# From GPIO to GPT

# Turning the Raspberry Pi into an AI Hub

Prof. Marcelo J. Rovai

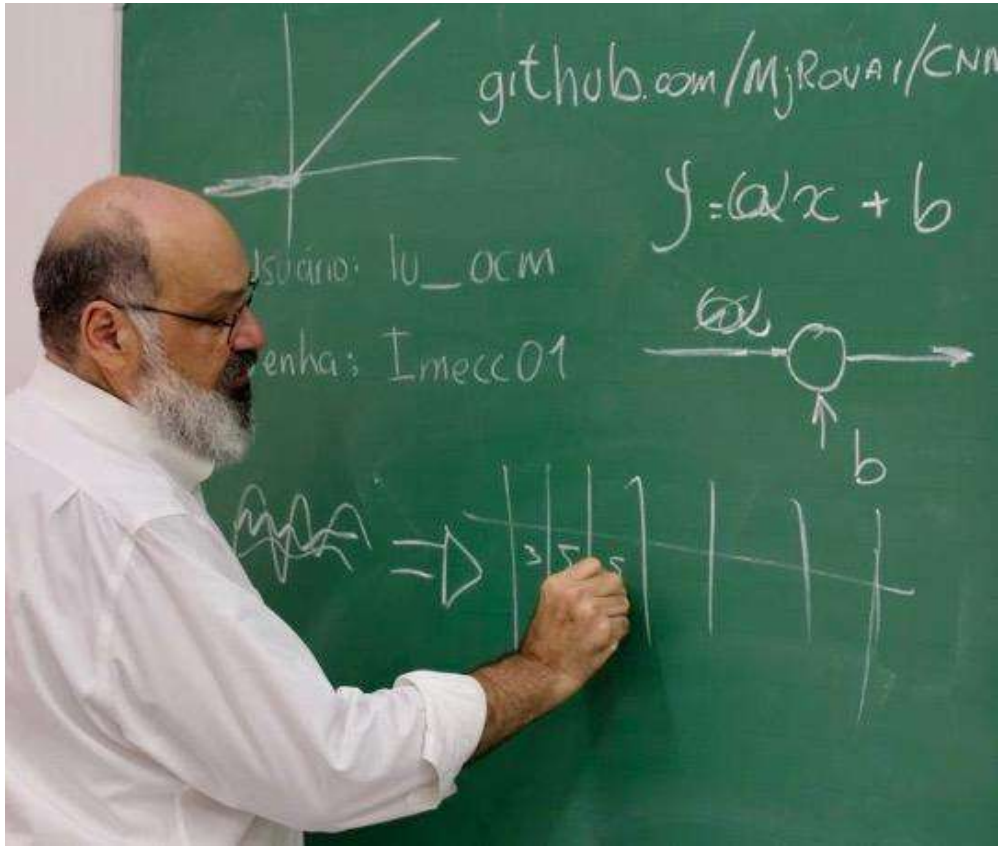
rovai@unifei.edu.br

UNIFEI - Federal University of Itajuba, Brazil

TinyML4D - Academic Network Co-Chair

EdgeAIP - Academia-Industry Partnership Co-Chair

**Marcelo Rovai** is an educator and professional in the field of engineering and technology, holding the title of Professor Honoris Causa from the Federal University of Itajubá, Brazil. His educational background includes an Engineering degree from UNIFEI and an advanced specialization from the Polytechnic School of São Paulo University (POLI/USP). Further enhancing his expertise, he earned an MBA from IBMEC (INSPER) and a Master's in Data Science from the Universidad del Desarrollo (UDD) in Chile.




With a career spanning several high-profile technology companies such as AVIBRAS Airspace, AT&T, NCR, and IGT, where he served as Vice President for Latin America, he brings a wealth of industry experience to his academic endeavors. He is a prolific writer on electronics-related topics and shares his knowledge through open platforms like Hackster.io.

In addition to his professional pursuits, he is dedicated to educational outreach, serving as a volunteer professor at the IESTI (UNIFEI) and engaging with the TinyML4D group and the EDGE AIP – the Academia-Industry Partnership of EDGEAI Foundation as a Co-Chair, promoting EdgeAI education in developing countries. His work underscores a commitment to leveraging technology for societal advancement.

# Content

MLSys: Machine Learning Systems

GenAI: Introduction and Demo

# Content

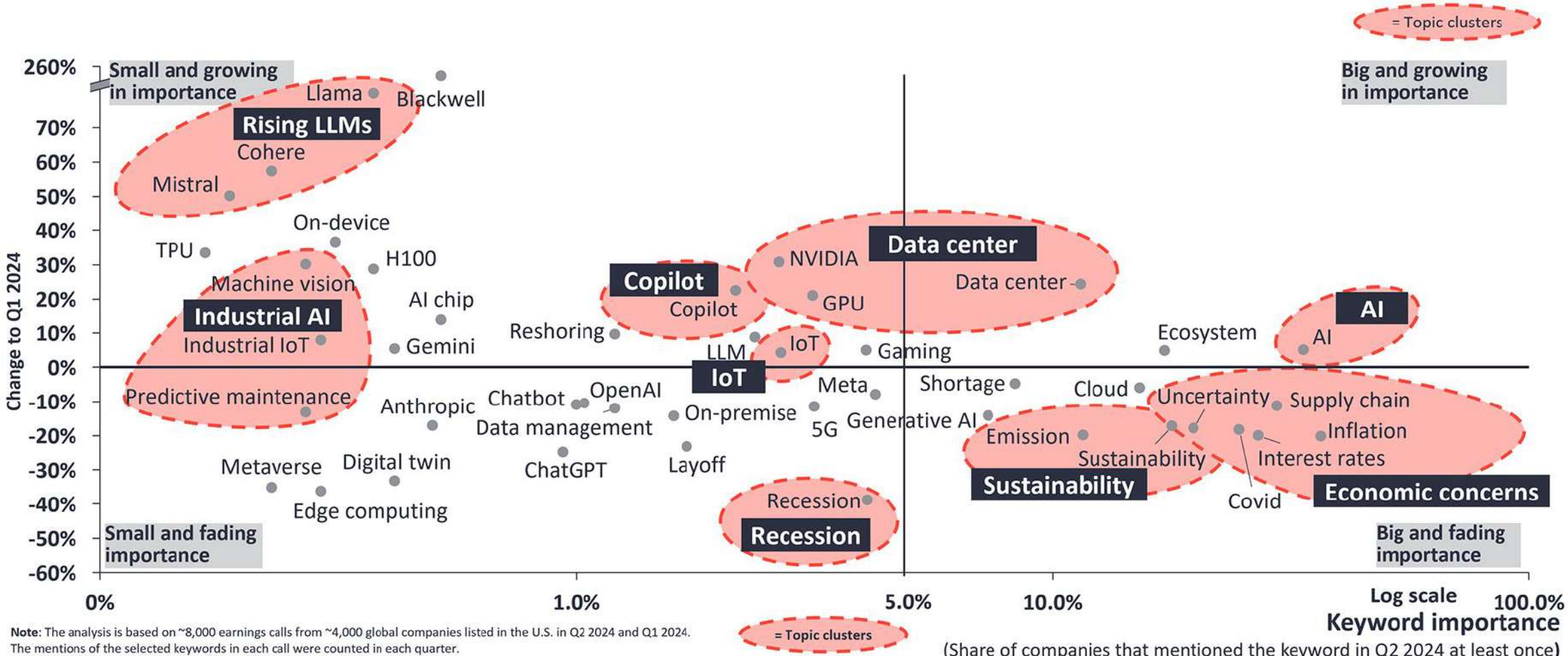
MLSys: Machine Learning Systems

GenAI: Introduction and Demo

IOT ANALYTICS

This analysis is part of the **CEO Insights** series

# What CEOs talked about in Q2 2024 (vs. Q1 2024)

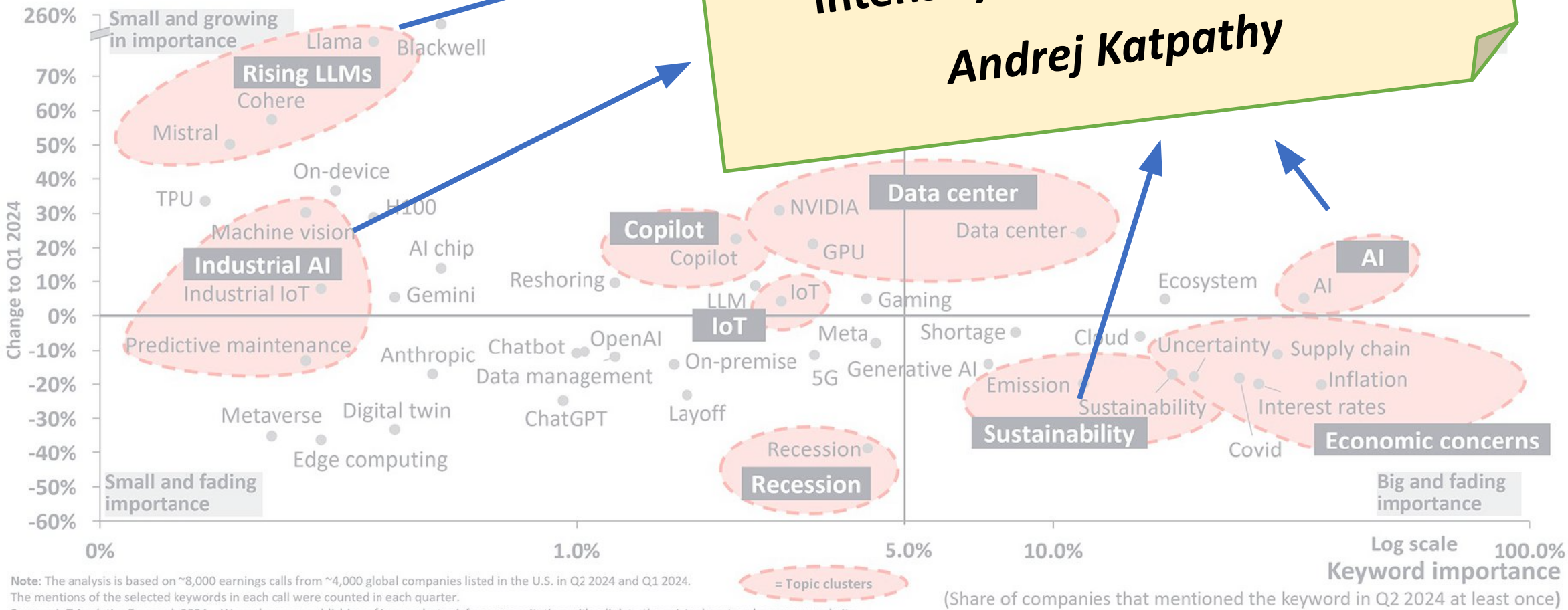



**Note**: The analysis is based on ~8,000 earnings calls from ~4,000 global companies listed in the U.S. in Q2 2024 and Q1 2024.
The mentions of the selected keywords in each call were counted in each quarter.
**Source:** IoT Analytics Research 2024 – We welcome republishing of images but ask for source citation with a link to the original post and company website.

https://iot-analytics.com/what-ceos-talked-about-in-q2-2024/

IOT ANALYTICS

# What CEOs talked about




Note: The analysis is based on ~8,000 earnings calls from ~4,000 global companies listed in the U.S. in Q2 2024 and Q1 2024.
The mentions of the selected keywords in each call were counted in each quarter.
Source: IoT Analytics Research 2024 – We welcome republishing of images but ask for source citation with a link to the original post and company website.

**"LLM model size competition is intensifying. … backwards!"**

***Andrej Katpathy***

https://iot-analytics.com/what-ceos-talked-about-in-q2-2024/

# What is AI?

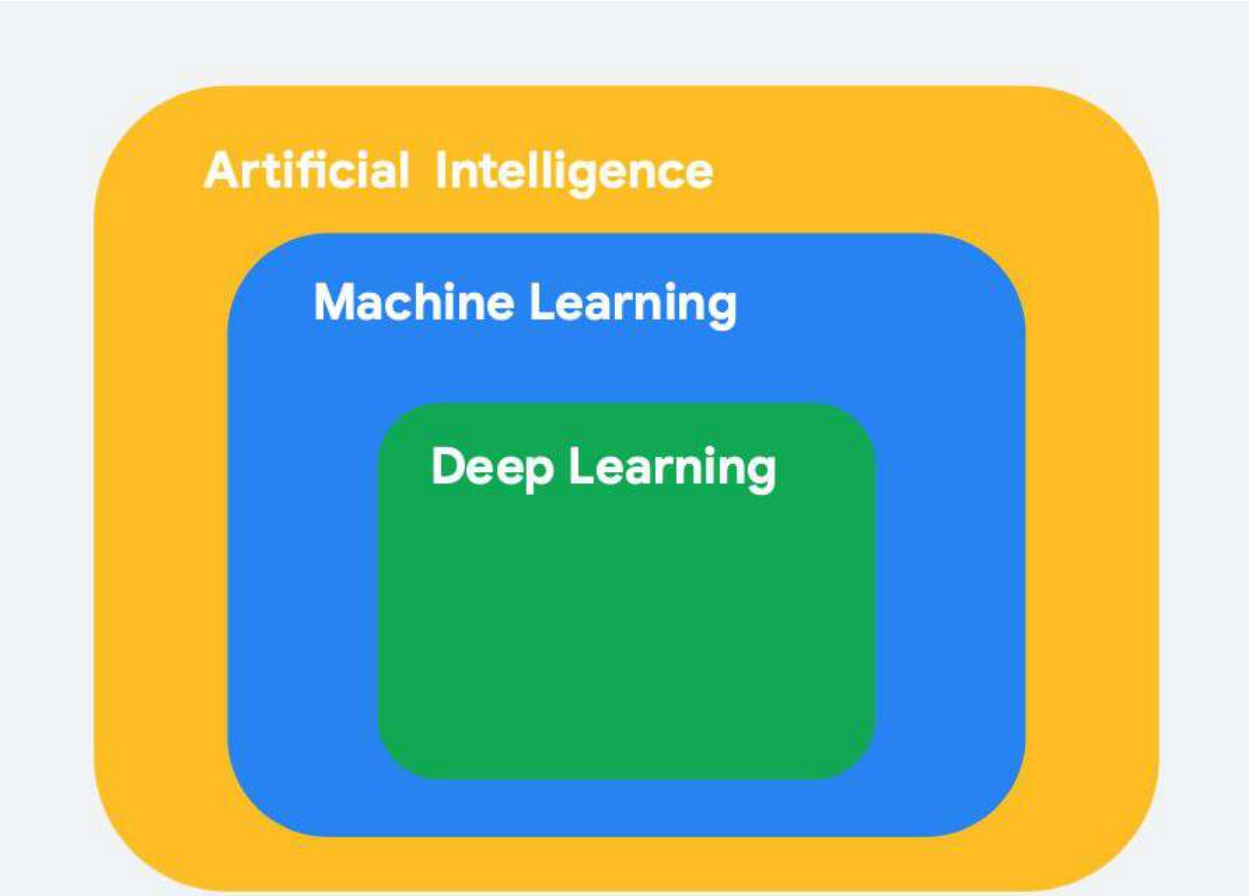
Artificial Intelligence (AI) is the simulation of human intelligence in machines that can perform tasks like learning, reasoning, and problem-solving.

Recommendation

Personal Assistants


LLMs (Chat Bots)

# Teaching computers how to learn a task directly from raw data



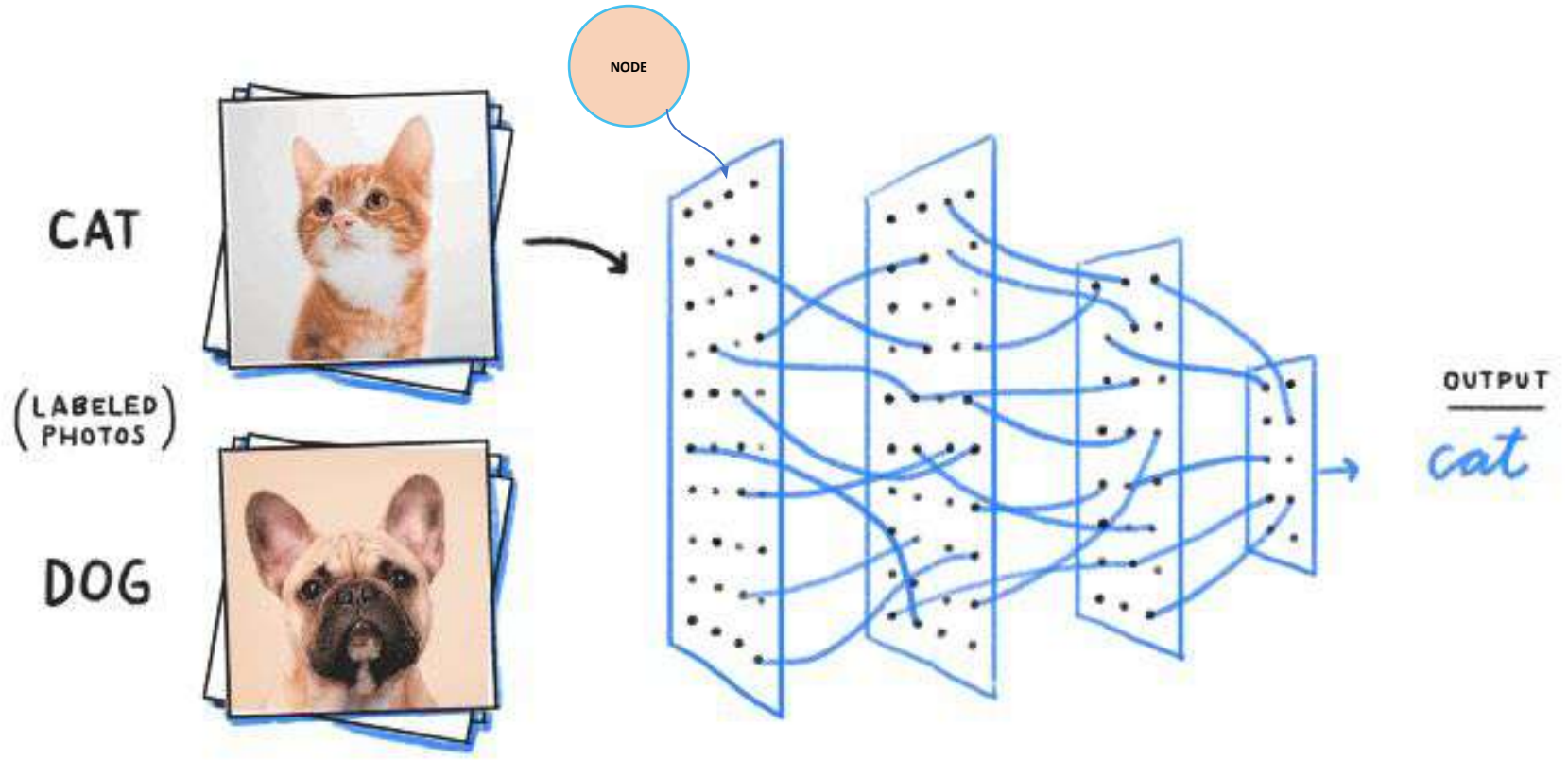

**AI: Any technique that enables computers to mimic human behavior**

**ML: Ability to learn without explicitly being programed**

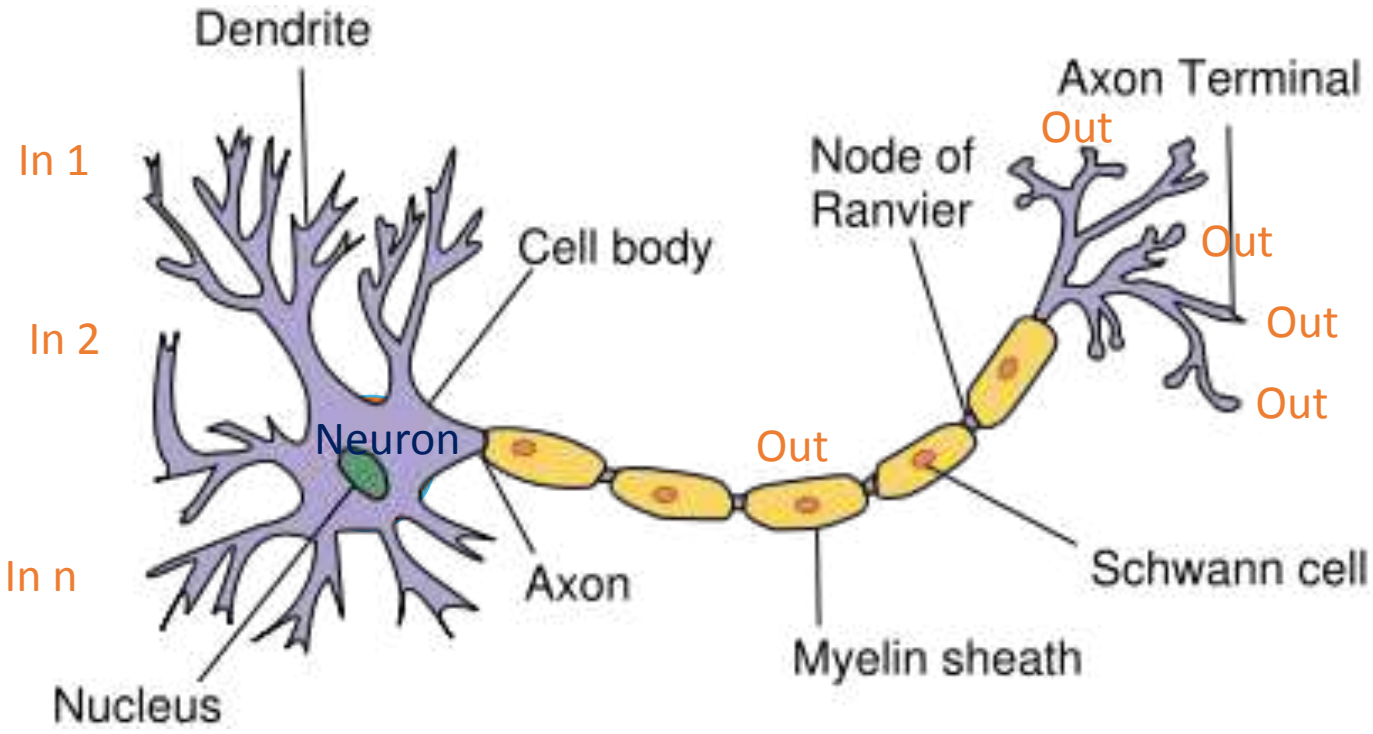
**DL: Extract patterns from data using neural networks**

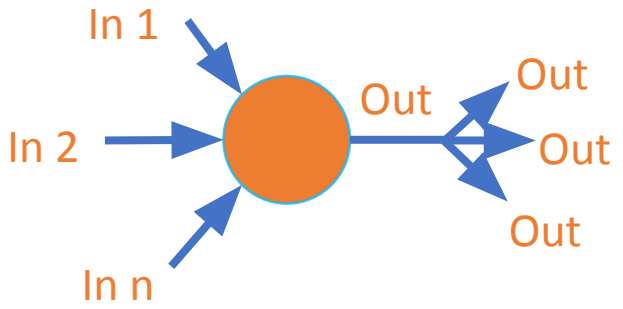
# (Deep) Machine Learning

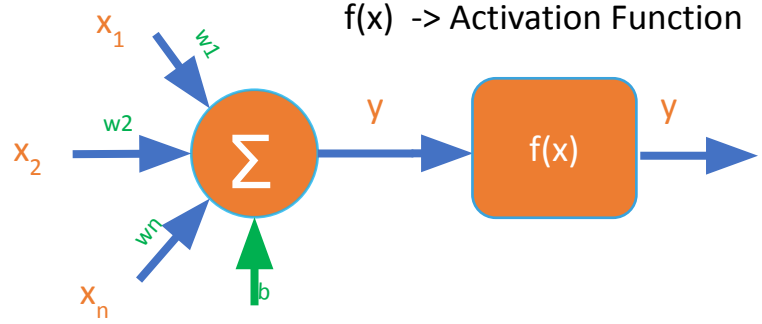
Deep Learning: Subset of Machine Learning in which multilayered neural networks learn from vast amounts of data



# Neuron (Perceptron)



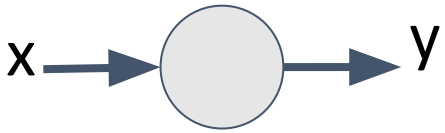




Parameters

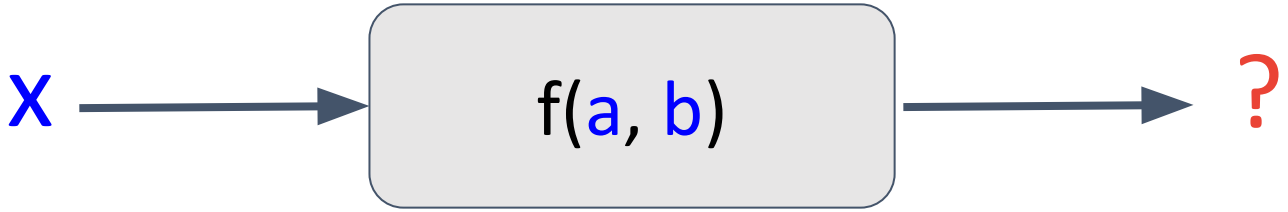
$$y = f(\sum_{i=1}^{n} x_i w_i + b)$$

$$y = ax + b$$

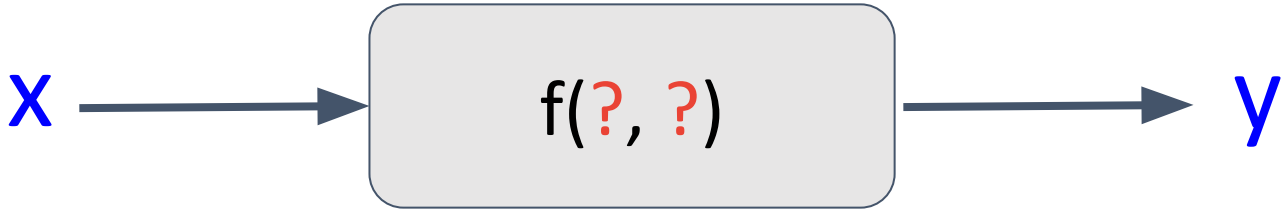

$y = ax + b$

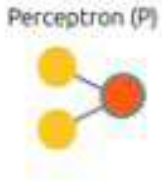



## Traditional Computation



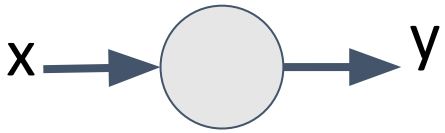

## Machine Learning

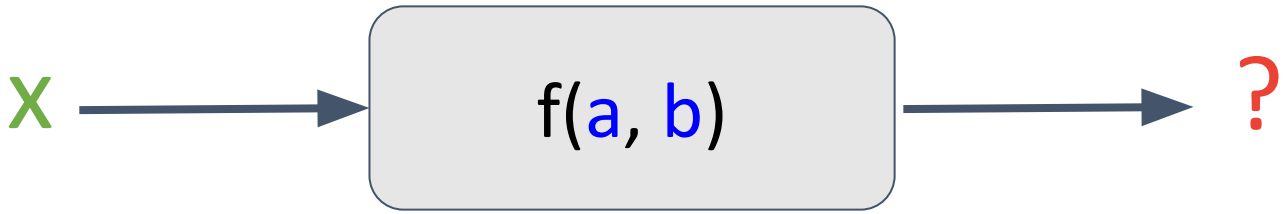





# Traditional Computation

$y = a\,x + b$

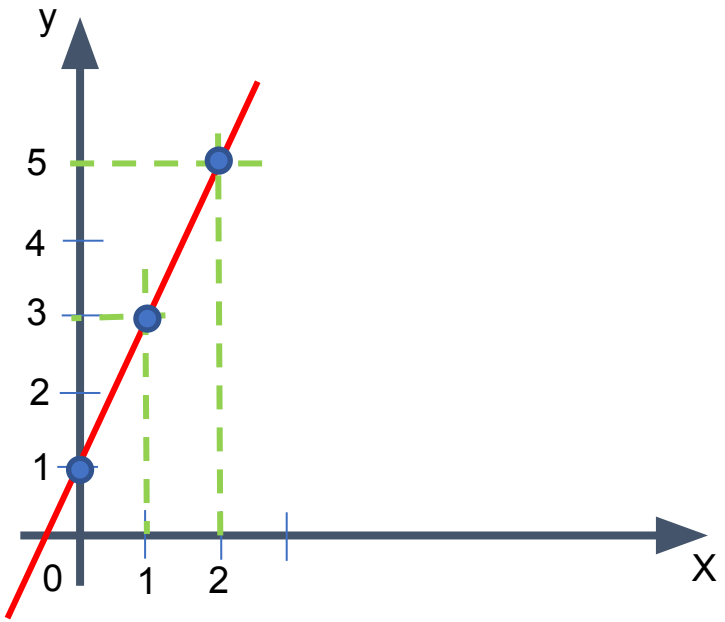


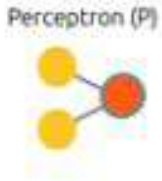




$y = a\,x + b$

$y = 2\,x + 1$

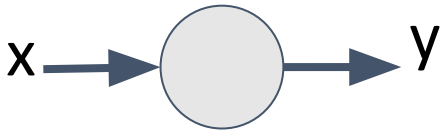
X = 0 → y = 1
X = 1 → y = 3
X = 2 → y = 5

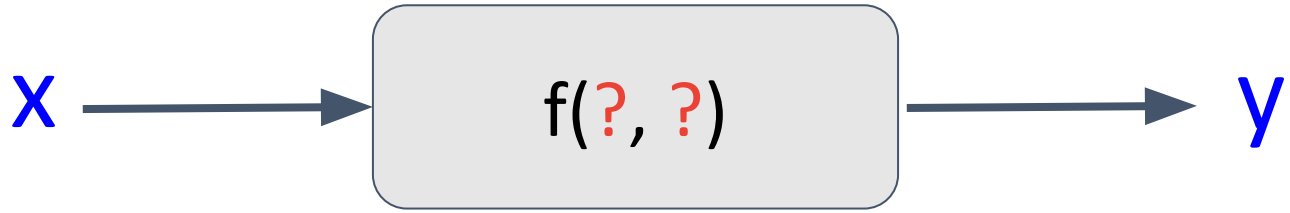





# Machine Learning

$y = a x + b$



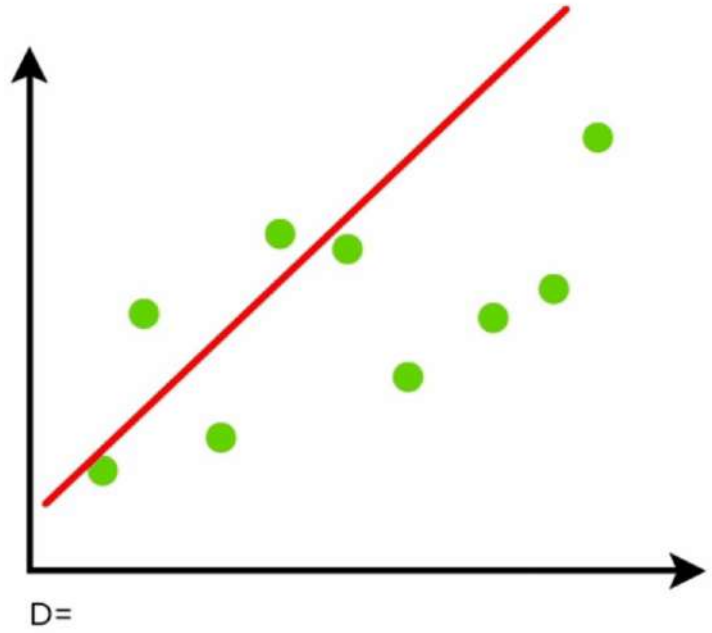




$y = a x + b$

X = 0 -> y = 1
X = 1 -> y = 3
X = 2 -> y = 5

a = 0 ; b = 0 -> error D1
a = 0 ; b = 1 -> error D2
a = 1; b = 2 -> error D3
.
.
.

