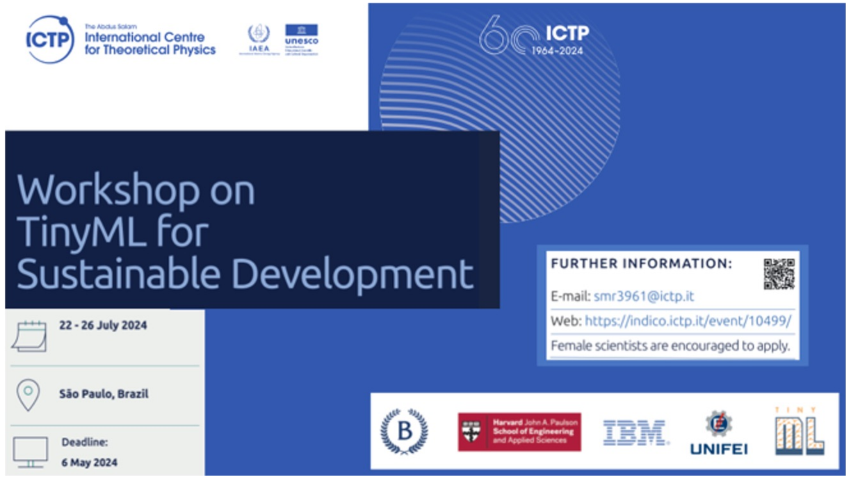


The Abdus Salam International Centre for Theoretical Physics

ICTP 1964-2024

Workshop on TinyML for Sustainable Development

22 - 26 July 2024

São Paulo, Brazil

Deadline: 6 May 2024

FURTHER INFORMATION:

E-mail: smr3961@ictp.it

Web: https://indico.ictp.it/event/10499/

Female scientists are encouraged to apply.

Large Language Models (LLMs) at the Edge

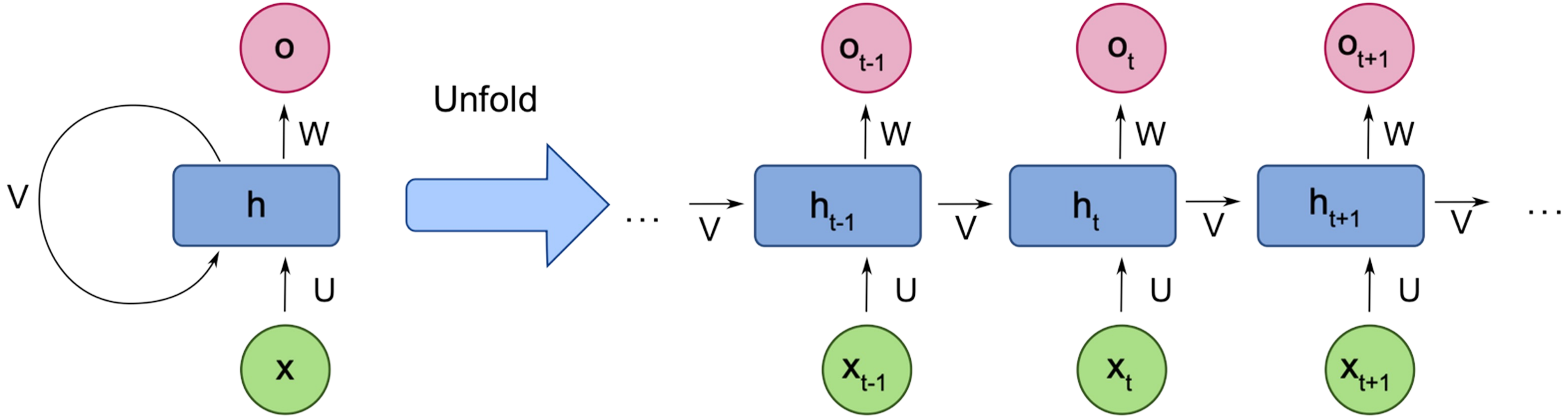
Prof. Marcelo J. Rovai
rovai@unifei.edu.br

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

TINYML4D

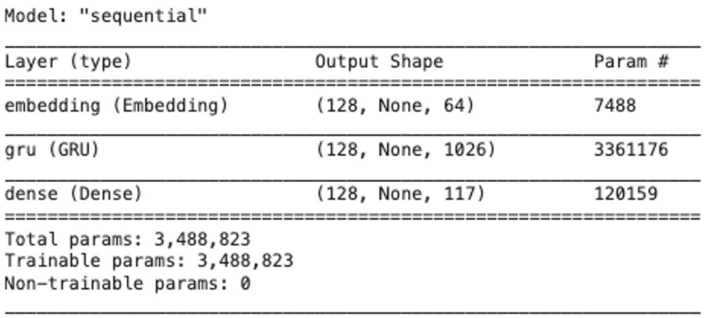
Deep Learning models (or artificial neural networks)

Recurrent Neural Networks (RNNs): Designed for **sequential data like time series or text**, these networks use their internal state (memory) to process sequences of inputs.

