IESTI05 – Edge AI

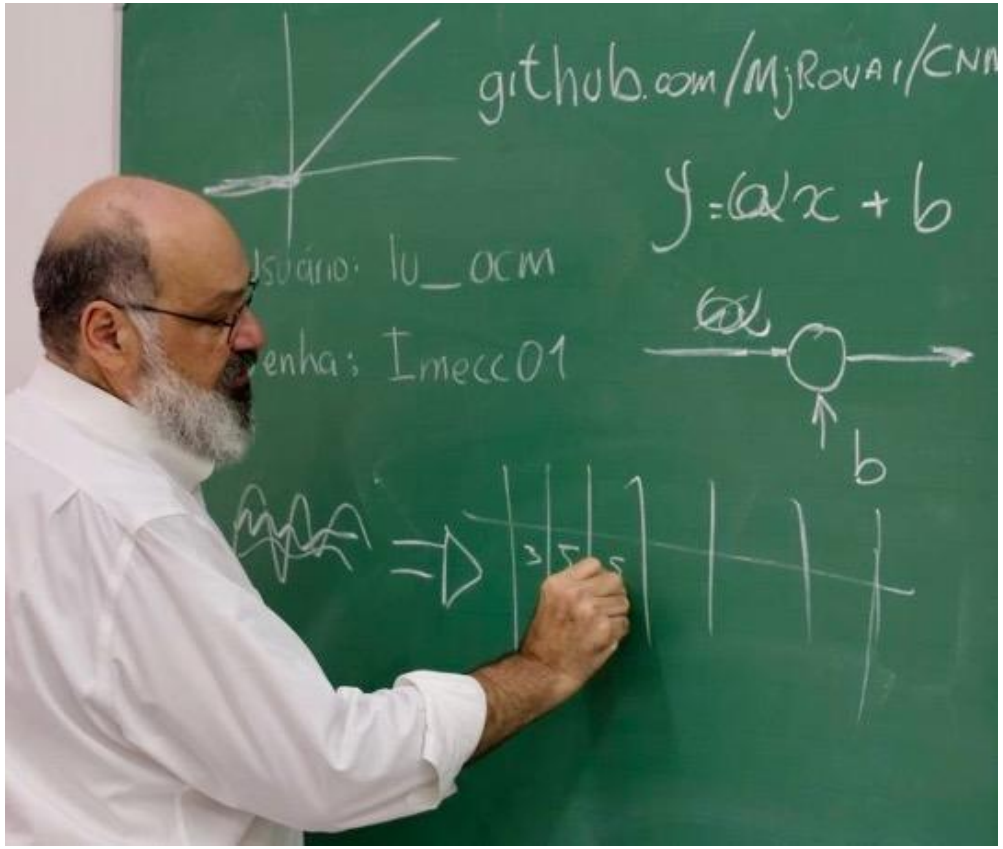
Machine Learning System Engineering

1. Introduction to Edge AI
   About the Course & Syllabus

UNIFEI

Marcelo Rovai
Professor Honoris Causa

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.

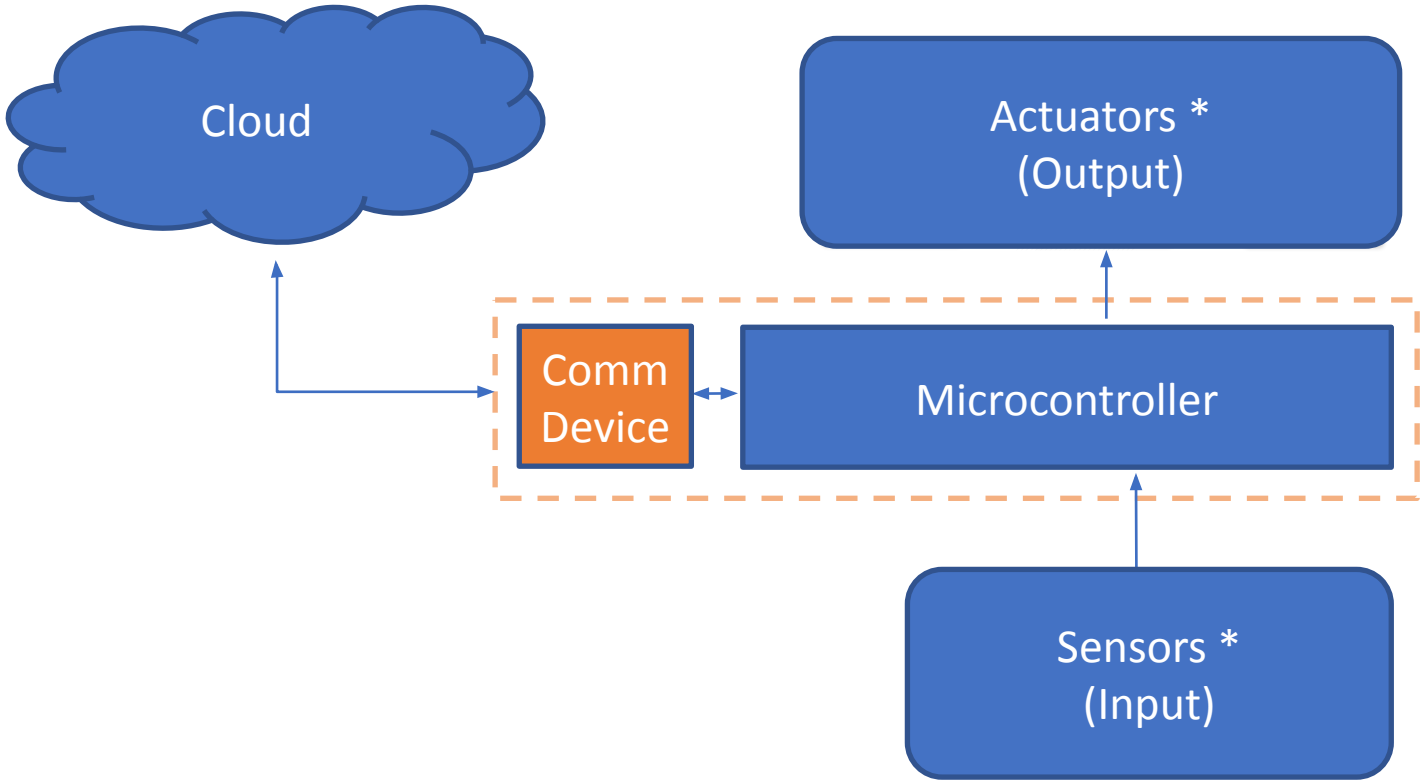


With a career spanning several high-profile technology companies such as AVIBRAS Airspace, AT&T, NCR, and IGT, where he served as Vice President for Latin America, he brings a wealth of industry experience to his academic endeavors. He is a prolific writer on electronics-related topics and shares his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