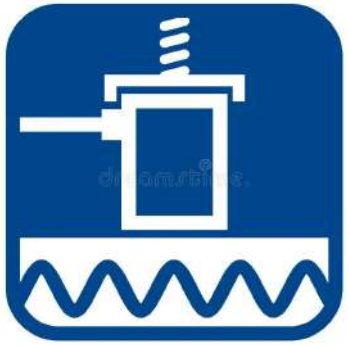
$$MSE = \frac{1}{n}\sum\left(y - \hat{y}\right)^2$$




The red line is the regression line, the green dots are the observed data value, and d1,d2,d3,….are the error values (y actual — y pred).


# Neural Network Architectures

## Vibration Analysis

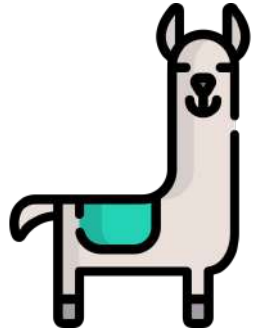


## Image Classification


## Text Generation

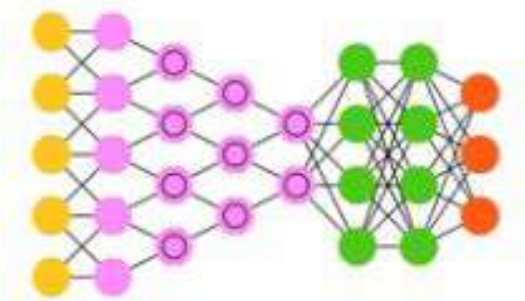
## Image Generation

## Large Language Models- LLMs

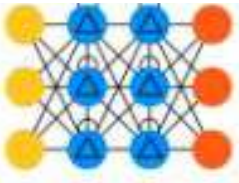


**DNN** - Deep Neural Network

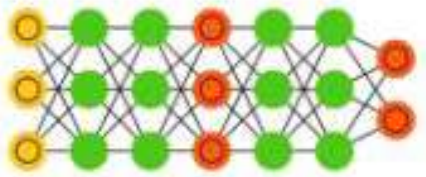
**CNN** - Convolutional NN



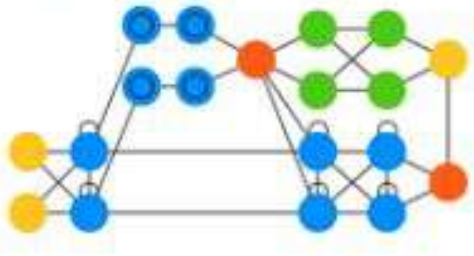
**RNN** - Recurrent NN (GRU/LSTM)



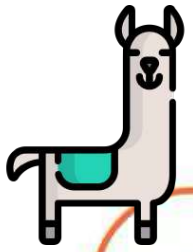
**GAN** - Generative Adversarial N.

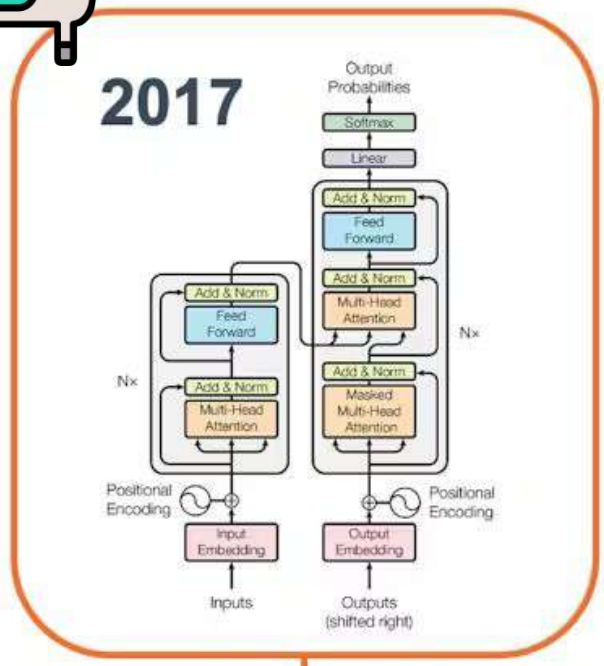


**AN** - Attention (Transformers)

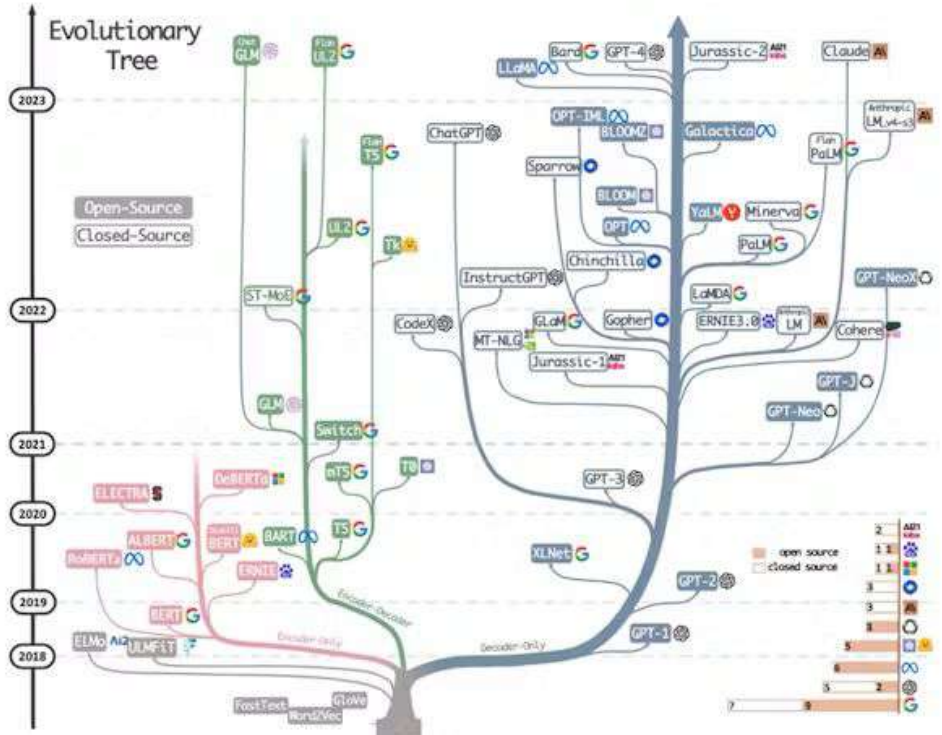


# LLMs (Large Language Models)









# The Evolution of AI

AI evolved from symbolic systems to deep learning:

1956: Dartmouth Conference defined AI.

1970s-1980s: Expert Systems

1990s: Statistical learning approaches emerged.

2012: Deep learning Neural Networks revolution with AlexNet.

Present: Large-scale systems like GPT-4 dominate.

# Deep Learning and AI: Why now?

Neural networks date back decades, so why do they dominate?


DATA


HW: Computer Infrastructure

SW: Algorithms & Tools

# MLSys & MLSysEng

ML systems (MLSys) integrate data, algorithms (SW), and computing infrastructure (HW) to transform theory into impactful real-world solutions.

Machine Learning Systems Engineering (MLSysEng) bridges the gap between Electrical Engineering's hardware expertise and Computer Science's focus on algorithms and software.

The illustration depicts the analogy between astronauts (algorithm developers) exploring new frontiers and rocket scientists (ML systems engineers) ensuring their success through robust systems

*(Created by DALL-E)*

# Why Machine Learning Systems Matter

AI impacts daily life: smart alarms, navigation apps, personalized playlists, and more.

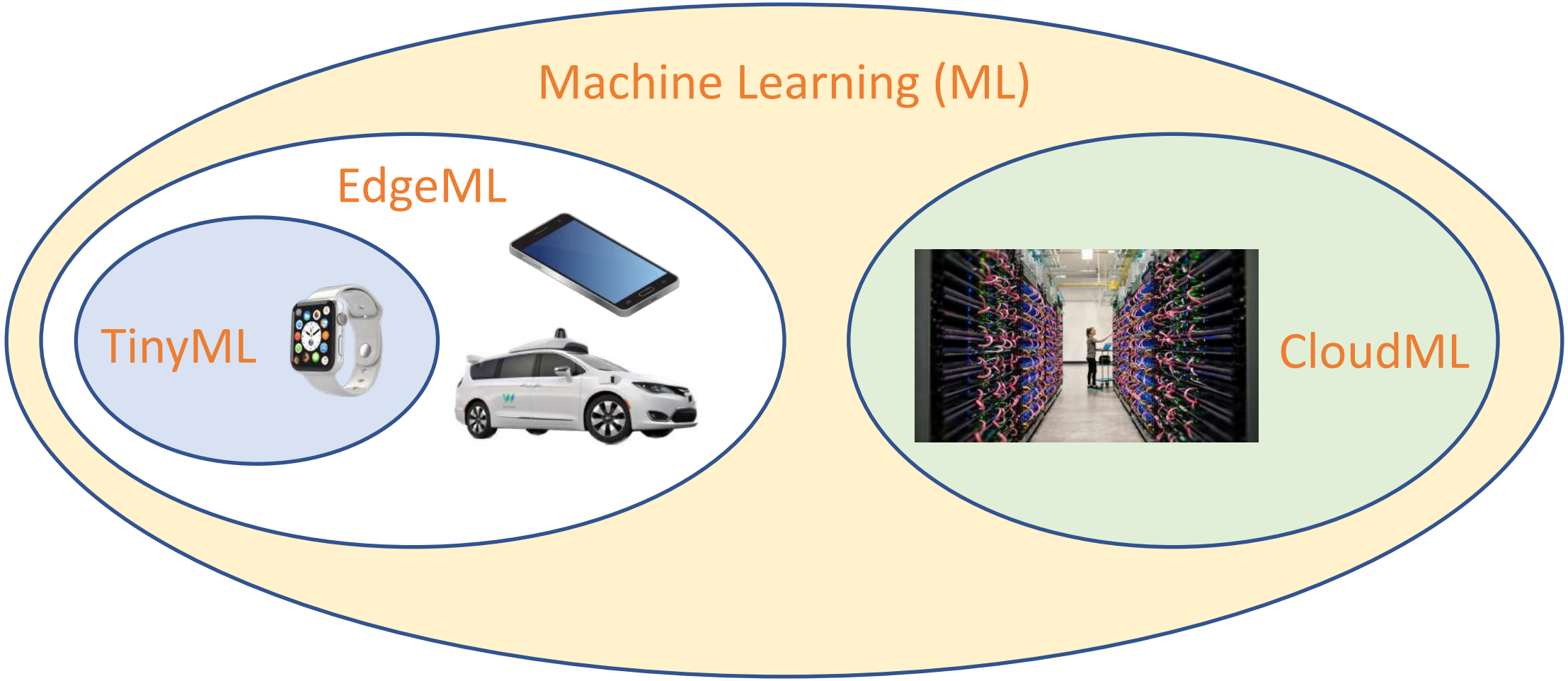
Beyond convenience, AI is revolutionizing industries: healthcare, weather prediction, autonomous vehicles, and scientific discoveries.

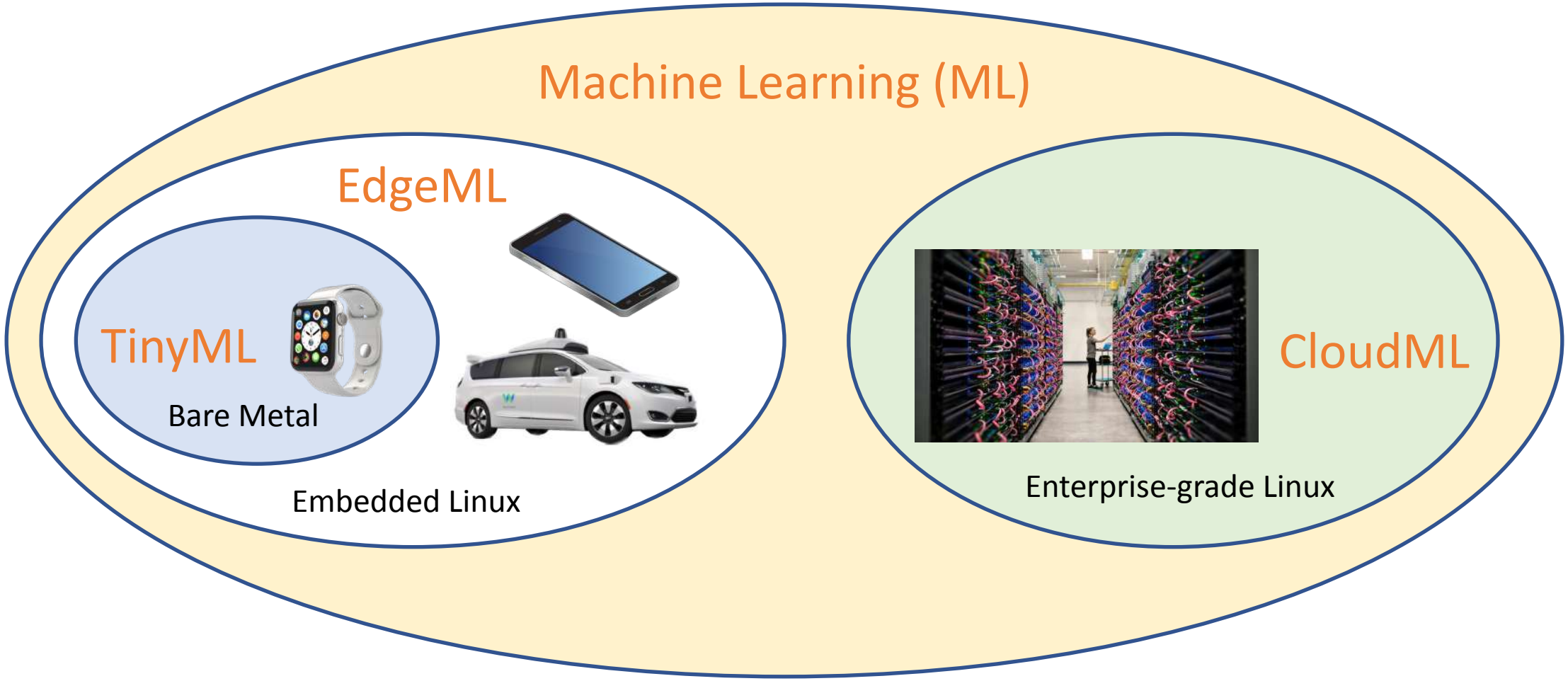
# Machine Learning

CloudML: is essential for tasks requiring massive computational power or large-scale data analysis.

EdgeML (or EdgeAI) is the processing of Artificial Intelligence algorithms on edge, that is, on users' devices.

TinyML is where sensors are generating data with ultra-low power consumption (batteries) ("always on devices")


Machine Learning (ML)
EdgeML
TinyML
CloudML

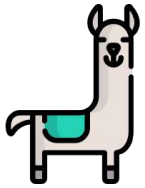

Machine Learning (ML)
EdgeML
TinyML
Bare Metal
Embedded Linux
CloudML
Enterprise-grade Linux

# Application Complexity vs. HW

Power

EdgeML

TinyML



LLMs/VLMs
1 GB+

Object Detection
Complex Voice
Processing
1 MB+

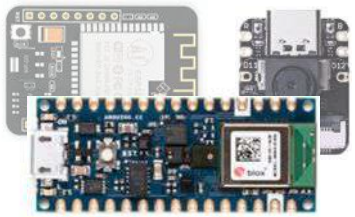
Image
Classification
250 KB+

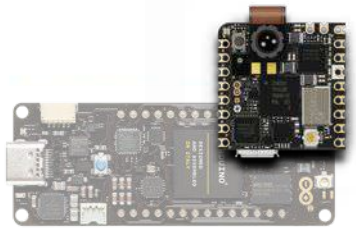
KeyWord Spotting
Audio Classification
50 KB

Anomaly Detection
Sensor Classification
20 KB

ESP32

XIAO





Rpi-Pico
(Cortex-M0+)

Arduino Nano
(Cortex-M4)

Arduino Pro
(Cortex-M7)

RaspberryPi SmartPhone
(Cortex-A)

Jetson Orin
(Cortex-A + GPU)

Application Complexity

CPU Power / Memory

# Application Complexity vs. HW

Power

EdgeML

TinyML

LLMs/VLMs
1 GB+

Object Detection
Complex Voice
Processing
1 MB+

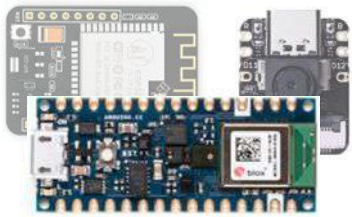
Image
Classification
250 KB+

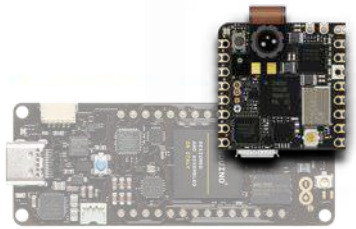
KeyWord Spotting
Audio Classification
50 KB

Anomaly Detection
Sensor Classification
20 KB

ESP32

XIAO





micro NPU

Rpi-Pico
(Cortex-M0+)

Arduino Nano
(Cortex-M4)

Arduino Pro
(Cortex-M7)

RaspberryPi SmartPhone
(Cortex-A)

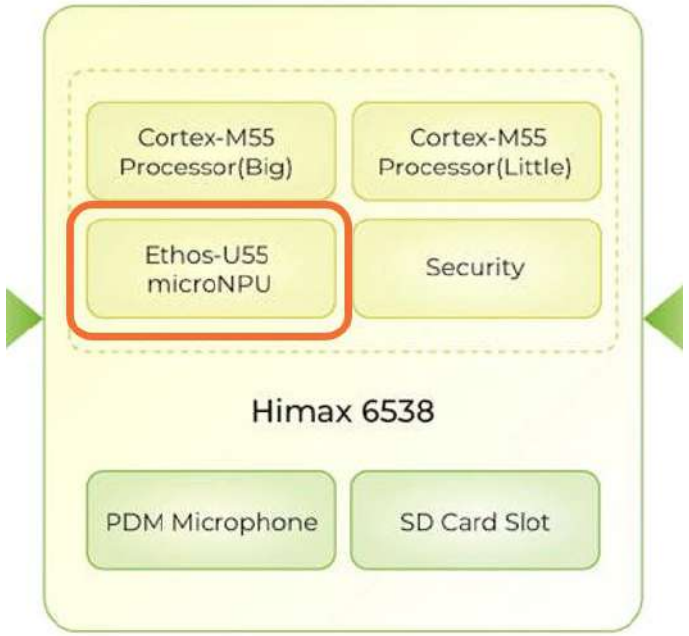
Jetson Orin
(Cortex-A + GPU)

Application Complexity

CPU Power / Memory

## microNPU

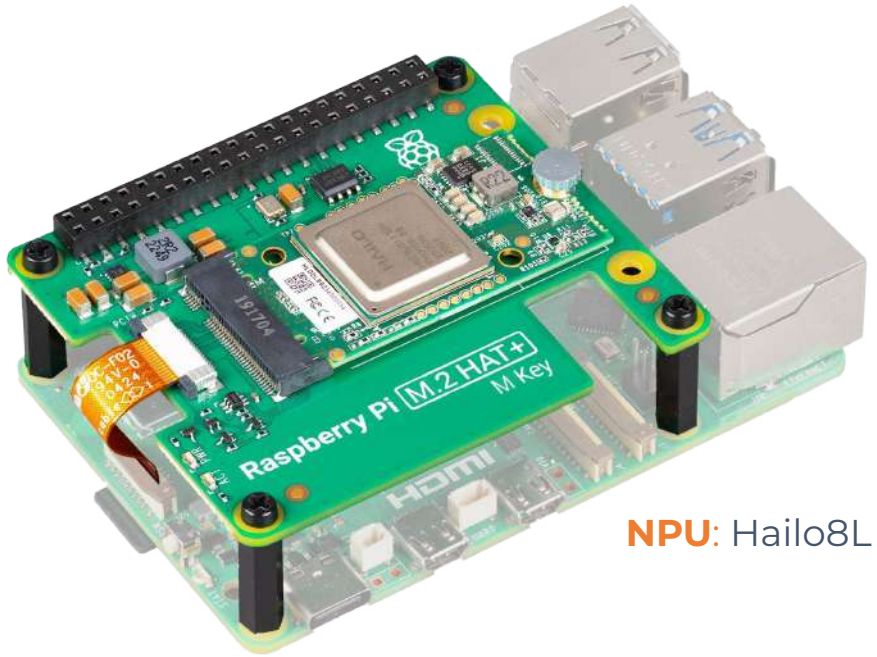
Grove Vision AI v2




0.5 TOPS

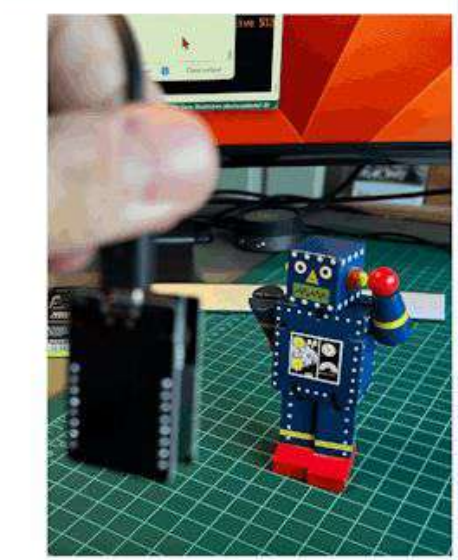
## NN Inference Accelerator

Raspberry Pi AI Kit




**NPU**: Hailo8L

13 - 25 TOPS



Classification: 687 ms

1.5 FPS

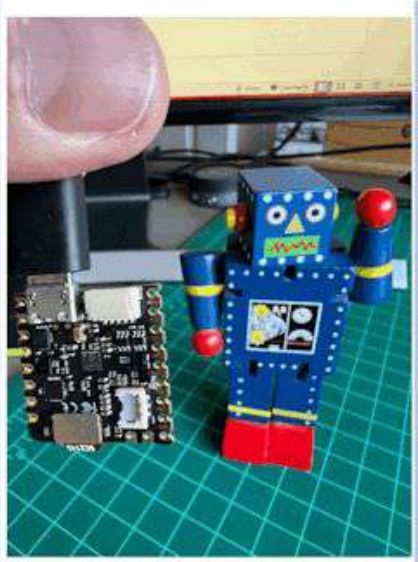
**ESP - CAM**
Xtensa LX6
240 MHz

Classification: 142 ms

7.0 FPS

**XIAO ESP32S3**
Xtensa LX7
240 MHz

450mW



Classification: 86 ms

11.6 FPS

**Nicla-Vision**
ARM M7
480 MHz

590mW

Classification: 83 ms

12.0 FPS

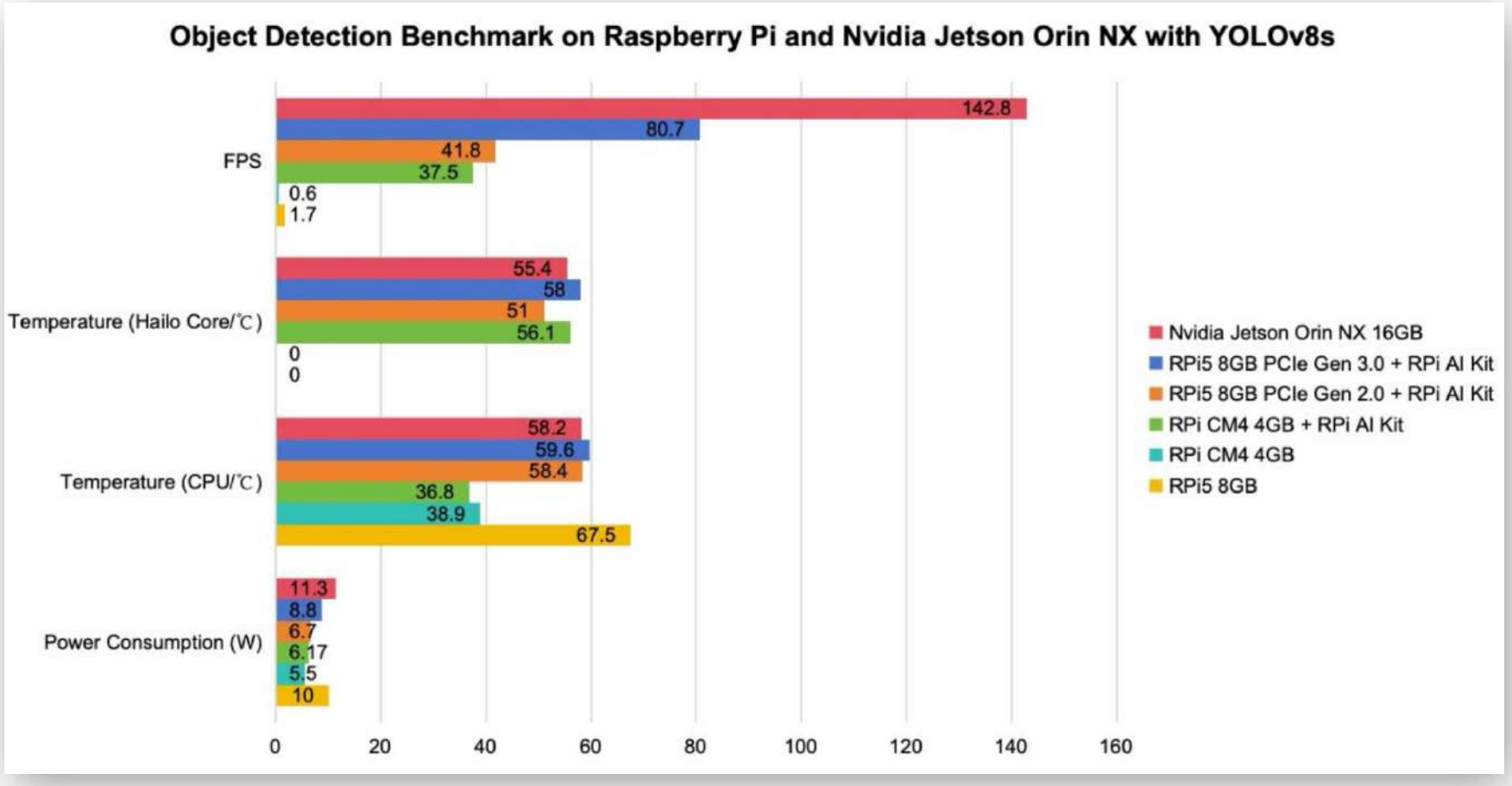
**Portenta H7**
ARM M7
480 MHz

Classification: 6 ms

167 FPS

**Grove Vision AI V2**
ARM Ethus-U55
400 MHz

350mW





**GPU**: NVIDIA Ampere 100 TOPS


**NPU**: Hailo8L*

* **Note:** Hailo-10H for LLMs is planned for future release

# The Five Pillars of MLSysEng

Data Engineering: Foundation of the system.

Model Training: Learning from data.

Model Deployment: Applying trained models in real-world settings.

Operation & Maintenance: Ensuring long-term reliability.

Ethics & Governance: Responsible and fair AI use.

# Real-World Applications

- Agriculture
- Healthcare
- Industry
- Environment

# Real-World Applications

- Agriculture
- Healthcare
- Industry
- Environment

# Classifying mosquito wingbeat sound using TinyML


Moez Altayeb
University of Khartoum, Sudan
ICTP, Trieste, Italy
mohedahmed@hotmail.com

Marcelo Rovai
Universidade Federal de Itajubá
Itajubá, Brazil
rovai@unifei.edu.br

Marco Zennaro
ICTP
Trieste, Italy
mzennaro@ictp.it

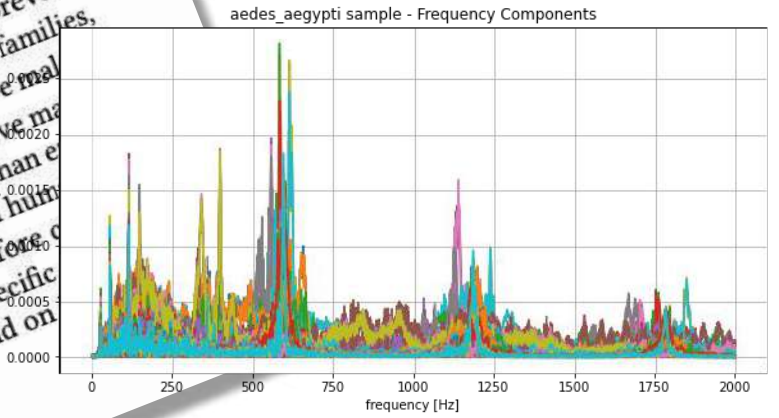


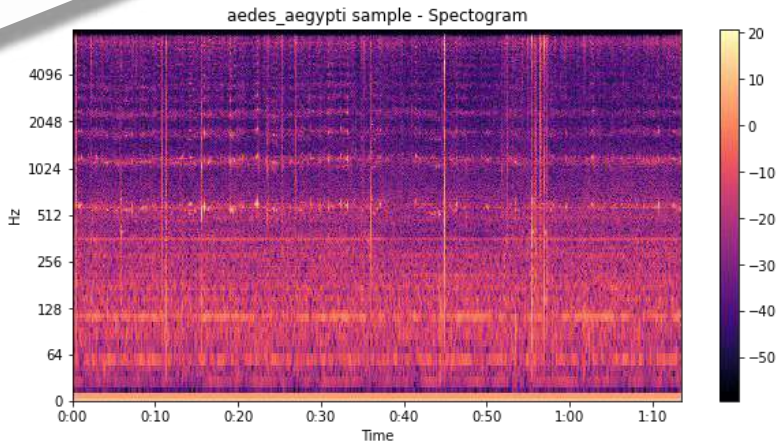
**ABSTRACT**

Every year more than one billion people are infected and more than one million people die from vector-borne diseases including malaria, dengue, zika and chikungunya. Mosquitoes are the best known disease vector and are geographically spread worldwide. It is important to raise awareness of mosquito proliferation by monitoring their incidence, especially in poor regions. Acoustic detection of mosquitoes has been studied for long and ML can be used to automatically identify mosquito species by their wingbeat. We present a prototype solution based on an openly available dataset, on the Edge Impulse platform and on three commercially-available TinyML devices. The proposed solution is low-power, low-cost and can run without human intervention in resource-constrained areas. This insect monitoring system can reach a global scale.



affected. People from poor communities with little access to health care and clean water sources are also at risk. Although anti-malarial drugs exist, there's currently no malaria vaccine. Vector-borne diseases also exacerbate poverty. Illness prevent people from working and supporting themselves and their families, impeding economic development. Countries with intensive mal have much lower income levels than those that don't have ma Countries affected by malaria turn to control rather than e disease carriers on an area-by-area basis. It is therefore c be able to detect the presence of mosquitoes in a specific paper presents an approach based on TinyML and on embedded devices.