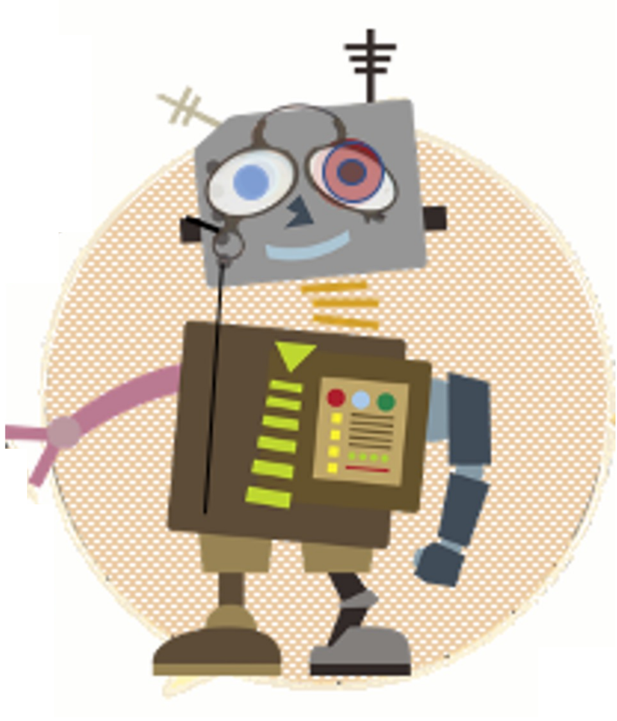


Machado de Assis Bot with RNN - GRU

The *robot writer* model is a **Recurrent Neural network (RNN —GRU)**. To obtain the final AI model, 3.5 million parameters were trained with a **120-letter 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)        (128, None, 64)           7488
_________________________________________________________________
gru (GRU)                    (128, None, 1026)         3361176
_________________________________________________________________
dense (Dense)                (128, None, 117)          120159
=================================================================
Total params: 3,488,823
Trainable params: 3,488,823
Non-trainable params: 0
_________________________________________________________________
```



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.

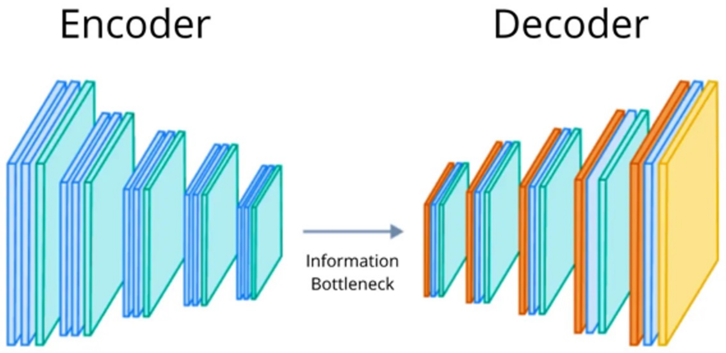
LLM / SLM

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 models like LSTMs, which surpass them in handling long-range dependencies and parallel processing efficiency.

Prof. Jesus's Presentation about IA: 🎬 📇

Transformers to LLMs and SLMs

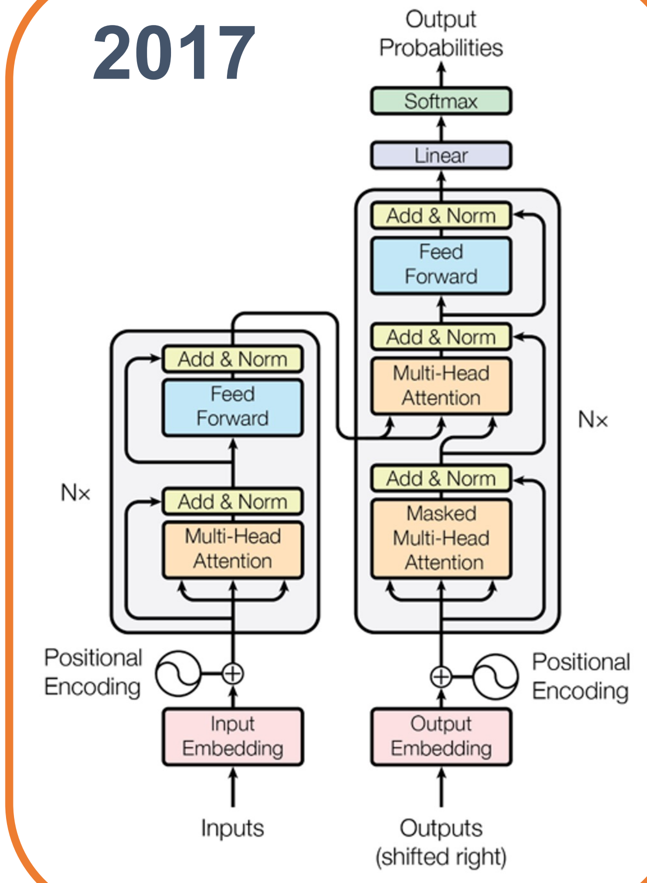
Open

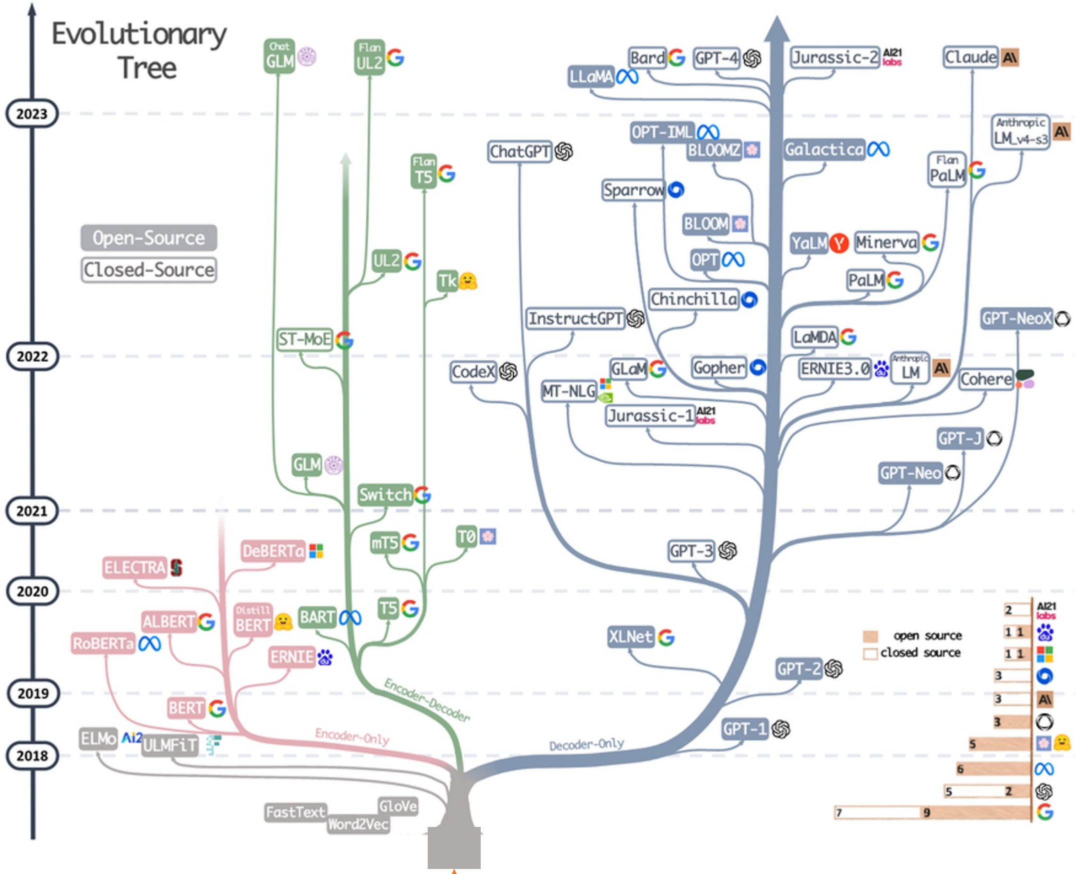
Phi | Mixtral | LlaMa

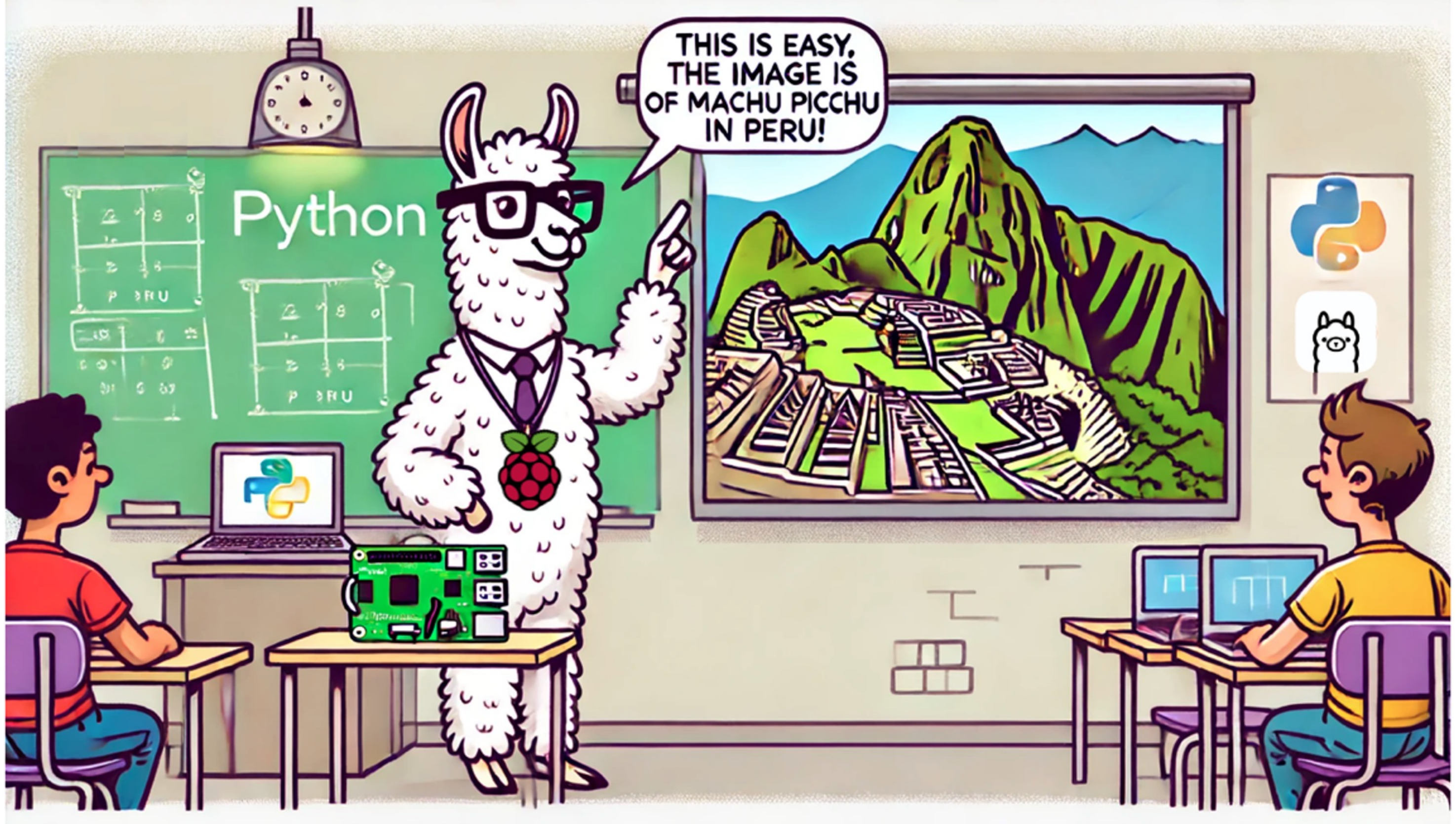
Closed

Gemini | GPT | Claude

2017

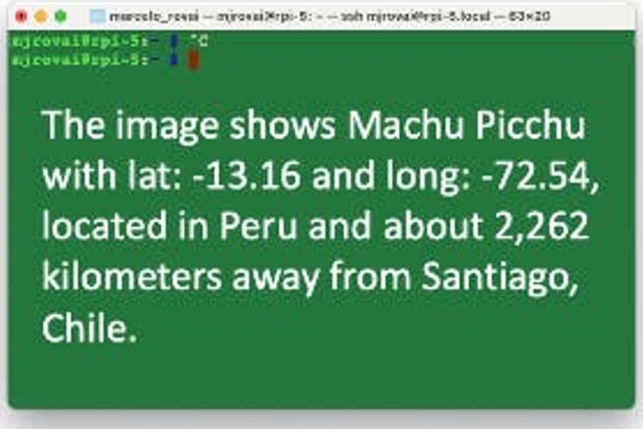






Running Large Language Models on Raspberry Pi at the Edge

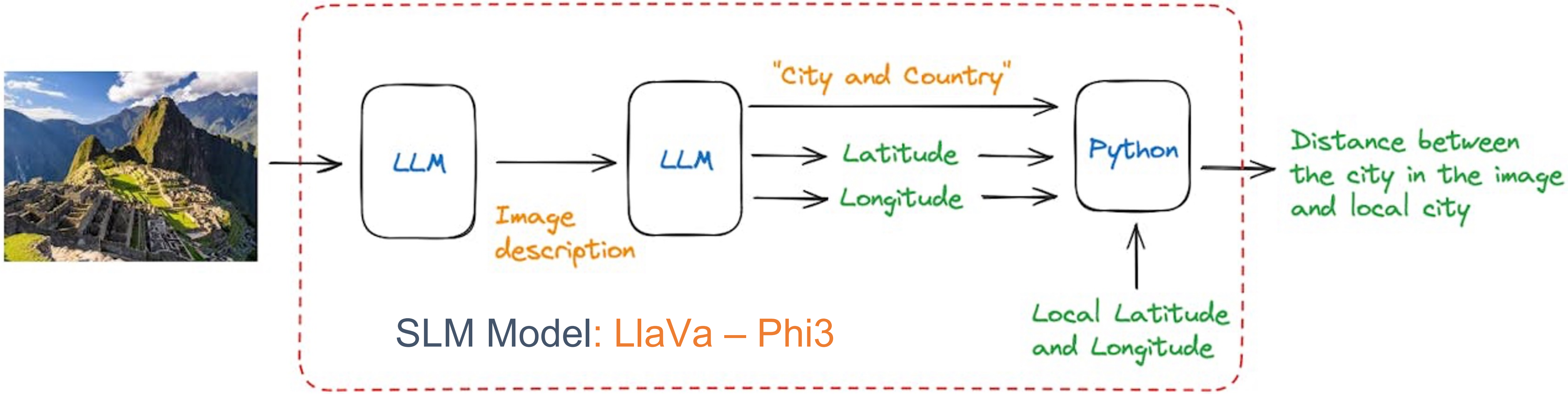
Transform a Raspberry Pi into a powerful AI hub, running SLMs for real-time, on-site data analysis and insights using Ollama and Python.



marcelo_rovai — mjrovai@rpi-5: ~ — ssh mjrovai@rpi-5.local — 63×20