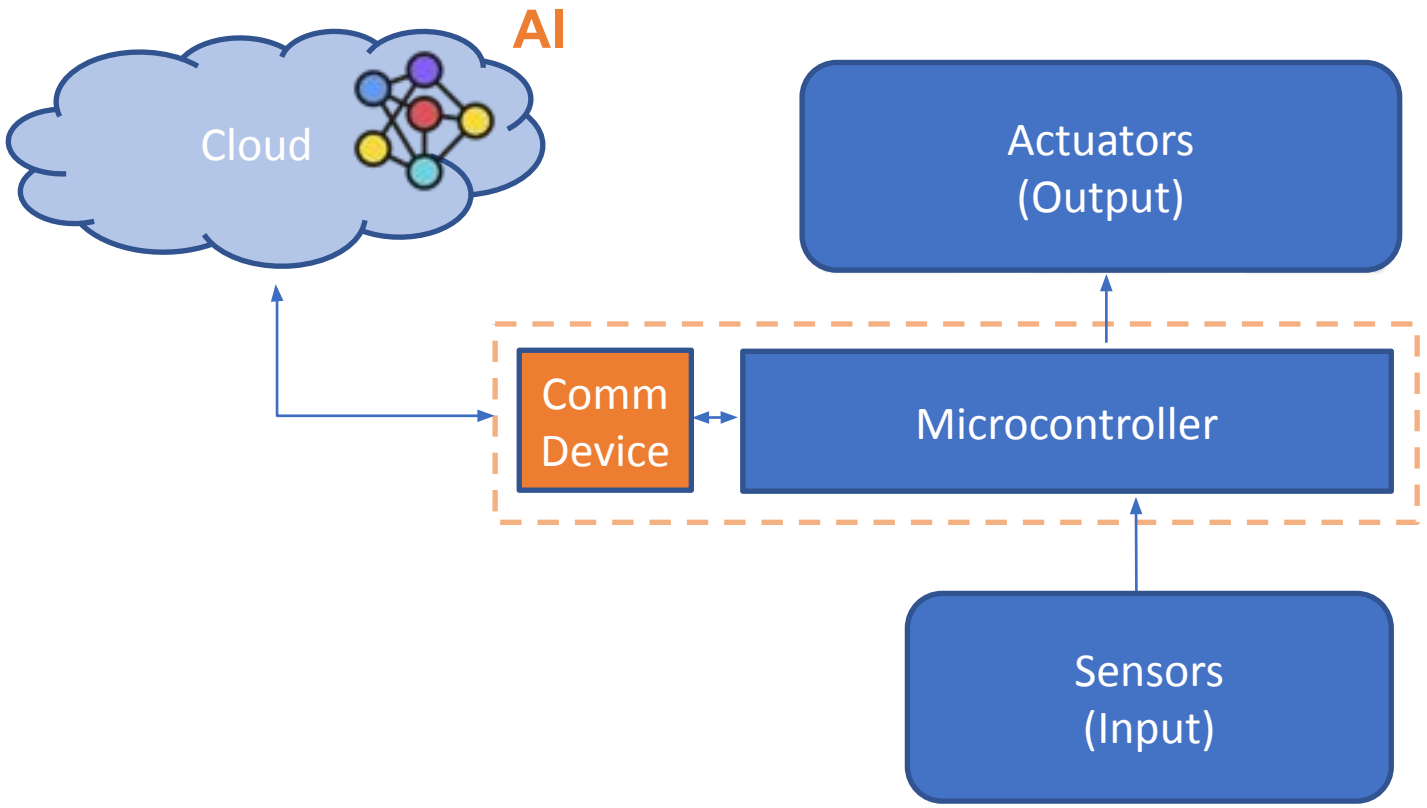
Internet of Things (IoT)

Typical IoT Project



* “Things”

Typical AIoT Project …



… Issues

Bandwidth

Latency

Energy

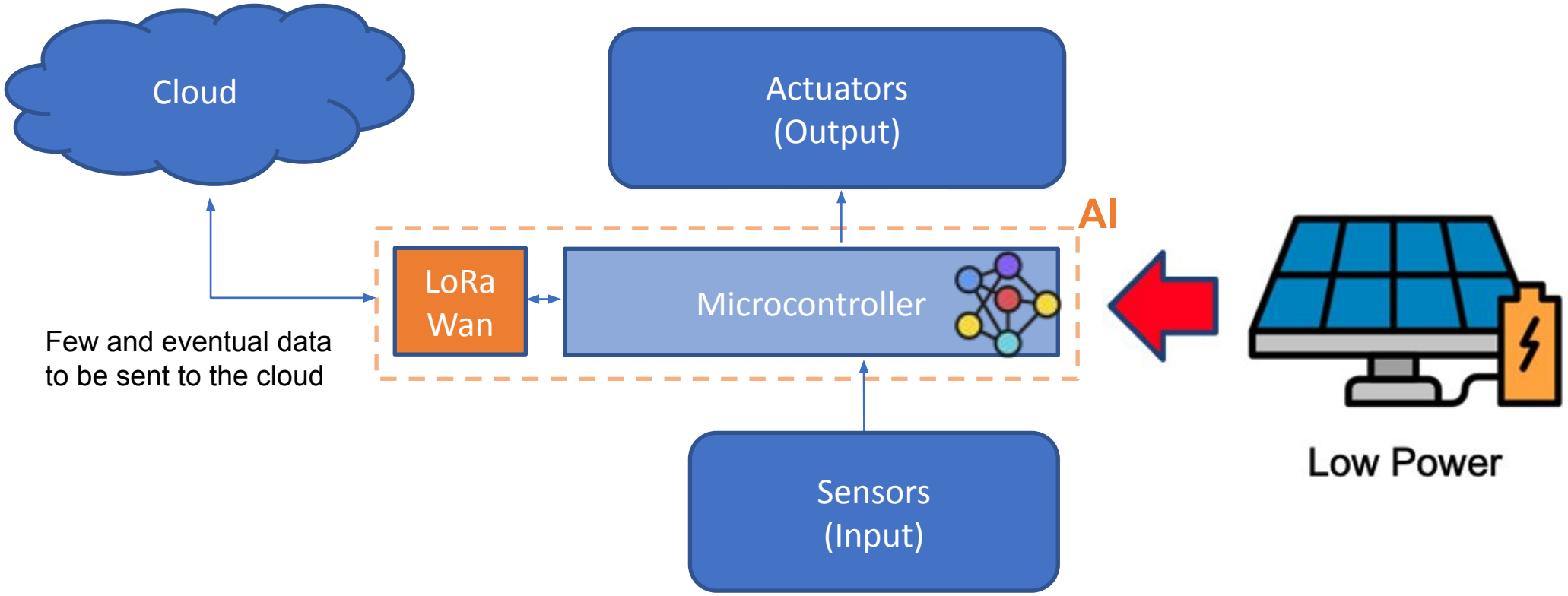
Reliability

Privacy

… Solution ?