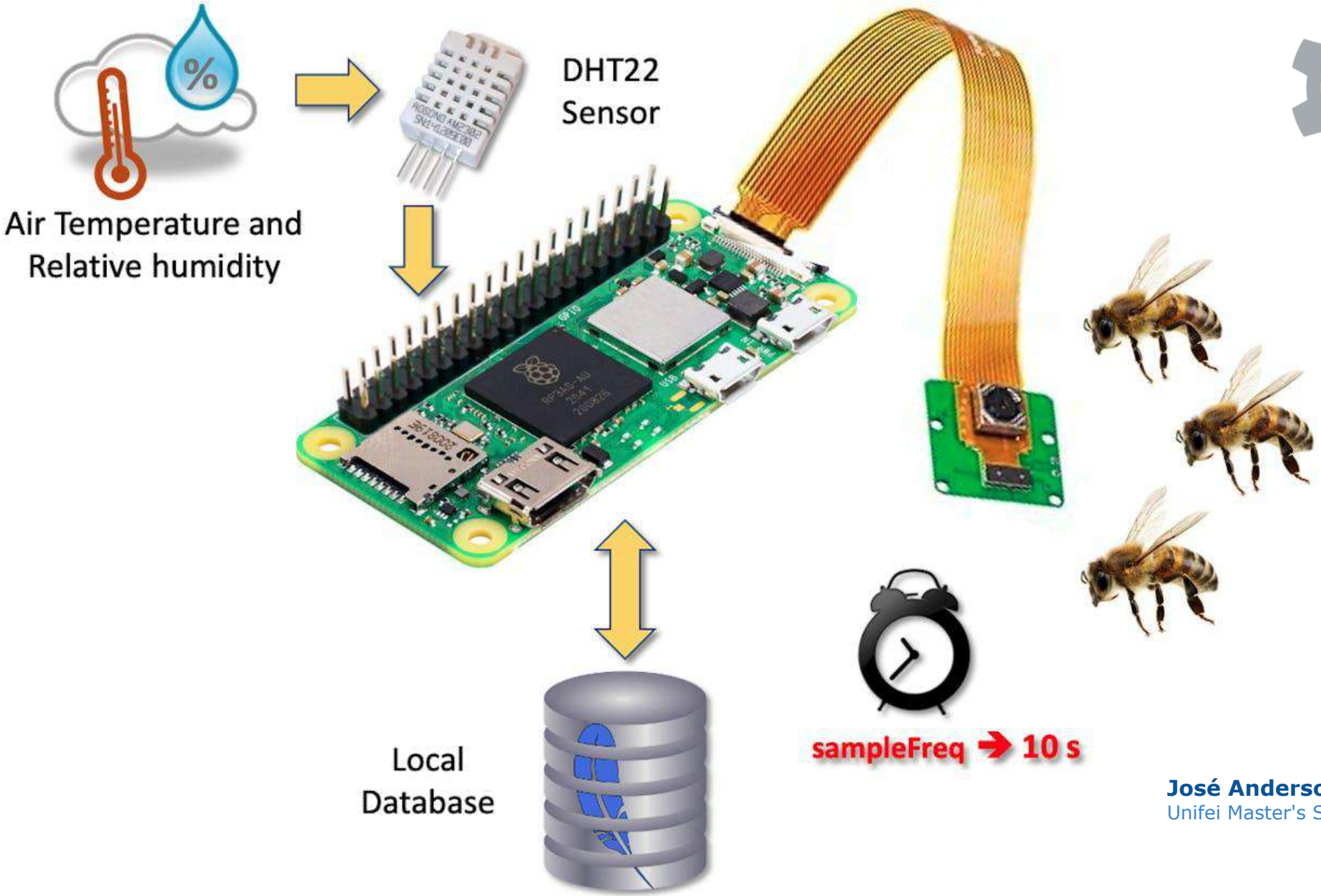




https://www.edgeimpulse.com/blog/the-world-is-all-abuzz-about-tinyml

# Coffee Disease Classification


**João Vitor Yukio Bordin Yamashita**
Engenheiro - UNIFEI

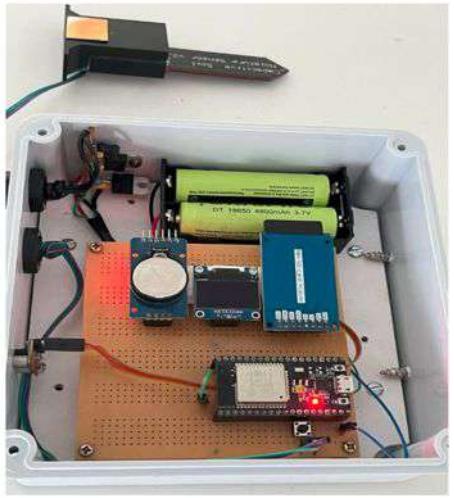
https://www.hackster.io/Yukio/coffee-disease-classification-with-ml-b0a3fc

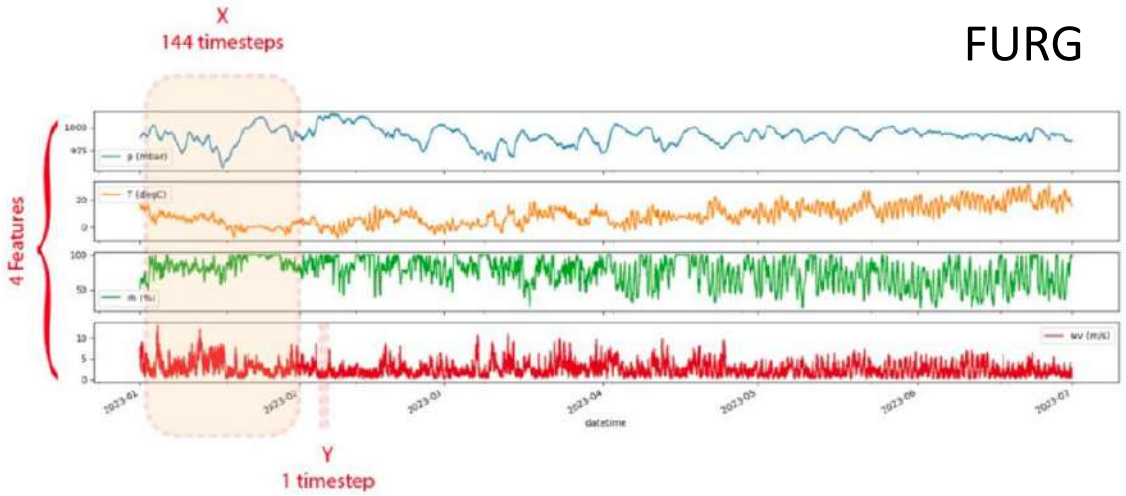


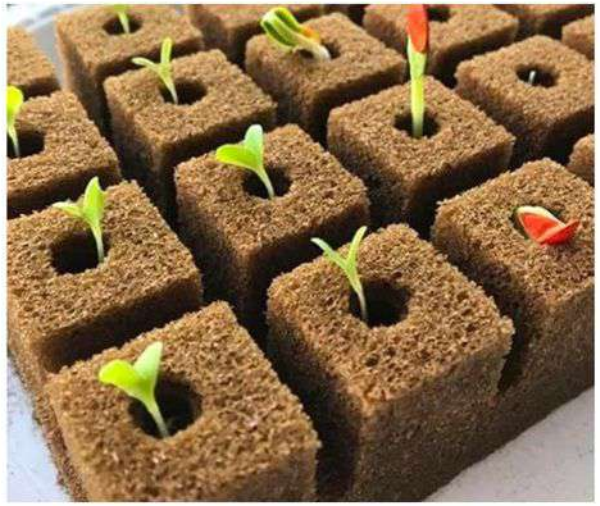

**José Anderson Reis**
Unifei Master's Student

# LSTM

FURG UTEC UNRaf






ESP32 LSTM Phenolic Sponge Moisture

# YOLO

FURG UTEC UNRaf

Ant Detection

# Real-World Applications

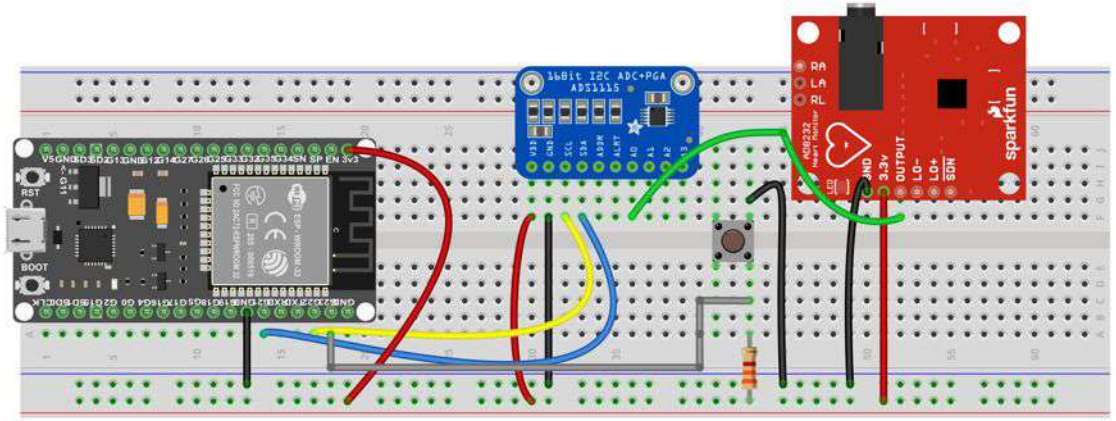
Agriculture

Healthcare

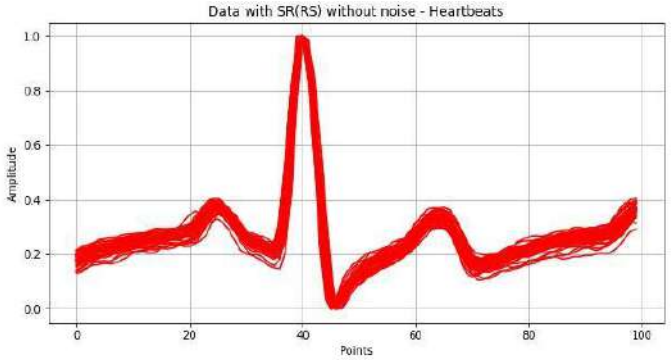
Industry

Environment

# AD8232 - Single Lead Heart Rate Monitor




fritzing

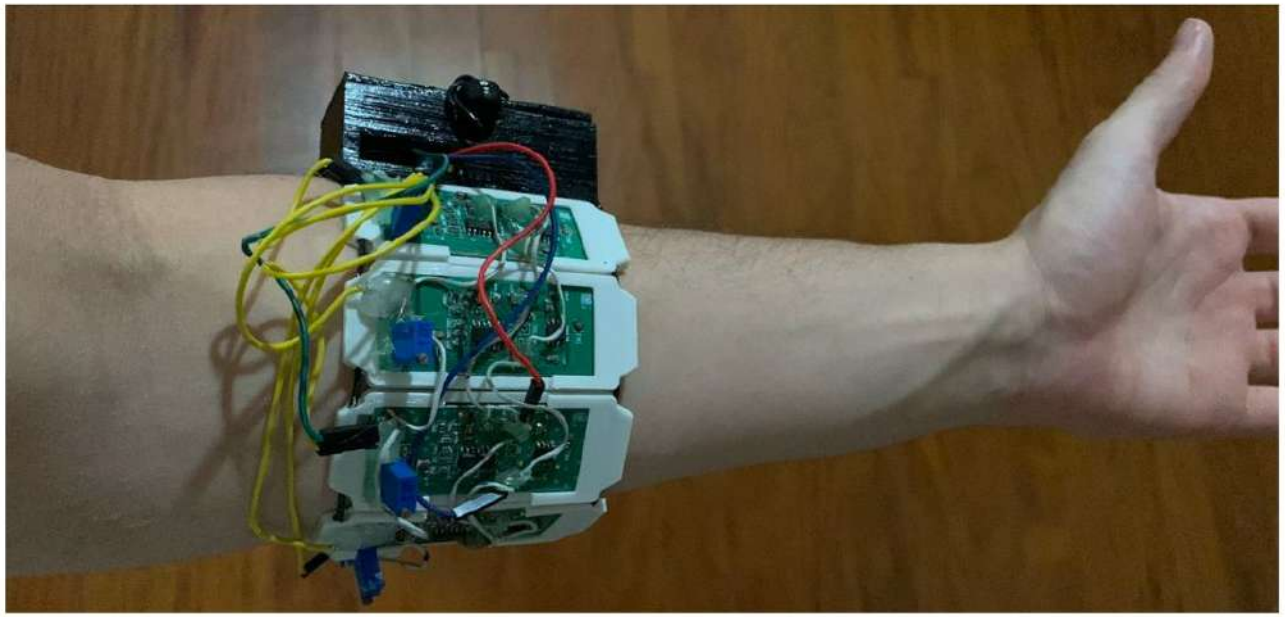


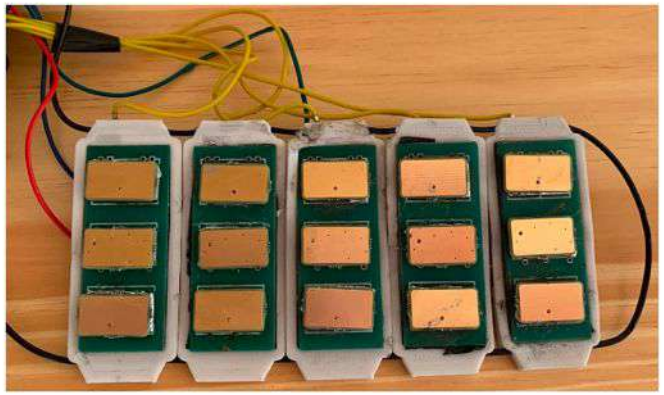


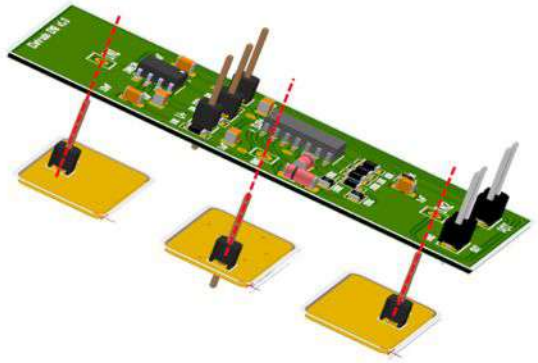
[Atrial Fibrillation Detection on ECG using TinyML
Silva et al. UNIFEI 2021]

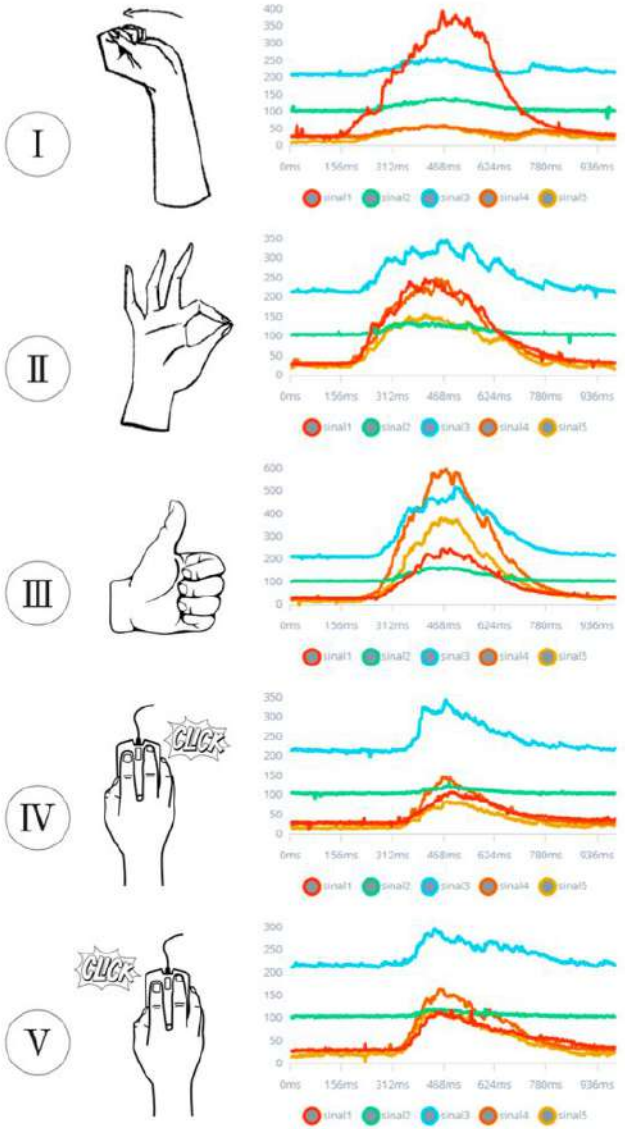
**Guilherme Silva**
**Matheus Lima**
Engenheiros - UNIFEI

# Surface electromyography



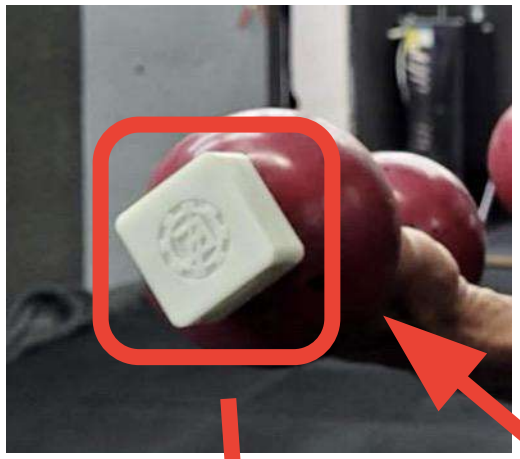


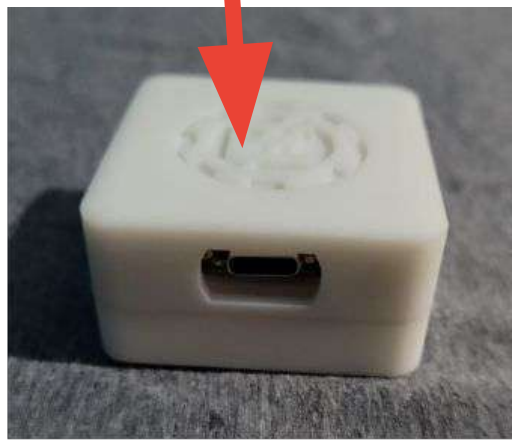


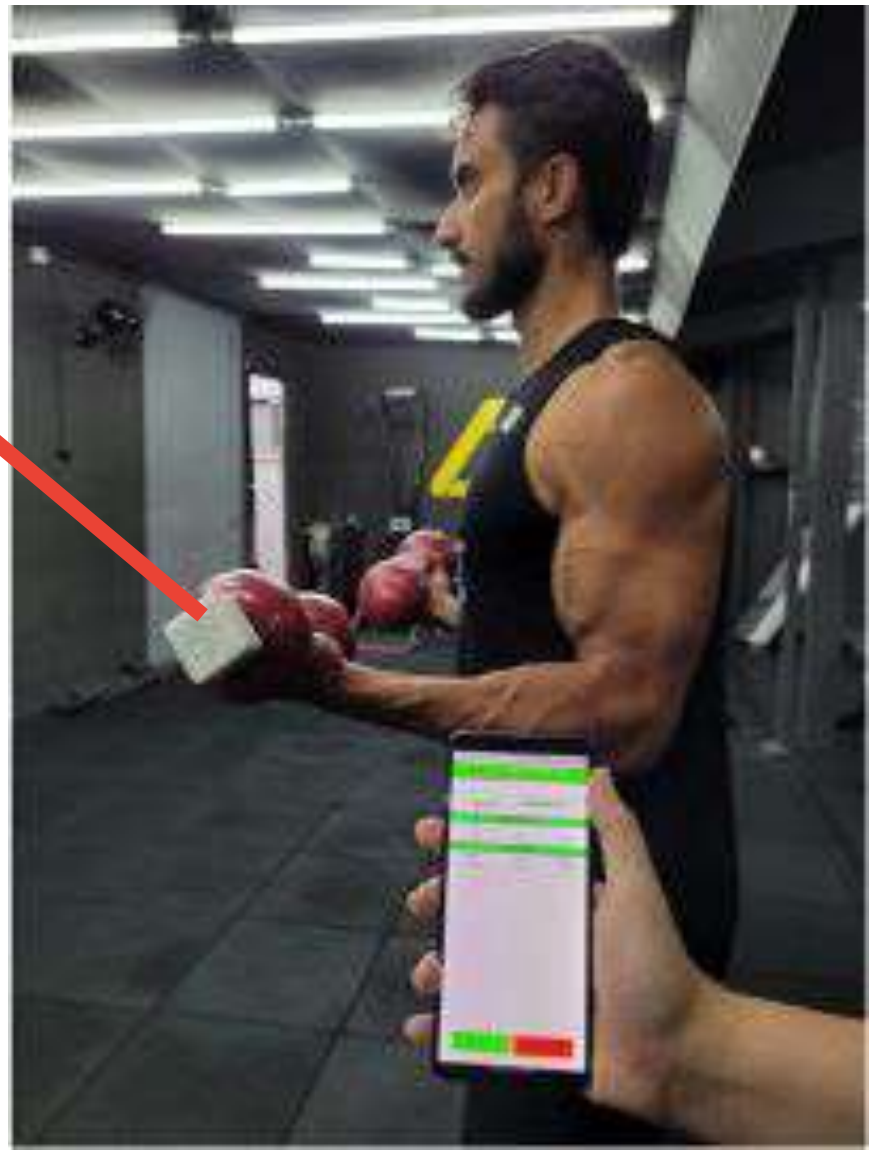



**Mateus F. Delangélica e Renato M. Neto,** UNIFEI

# Personal Trainer

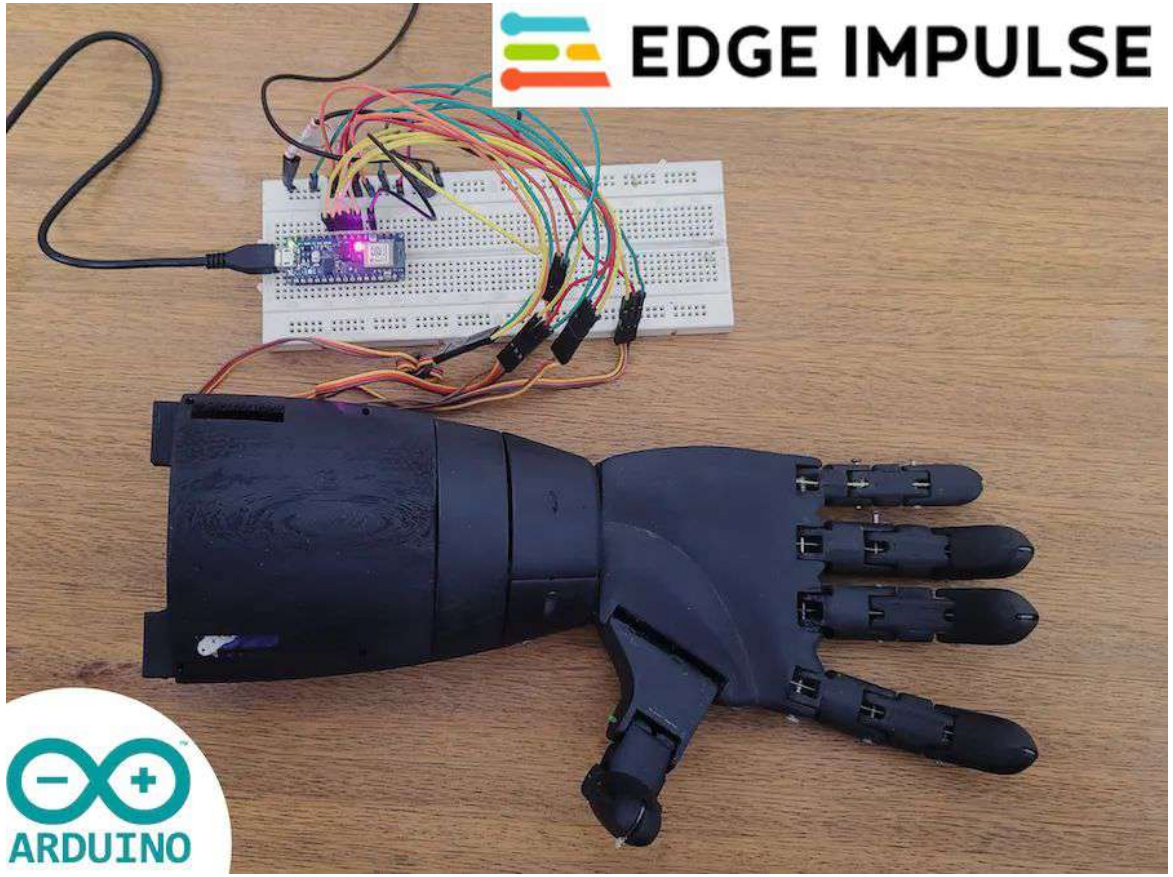






**Ricardo Magno C. Junior**
**Luiz Fernando Kikuchi**
UNIFEI

# Bionic Hand Voice Commands





https://www.hackster.io/ex-machina/bionic-hand-voice-commands-module-w-edge-impulse-arduino-aa97e3

# Real-World Applications

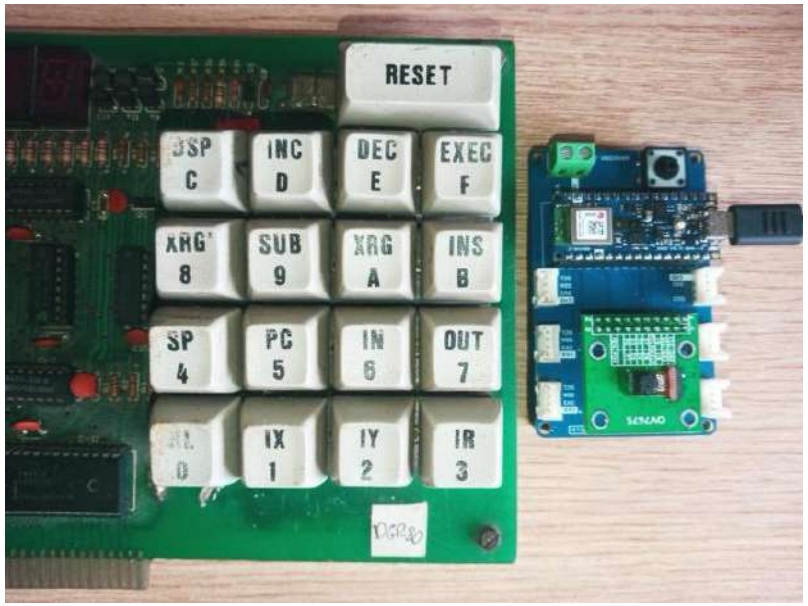
Agriculture

Healthcare

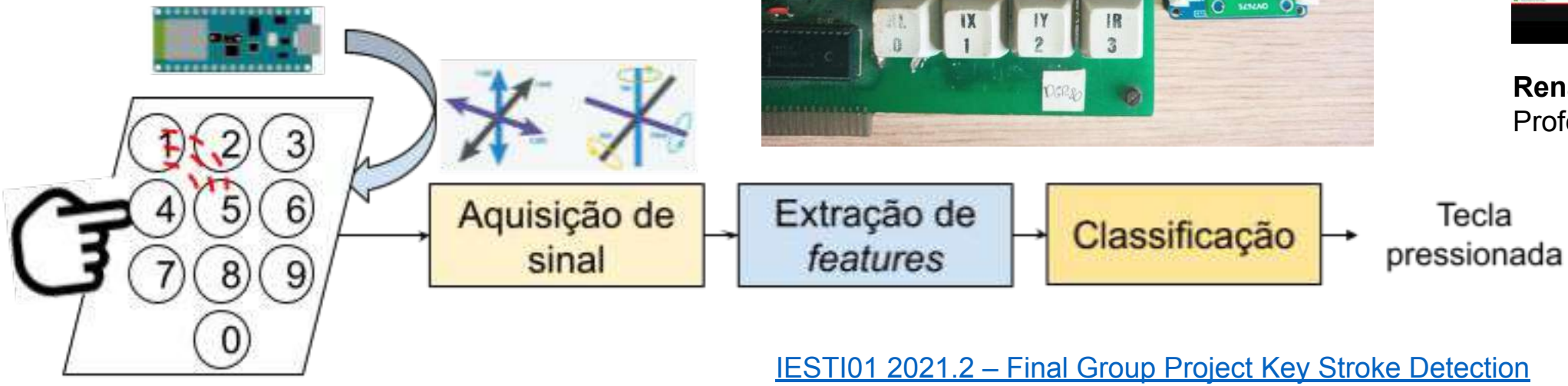
Industry

Environment

# Keystroke **Sound** Detection

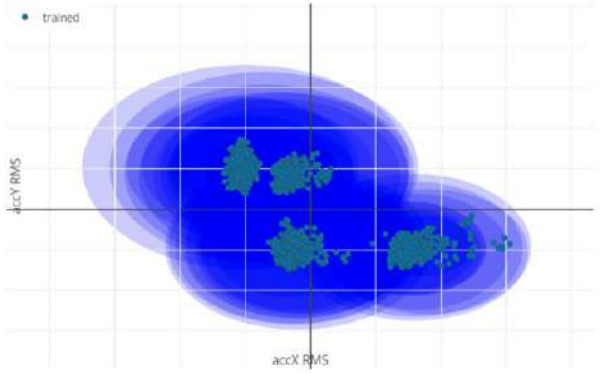


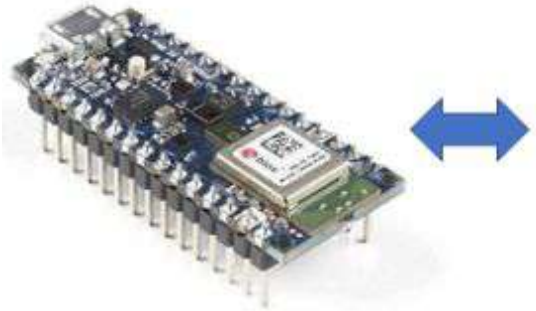

**Renam Castro**
Professor IFESP




[IESTI01 2021.2 – Final Group Project Key Stroke Detection]

# Industrial – Anomaly Detection



[IESTI01 2021.2 - Final Group Project: Bearing Failure Detection]