```
mjrovai@rpi-5:~ $ ^C
mjrovai@rpi-5:~ $
```

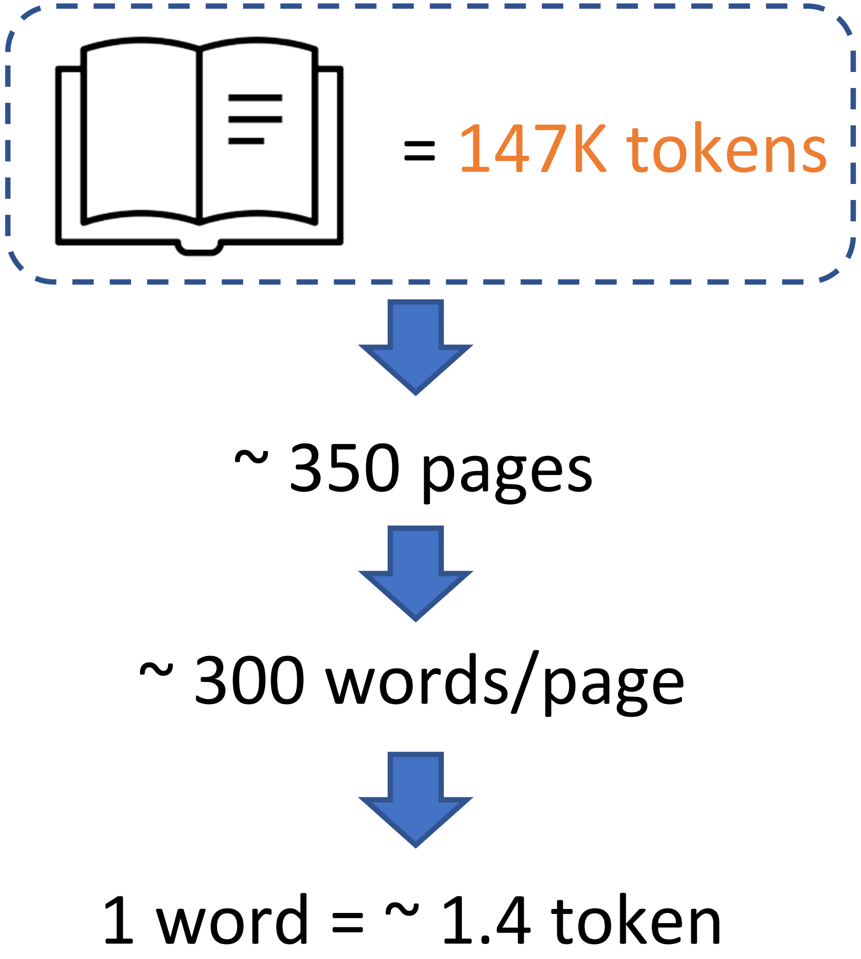
The image shows Machu Picchu with lat: -13.16 and long: -72.54, located in Peru and about 2,262 kilometers away from Santiago, Chile.



LLM
Image description
LLM
"City and Country"
Latitude
Longitude
Python
Local Latitude and Longitude
Distance between the city in the image and local city

SLM Model: LlaVa – Phi3

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

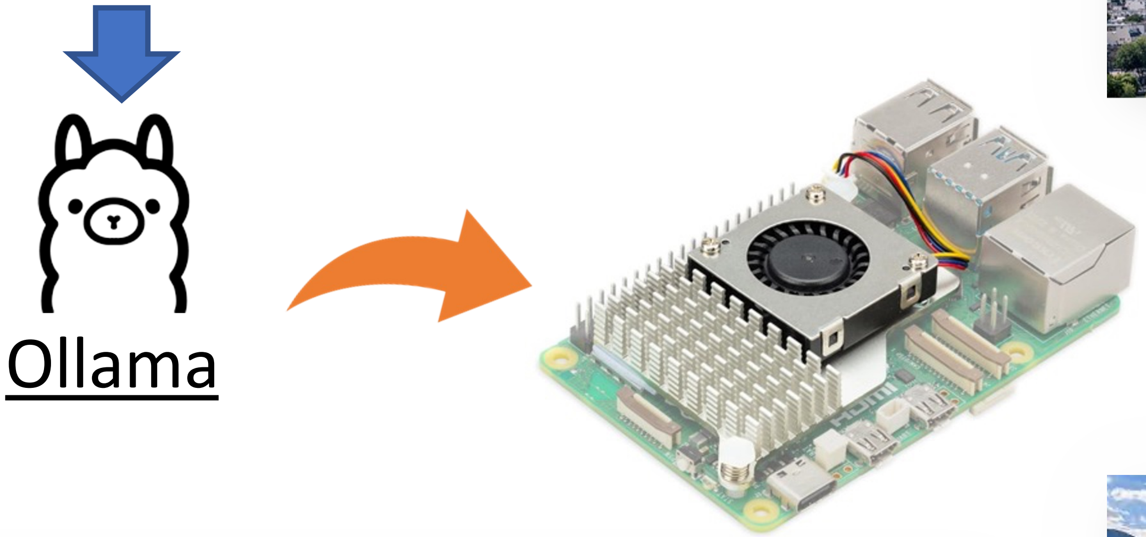
1 word = ~ 1.4 token

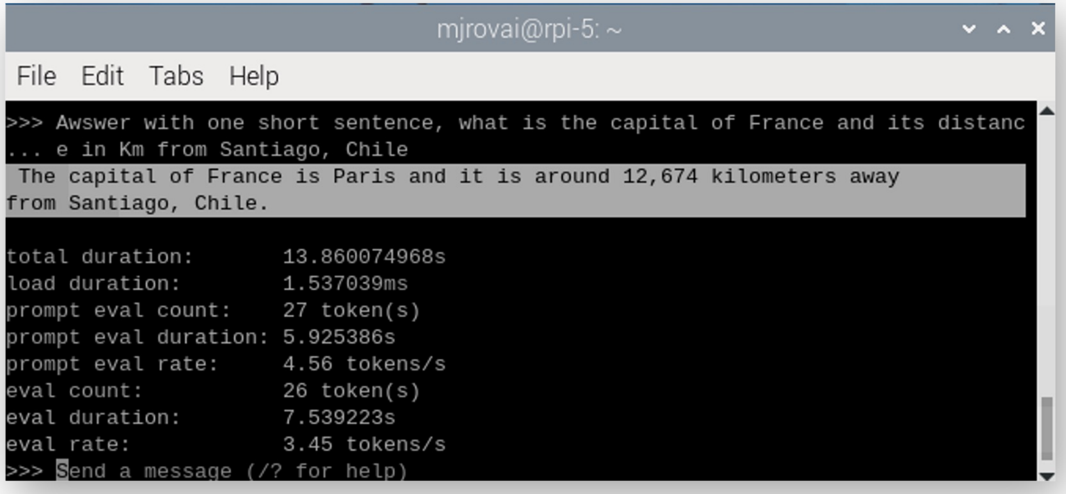
A **4-bit** quantized 3.8 billion parameter * language model trained on 3.3 trillion tokens**, whose overall performance, as measured by both academic benchmarks and internal testing, rivals that of models such as Mixtral 8x7B and GPT-3.5

* 2.4 GB

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

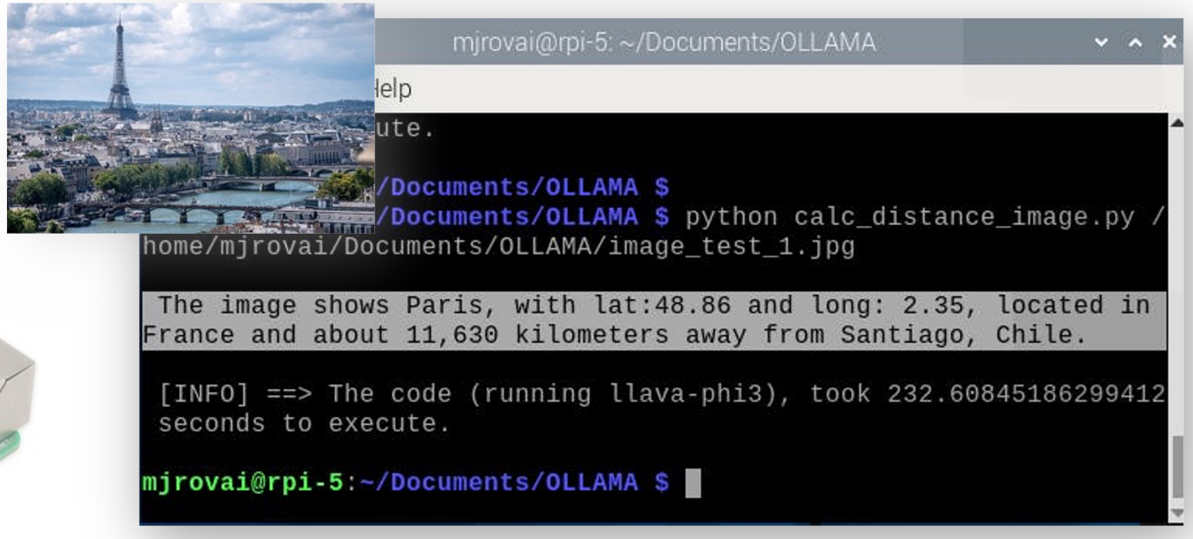
llava-phi-3 (2.9 GB)



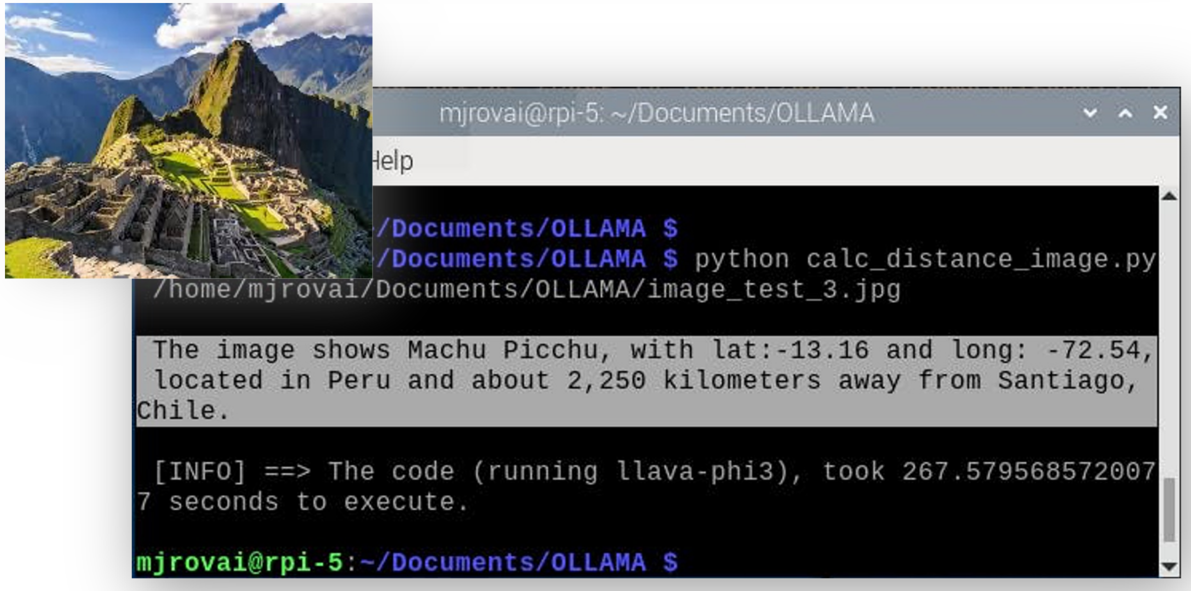


```
>>> Awswer with one short sentence, what is the capital of France and its distanc
... e in Km from Santiago, Chile
 The capital of France is Paris and it is around 12,674 kilometers away
from Santiago, Chile.