IoT 2.0 * – Edge AI/ML

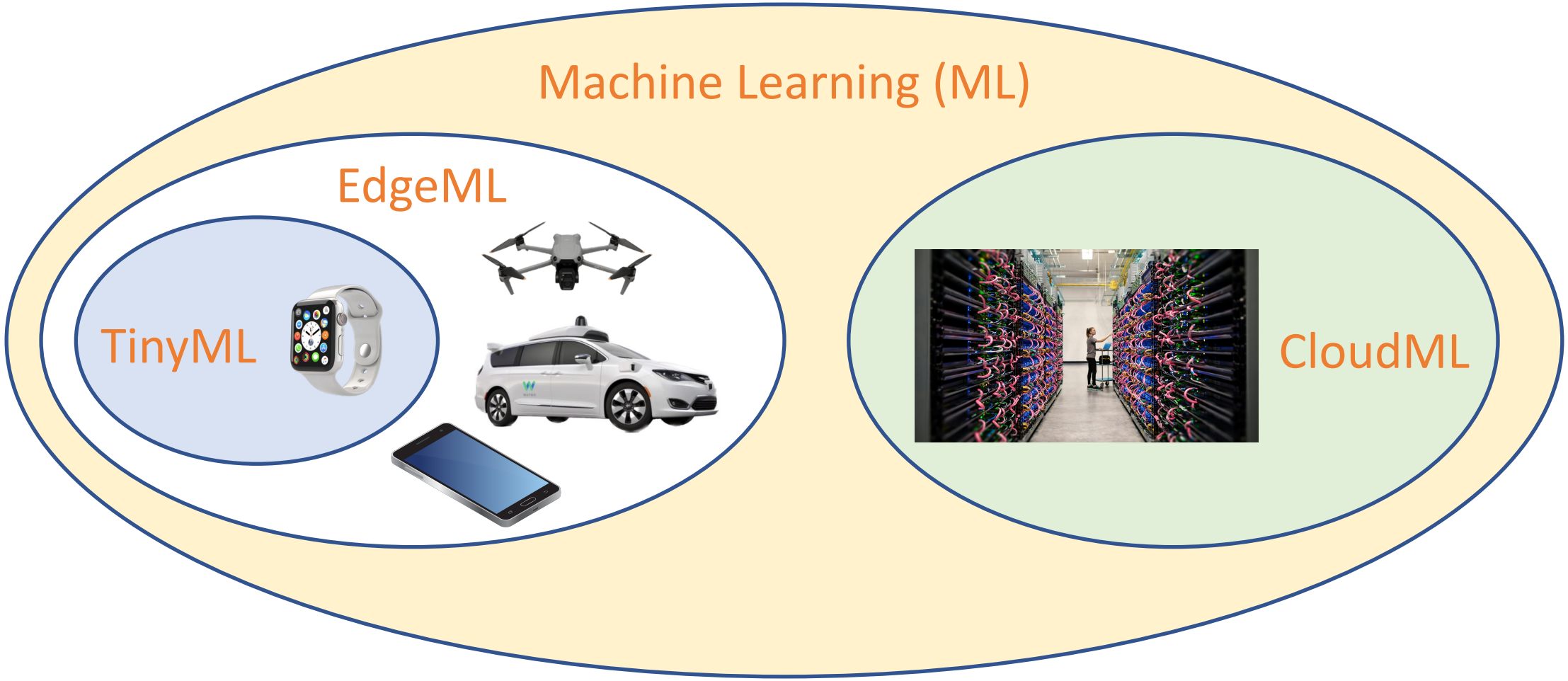
* Intelligence of Things



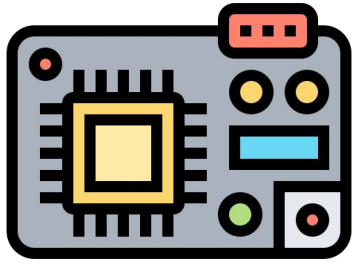
… **Solution ->** ML goes close to data

Embedded ML

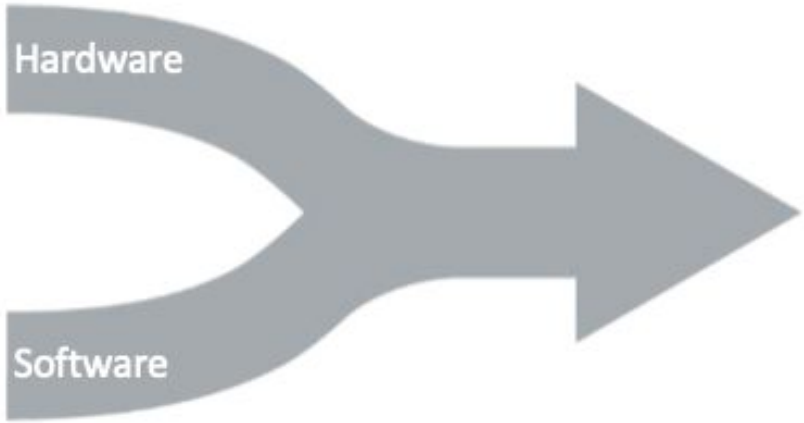
EdgeAI & TinyML


Machine Learning (ML)
EdgeML
TinyML
CloudML

What Makes Edge AI (ML)?