# Real-World Applications

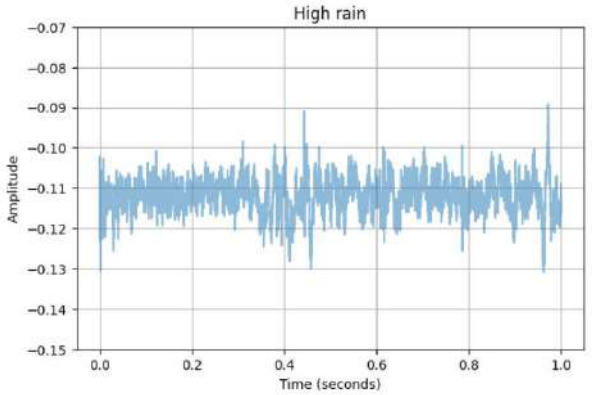
Agriculture

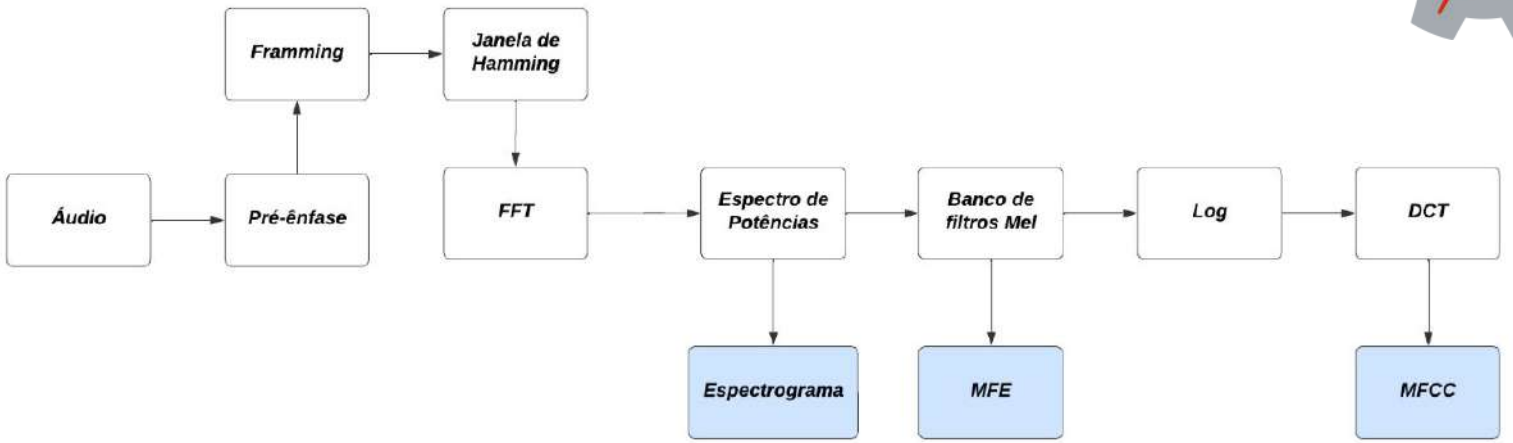
Healthcare

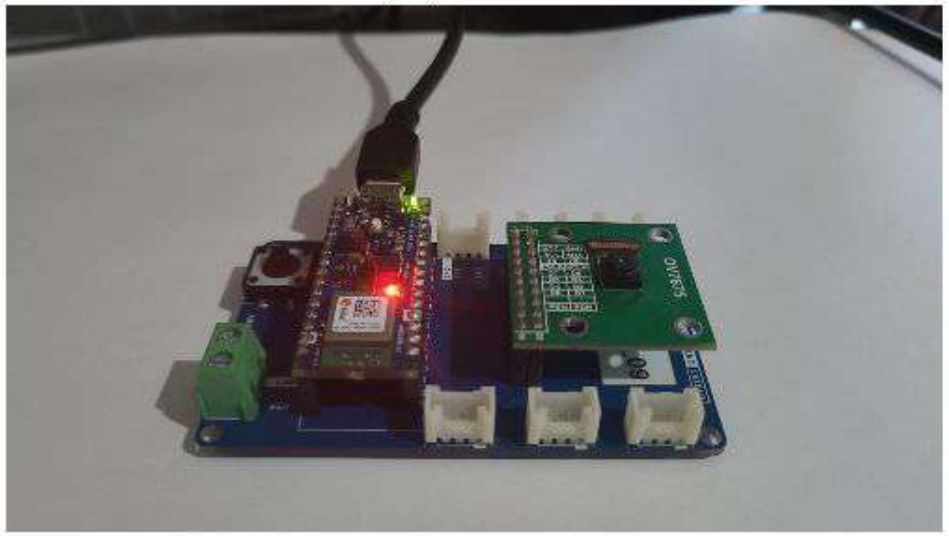
Industry

Environment

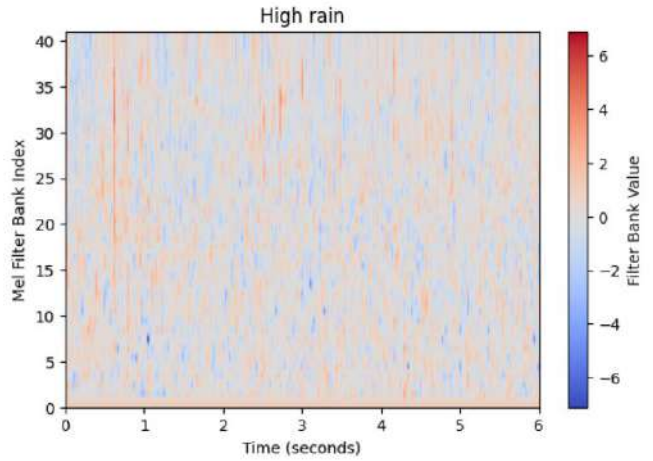
# Measure rainfall using sound detection



```
Edge Impulse standalone inferencing
run_classifier returned: 0
Timing: DSP 630 ms, inference 47 ms,
Predictions:
  HighRain: 0.99609
  LowRain: 0.00000
  MediumRain: 0.00000
  NoRain: 0.00000
```



# Challenges in MLSysEng

Managing messy and large-scale data.

Addressing 'data drift' in changing environments.

Ensuring real-world reliability and fairness.

Balancing transparency, privacy, and operational scalability.

# Trends and Opportunities

AutoML: Automating ML workflows for accessibility.

Explainable AI: Building trust through interpretability.

Efficiency: Developing energy-aware computing systems.

Democratization: Lowering barriers to AI adoption.

# Learning MLSysEng Hands-On with Labs

TinyML and EdgeAI are used to teach MLSysEng concepts practically.

Leveraging small, low-power devices such as Arduino Nicla Vision, Seeed XIAO ESP32S3, and Raspberry Pi.

Focuses on integrating hardware and software for real-world applications.

# Why TinyML and EdgeAI are a Game-Changer for Learning

Simplifies complex MLSysEng concepts

Enables hands-on experimentation with full workflows.

Demonstrates scalability and deployment on edge devices.

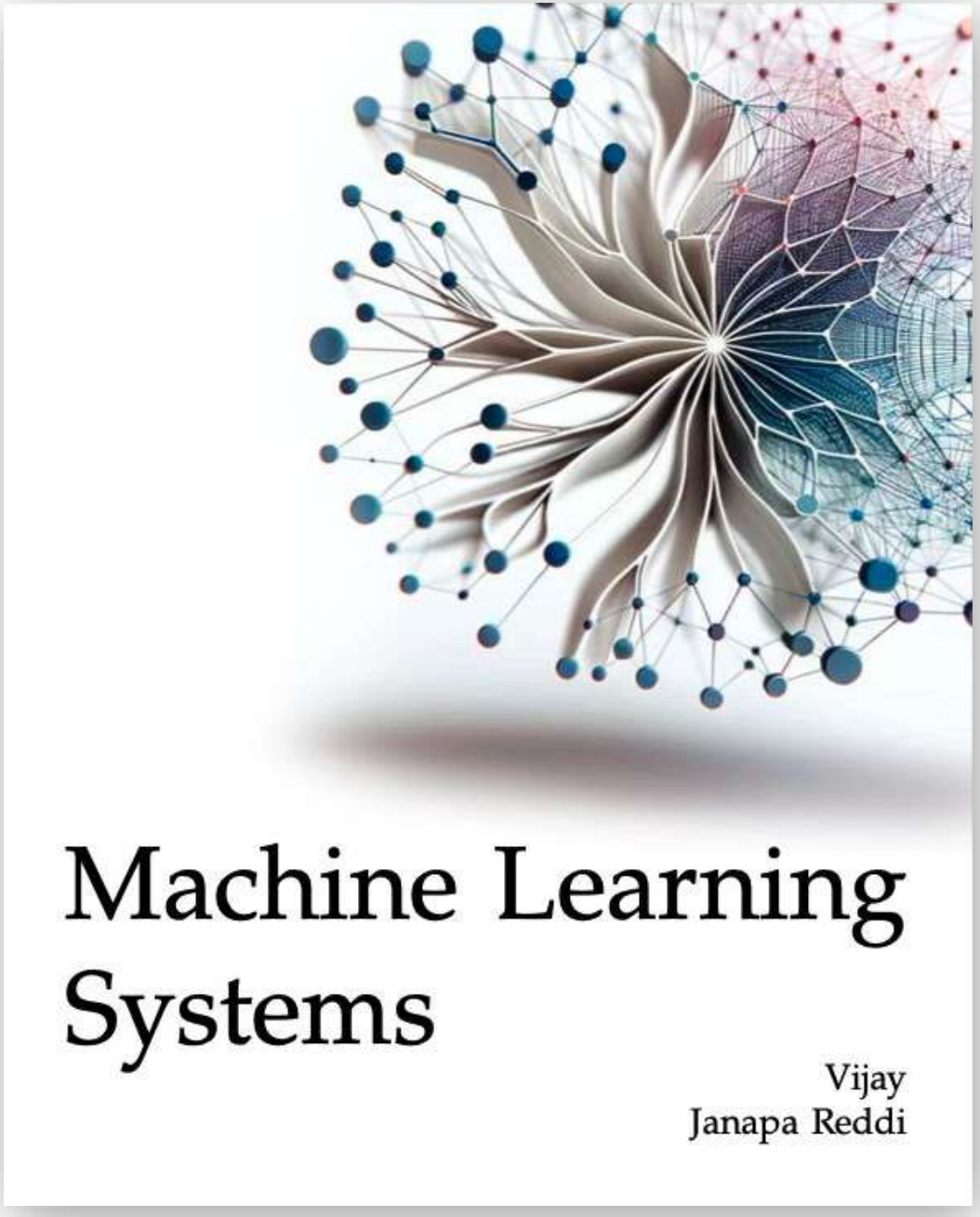
# Closing Thoughts

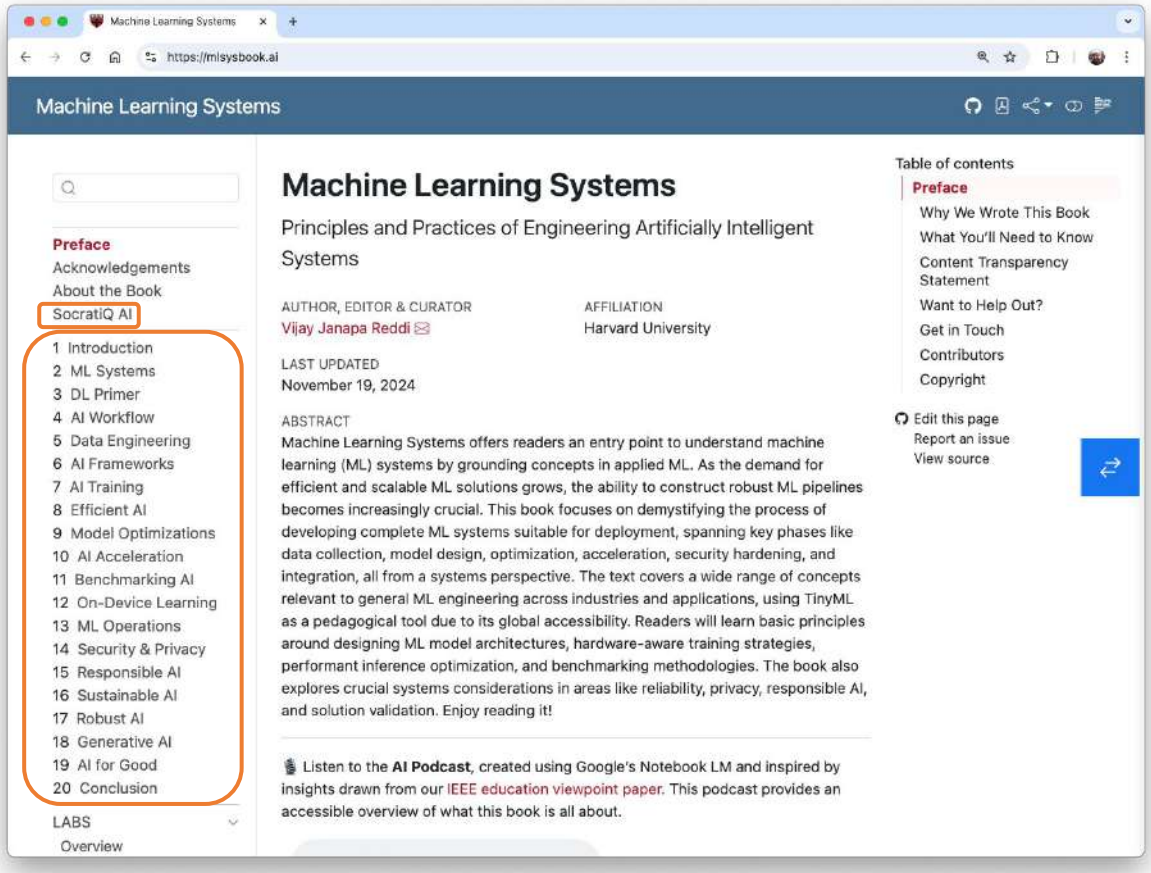
Machine Learning Systems Engineering integrates diverse skills and tools to turn AI algorithms into impactful real-world solutions.

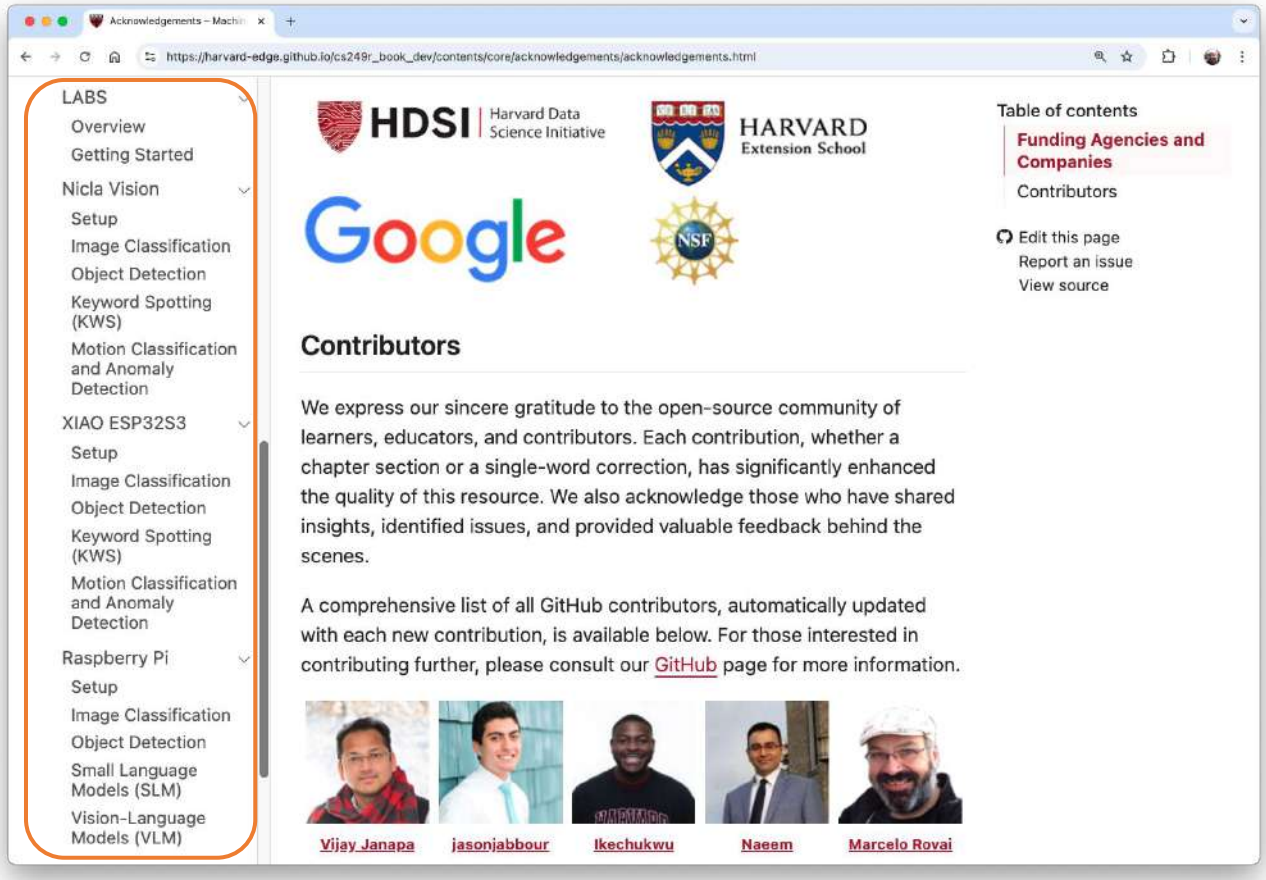
Hands-on labs make learning accessible and engaging.

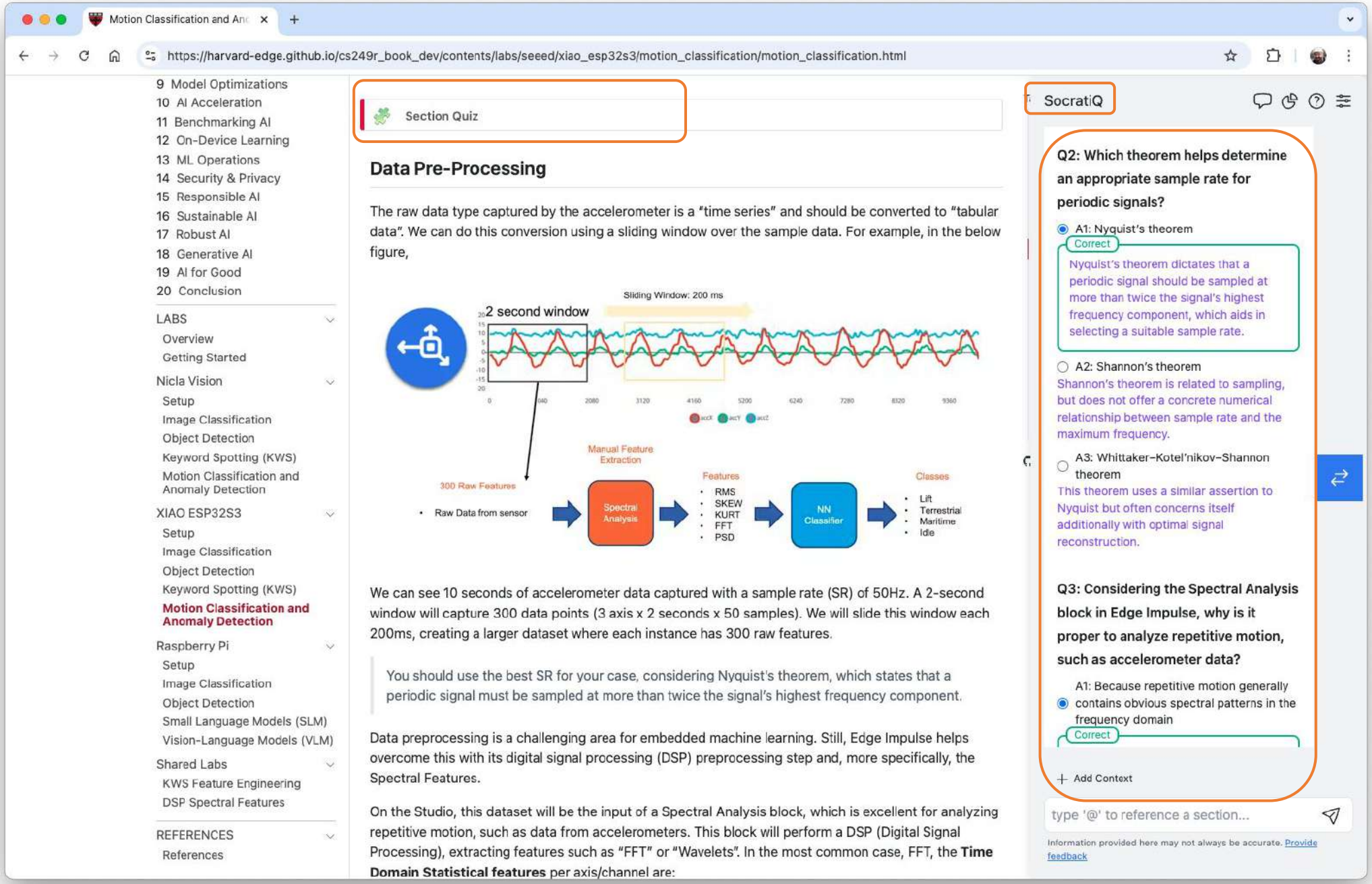
Resources like MLSysBook.ai provide comprehensive insights and hands-on tools to explore this transformative field.

https://mlsysbook.ai/




Machine Learning Systems
https://mlsysbook.ai
Machine Learning Systems
Preface
Acknowledgements
About the Book
SocratiQ AI
1 Introduction
2 ML Systems
3 DL Primer
4 AI Workflow
5 Data Engineering
6 AI Frameworks
7 AI Training
8 Efficient AI
9 Model Optimizations
10 AI Acceleration
11 Benchmarking AI
12 On-Device Learning
13 ML Operations
14 Security & Privacy
15 Responsible AI
16 Sustainable AI
17 Robust AI
18 Generative AI
19 AI for Good
20 Conclusion
LABS
Overview
Machine Learning Systems
Principles and Practices of Engineering Artificially Intelligent Systems
AUTHOR, EDITOR & CURATOR
Vijay Janapa Reddi
AFFILIATION
Harvard University
LAST UPDATED
November 19, 2024
ABSTRACT
Machine Learning Systems offers readers an entry point to understand machine learning (ML) systems by grounding concepts in applied ML. As the demand for efficient and scalable ML solutions grows, the ability to construct robust ML pipelines becomes increasingly crucial. This book focuses on demystifying the process of developing complete ML systems suitable for deployment, spanning key phases like data collection, model design, optimization, acceleration, security hardening, and integration, all from a systems perspective. The text covers a wide range of concepts relevant to general ML engineering across industries and applications, using TinyML as a pedagogical tool due to its global accessibility. Readers will learn basic principles around designing ML model architectures, hardware-aware training strategies, performant inference optimization, and benchmarking methodologies. The book also explores crucial systems considerations in areas like reliability, privacy, responsible AI, and solution validation. Enjoy reading it!
Listen to the AI Podcast, created using Google's Notebook LM and inspired by insights drawn from our IEEE education viewpoint paper. This podcast provides an accessible overview of what this book is all about.
Table of contents
Preface
Why We Wrote This Book
What You'll Need to Know
Content Transparency Statement
Want to Help Out?
Get in Touch
Contributors
Copyright
Edit this page
Report an issue
View source


https://harvard-edge.github.io/cs249r_book_dev/contents/core/acknowledgements/acknowledgements.html
LABS
Overview
Getting Started
Nicla Vision
Setup
Image Classification
Object Detection
Keyword Spotting (KWS)
Motion Classification and Anomaly Detection
XIAO ESP32S3
Setup
Image Classification
Object Detection
Keyword Spotting (KWS)
Motion Classification and Anomaly Detection
Raspberry Pi
Setup
Image Classification
Object Detection
Small Language Models (SLM)
Vision-Language Models (VLM)
HDSI Harvard Data Science Initiative
HARVARD Extension School
Google
NSF
Contributors
We express our sincere gratitude to the open-source community of learners, educators, and contributors. Each contribution, whether a chapter section or a single-word correction, has significantly enhanced the quality of this resource. We also acknowledge those who have shared insights, identified issues, and provided valuable feedback behind the scenes.
A comprehensive list of all GitHub contributors, automatically updated with each new contribution, is available below. For those interested in contributing further, please consult our GitHub page for more information.
Vijay Janapa
jasonjabbour
Ikechukwu
Naeem
Marcelo Rovai
Table of contents
Funding Agencies and Companies
Contributors
Edit this page
Report an issue
View source


Motion Classification and An
https://harvard-edge.github.io/cs249r_book_dev/contents/labs/seeed/xiao_esp32s3/motion_classification/motion_classification.html
9 Model Optimizations
10 AI Acceleration
11 Benchmarking AI
12 On-Device Learning
13 ML Operations
14 Security & Privacy
15 Responsible AI
16 Sustainable AI
17 Robust AI
18 Generative AI
19 AI for Good
20 Conclusion
LABS
Overview
Getting Started
Nicla Vision
Setup
Image Classification
Object Detection
Keyword Spotting (KWS)
Motion Classification and Anomaly Detection
XIAO ESP32S3
Setup
Image Classification
Object Detection
Keyword Spotting (KWS)
Motion Classification and Anomaly Detection
Raspberry Pi
Setup
Image Classification
Object Detection
Small Language Models (SLM)
Vision-Language Models (VLM)
Shared Labs
KWS Feature Engineering
DSP Spectral Features
REFERENCES
References
Section Quiz
Data Pre-Processing
The raw data type captured by the accelerometer is a "time series" and should be converted to "tabular data". We can do this conversion using a sliding window over the sample data. For example, in the below figure,
Sliding Window: 200 ms
2 second window
Manual Feature Extraction
300 Raw Features
Raw Data from sensor
Spectral Analysis
Features
RMS
SKEW
KURT
FFT
PSD
NN Classifier
Classes
Lift
Terrestrial
Maritime
Idle
We can see 10 seconds of accelerometer data captured with a sample rate (SR) of 50Hz. A 2-second window will capture 300 data points (3 axis x 2 seconds x 50 samples). We will slide this window each 200ms, creating a larger dataset where each instance has 300 raw features.
You should use the best SR for your case, considering Nyquist's theorem, which states that a periodic signal must be sampled at more than twice the signal's highest frequency component.
Data preprocessing is a challenging area for embedded machine learning. Still, Edge Impulse helps overcome this with its digital signal processing (DSP) preprocessing step and, more specifically, the Spectral Features.
On the Studio, this dataset will be the input of a Spectral Analysis block, which is excellent for analyzing repetitive motion, such as data from accelerometers. This block will perform a DSP (Digital Signal Processing), extracting features such as "FFT" or "Wavelets". In the most common case, FFT, the Time Domain Statistical features per axis/channel are:
SocratiQ
Q2: Which theorem helps determine an appropriate sample rate for periodic signals?
A1: Nyquist's theorem
Correct
Nyquist's theorem dictates that a periodic signal should be sampled at more than twice the signal's highest frequency component, which aids in selecting a suitable sample rate.
A2: Shannon's theorem
Shannon's theorem is related to sampling, but does not offer a concrete numerical relationship between sample rate and the maximum frequency.
A3: Whittaker–Kotel'nikov–Shannon theorem
This theorem uses a similar assertion to Nyquist but often concerns itself additionally with optimal signal reconstruction.
Q3: Considering the Spectral Analysis block in Edge Impulse, why is it proper to analyze repetitive motion, such as accelerometer data?
A1: Because repetitive motion generally contains obvious spectral patterns in the frequency domain
Correct
Add Context
type '@' to reference a section...
Information provided here may not always be accurate. Provide feedback

# Content

MLSys: Machine Learning Systems

GenAI: Introduction and Demo

# Generative AI (GenAI)

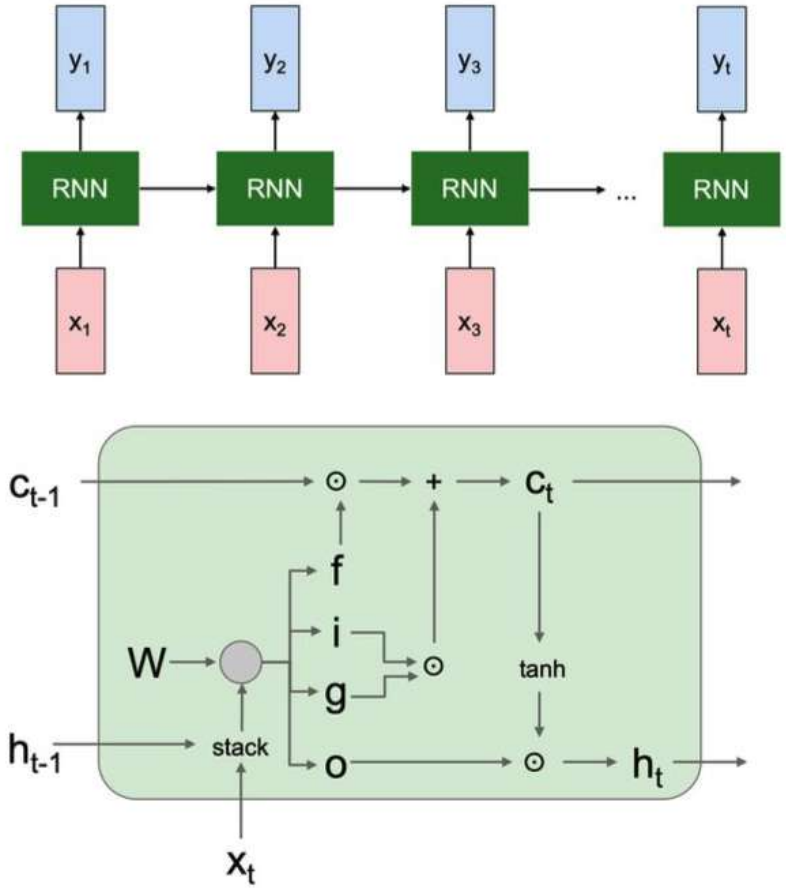
Generative AI is an artificial intelligence system capable of creating new, original content across various mediums such as text, images, audio, and video. These systems learn patterns from existing data and use that knowledge to generate novel outputs that didn't previously exist.

When used for generative tasks, Large Language Models (LLMs), Small Language Models (SLMs), and Visual-Language Models (VLMs) can all be considered types of GenAI.