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)



```
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 $
```



```
/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)

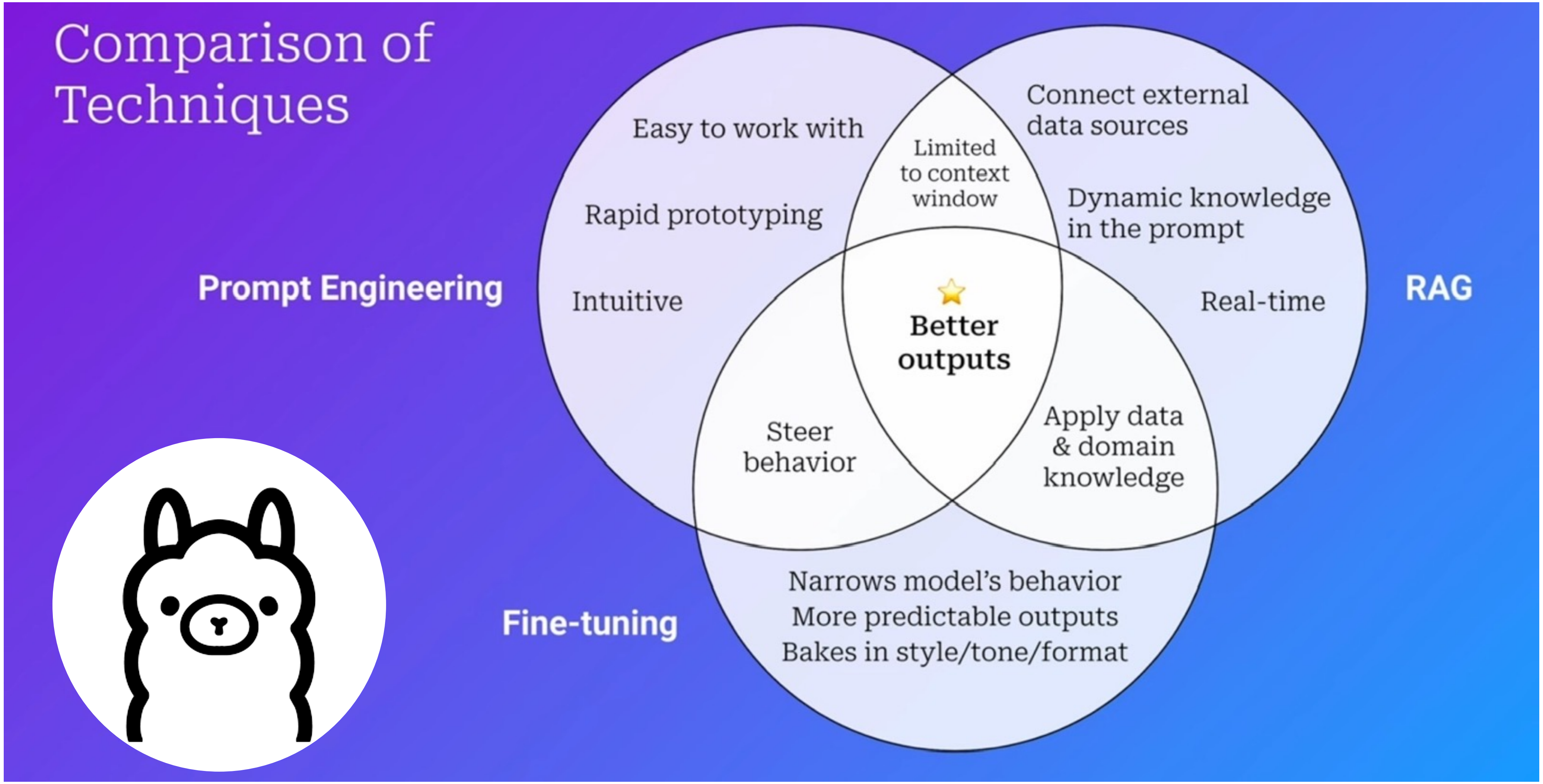
LLMs: Optimization Techniques

LLMs: Optimization Techniques

1. **Prompt Engineering**: Tailor your interactions.

2. **RAG**: Enhance with relevant data.

3. **Fine-tuning**: Perfect the model's tasks.



Comparison of Techniques

Prompt Engineering

- Easy to work with
- Rapid prototyping
- Intuitive

RAG

- Connect external data sources
- Dynamic knowledge in the prompt
- Real-time

Fine-tuning

- Narrows model's behavior
- More predictable outputs
