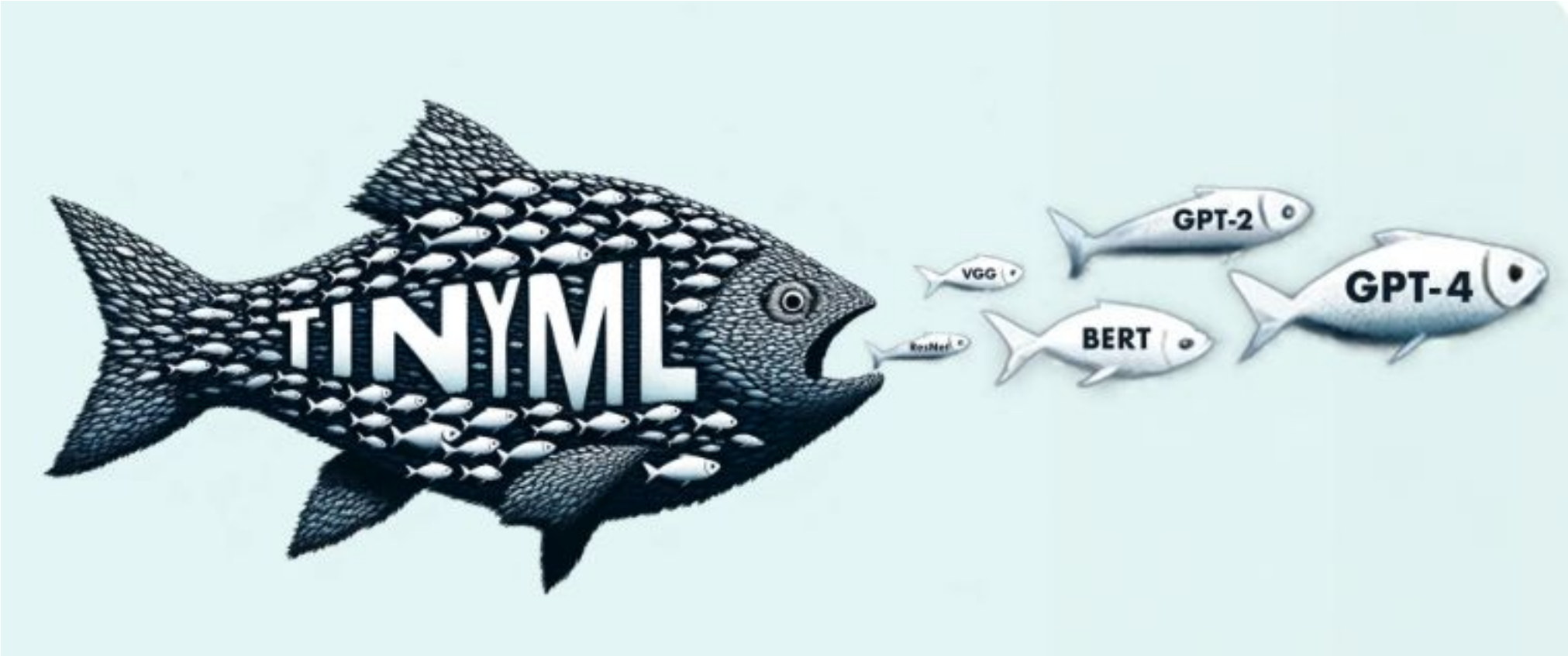
Official inference framework for 1-bit LLMs

Bitnet.cpp employs one-bit quantization, representing values with a ternary system (+1, -1, 0). This approach simplifies calculations by replacing complex multiplications with additions and subtractions, eliminating the need for GPUs.

- Speedups range from 1.37x to 6.1x on various CPUs.
- Power consumption reductions between 55.4% and 82.2% compared to traditional GPU-based inference.

bitnet.cpp

TinyML: Why the Future of Machine Learning is Tiny and Bright



Shvetank Prakash, Emil Njor, Colby Banbury, Matthew Stewart, Vijay Janapa Reddi

Questions?