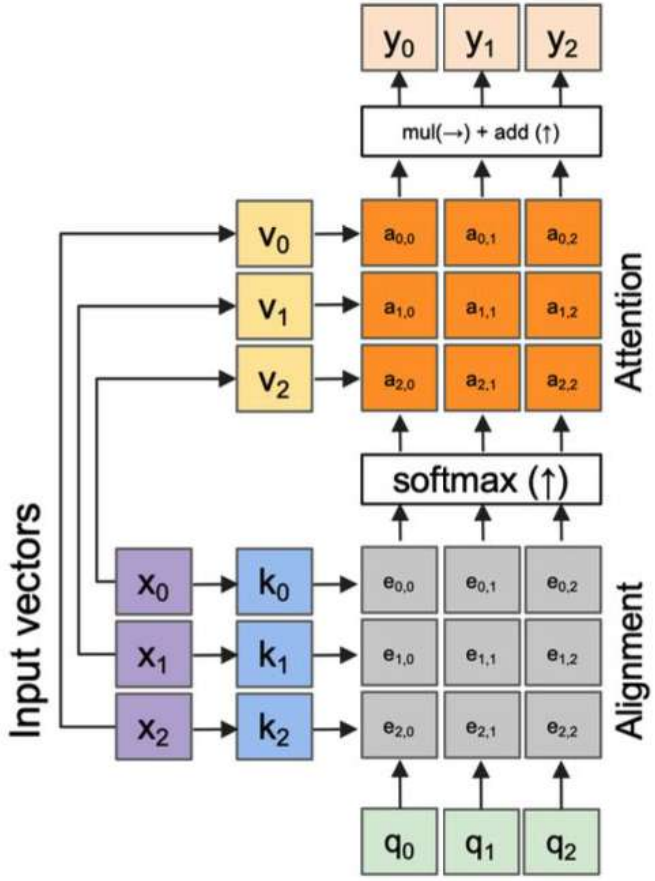
# LLM / SLM

## Large Language Model / Small Language Models

# Recap: Models Beyond DNN and CNN




Recurrent neural network




Attention mechanism / Transformers

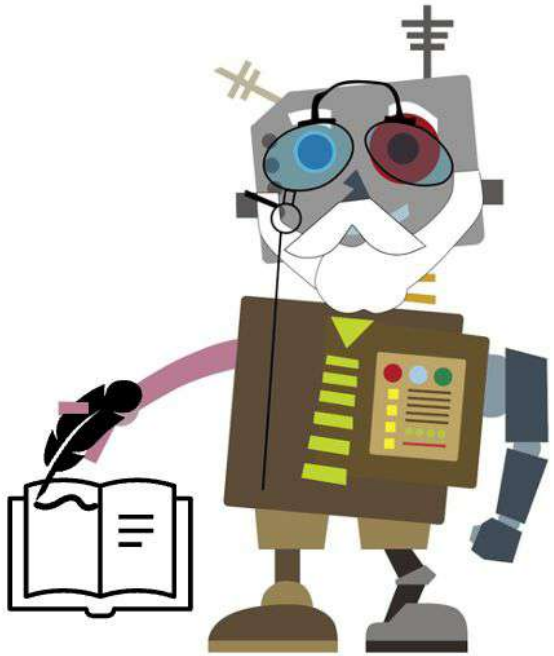
# Machado de Assis Bot with RNN - GRU


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**

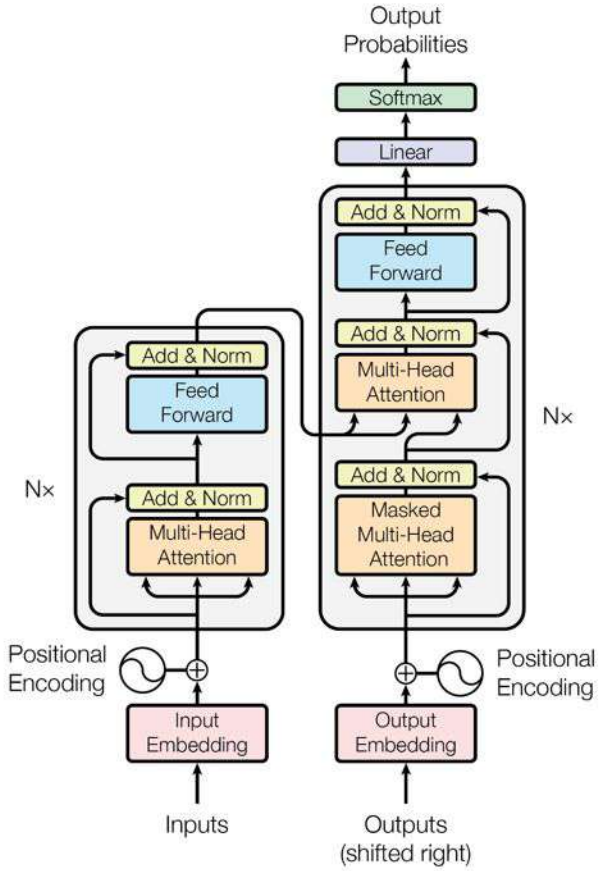
*A missa do coupé e um presente e o governo devia cazar logo no papel, a morte do autor, e todos os seus considerados de alegria. Era um espirito de vinte e cinco annos, e eu não estou alguns passos no cerebro, como de outra cousa. Deus me disse:*

*--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.*

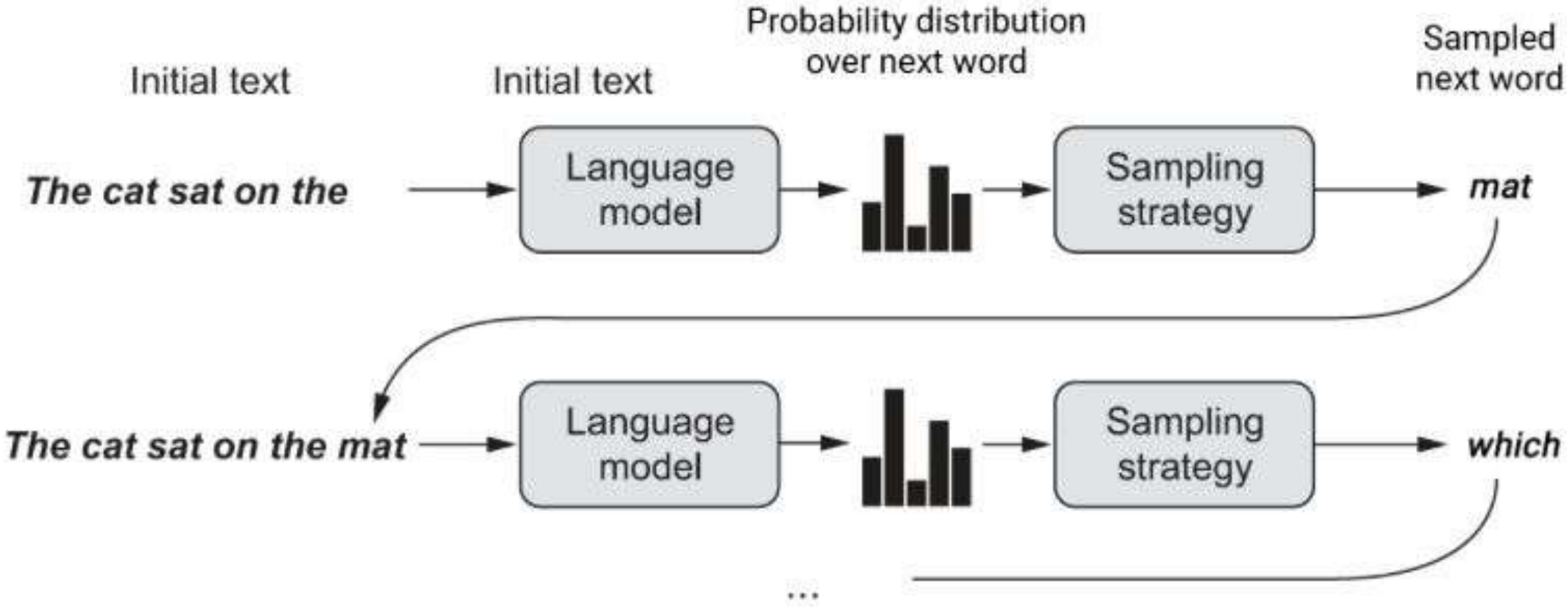
# LLM/SLM – Large /Small Language Model

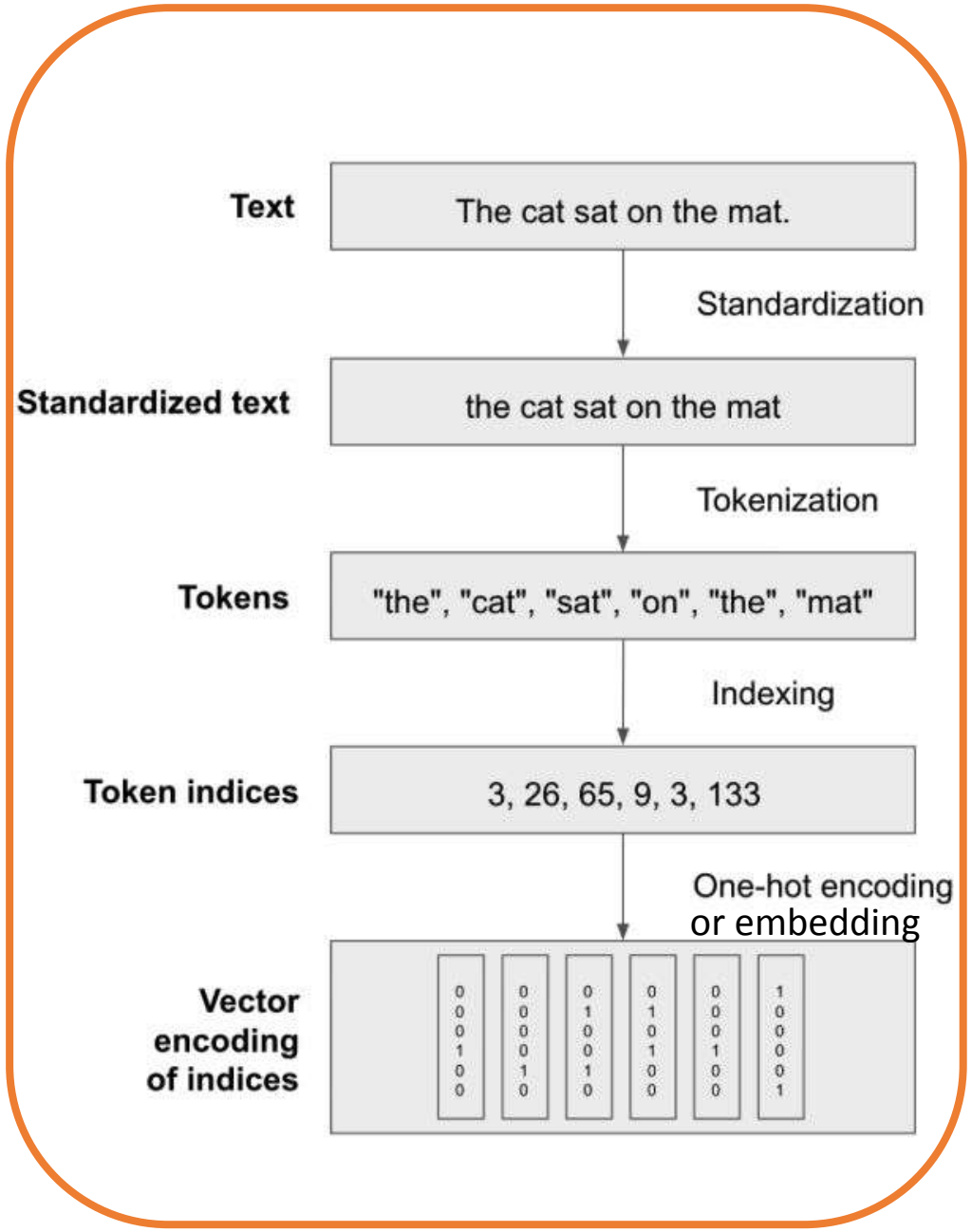
Large Language Models (LLMs) and SLMs are advanced neural networks based on the Transformer architecture that excel in understanding and generating human language. They represent a significant evolution from earlier sequence-based mode like RNNs, which surpass them in handling long-range dependencies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