- Bakes in style/tone/format

Limited to context window

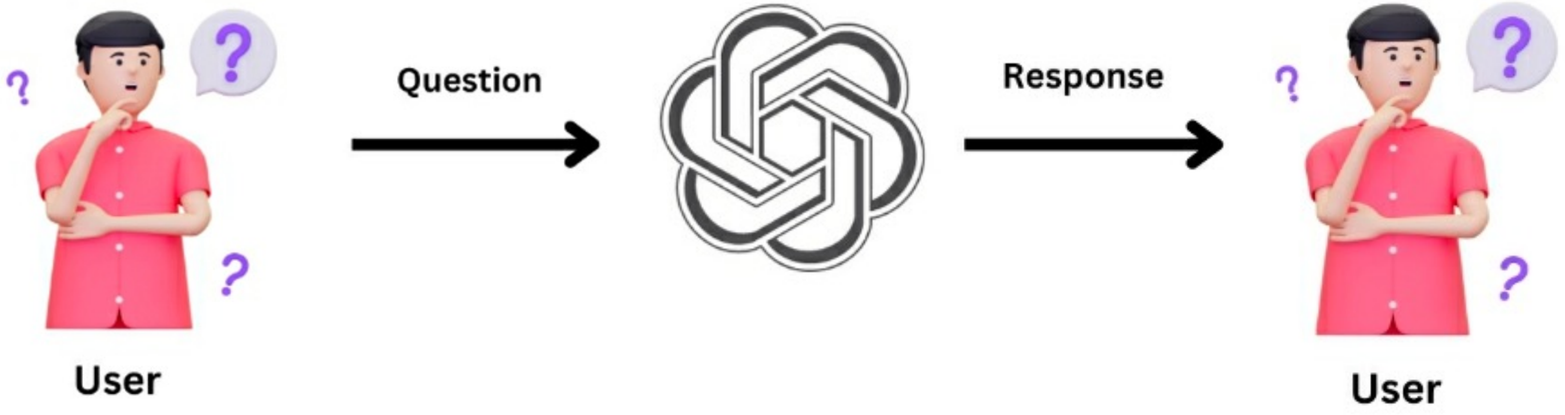
Steer behavior

Apply data & domain knowledge

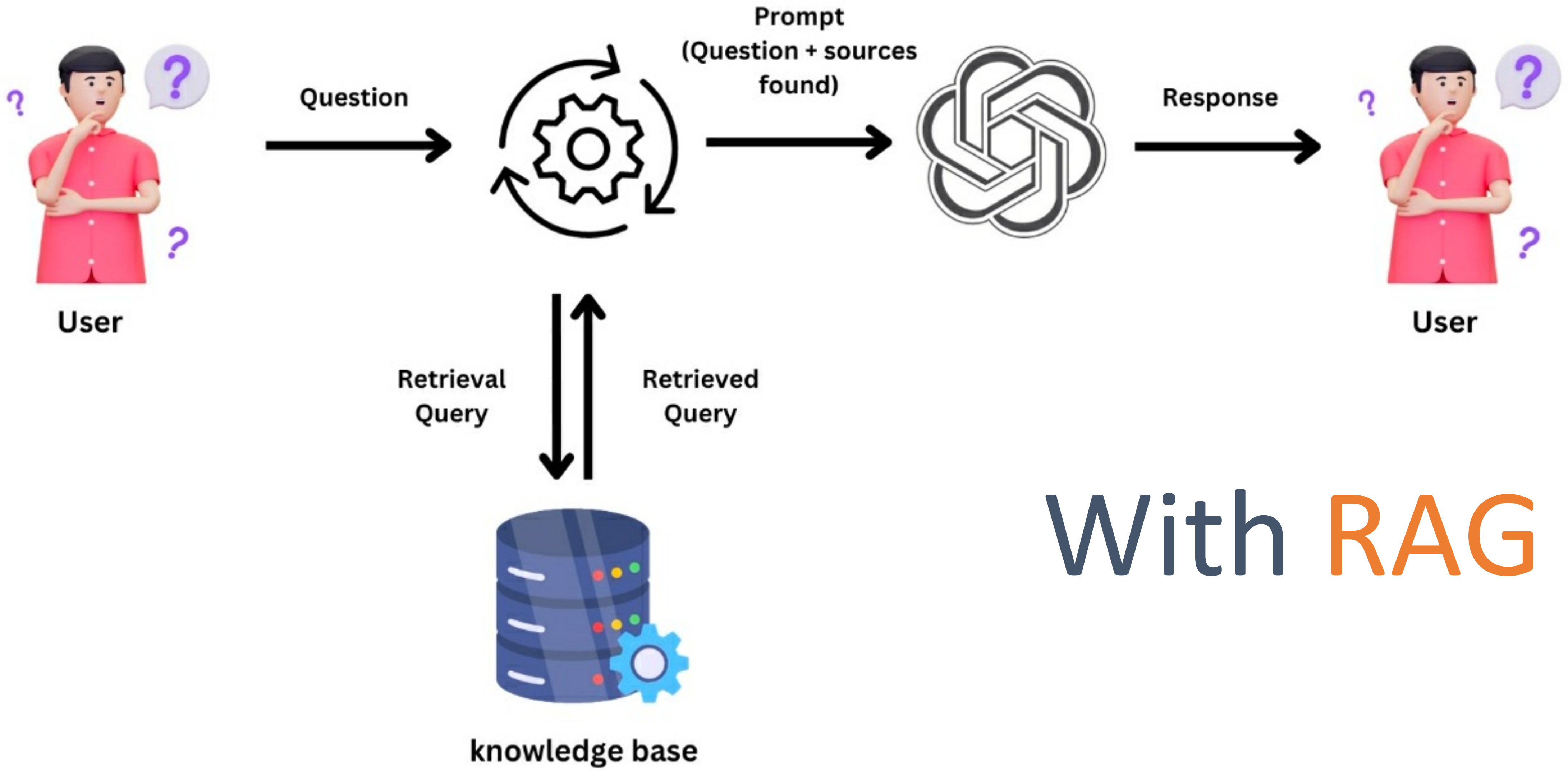
⭐ Better outputs

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.”

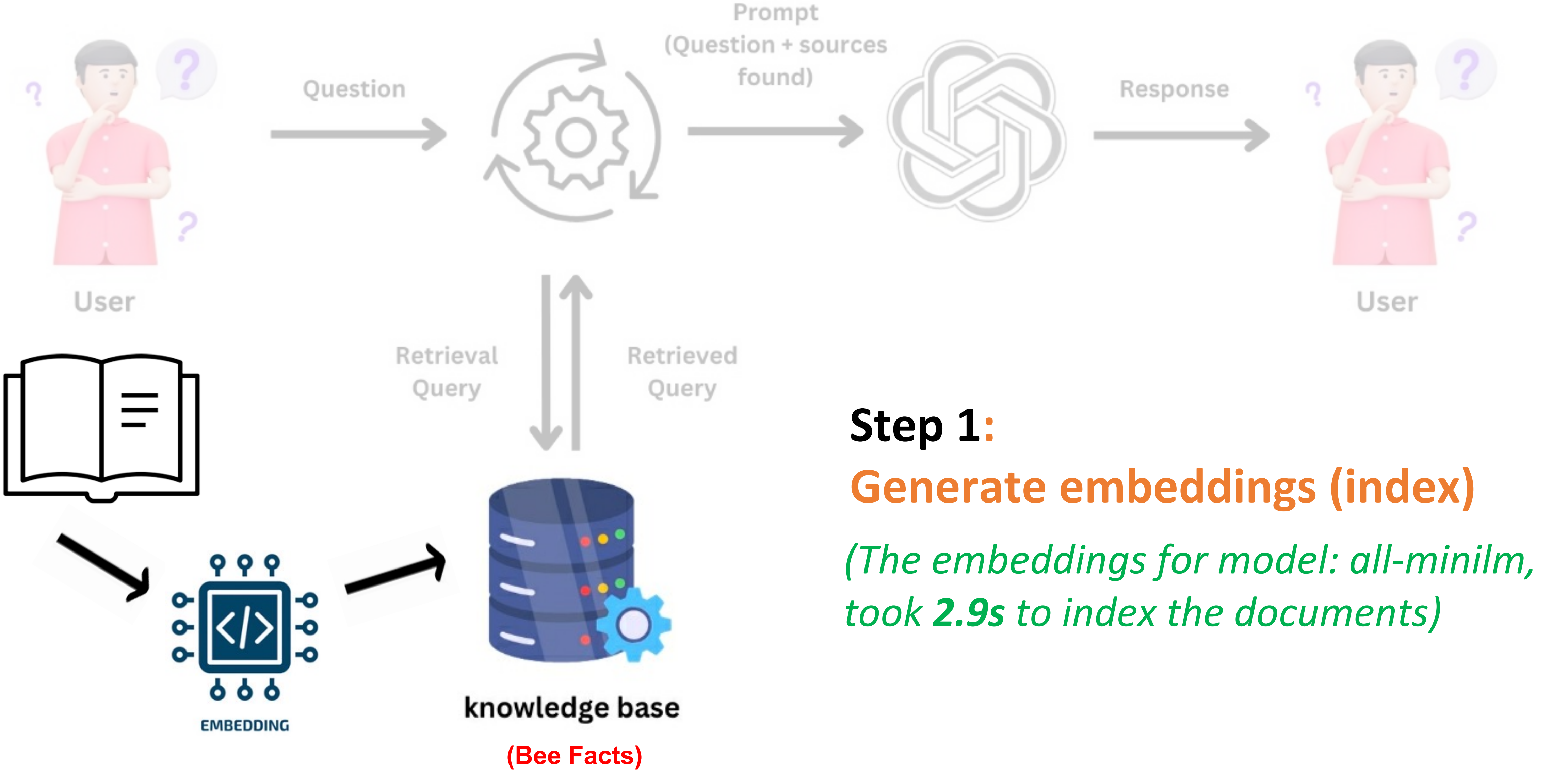


Usual Prompt



With RAG

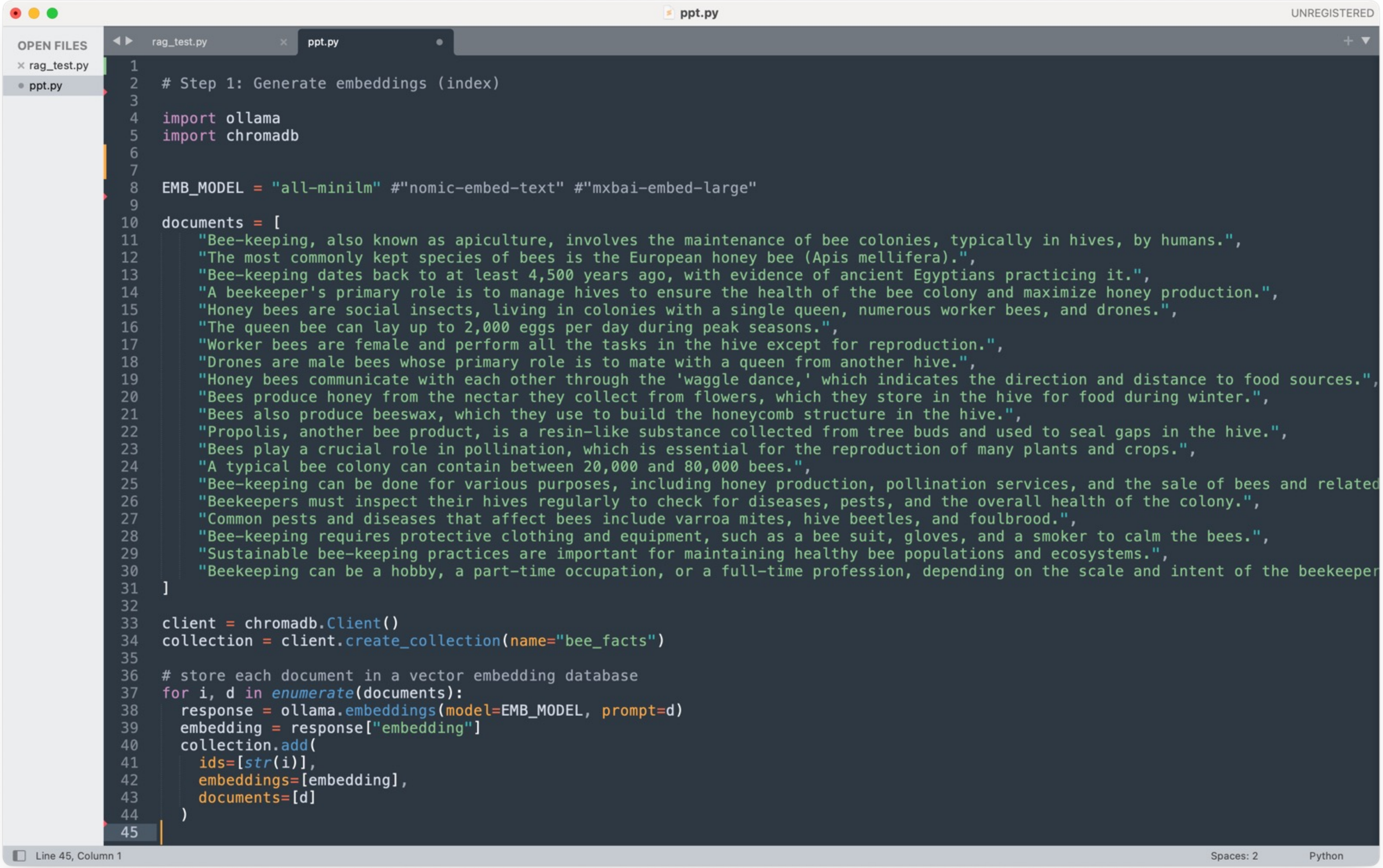
Bouchard, Louis-François; Peters, Louie . Building LLMs for Production: Enhancing LLM Abilities and Reliability with Prompting, Fine-Tuning, and RAG (p. 282). Kindle Edition.



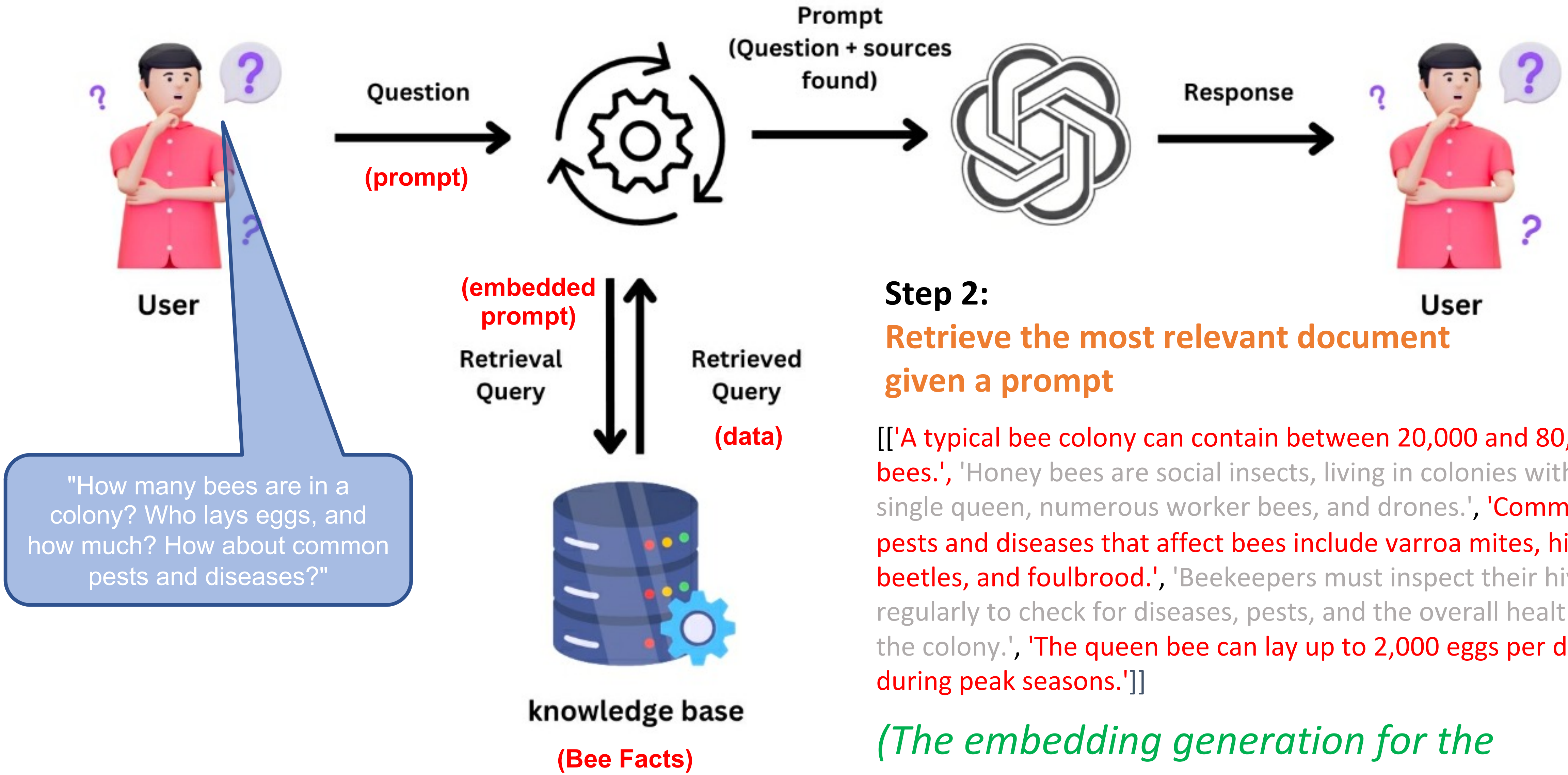
Step 1:

Generate embeddings (index)

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