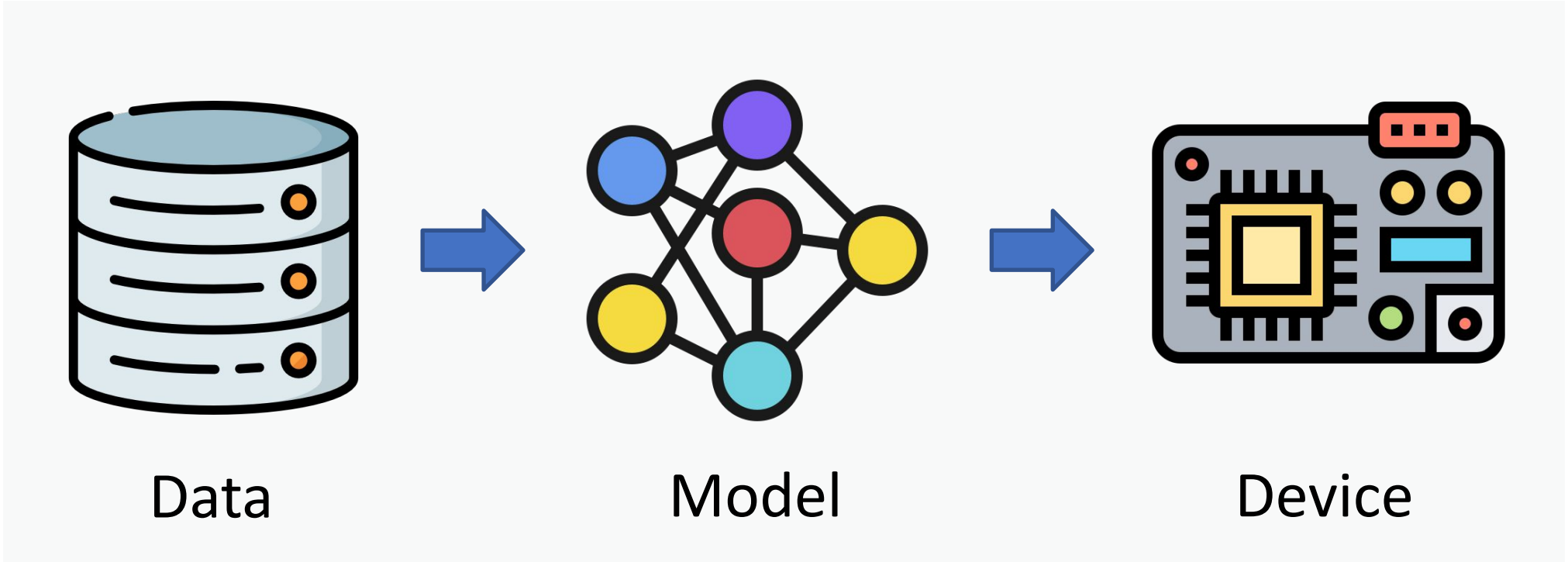


TensorFlow Lite

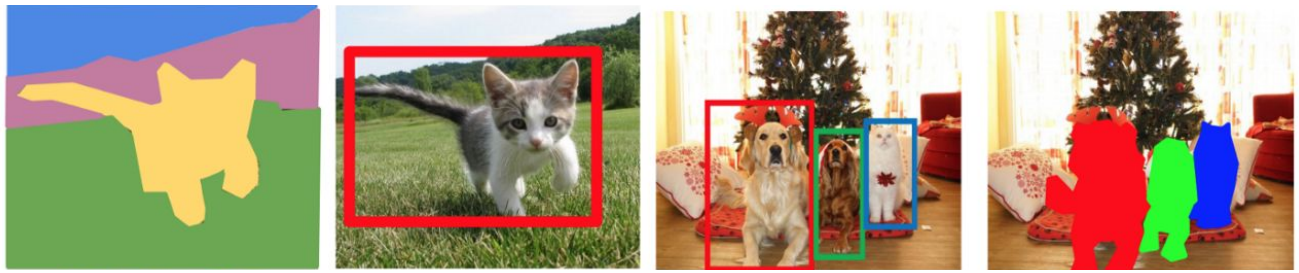


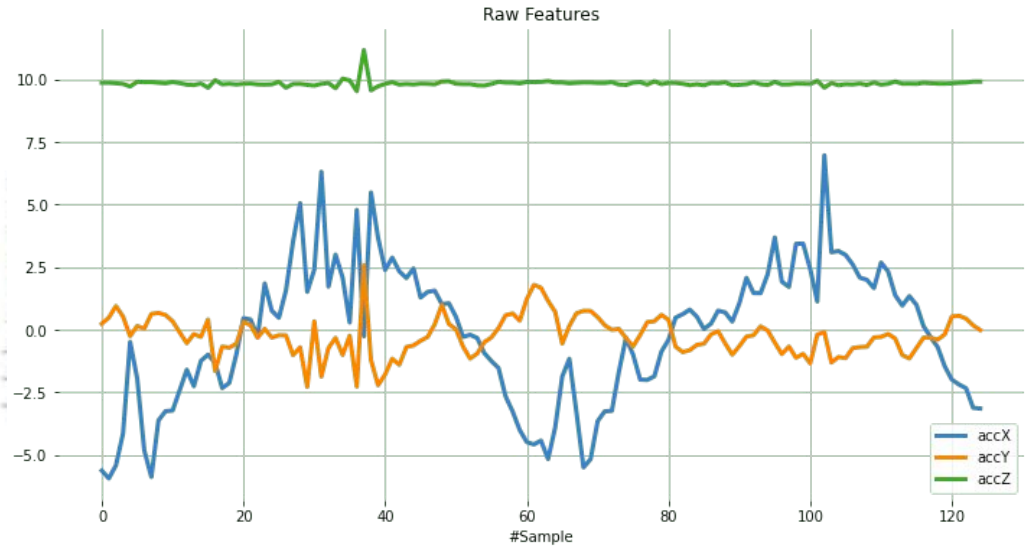
Edge AI (ML)

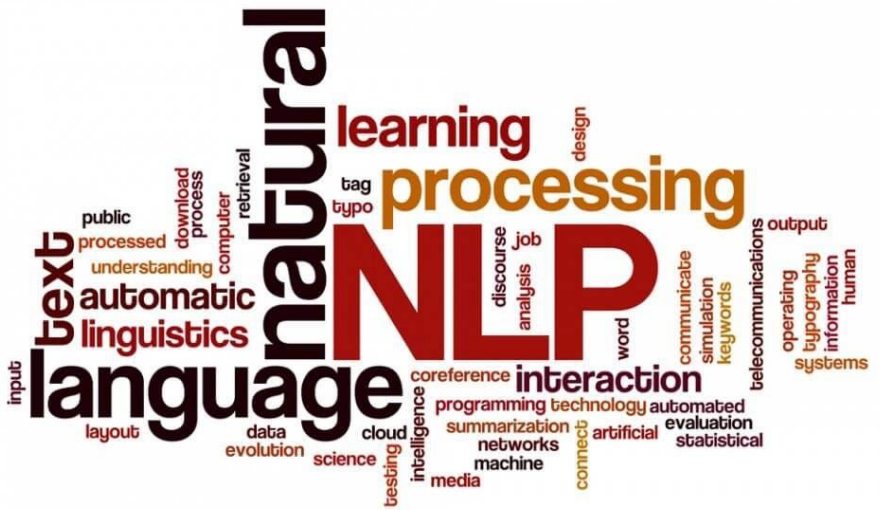
ML Deployment Pipeline

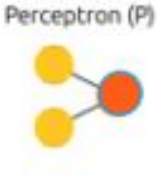


Unstructured Data









Neural Network Architectures

Vibration Analysis

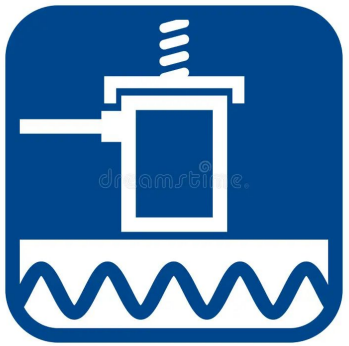
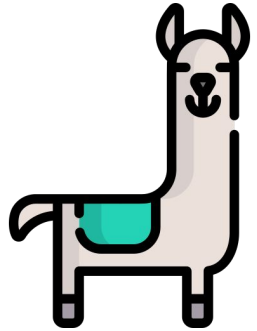


Image Classification

Text Generation

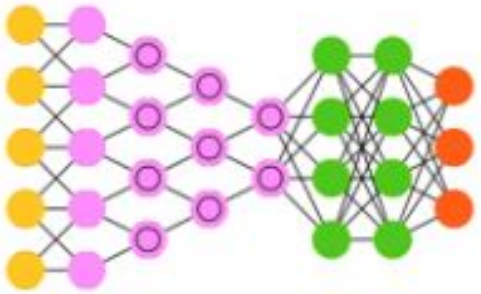
Image Generation

Large Language Models- LLMs



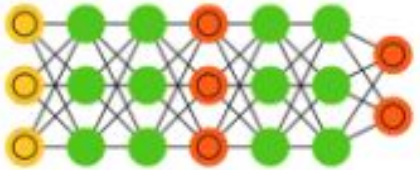
MLP - Deep Neural Network

CNN - Convolutional NN

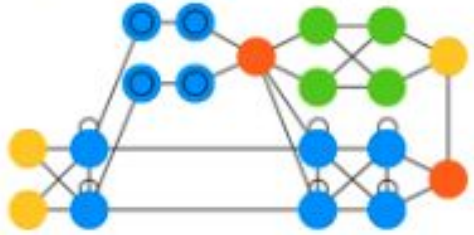


RNN - Recurrent NN (GRU/LSTM)

GAN - Generative Adversarial N.



AN - Attention (Transformers)



Datasets Preprocessing

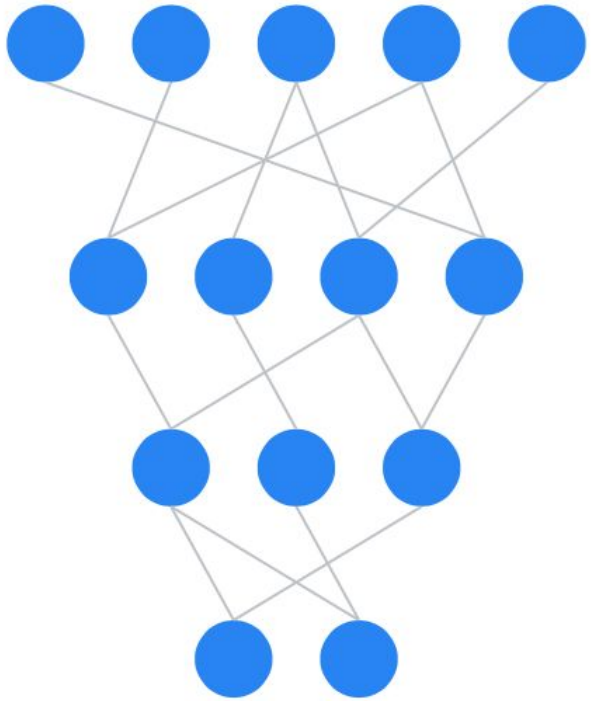
Sound

Vision

Vibration

Text

Quantization Pruning, Distillation



Resource constraints

End-to-end TinyML/EdgeAI Application

Application Complexity vs. HW

Power

EdgeML

IEST05

TinyML

LLMs/VLMs
1 GB+

IEST01

Object Detection
Complex Voice
Processing
1 MB+

Image
Classification
250 KB+

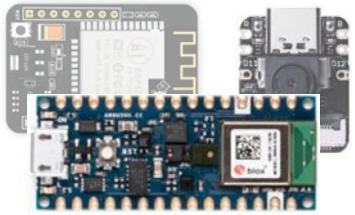
KeyWord Spotting
Audio Classification
50 KB