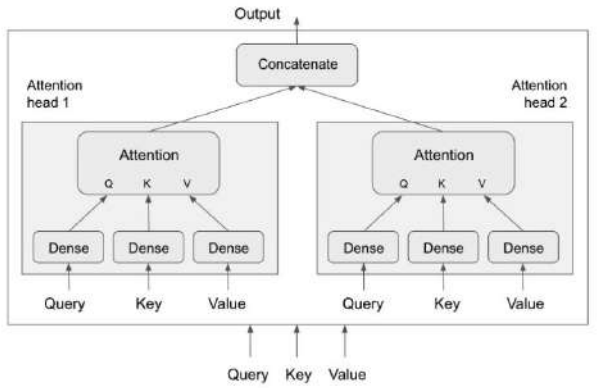



The Illustrated Transformer

# LLM/SLM – Large /Small Language Model





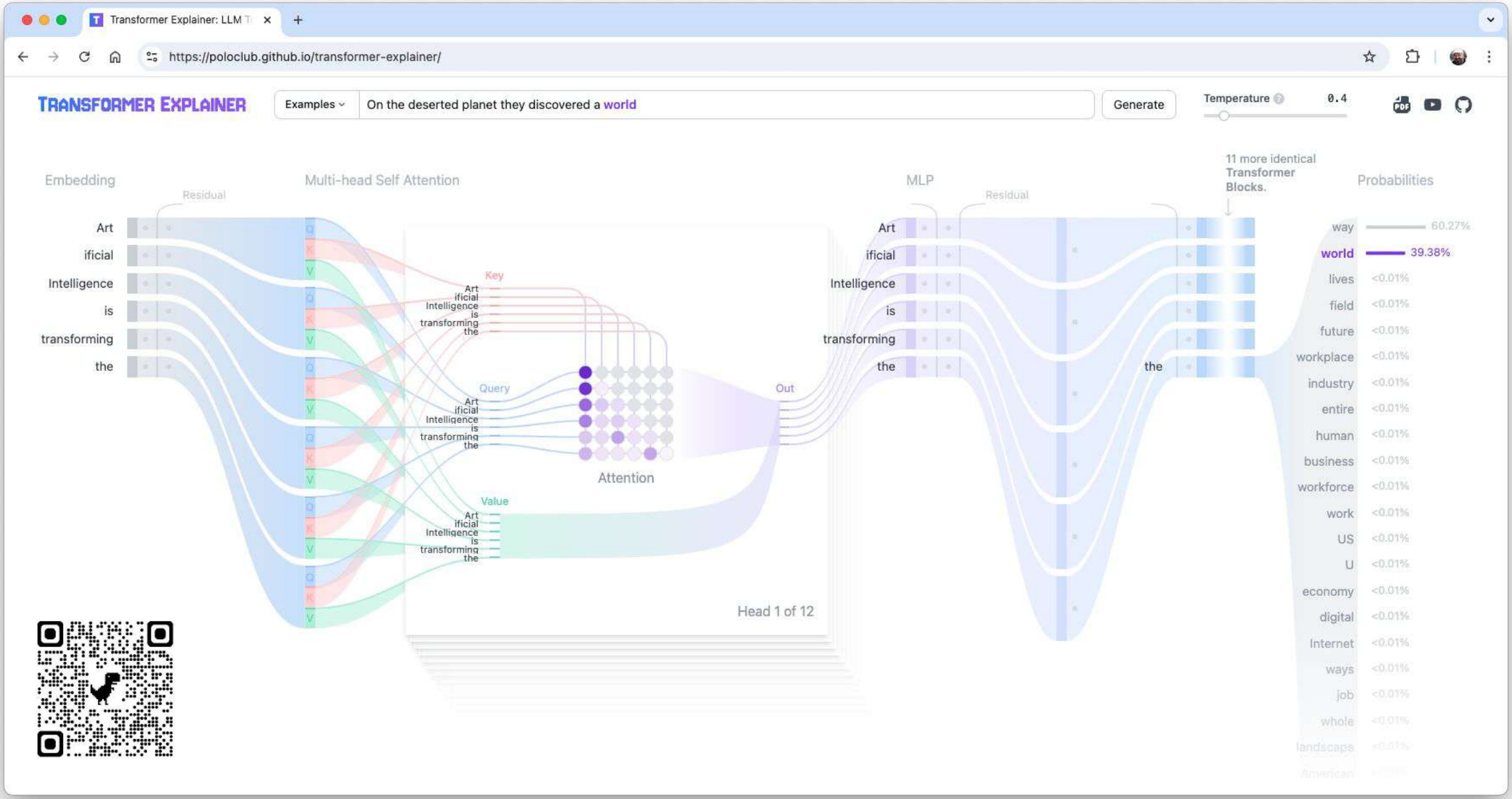




Encoder Decoder

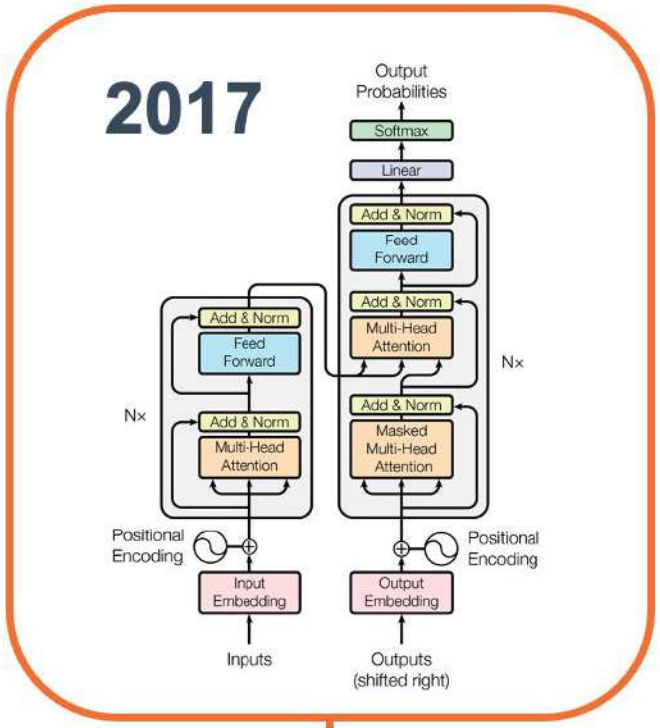
# LLM/SLM – Large /Small Language Model

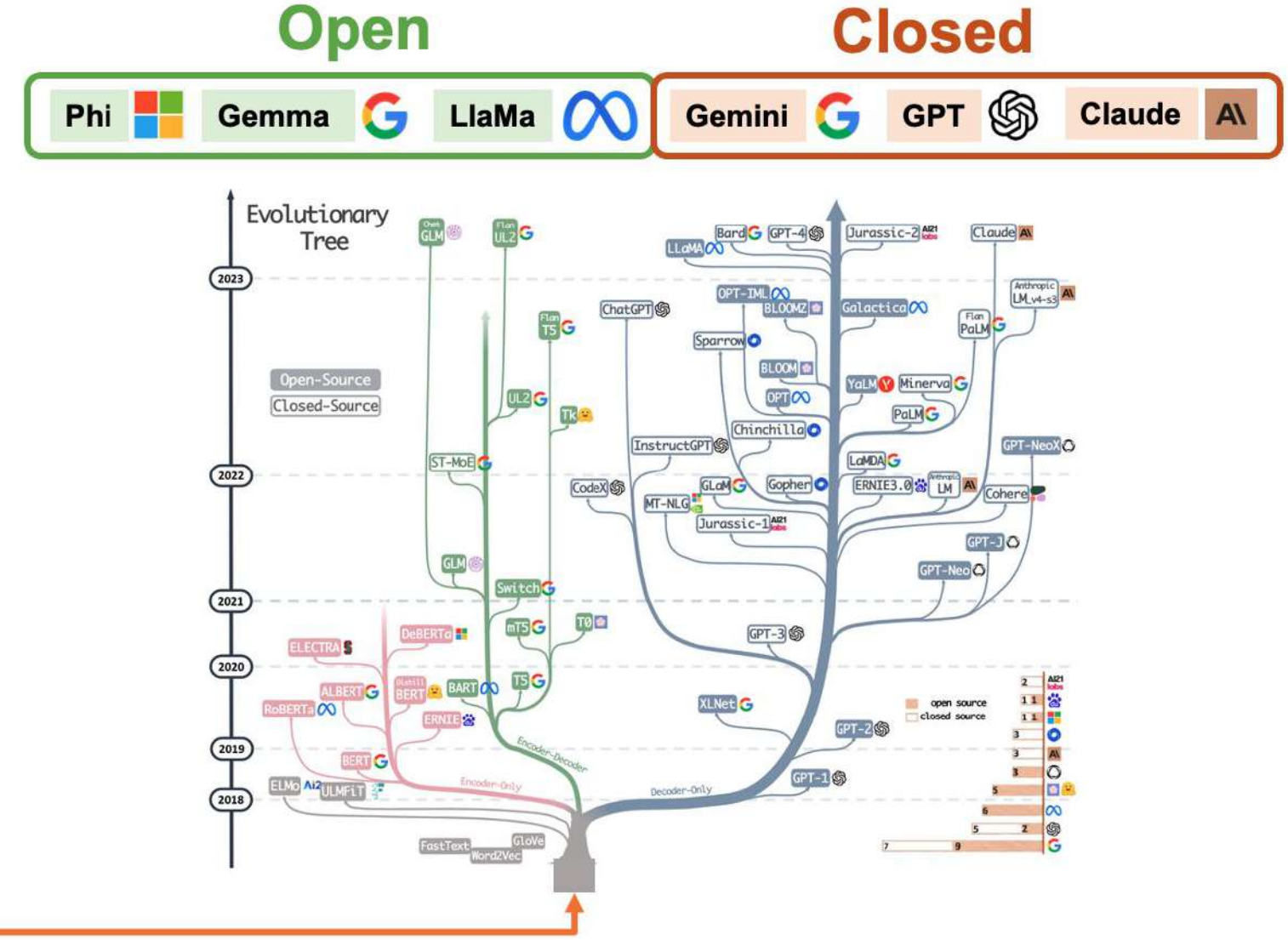

How large language models work

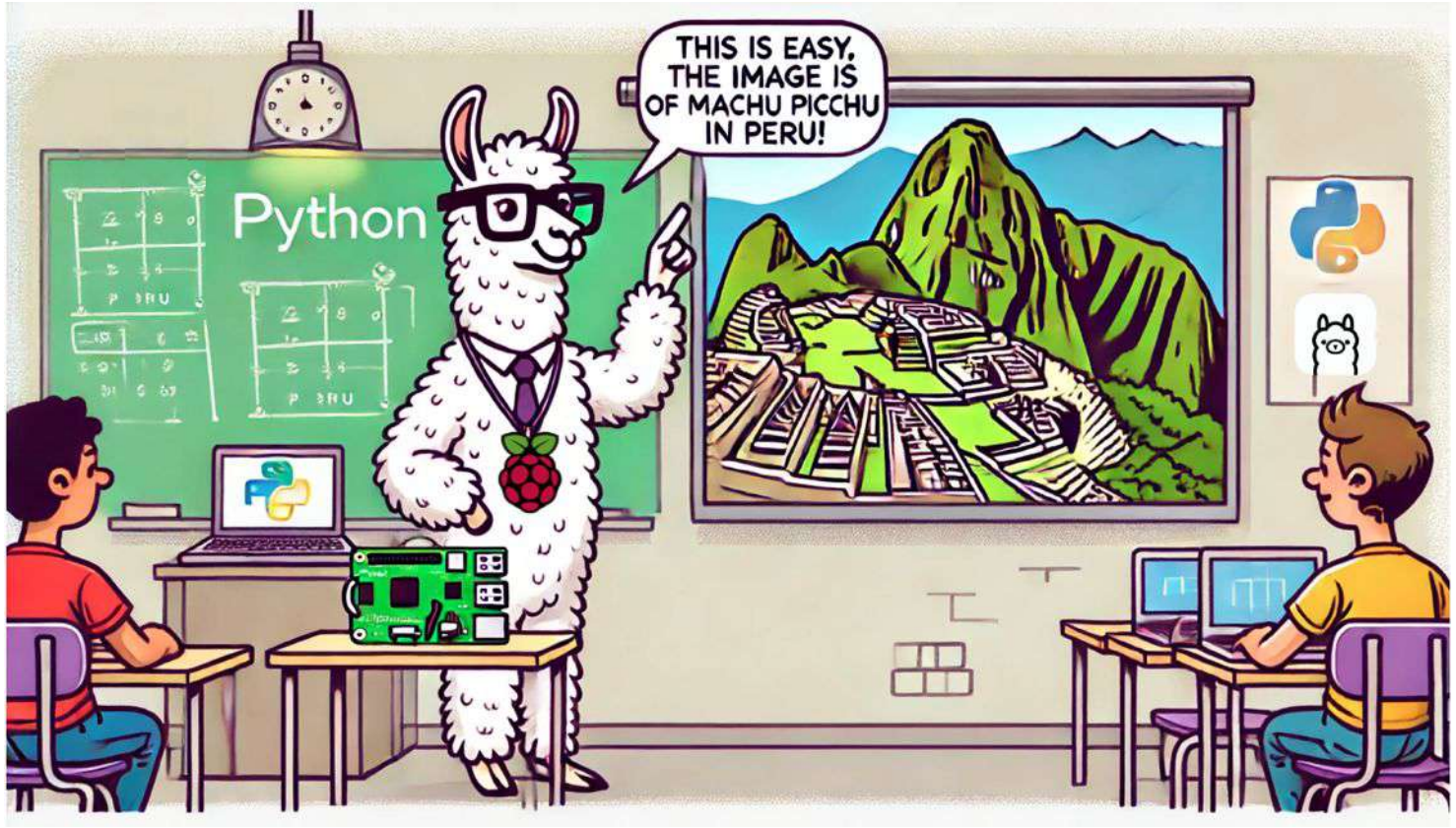



https://poloclub.github.io/transformer-explainer/

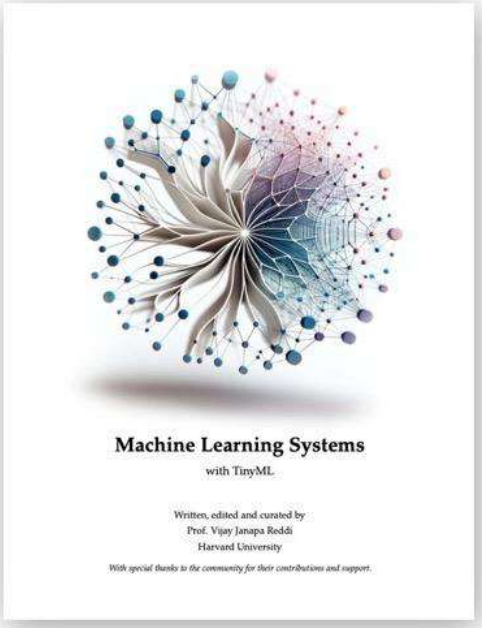
# Transformers to LLMs and SLMs

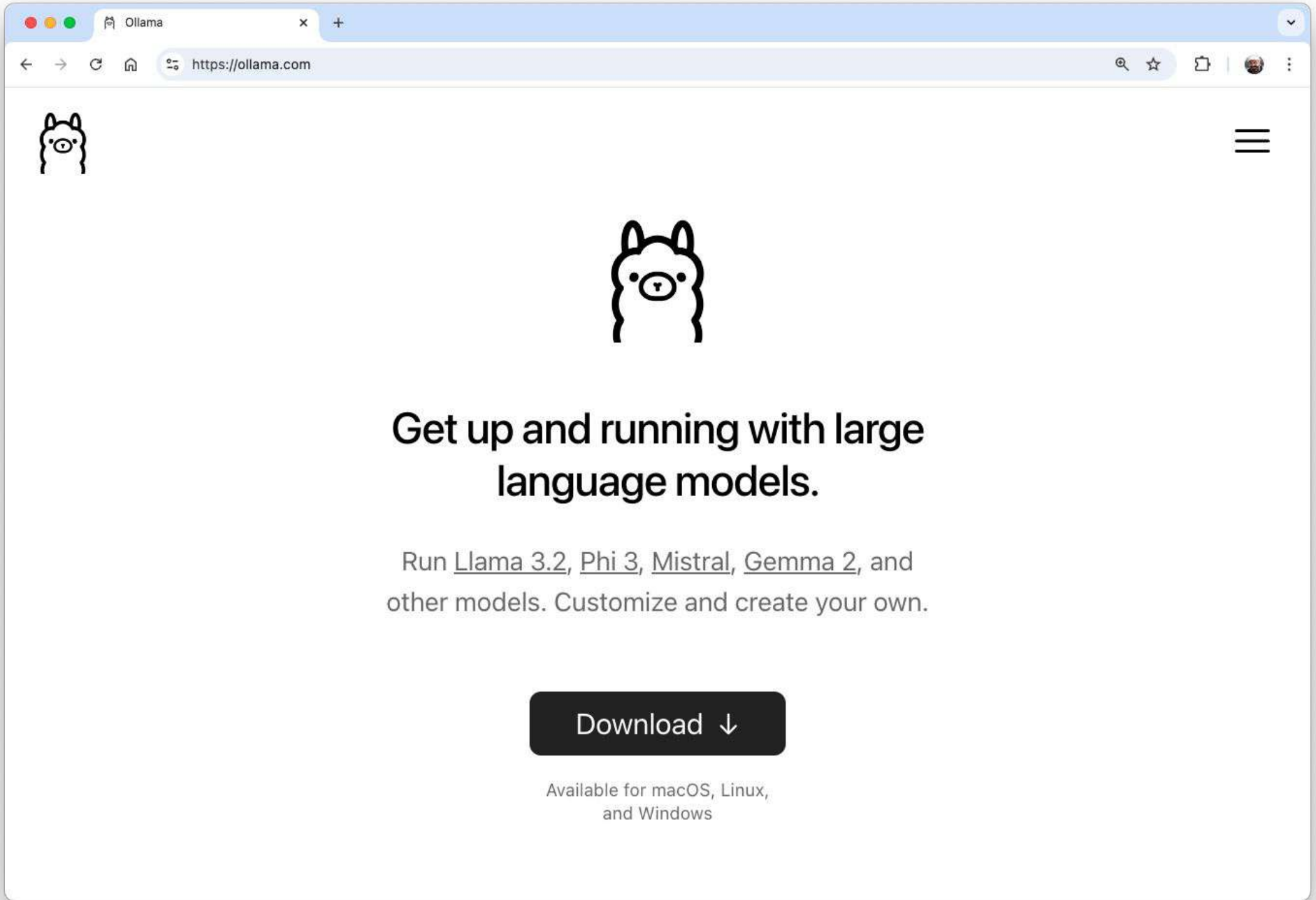


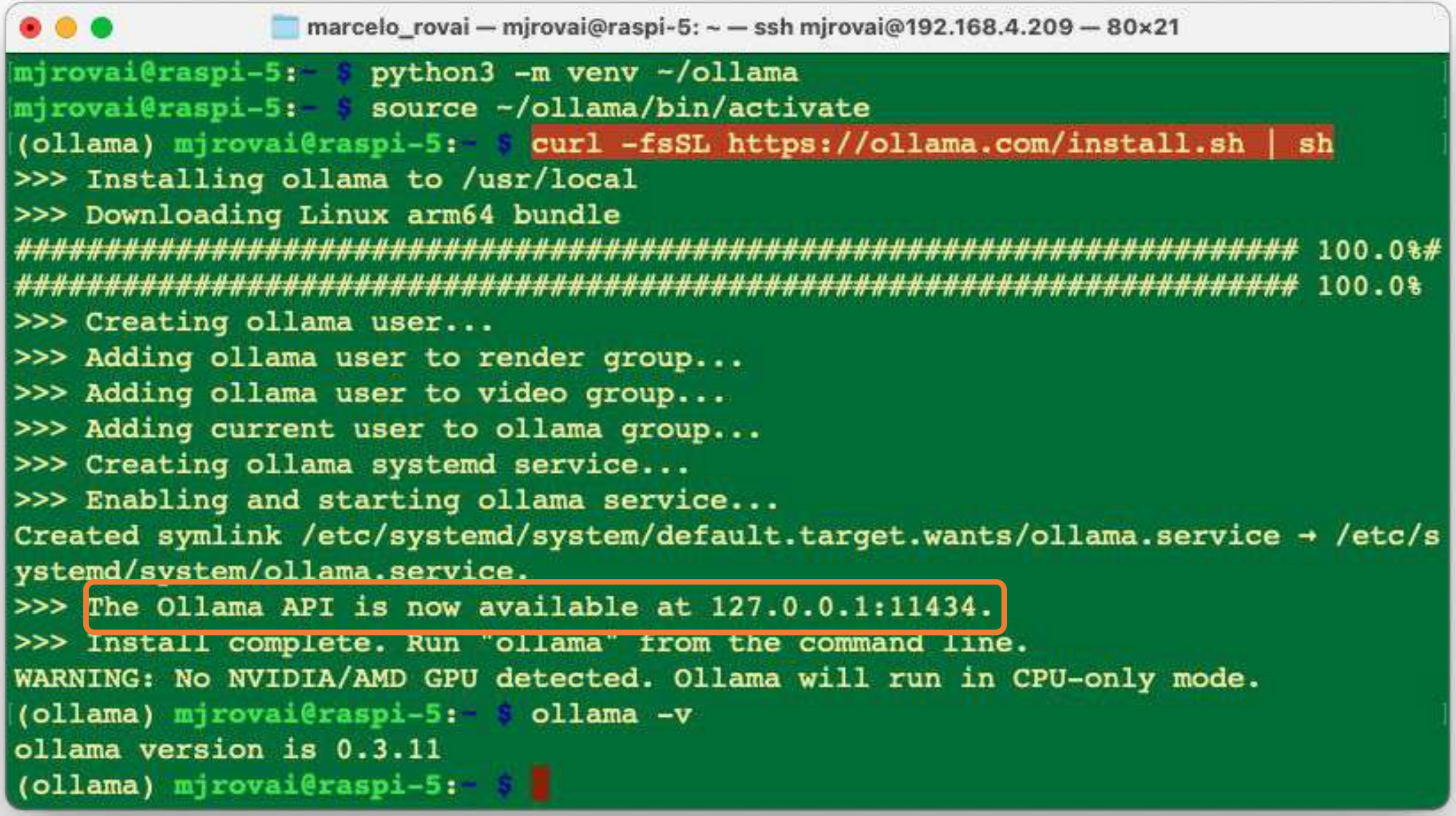



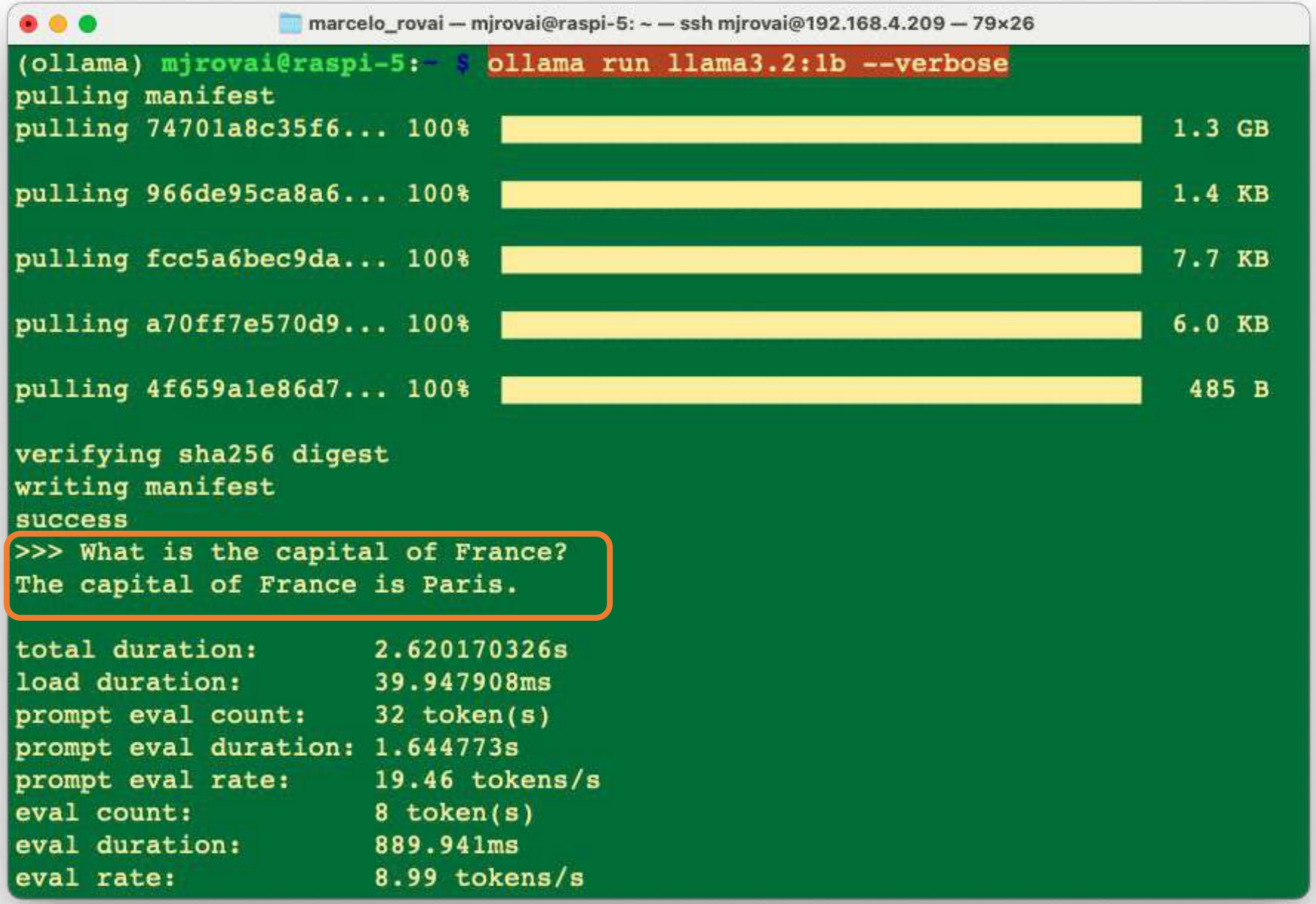



# Running Large Language Models on Raspberry Pi at the Edge




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:~ $

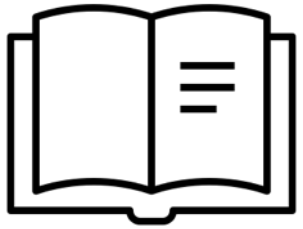

marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 79×26
(ollama) mjrovai@raspi-5:~ $ ollama run llama3.2:1b --verbose
pulling manifest
pulling 74701a8c35f6... 100% 1.3 GB
pulling 966de95ca8a6... 100% 1.4 KB
pulling fcc5a6bec9da... 100% 7.7 KB
pulling a70ff7e570d9... 100% 6.0 KB
pulling 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 $
```

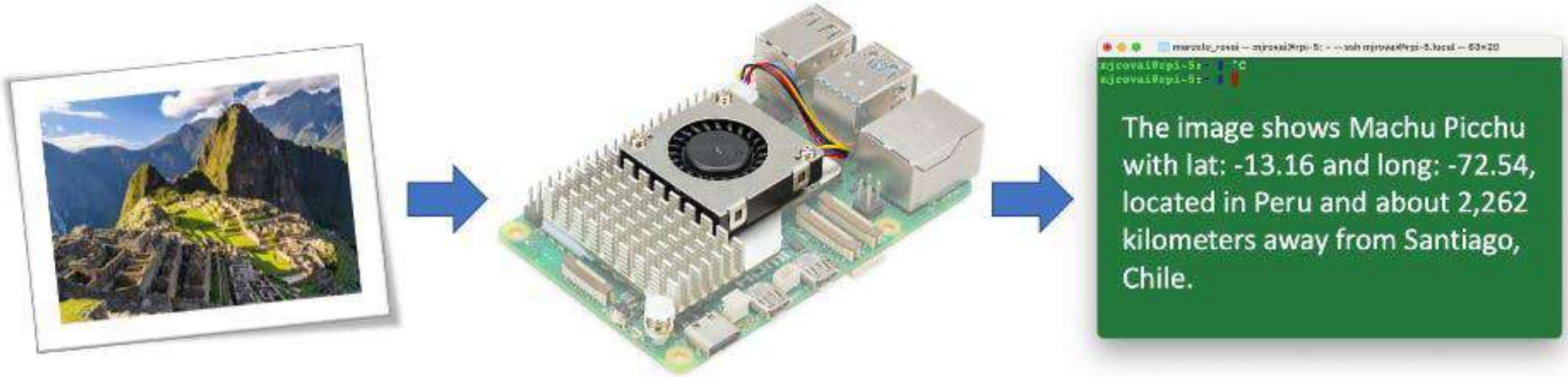
```
mjrovai@rpi-5: ~
File  Edit  Tabs  Help
>>> Awswer with one short sentence, what is the capital of France and its distanc
... e in Km from Santiago, Chile
 The capital of France is Paris and it is around 12,674 kilometers away
from Santiago, Chile.

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)
```

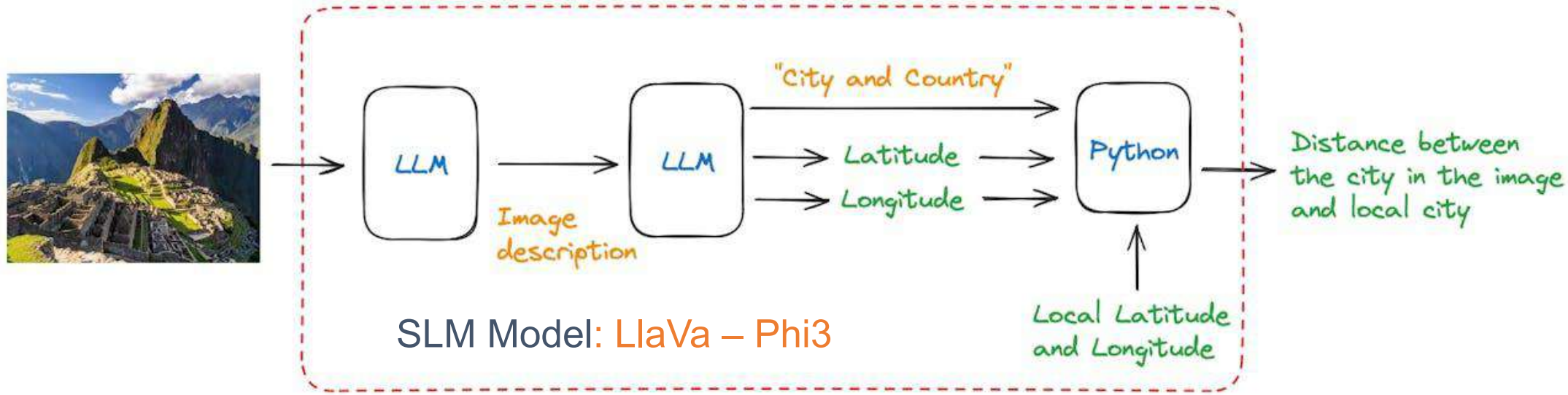
(13 seconds)

```
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



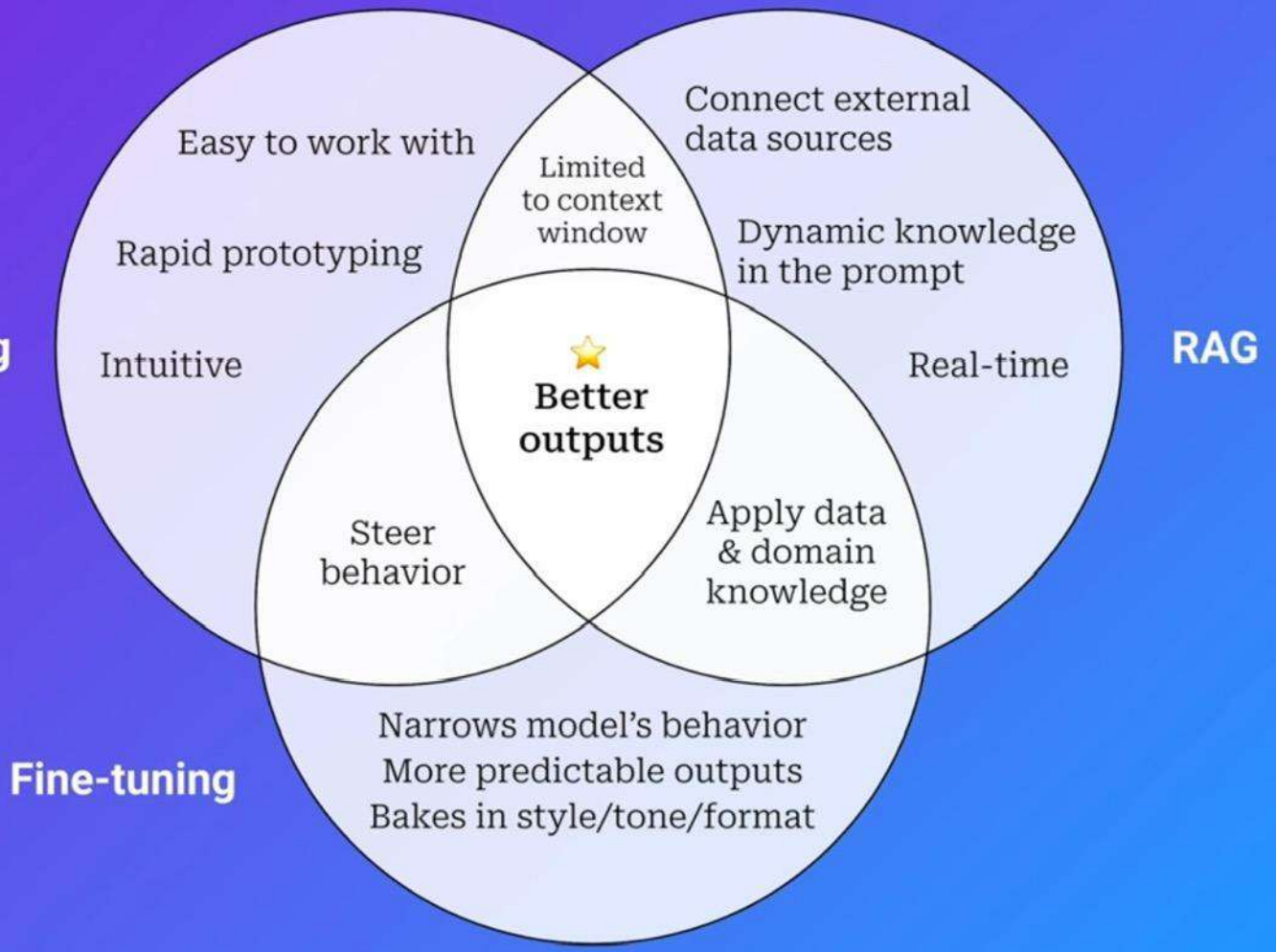



# LLMs: Optimization Techniques

# LLMs: Common Optimization Techniques

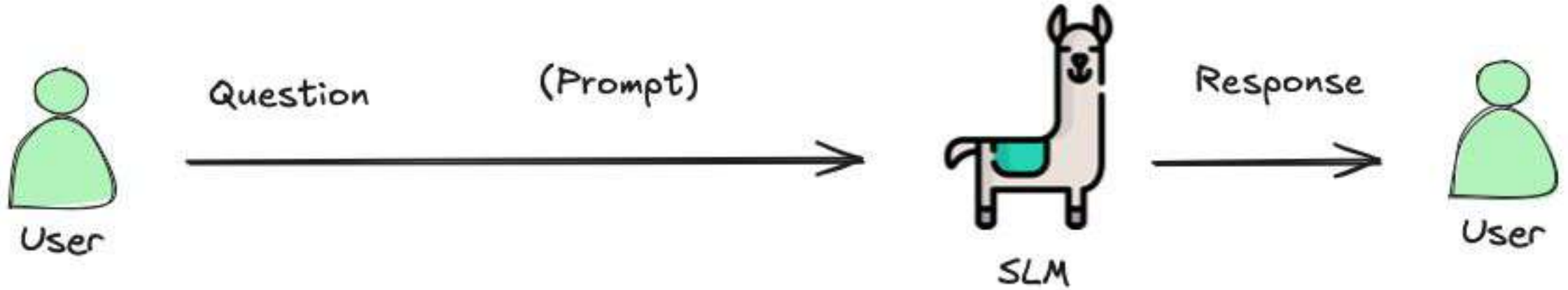
1. **Prompt Engineering**: Tailor your interactions.
2. **Fine-tuning**: Perfect the model's tasks.
3. **RAG**: Enhance with relevant data.

# Comparison of Techniques

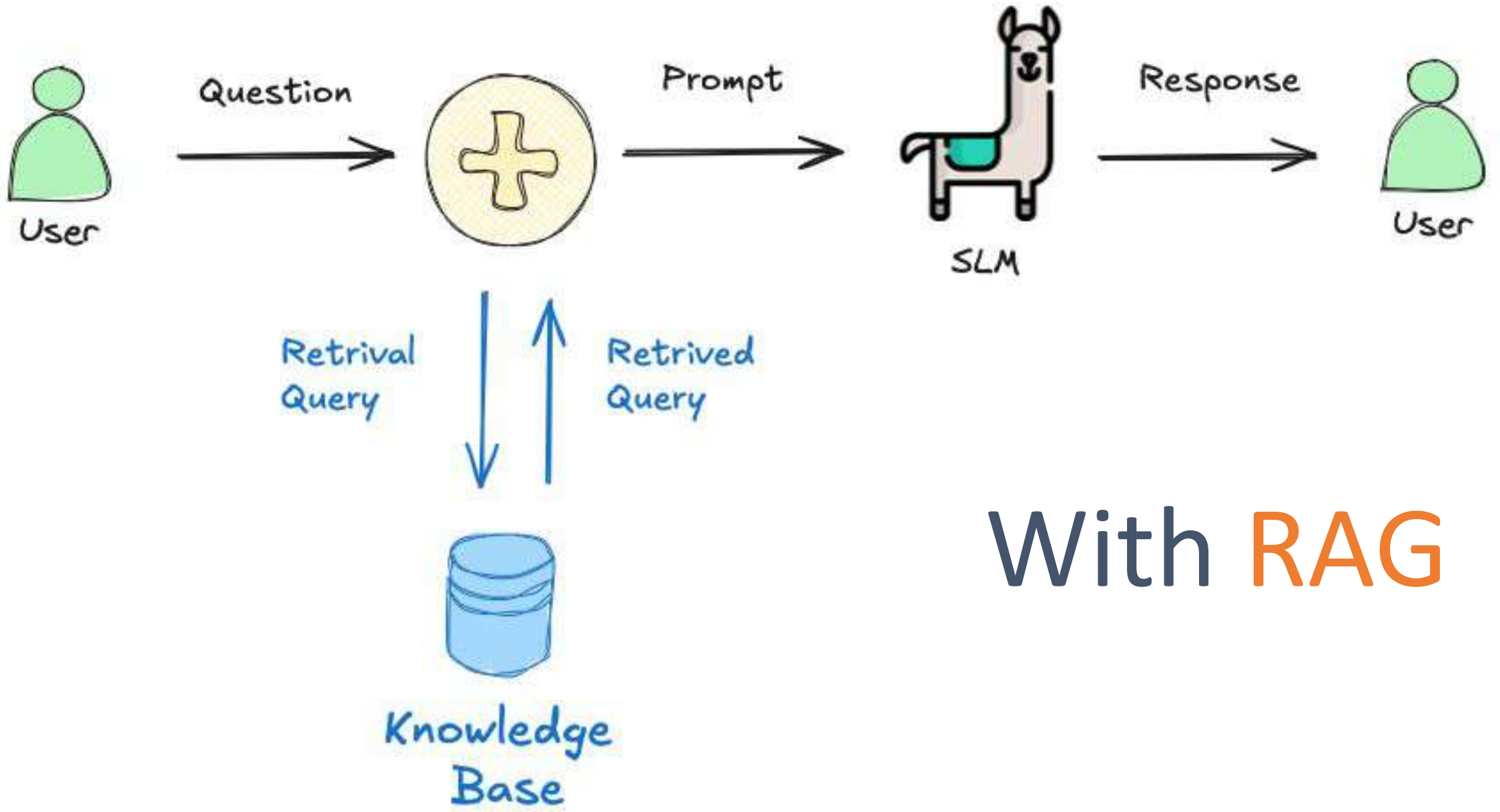


# Retrieval-Augmented Generation (RAG)

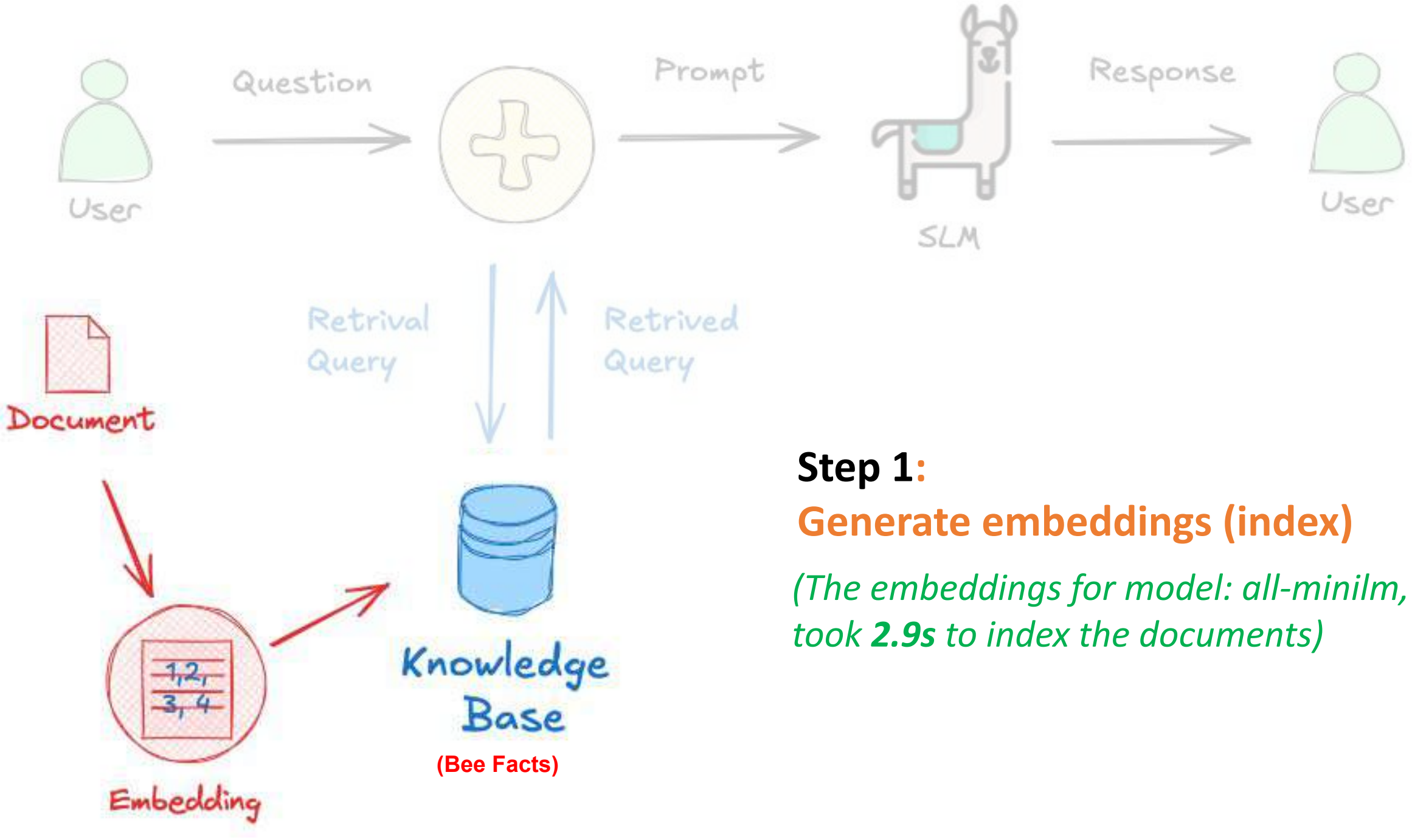
“A method created by the FAIR team at Meta to enhance the accuracy of Large Language Models (LLMs) and reduce false information or “hallucinations.”




# Usual Prompt


User
Question
Prompt
SLM
Response
User
Retrival Query
Retrived Query
Knowledge Base