```python

# 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]
  )

```



User

Question

(prompt)

Prompt (Question + sources found)

Response

User

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

(embedded prompt)

Retrieval Query

Retrieved Query

(data)

knowledge base

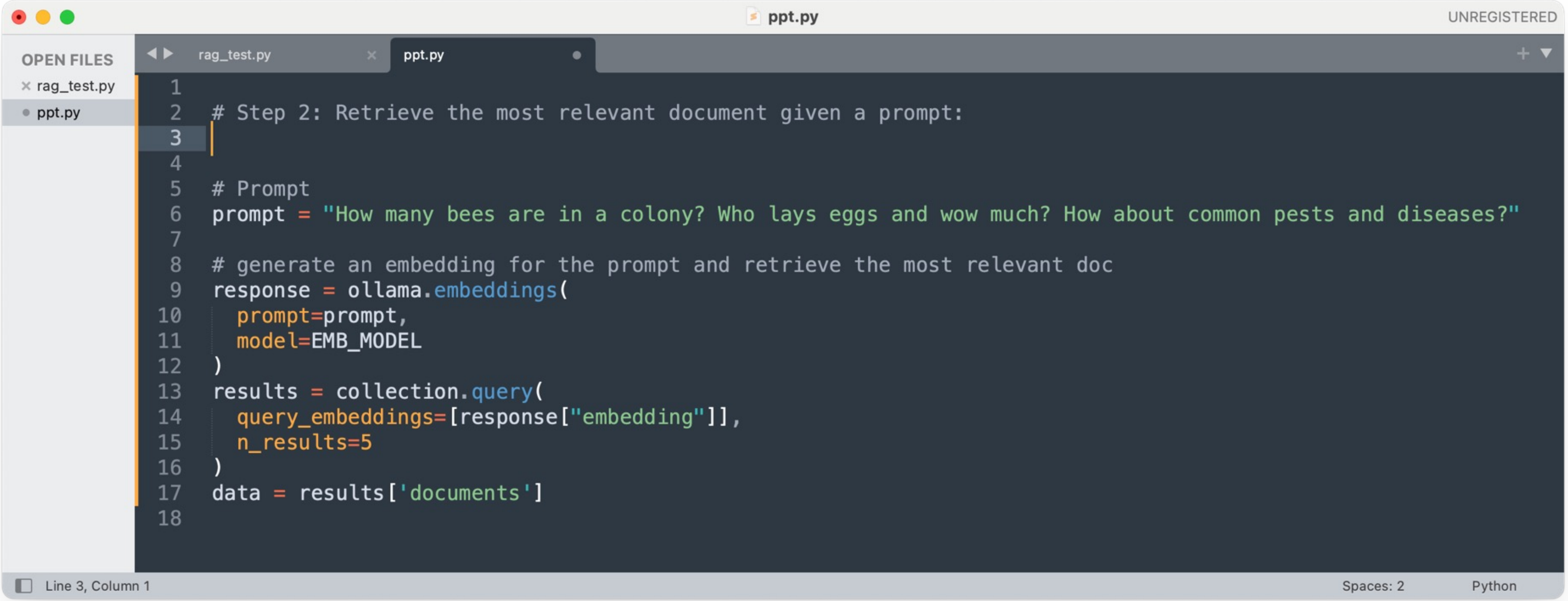
(Bee Facts)

Step 2:

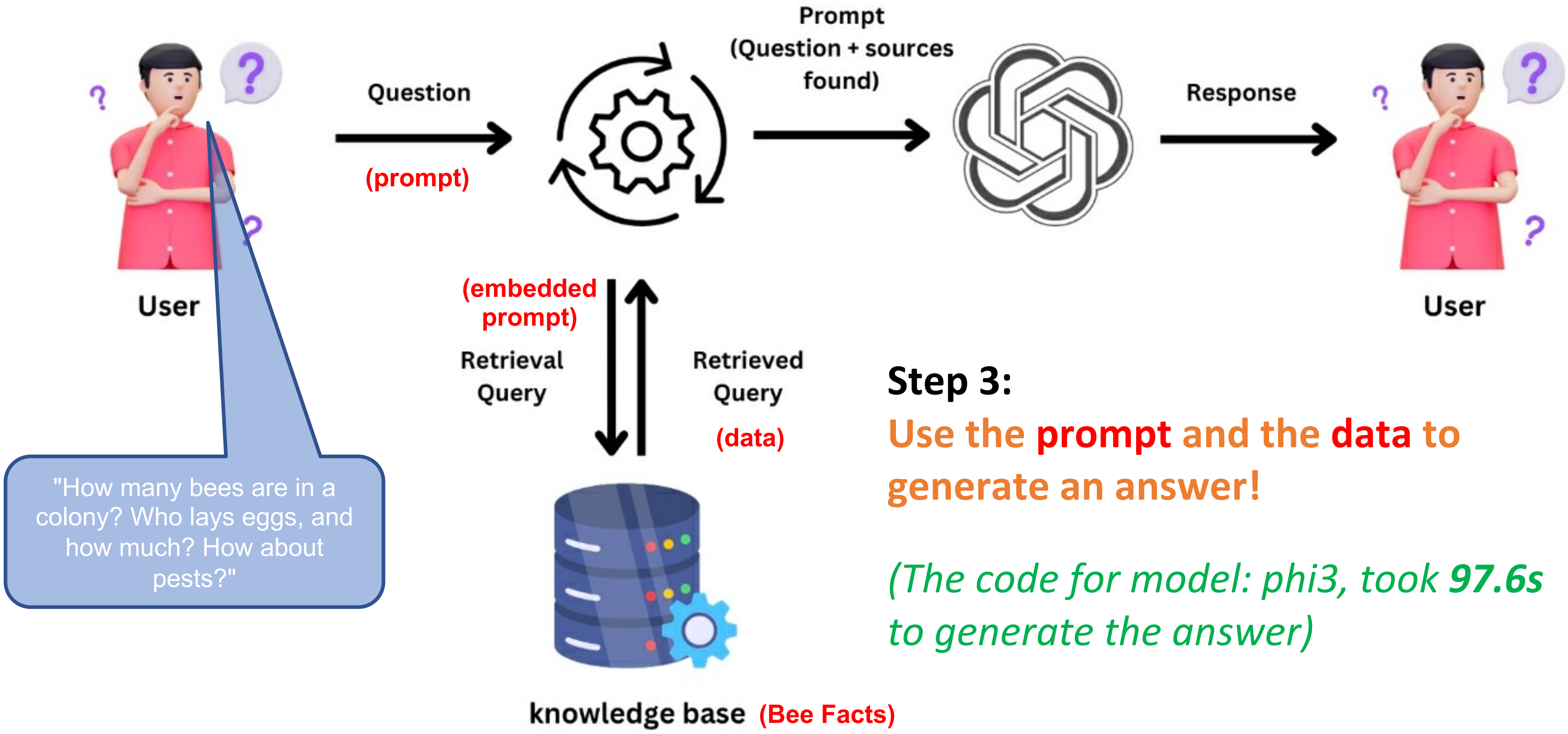
Retrieve the most relevant document given a prompt

[['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**)*