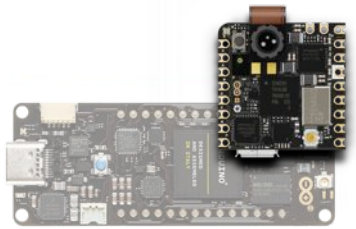
Anomaly Detection
Sensor Classification
20 KB

ESP32

XIAO

micro NPU





Rpi-Pico
(Cortex-M0+)

Arduino Nano
(Cortex-M4)

Arduino Pro
(Cortex-M7)

RaspberryPi SmartPhone
(Cortex-A)

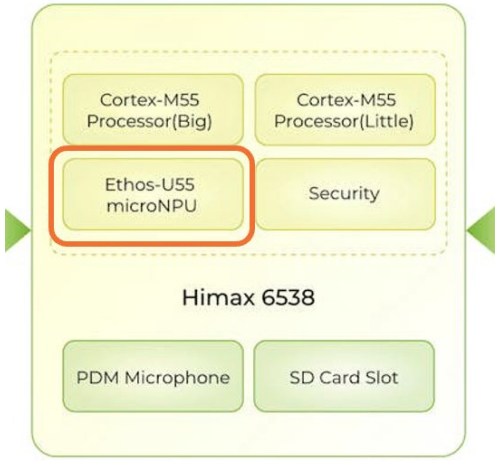
Jetson Orin
(Cortex-A + GPU)