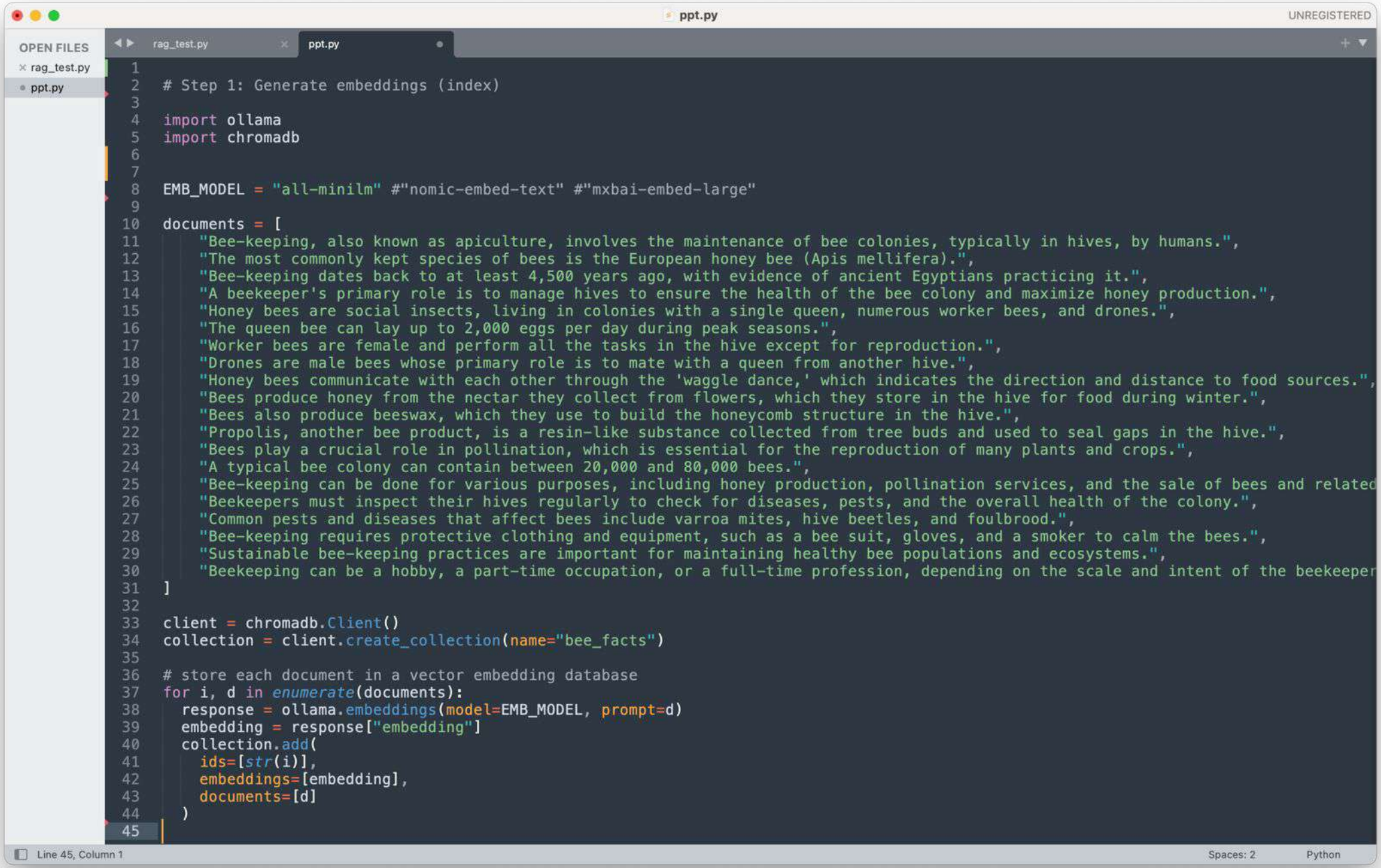
# With RAG

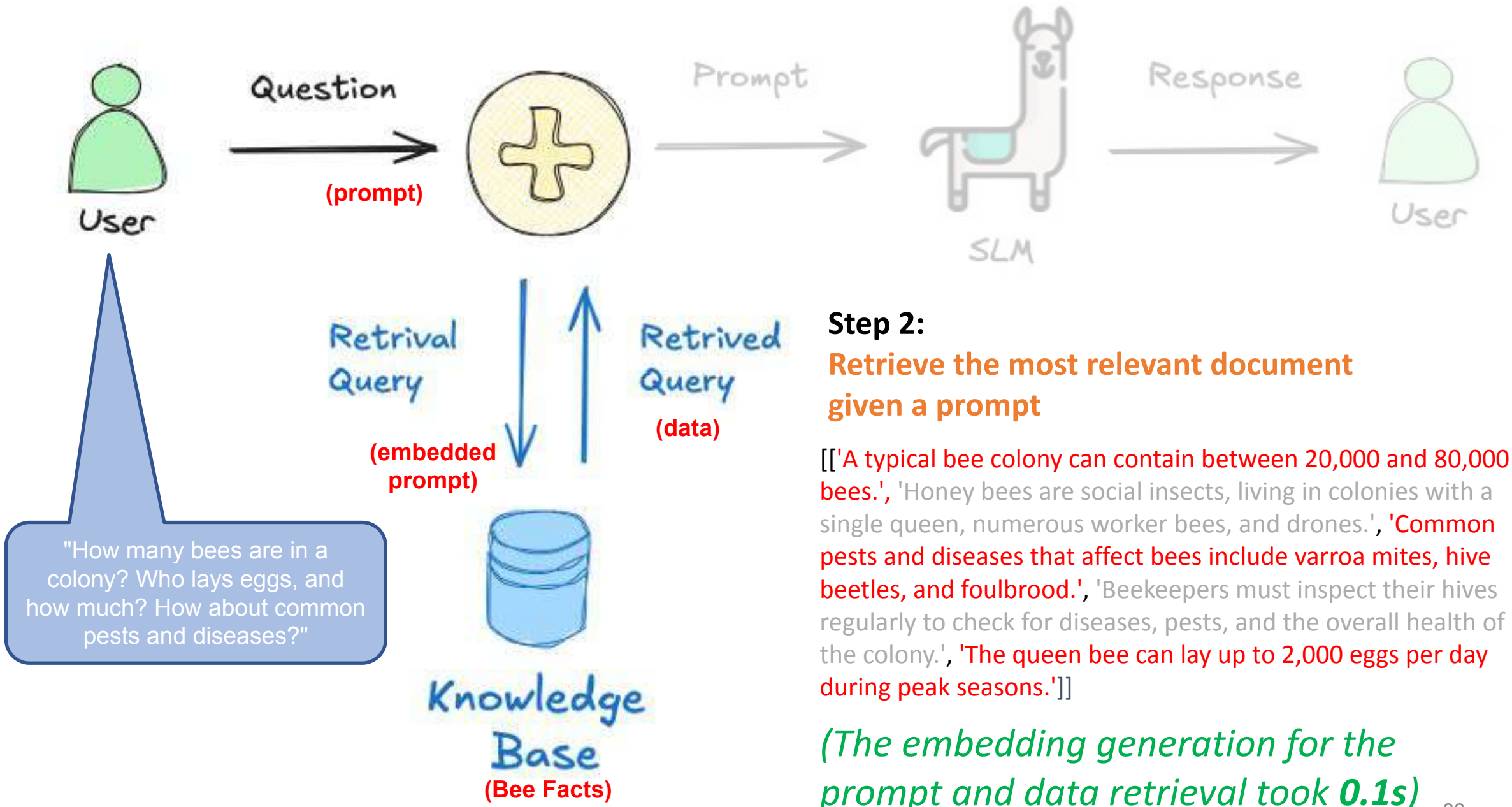



**Step 1:**

**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

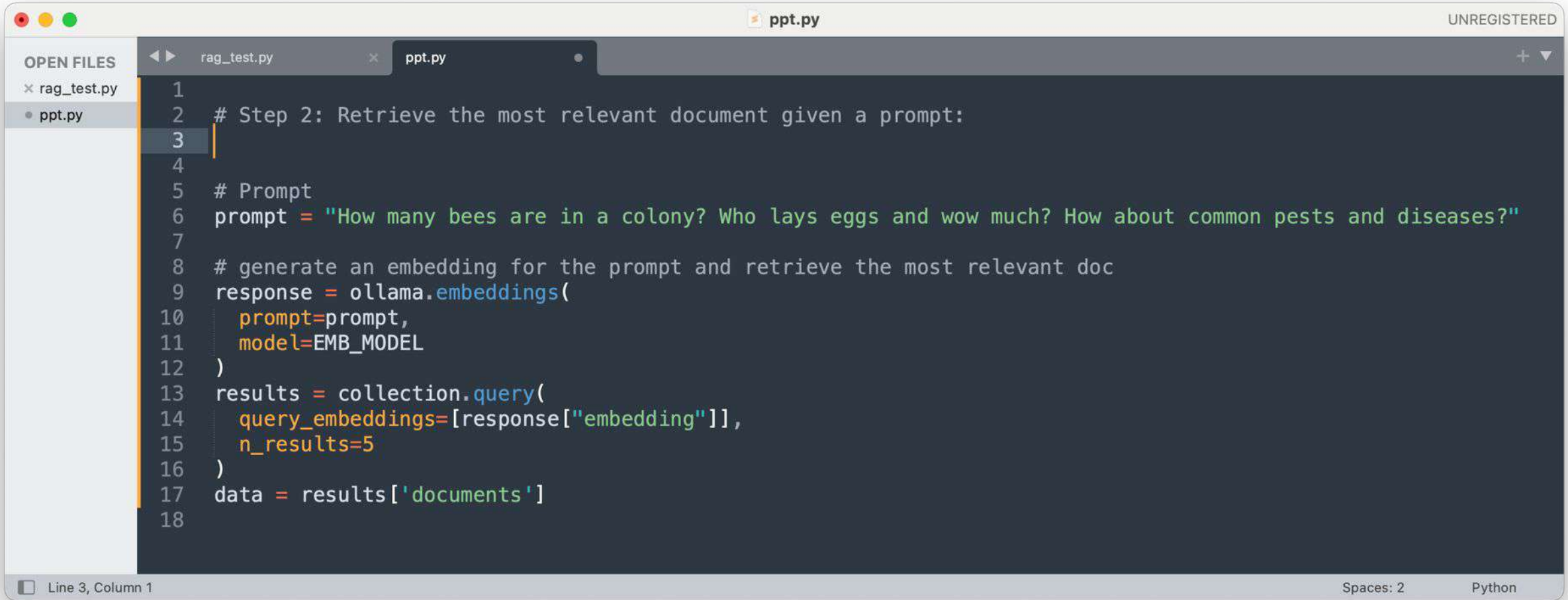


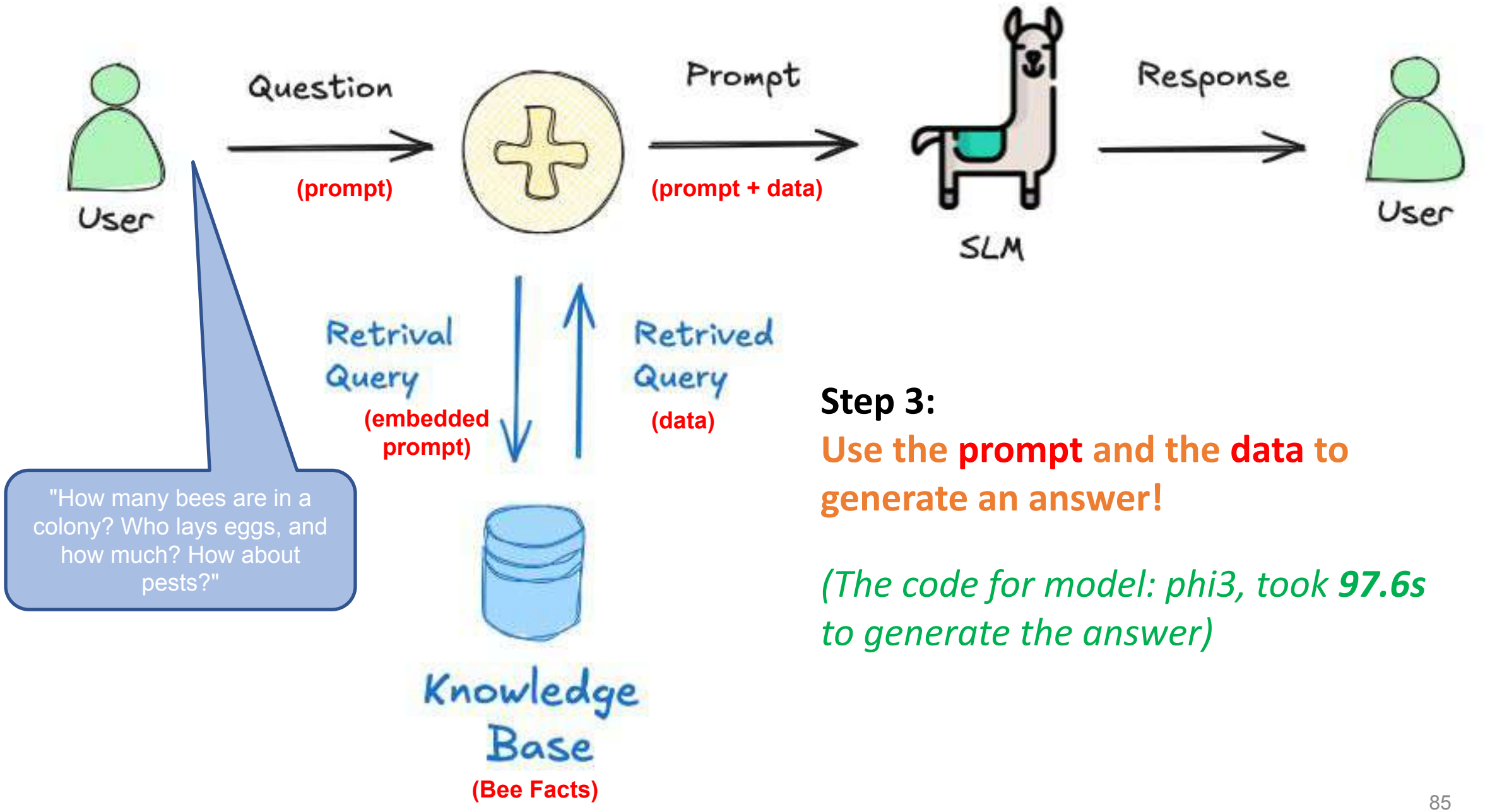

**Step 2:**

**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**)*

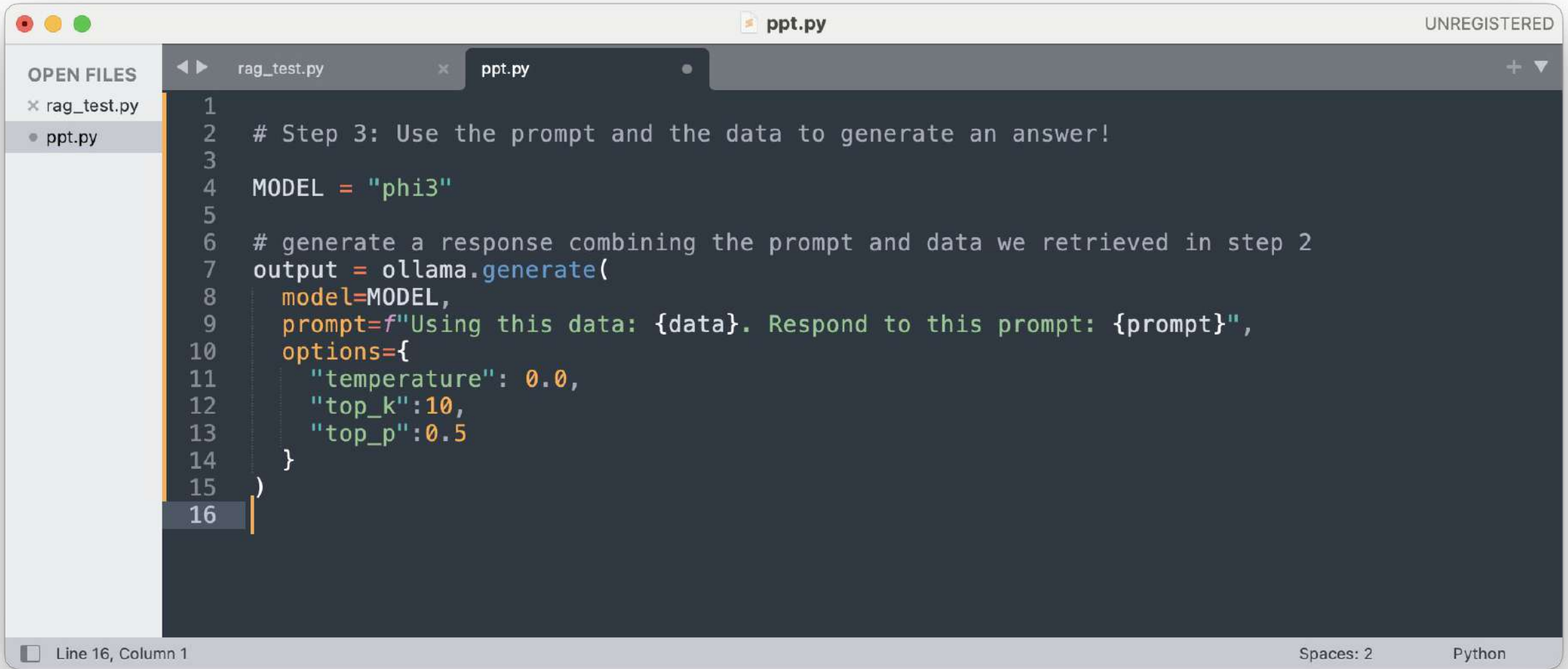

ppt.py
UNREGISTERED
OPEN FILES
rag_test.py
ppt.py
rag_test.py
ppt.py
1
2 # Step 2: Retrieve the most relevant document given a prompt:
3
4
5 # Prompt
6 prompt = "How many bees are in a colony? Who lays eggs and wow much? How about common pests and diseases?"
7
8 # generate an embedding for the prompt and retrieve the most relevant doc
9 response = ollama.embeddings(
10 prompt=prompt,
11 model=EMB_MODEL
12 )
13 results = collection.query(
14 query_embeddings=[response["embedding"]],
15 n_results=5
16 )
17 data = results['documents']
18
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)*

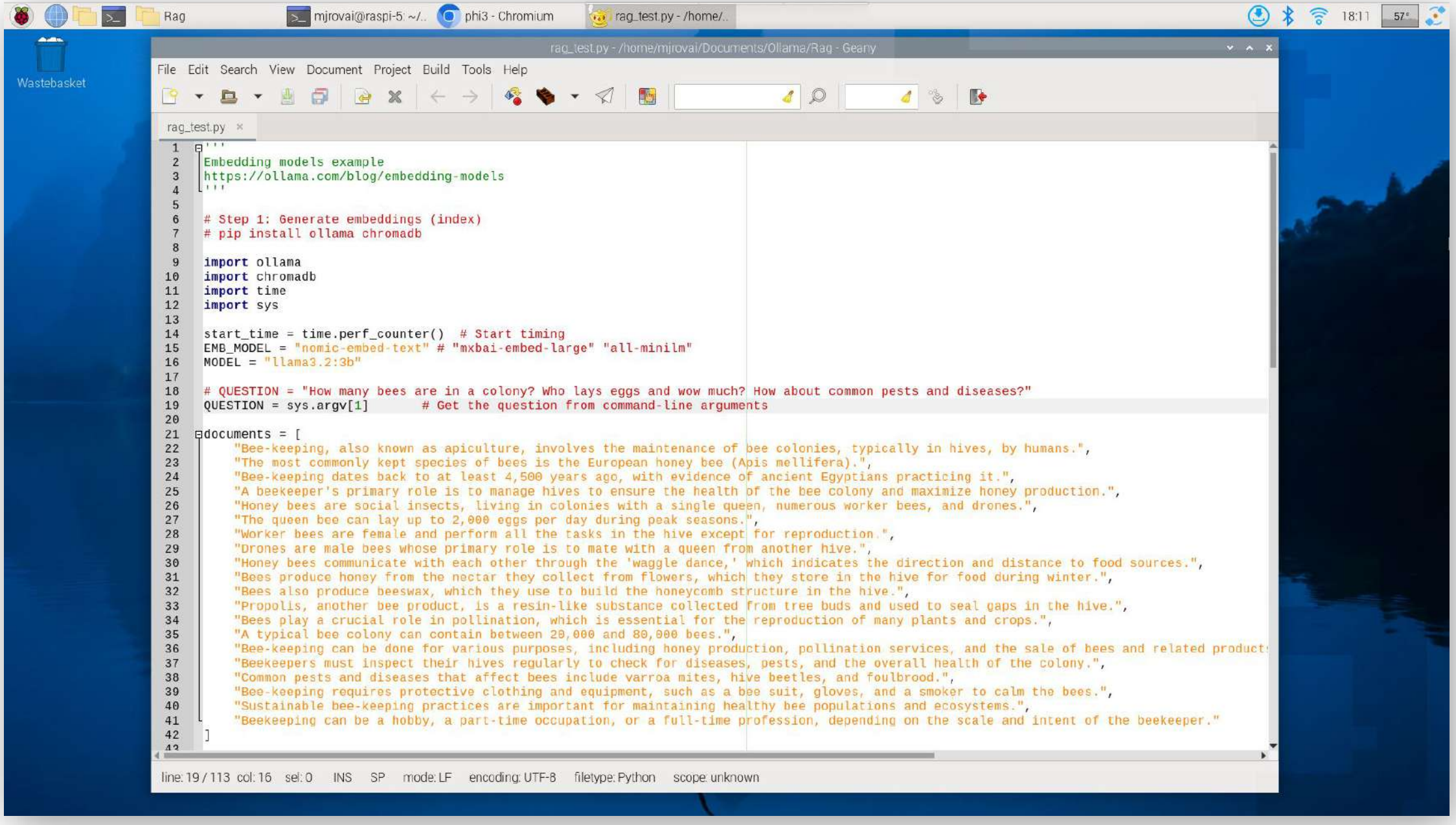

ppt.py
UNREGISTERED
OPEN FILES
rag_test.py
ppt.py
rag_test.py
ppt.py
# Step 3: Use the prompt and the data to generate an answer!
MODEL = "phi3"
# generate a response combining the prompt and data we retrieved in step 2
output = ollama.generate(
model=MODEL,
prompt=f"Using this data: {data}. Respond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

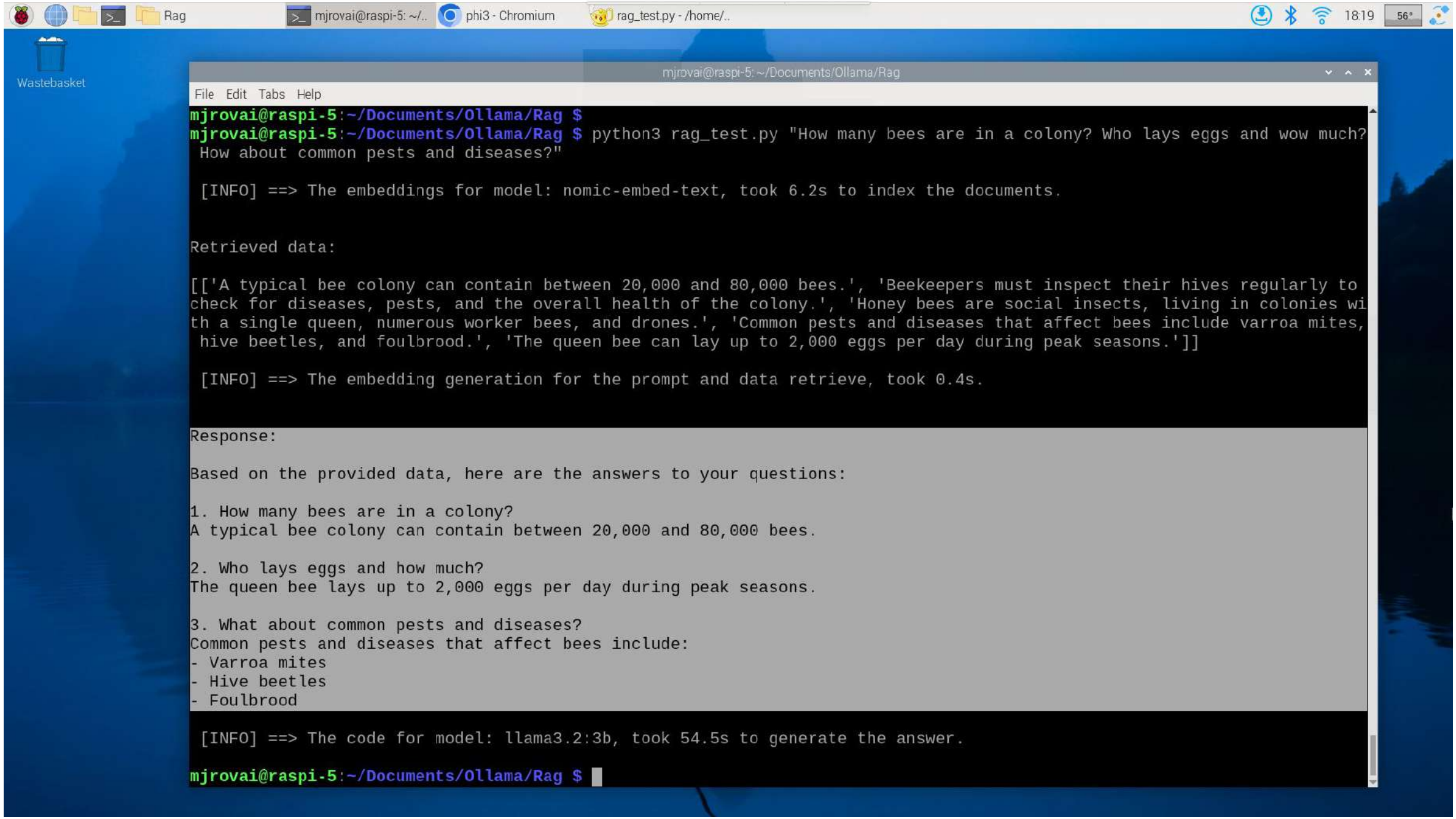
## Question:

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

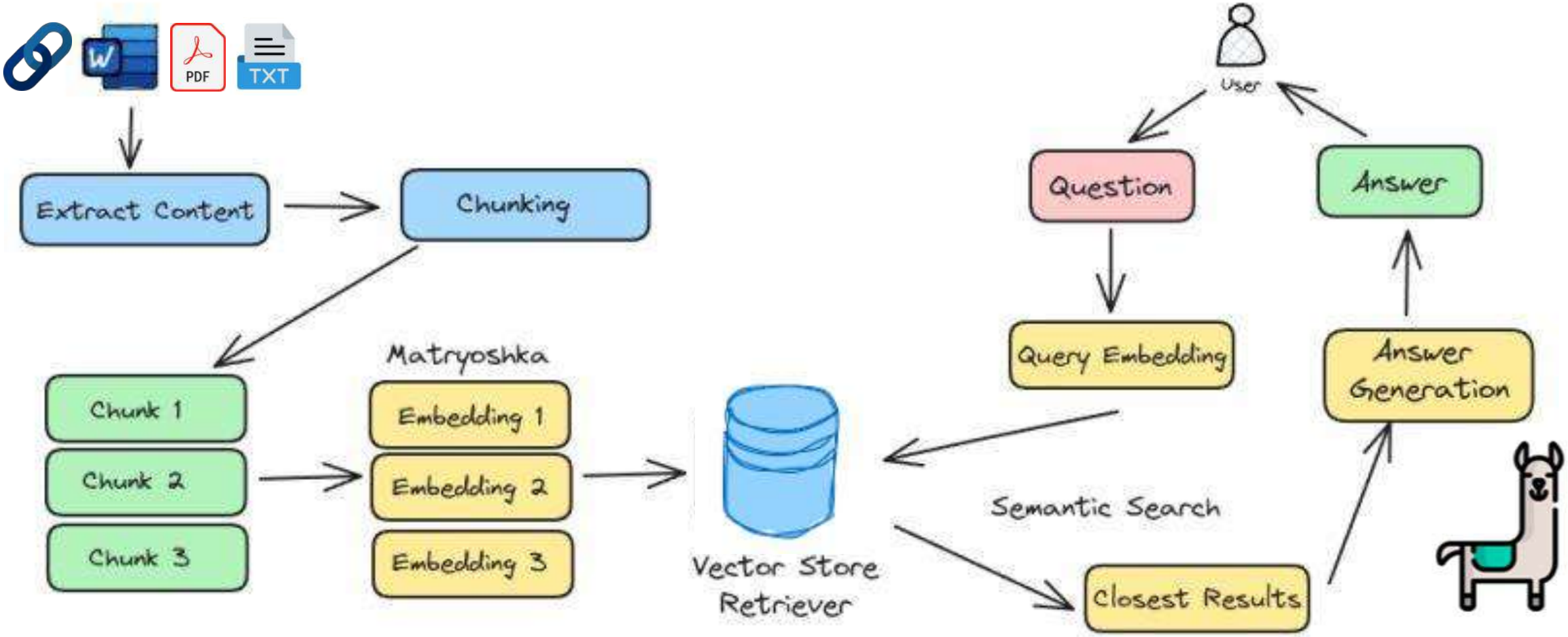
## Response

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.

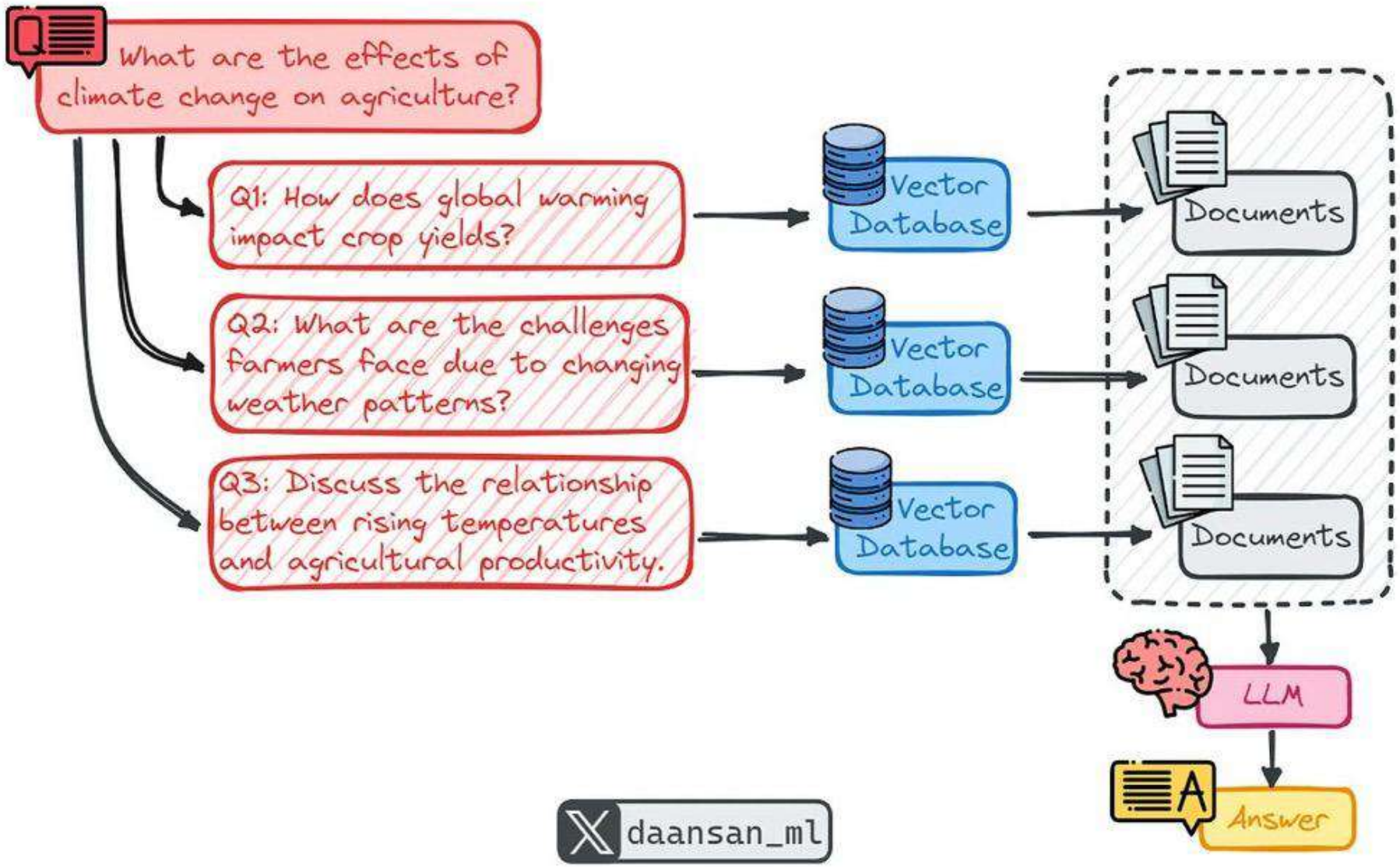

Rag
mjrovai@raspi-5: ~/..
phi3 - Chromium
rag_test.py - /home/..
18:11
57°
Wastebasket
rag_test.py - /home/mjrovai/Documents/Ollama/Rag - Geany
File Edit Search View Document Project Build Tools Help
rag_test.py
'''
Embedding models example
https://ollama.com/blog/embedding-models
'''
# Step 1: Generate embeddings (index)
# pip install ollama chromadb
import ollama
import chromadb
import time
import sys
start_time = time.perf_counter()  # Start timing
EMB_MODEL = "nomic-embed-text" # "mxbai-embed-large" "all-minilm"
MODEL = "llama3.2:3b"
# QUESTION = "How many bees are in a colony? Who lays eggs and wow much? How about common pests and diseases?"
QUESTION = sys.argv[1]       # Get the question from command-line arguments
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 product
"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."
]
line: 19 / 113 col: 16 sel: 0 INS SP mode: LF encoding: UTF-8 filetype: Python scope: unknown


Rag
mjrovai@raspi-5: ~/..
phi3 - Chromium
rag_test.py - /home/..
18:19
56°
Wastebasket
mjrovai@raspi-5: ~/Documents/Ollama/Rag
File Edit Tabs Help
mjrovai@raspi-5:~/Documents/Ollama/Rag $
mjrovai@raspi-5:~/Documents/Ollama/Rag $ python3 rag_test.py "How many bees are in a colony? Who lays eggs and wow much? How about common pests and diseases?"
[INFO] ==> The embeddings for model: nomic-embed-text, took 6.2s to index the documents.
Retrieved data:
[['A typical bee colony can contain between 20,000 and 80,000 bees.', 'Beekeepers must inspect their hives regularly to check for diseases, pests, and the overall health of the colony.', '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.', 'The queen bee can lay up to 2,000 eggs per day during peak seasons.']]
[INFO] ==> The embedding generation for the prompt and data retrieve, took 0.4s.
Response:
Based on the provided data, here are the answers to your questions:
1. How many bees are in a colony?
A typical bee colony can contain between 20,000 and 80,000 bees.
2. Who lays eggs and how much?
The queen bee lays up to 2,000 eggs per day during peak seasons.
3. What about common pests and diseases?
Common pests and diseases that affect bees include:
- Varroa mites
- Hive beetles
- Foulbrood
[INFO] ==> The code for model: llama3.2:3b, took 54.5s to generate the answer.
mjrovai@raspi-5:~/Documents/Ollama/Rag $