```python

# 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']

```

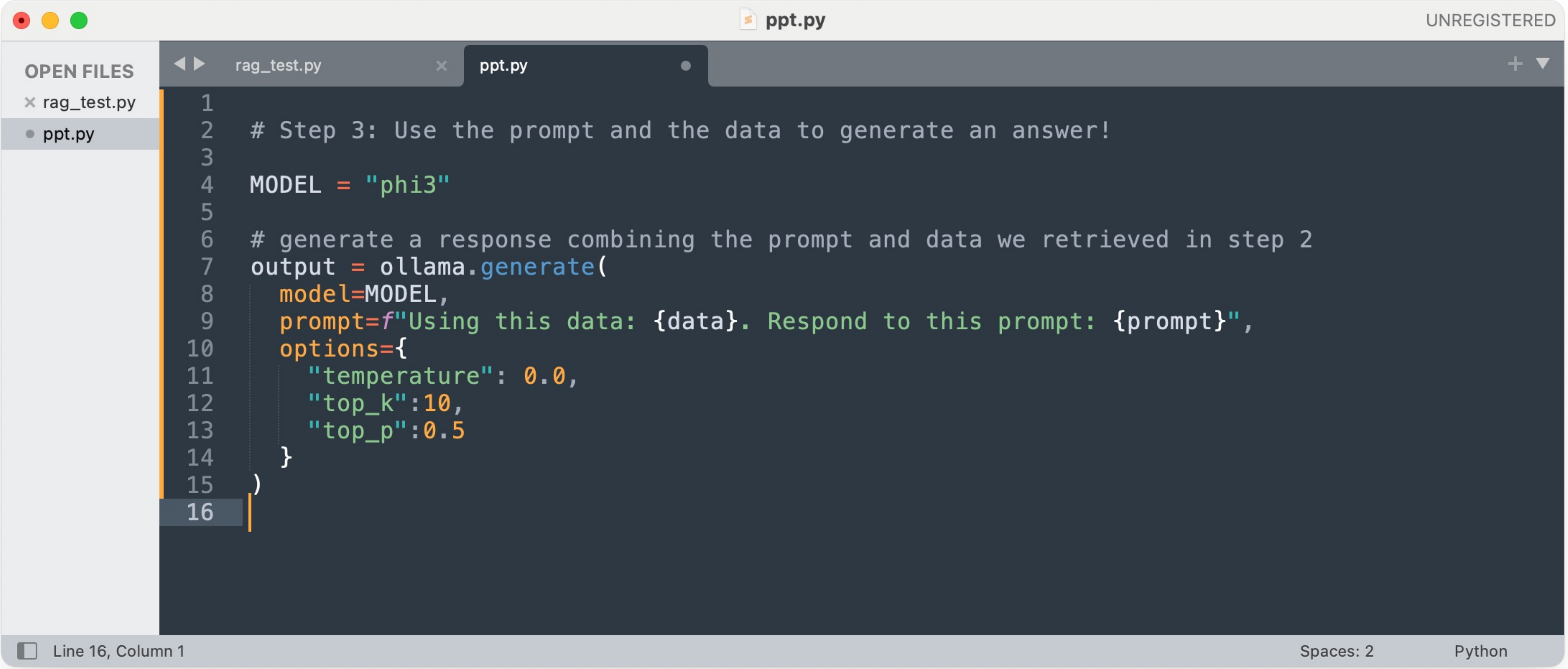


"How many bees are in a colony? Who lays eggs, and how much? How about pests?"

Step 3:

Use the prompt and the data to generate an answer!

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



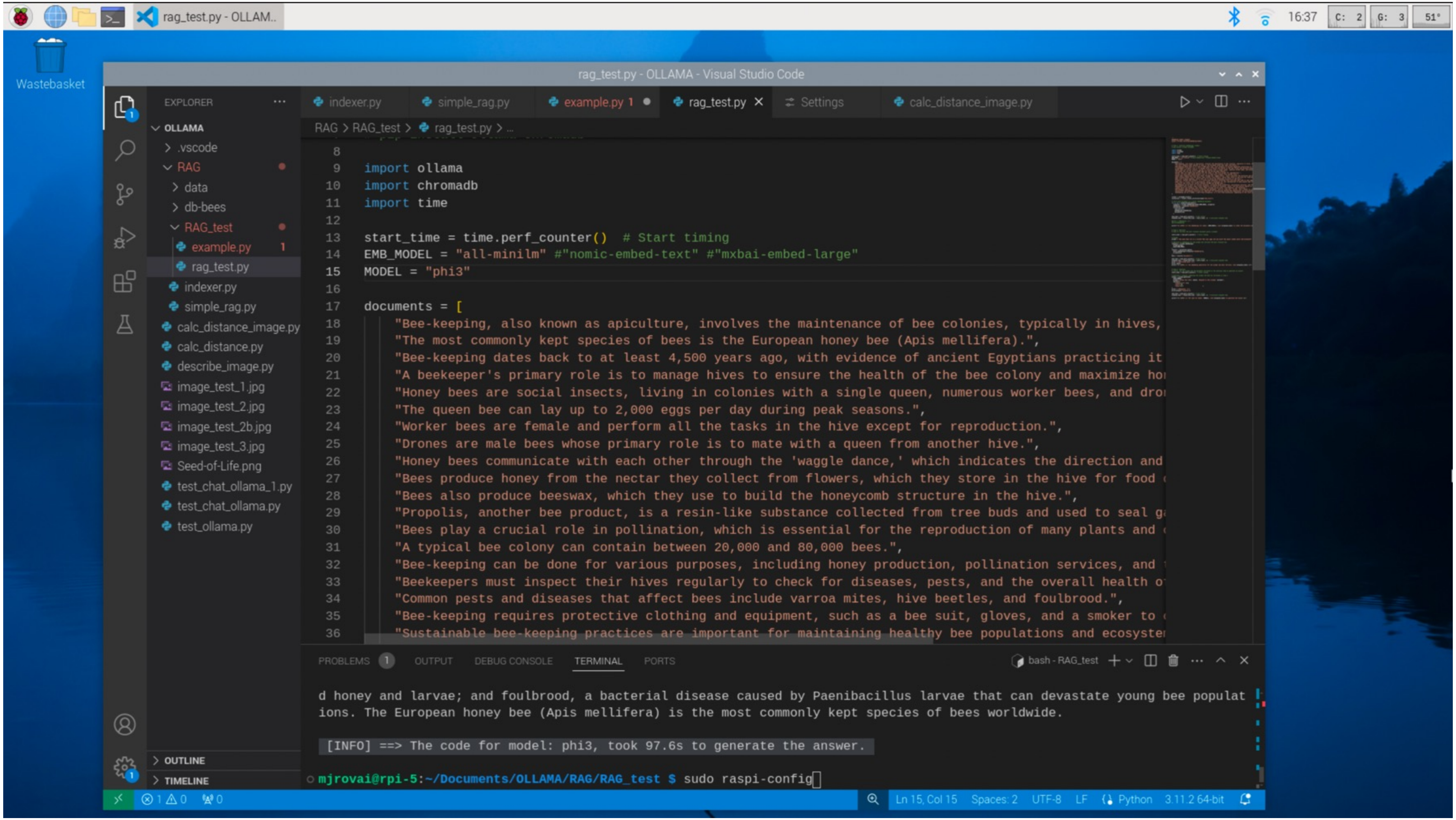
```python
# Step 3: Use the prompt and the data to generate an answer!

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
  }
)
```

Question:

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

Response

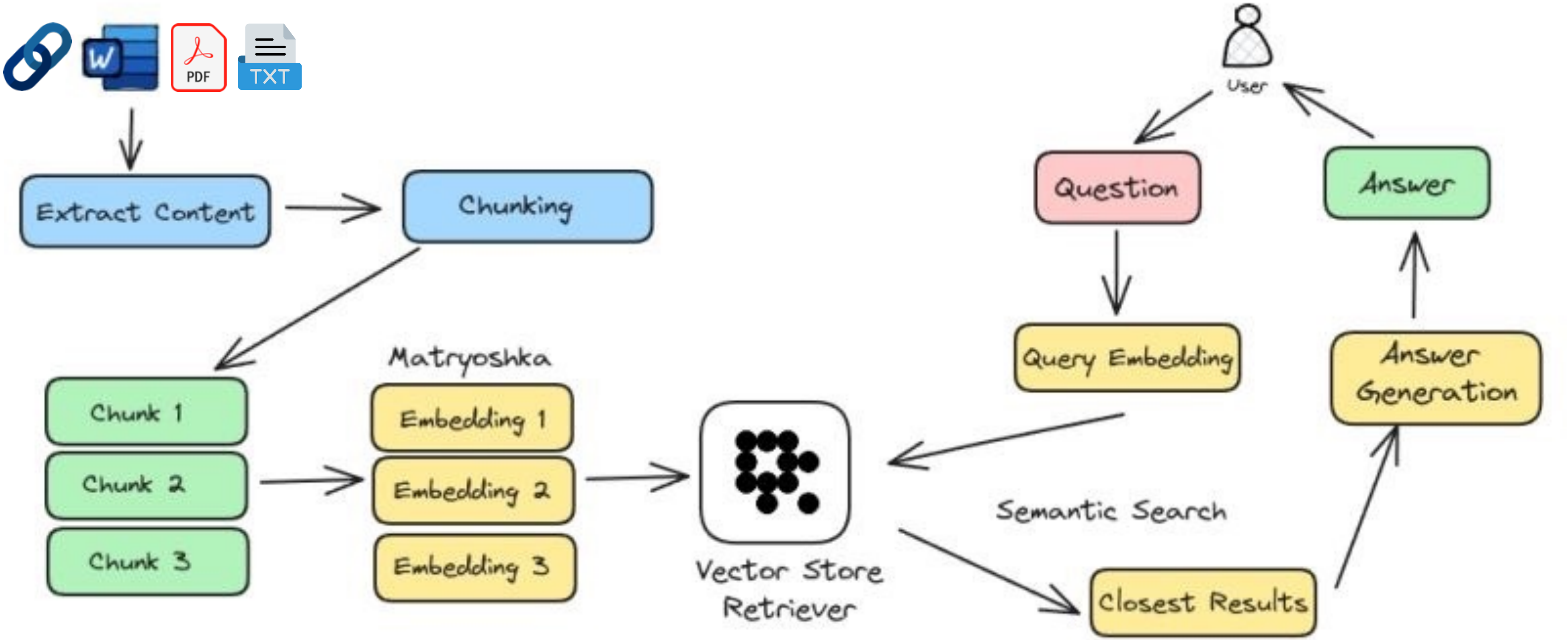
A typical bee colony contains between 20,000 and 80,000 bees. The queen bee is responsible for laying the majority of these eggs; she can produce up to 2,000 eggs per day during peak seasons. Beekeepers must regularly inspect their hives not only to monitor egg-laying but also to check for common pests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



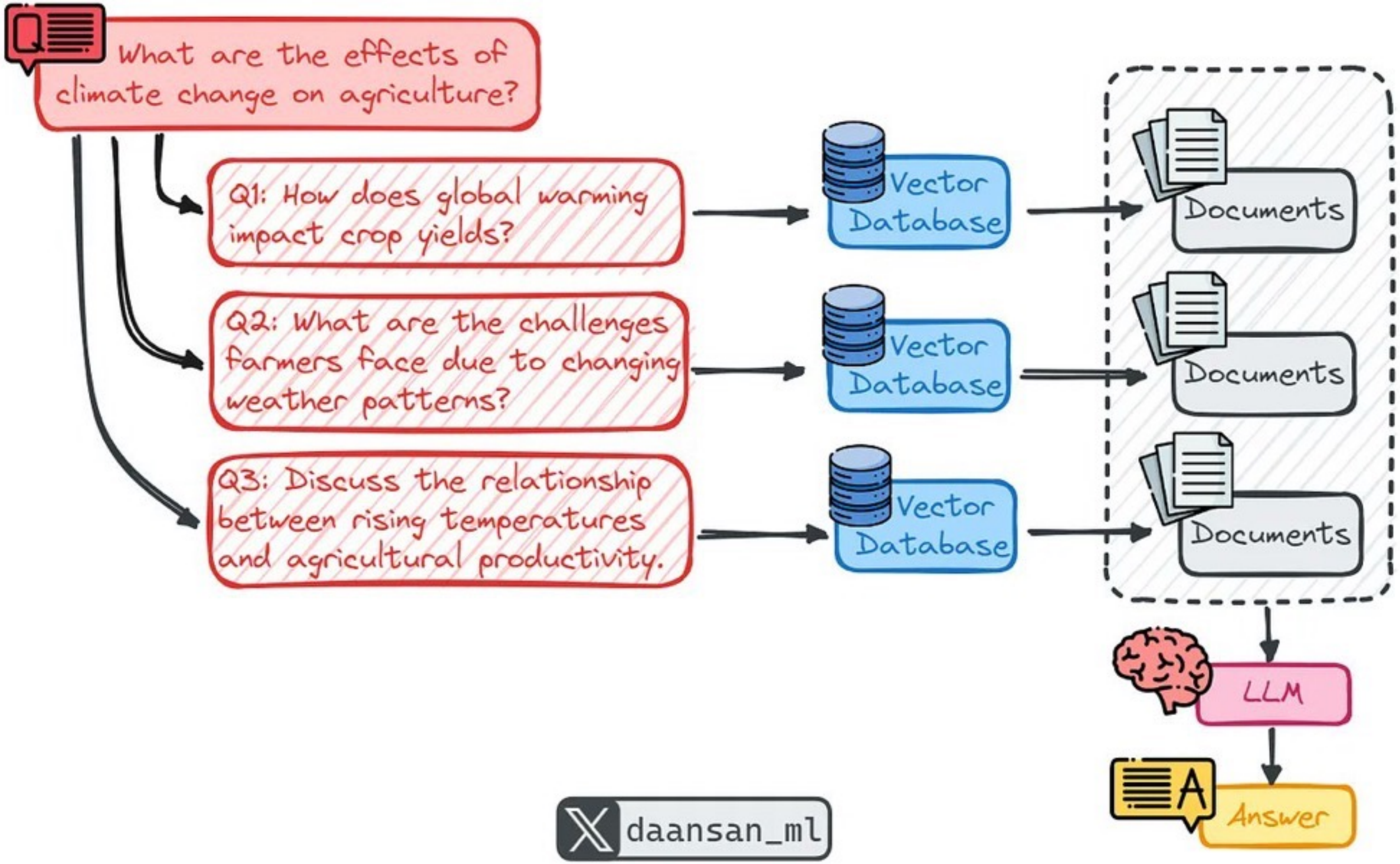
```python
import ollama
import chromadb
import time