Application Complexity

CPU Power / Memory

microNPU

Grove Vision AI v2

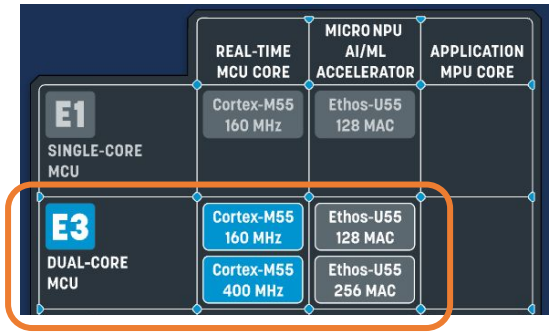


350mW

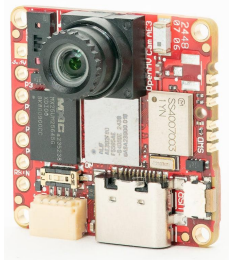
0.5 TOPS

seeed

OpenMV AE3



Alif E3

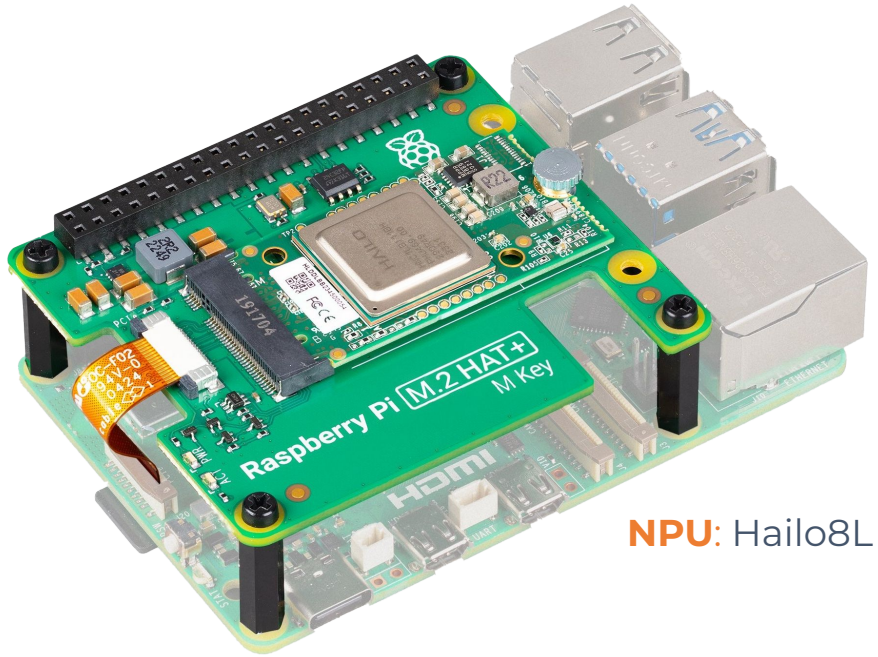


150mW

0.25 TOPS

NN Inference Accelerator

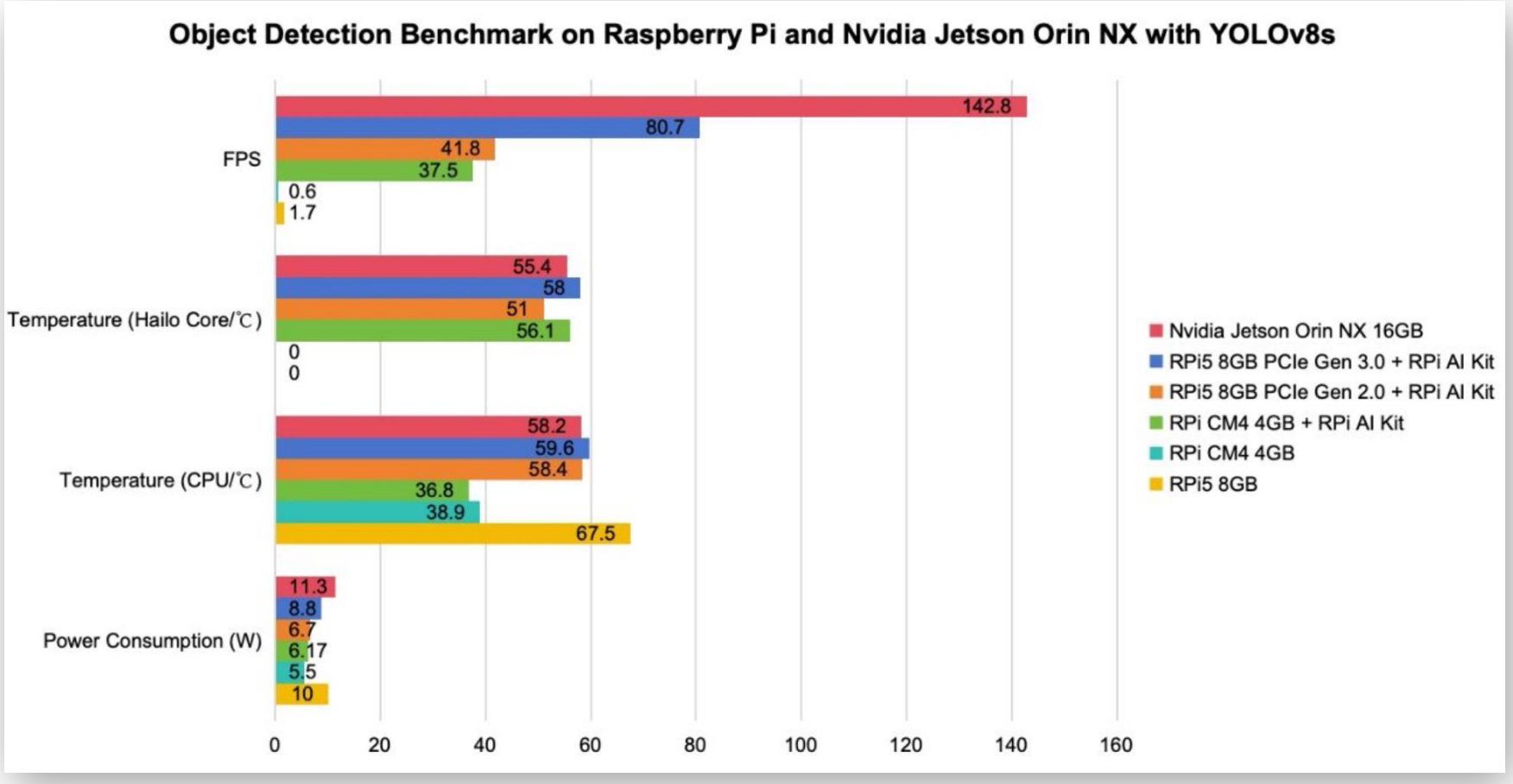
Raspberry Pi AI Kit



NPU: Hailo8L

30 W

13 - 25 TOPS

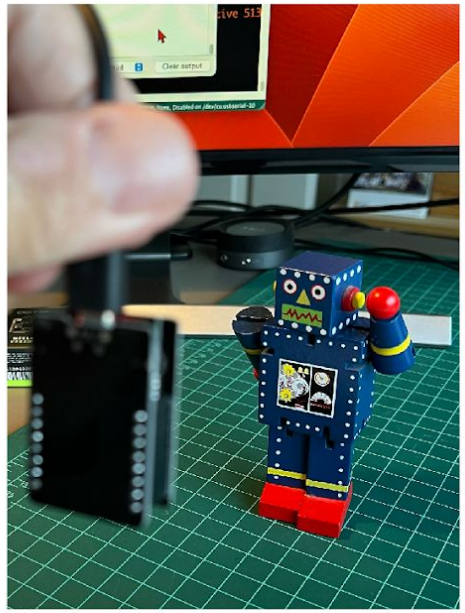


GPU: NVIDIA Ampere
100 TOPS

NPU: Hailo8L*

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

Tiny Image Classification Benchmark (MobileNetV2 96x96 0.1)