# RAG: Simple Query



# Advanced RAG: Multi Query

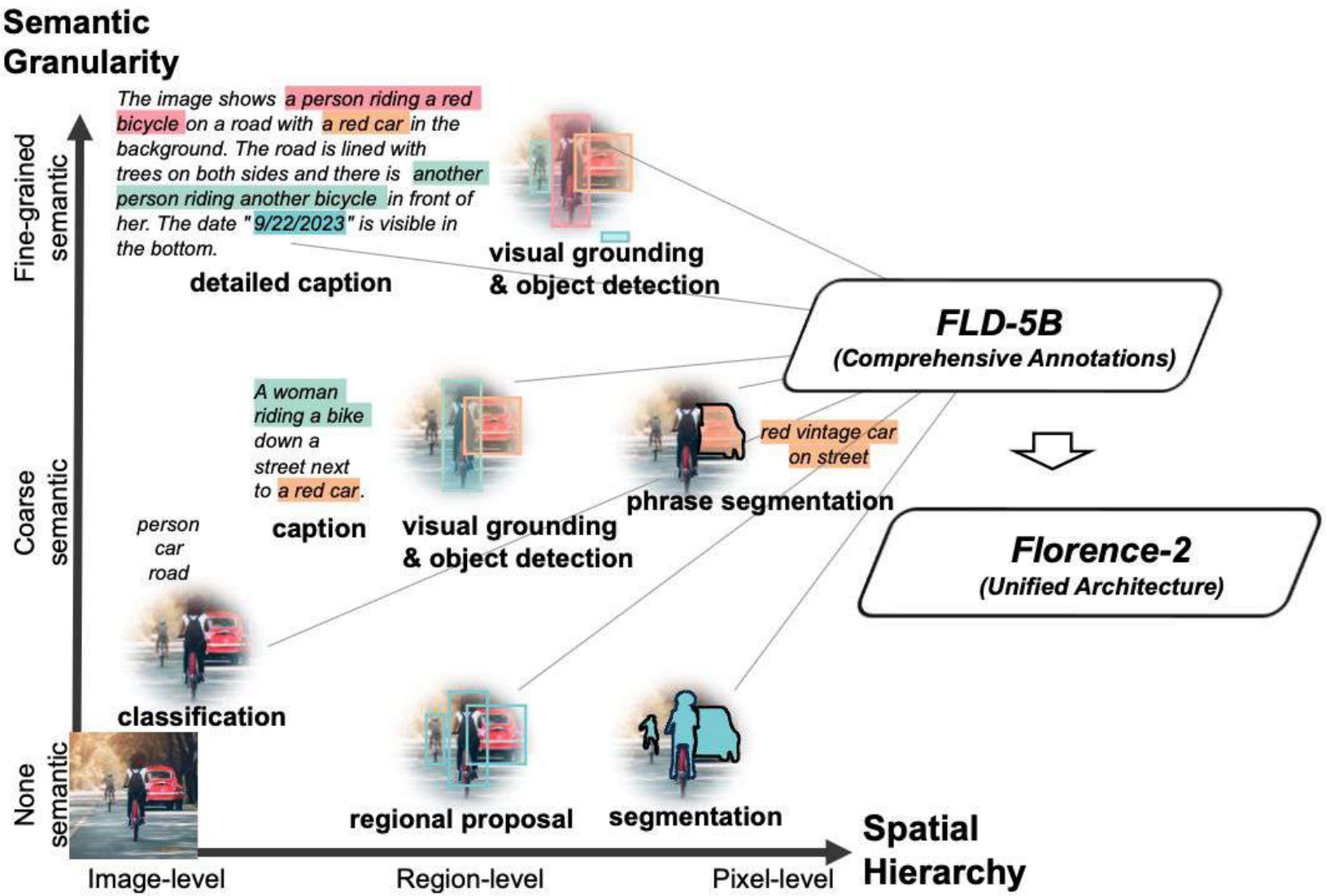


# VLM

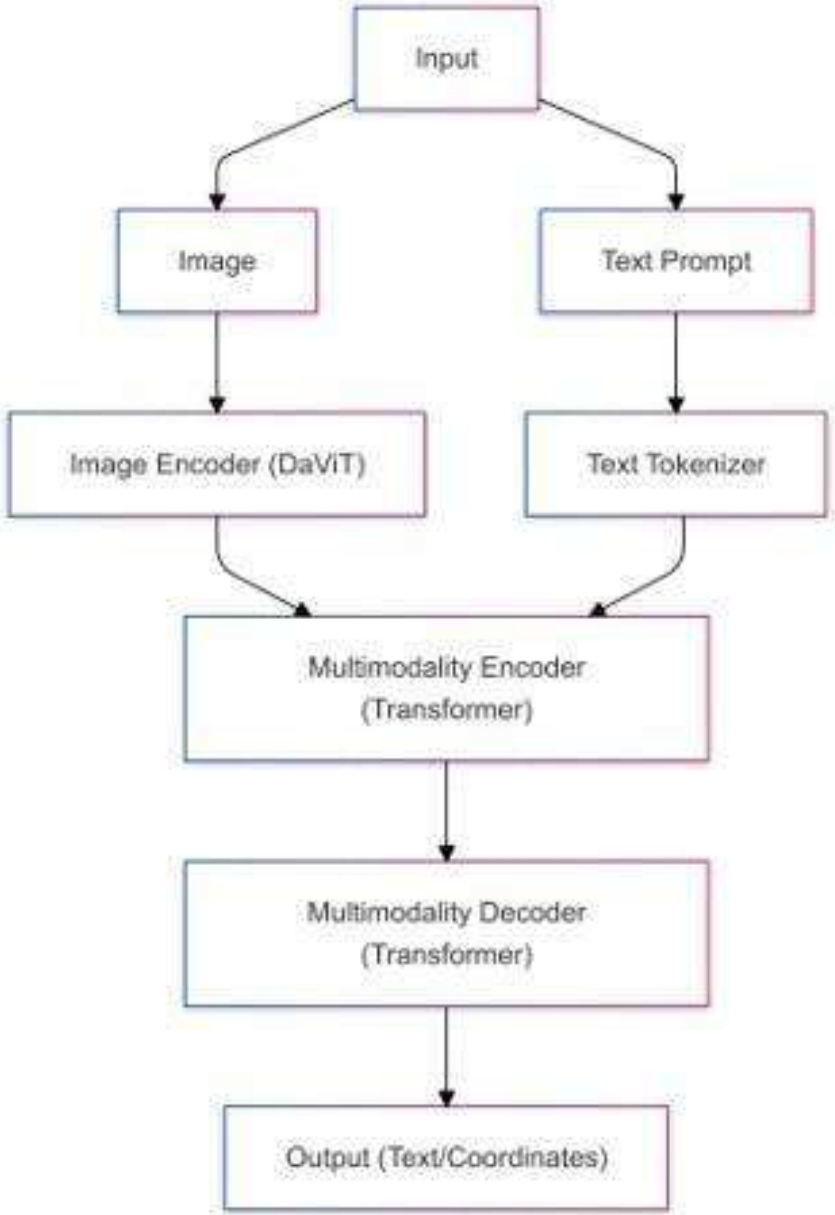
## Visual-Language Models

# Florence-2

Advancing a Unified Representation for a Variety of Vision Tasks




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


Input
Image
Text Prompt
Image Encoder (DaViT)
Text Tokenizer
Multimodality Encoder
(Transformer)
Multimodality Decoder
(Transformer)
Output (Text/Coordinates)

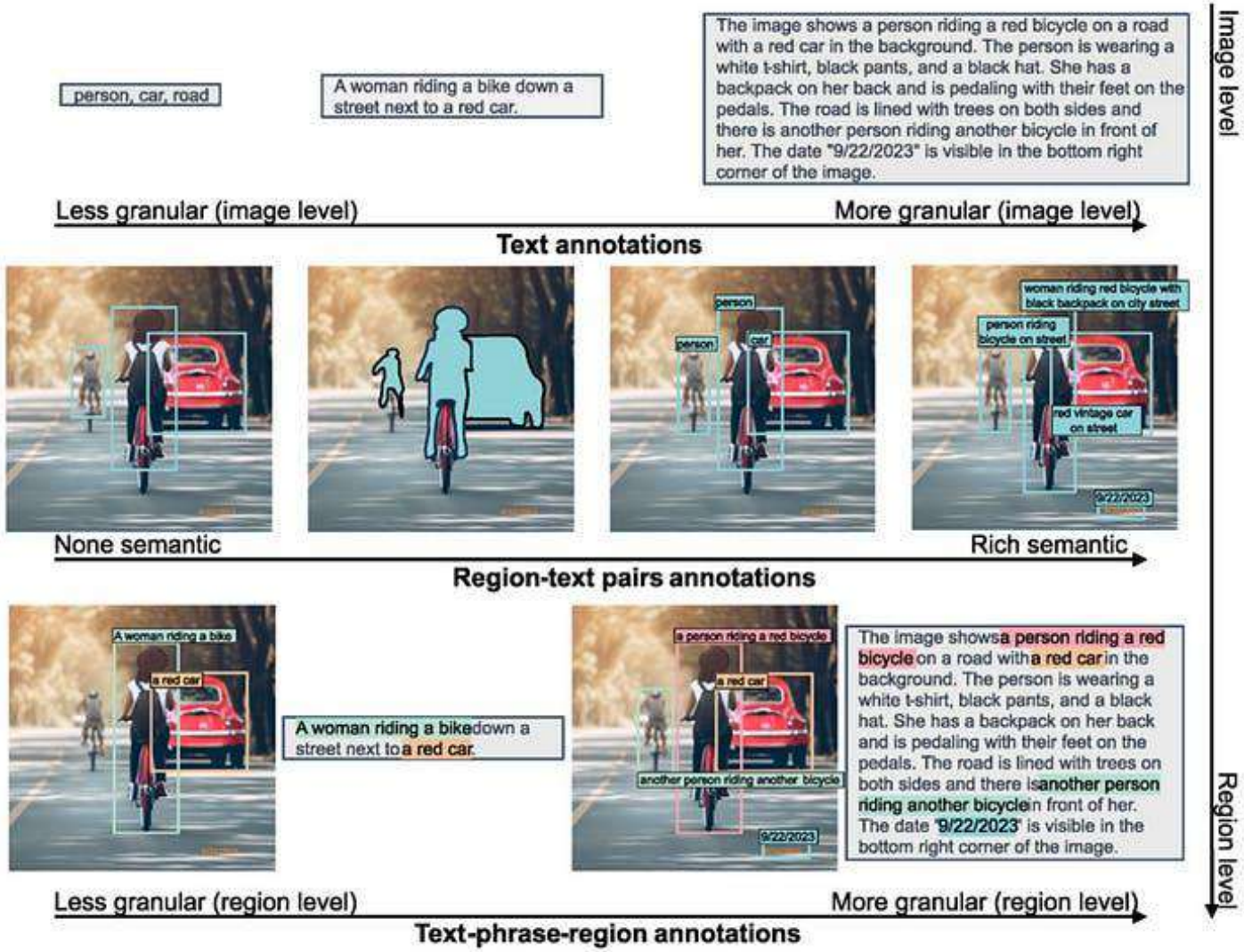

person, car, road
A woman riding a bike down a street next to a red car.
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the 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.
Image level
Less granular (image level)
More granular (image level)
Text annotations
person
car
woman riding red bicycle with black backpack on city street
person riding bicycle on street
red vintage car on street
9/22/2023
None semantic
Rich semantic
Region-text pairs annotations
A woman riding a bike
a red car
A woman riding a bike down a street next to a red car.
a person riding a red bicycle
another person riding another bicycle
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.
Region level
Less granular (region level)
More granular (region level)
Text-phrase-region annotations

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

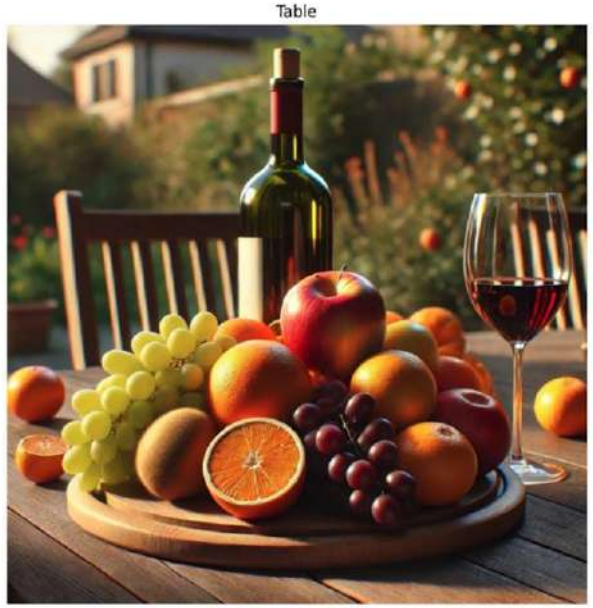


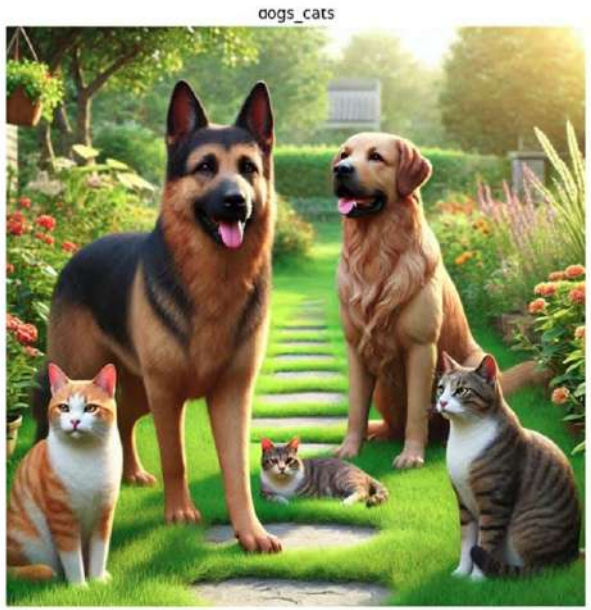

person riding red bicycle on road

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

# Caption

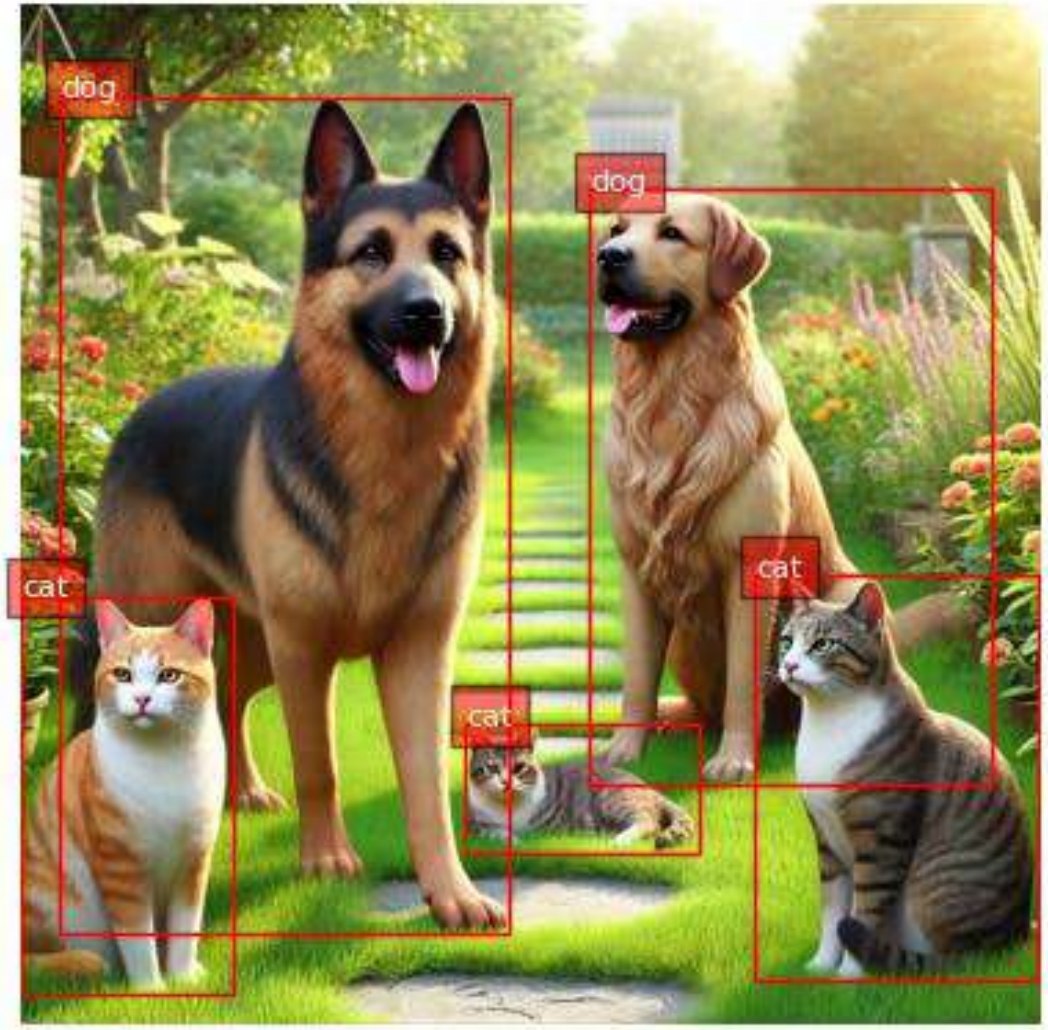
'The image shows a wooden table with a wooden tray on it. On the tray, there are various fruits such as grapes, oranges, apples, and grapes. There is also a bottle of red wine on the table. The background shows a garden with trees and a house. The overall mood of the image is peaceful and serene.'



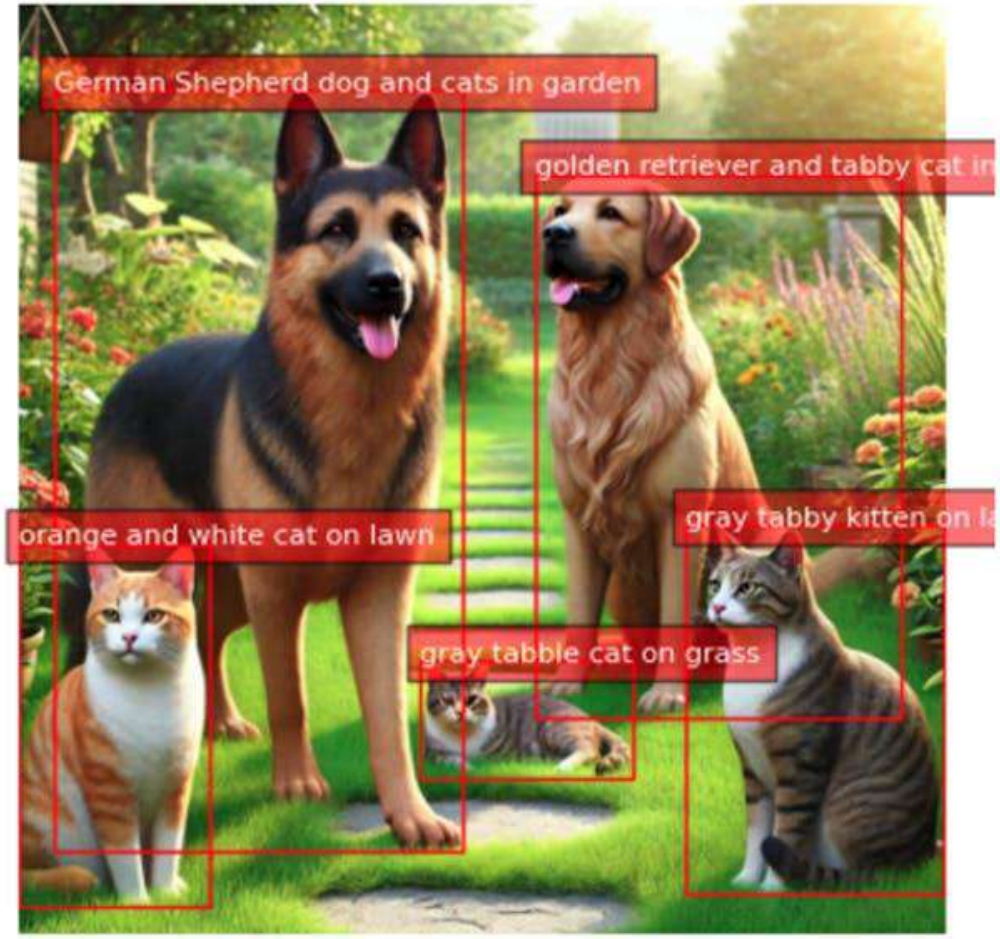



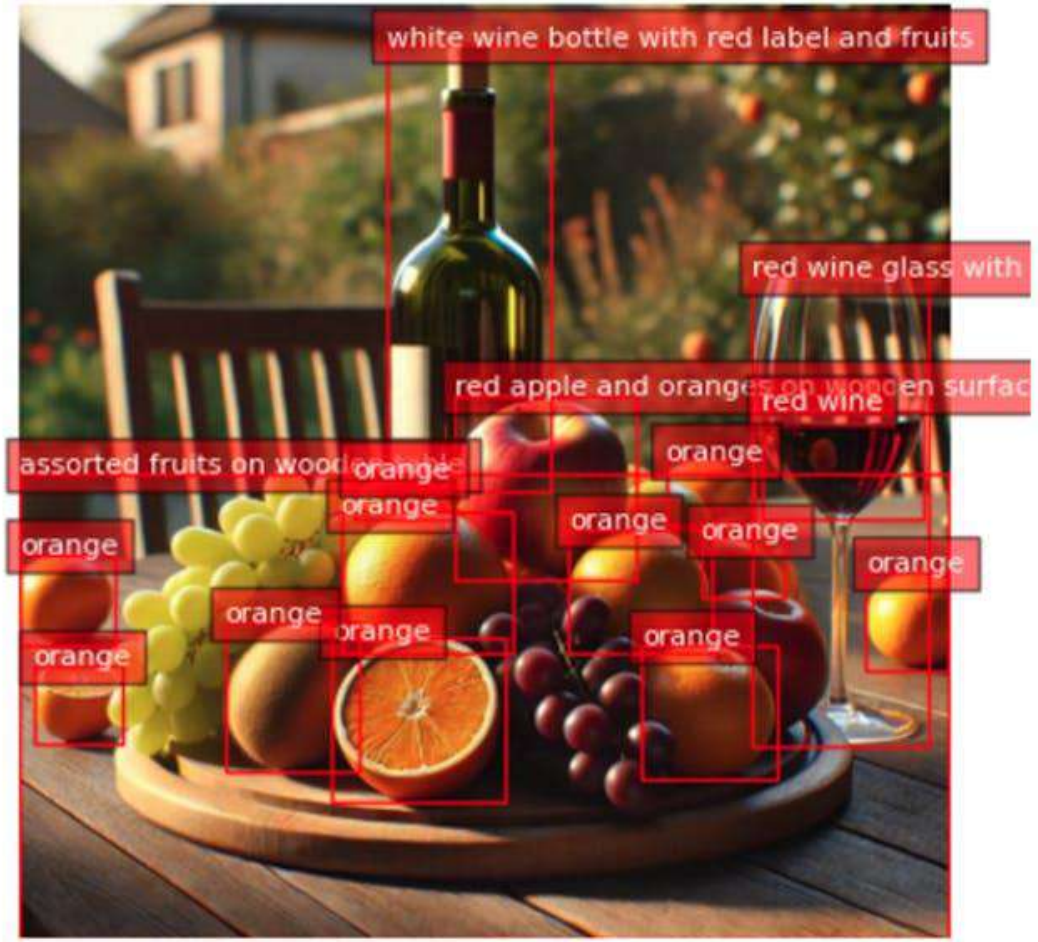

'The image shows a group of four cats and a dog in a garden. The garden is filled with colorful flowers and plants, and there is a pathway leading up to a house in the background. The main focus of the image is a large German Shepherd dog standing on the left side of the garden, with its tongue hanging out and its mouth open, as if it is panting or panting. On the right side, there are two smaller cats, one orange and one gray, sitting on the grass. In the background, there is another golden retriever dog sitting and looking at the camera. The sky is blue and the sun is shining, creating a warm and inviting atmosphere.'}

# Object Detection

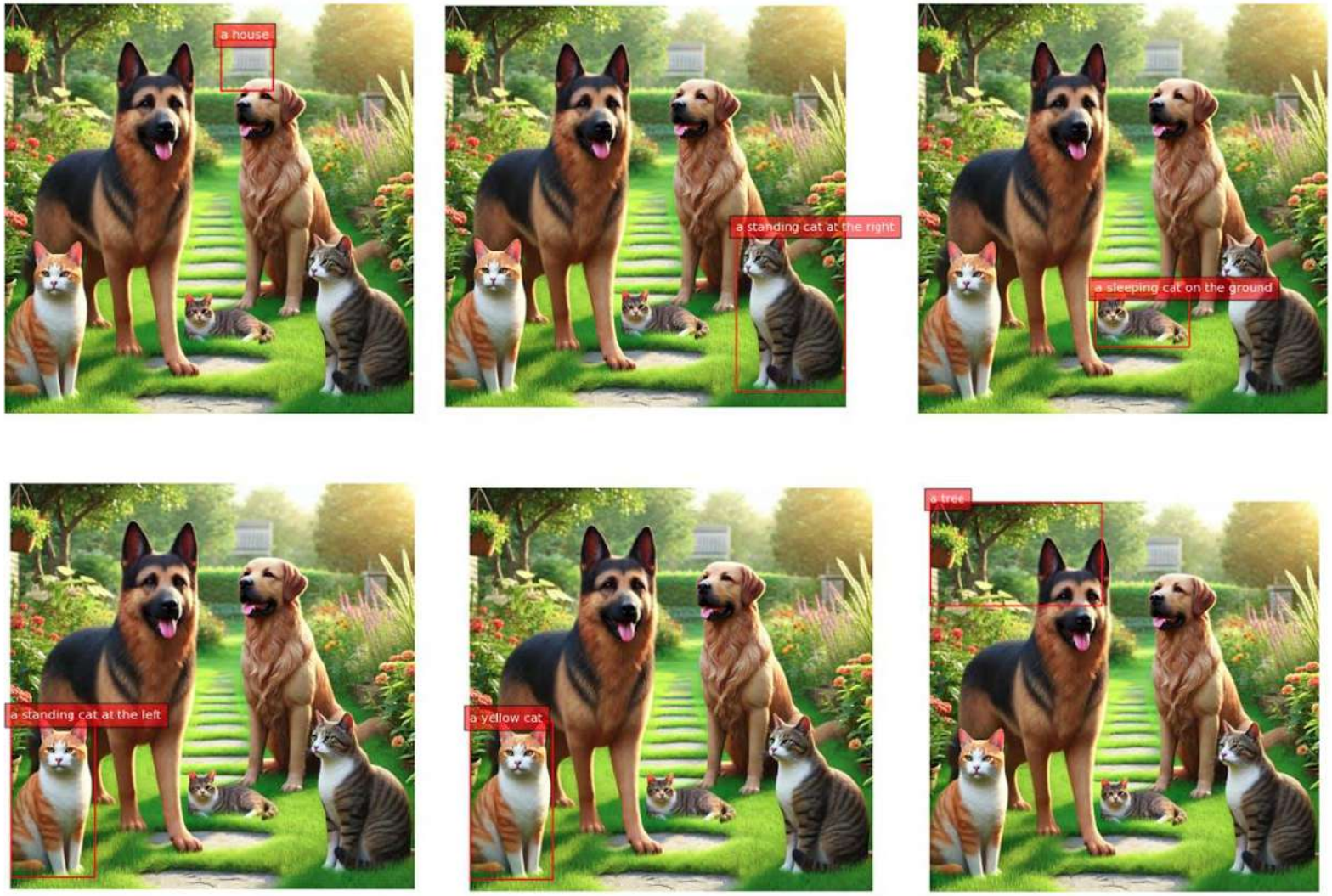



# Dense Region Caption



# Open Vocabulary Detection



# Segmentation

# OCR



```
results['<OCR_WITH_REGION>']['labels']
```

```
['</s>Machine Learning',
 'Café',
 'com',
 'Embarcado',
 'Embarcados',
 'Democratizando a Inteligência',
 'Artificial para Paises em',
 '25 de Setembro ás 17h',
 'Desenvolvimento',
 'Toda quarta-feira',
 'Marcelo Roval',
 'Professor na UNIFIEI e',
 'Transmissão via',
 'in',
 'Co-Director do TinyML4D']
```

# Fine-Tunning


{"<OD>": {"bboxes": [[0.1599999964237213, 133.59999084472656, 78.23999786376953, 232.1599884033203], [117.27999877929688, 139.0399932861328, 196.63999938964844, 243.67999267578125], [190.239990234375, 193.1199951171875, 270.239990234375, 319.5199890136719], [248.1599884033203, 91.04000091552734, 319.5199890136719, 189.27999877929688], [160.8000030517578, 27.680000305175781, 221.27999877929688, 118.23999786376953], [0.1599999964237213, 0.1599999964237213, 86.23999786376953, 57.119998931884766], [35.360000610351562, 36.31999969482422, 104.15999603271484, 112.15999603271484], [0.1599999964237213, 0.47999998927116394, 319.5199890136719, 319.5199890136719]], "labels": ["wheel", "wheel", "box", "box", "box", "box", "wheel", "box"]}}

# The Future...

# microsoft/BitNet

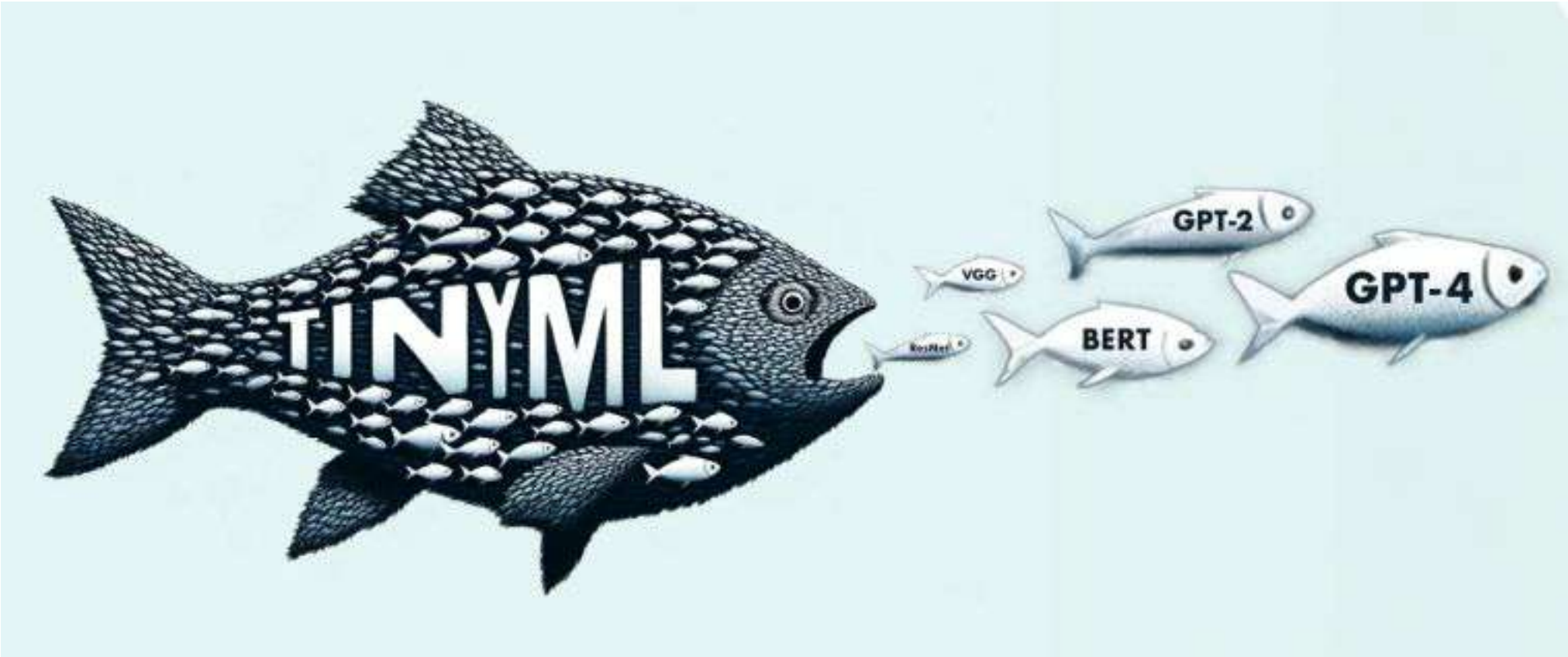
Official inference framework for 1-bit LLMs

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

# To learn more …

**Online Courses**

Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning
Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
Introduction to Embedded Machine Learning - Coursera/Edge Impulse
Computer Vision with Embedded Machine Learning - Coursera/Edge Impulse
UNIFEI-IESTI01 TinyML: “Machine Learning for Embedding Devices”

**Books**

“Python for Data Analysis” by Wes McKinney
“Deep Learning with Python” by François Chollet - GitHub Notebooks
“TinyML” by Pete Warden and Daniel Situnayake
“TinyML Cookbook 2nd Edition” by Gian Marco Iodice
“Technical Strategy for AI Engineers, In the Era of Deep Learning” by Andrew Ng
“AI at the Edge” book by Daniel Situnayake and Jenny Plunkett
“XIAO: Big Power, Small Board” by Lei Feng and Marcelo Rovai
“MACHINE LEARNING SYSTEMS” by a collaborative effort

**Projects Repository**

Edge Impulse Expert Network

**On the TinyML4D website, You can find lots of educational materials on TinyML. They are all free and open-source for educational uses – we ask that if you use the material, please cite them! TinyML4D is an initiative to make TinyML education available to everyone globally.**

# Questions?