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
```

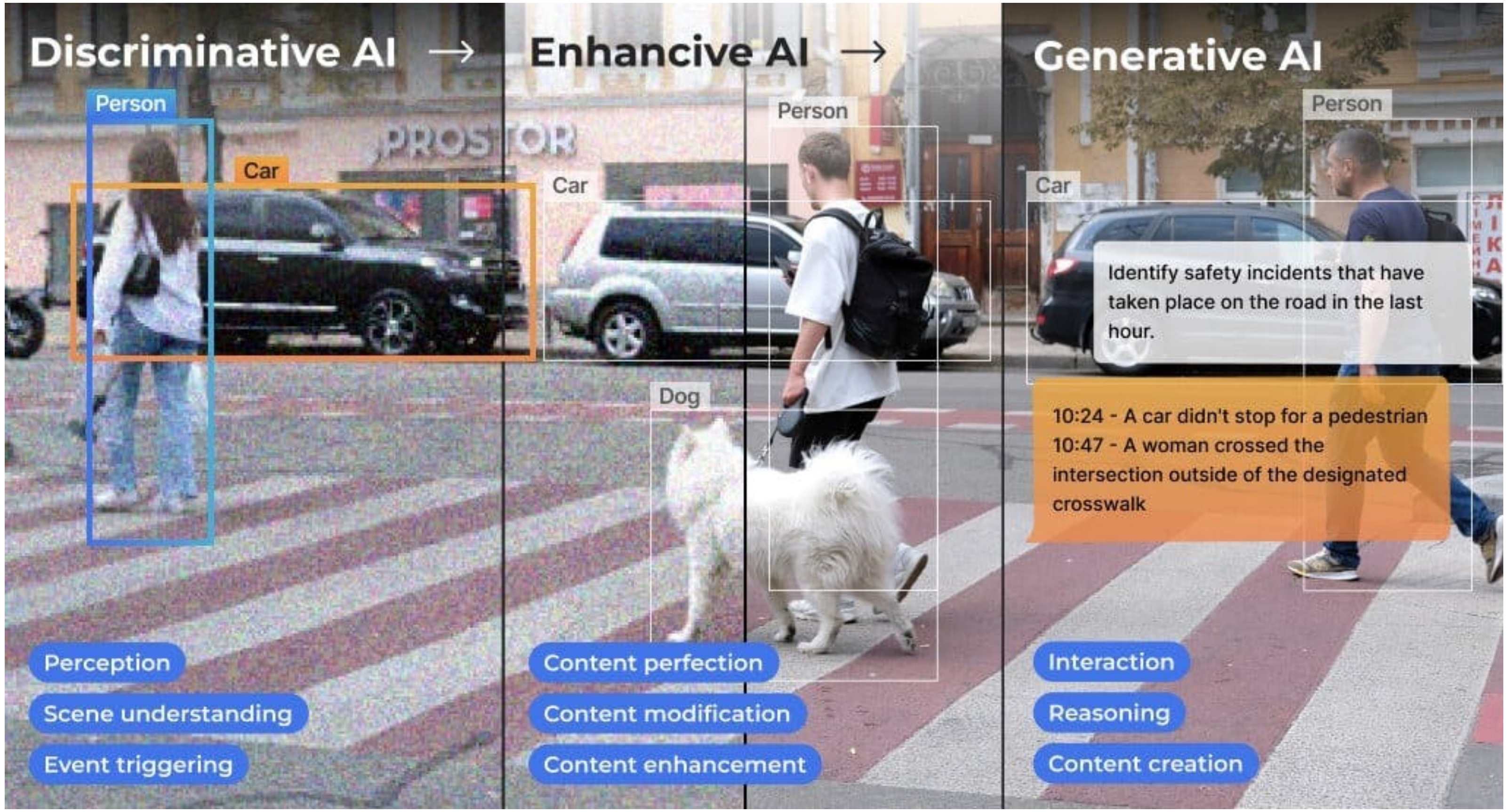
```
d honey and larvae; and foulbrood, a bacterial disease caused by Paenibacillus larvae that can devastate young bee populat
ions. The European honey bee (Apis mellifera) is the most commonly kept species of bees worldwide.

 [INFO] ==> The code for model: phi3, took 97.6s to generate the answer.

mjrovai@rpi-5:~/Documents/OLLAMA/RAG/RAG_test $ sudo raspi-config
```

RAG: Simple Query



Advanced RAG: Multi Query

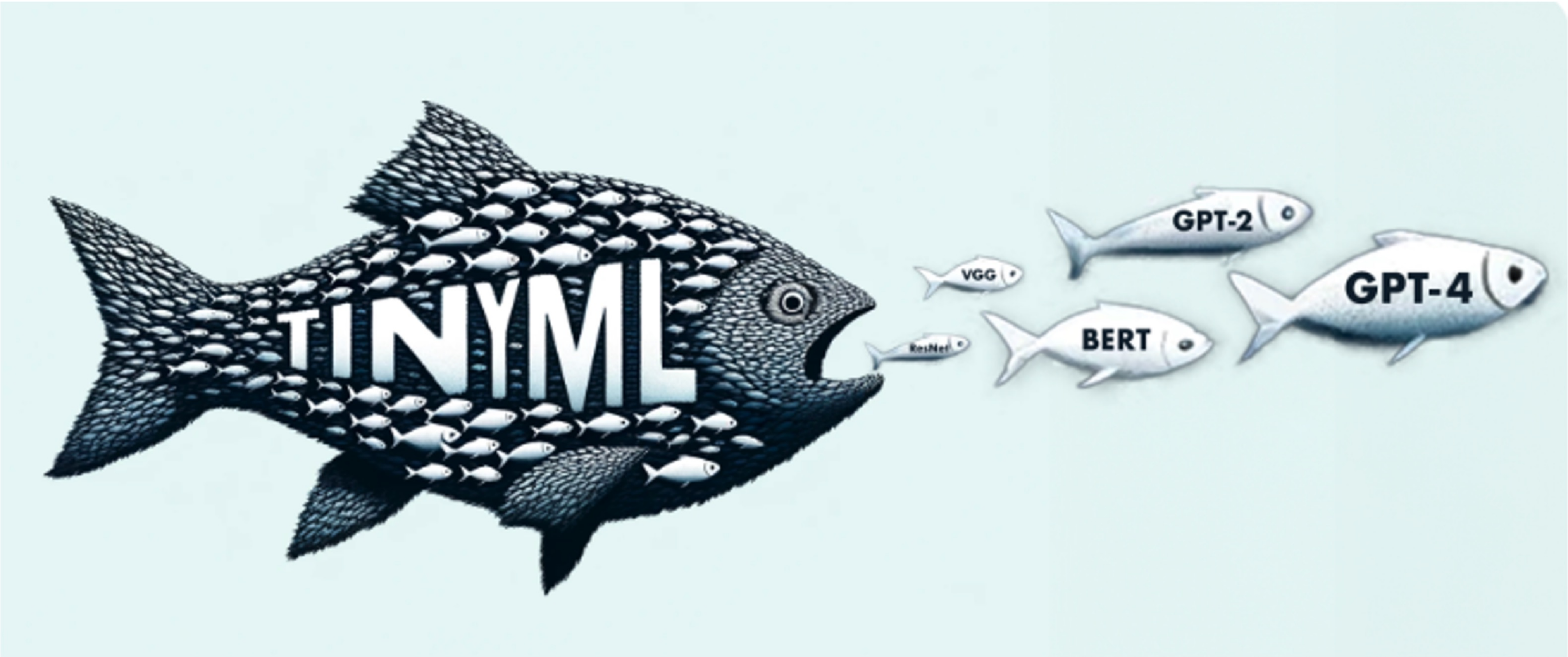

What are the effects of climate change on agriculture?
Q1: How does global warming impact crop yields?
Q2: What are the challenges farmers face due to changing weather patterns?
Q3: Discuss the relationship between rising temperatures and agricultural productivity.
Vector Database
Vector Database
Vector Database
Documents
Documents
Documents
LLM
Answer
daansan_ml



“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

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



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

Thanks

TINYML4D