1.5 FPS

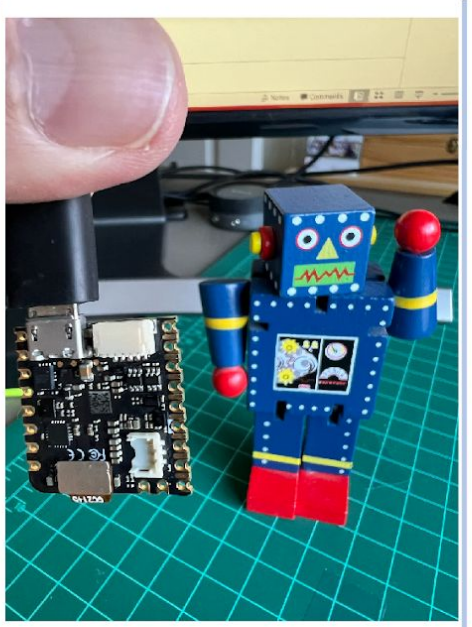
ESP - CAM
Xtensa LX6
240 MHz

300 mW

7.0 FPS

XIAO ESP32S3
Xtensa LX7
240 MHz

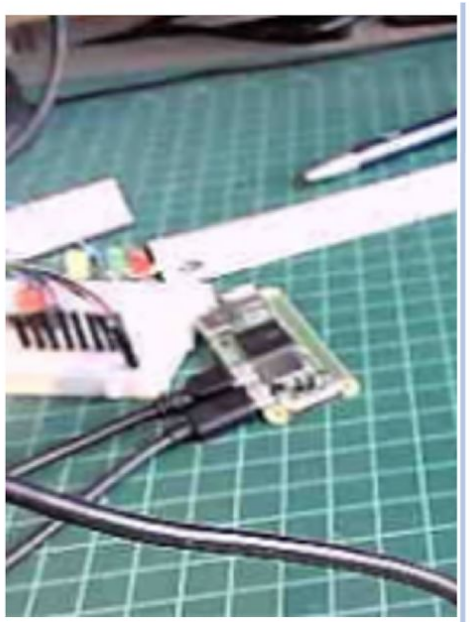
525 mW



11.6 FPS

Nicla-Vision
ARM M7
480 MHz

600 mW



91.0 FPS

Raspi Zero W2
ARM A53
1 GHz

1,500 mW

167 FPS

Grove Vision AI V2
ARM Ethus-U55
400 MHz

420 mW

Generative AI at the Edge

Transformers to LLMs and SLMs

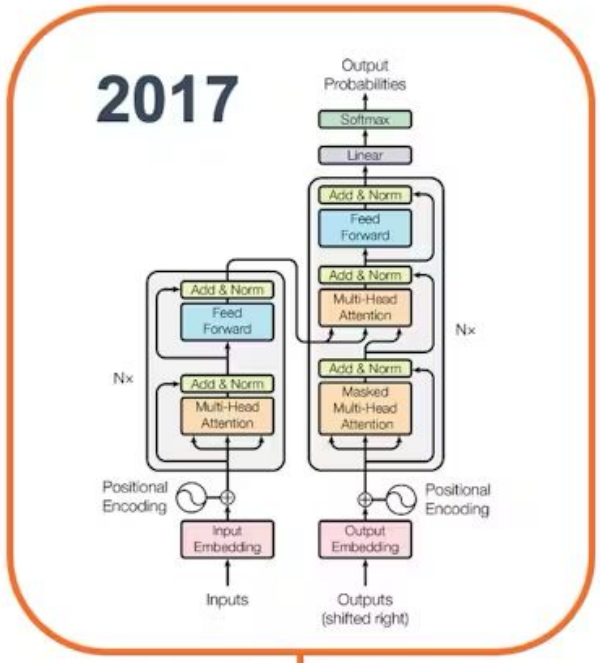
2025

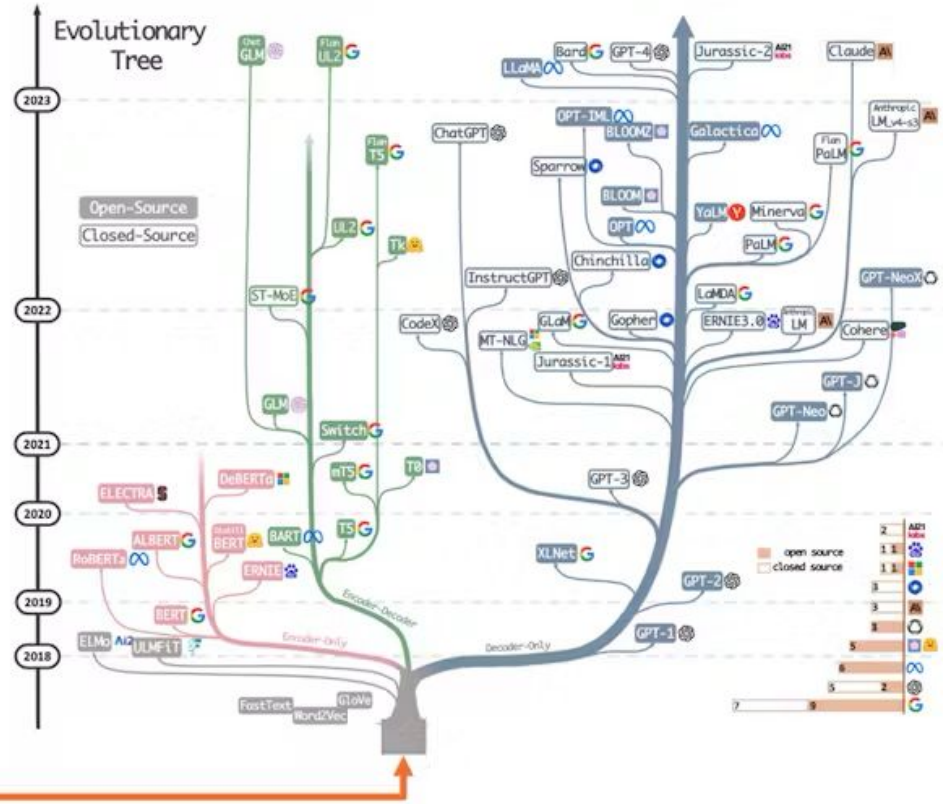
Open

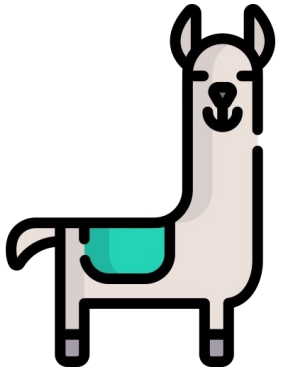
Qwen DeepSeek Phi Gemma LlaMa

Closed

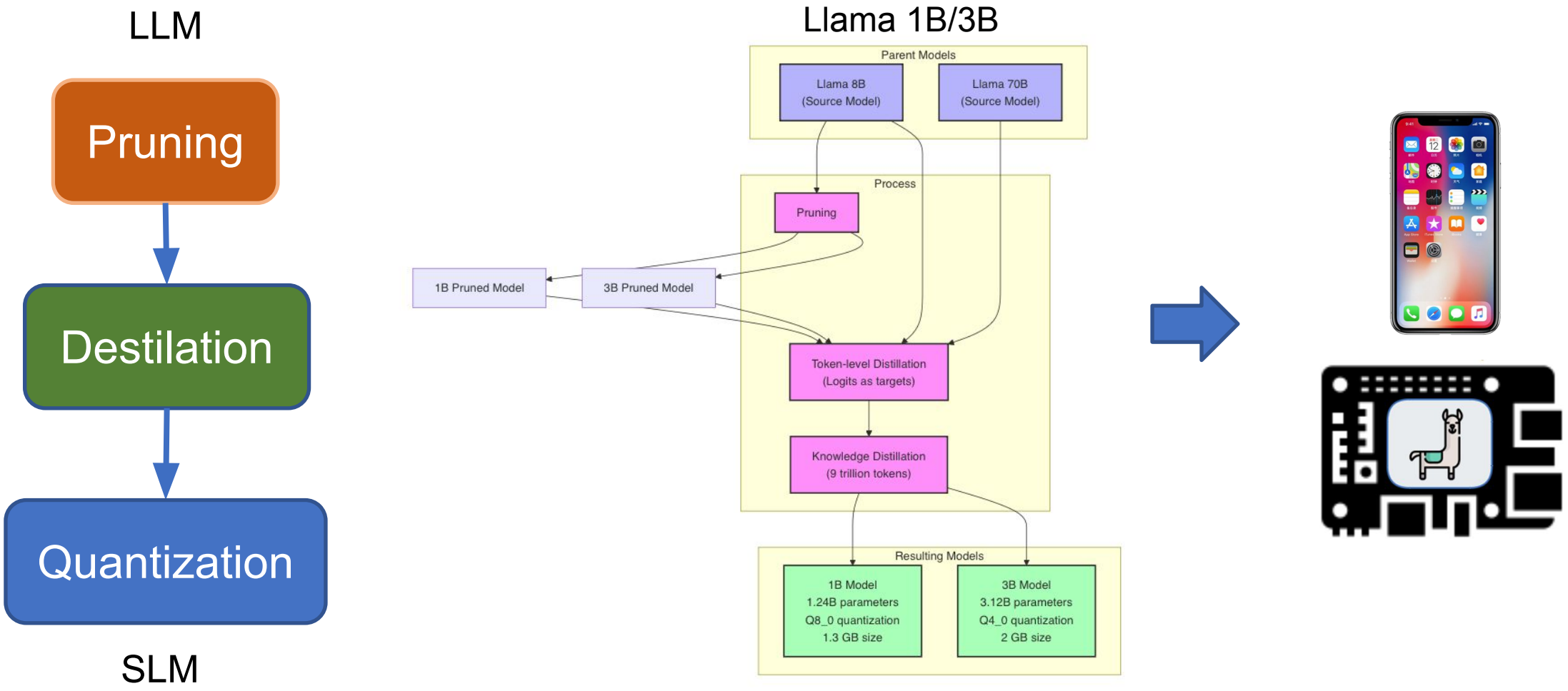
Gemini GPT Claude Grok







Small Language Models (SLMs)



Small Models

LlaMa

Gemma

```
marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 74×12
(ollama) mjrovai@raspi-5:~ $ ollama run llama3.2:3b --verbose
>>> What is the capital of France?
The capital of France is Paris.

total duration:       1.808927736s
load duration:        39.854862ms
prompt eval count:    32 token(s)
prompt eval duration: 221.506ms
prompt eval rate:     144.47 tokens/s
eval count:           8 token(s)
eval duration:        1.506376s
eval rate:            5.31 tokens/s
```

```
marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 67×13
(ollama) mjrovai@raspi-5:~ $ ollama run gemma2:2b --verbose
>>> What is the capital of France?
The capital of France is **Paris**.


total duration:       4.373339337s
load duration:        48.129697ms
prompt eval count:    16 token(s)
prompt eval duration: 1.968114s
prompt eval rate:     8.13 tokens/s
eval count:           13 token(s)
eval duration:        2.313284s
eval rate:            5.62 tokens/s
```

GenAi at the Edge: Llama3.2:1B

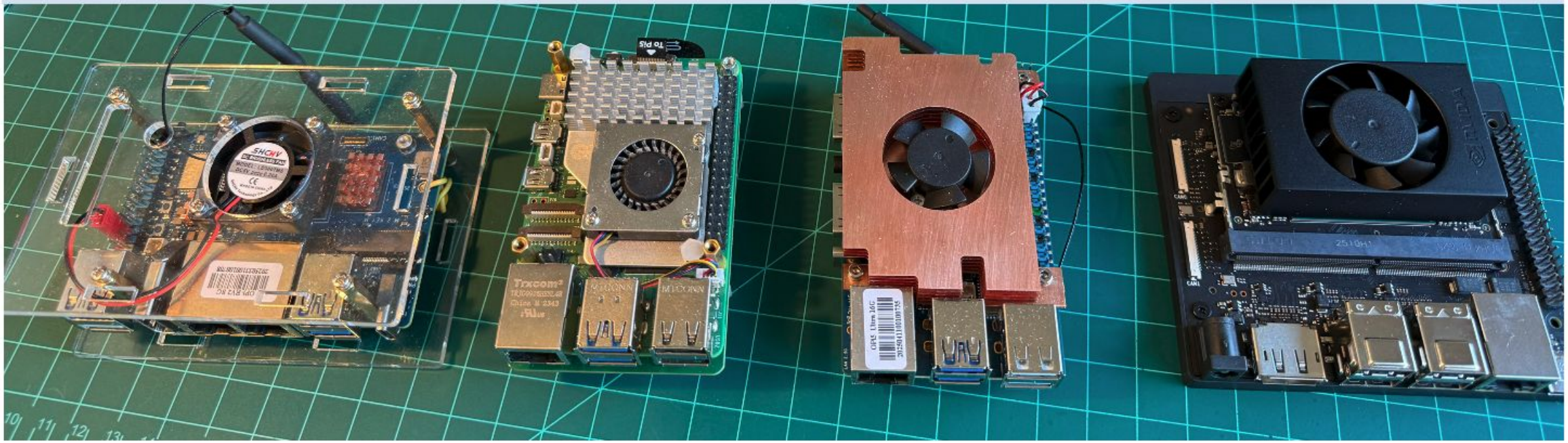
Desktop Reference

PC Linux (i7): 20 tokens/s

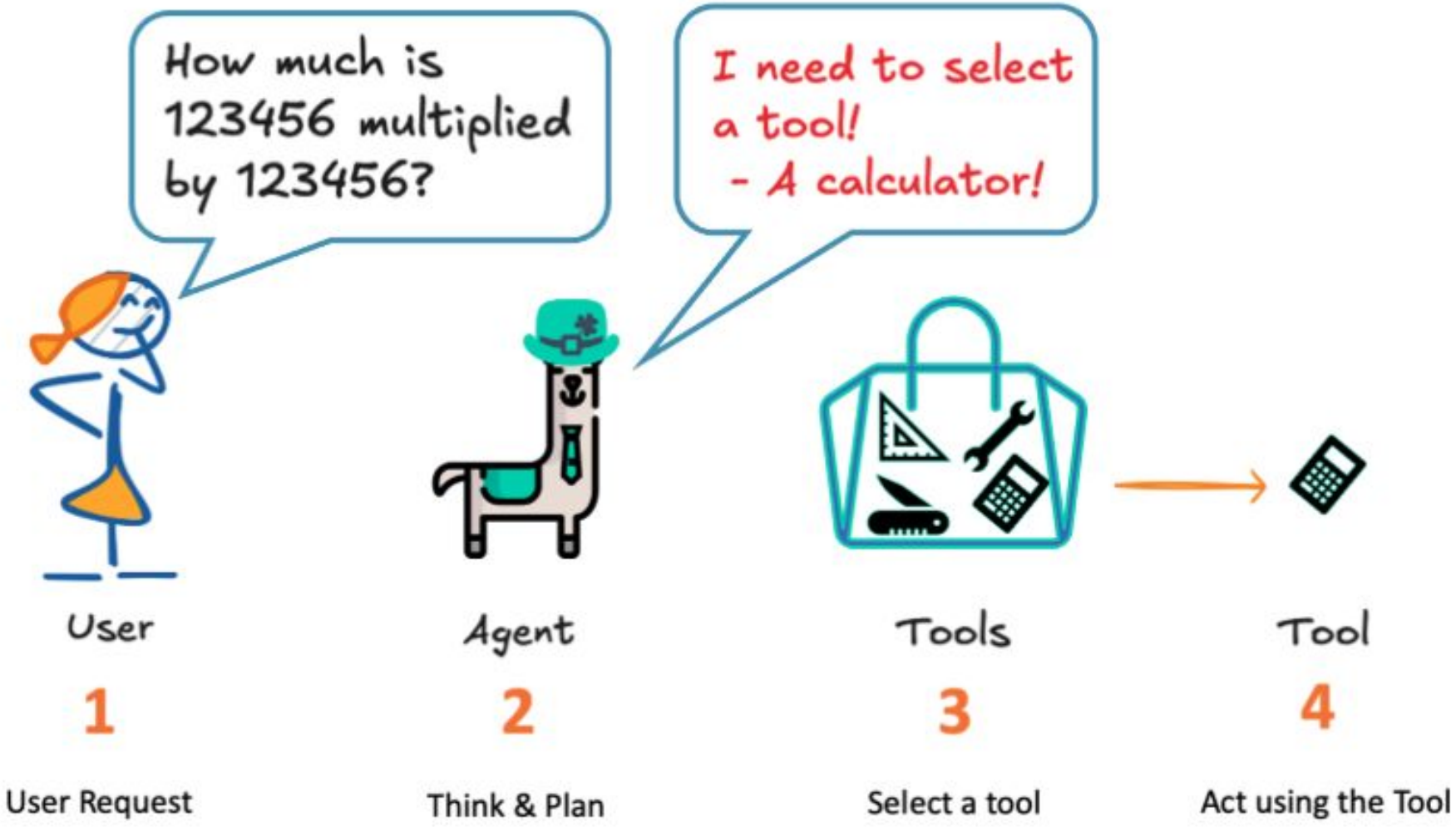
Mac (M1-Pro): 111 tokens/s

(*) Running on GPU

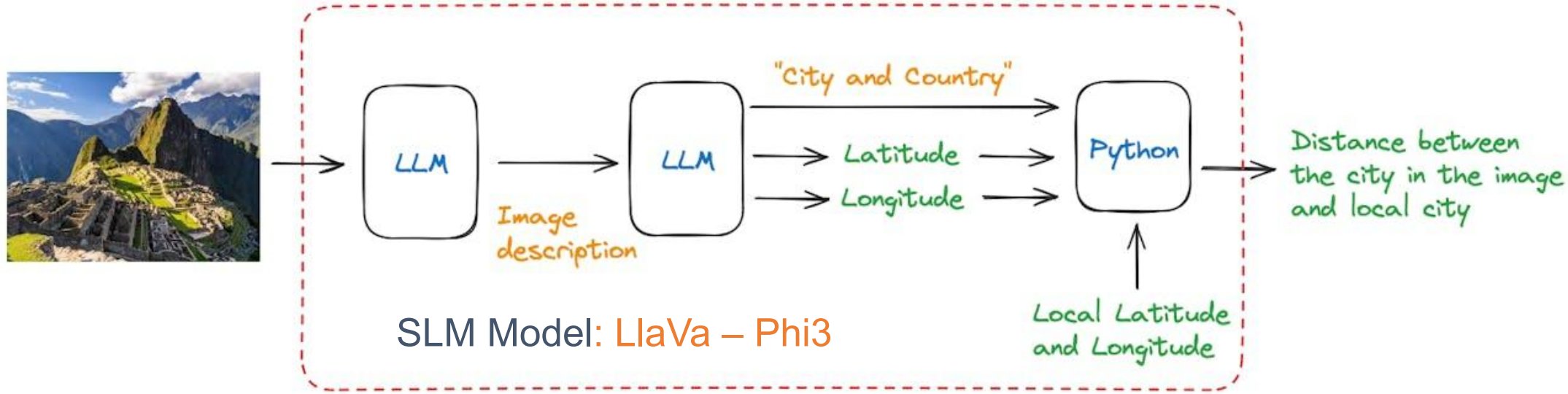
Orange Pi RV2	Raspi-5	Orange Pi 5 Ultra	Jetson Orin Nano
1 token/s	7.5 tokens/s	12 tokens/s	26 tokens/s (*)

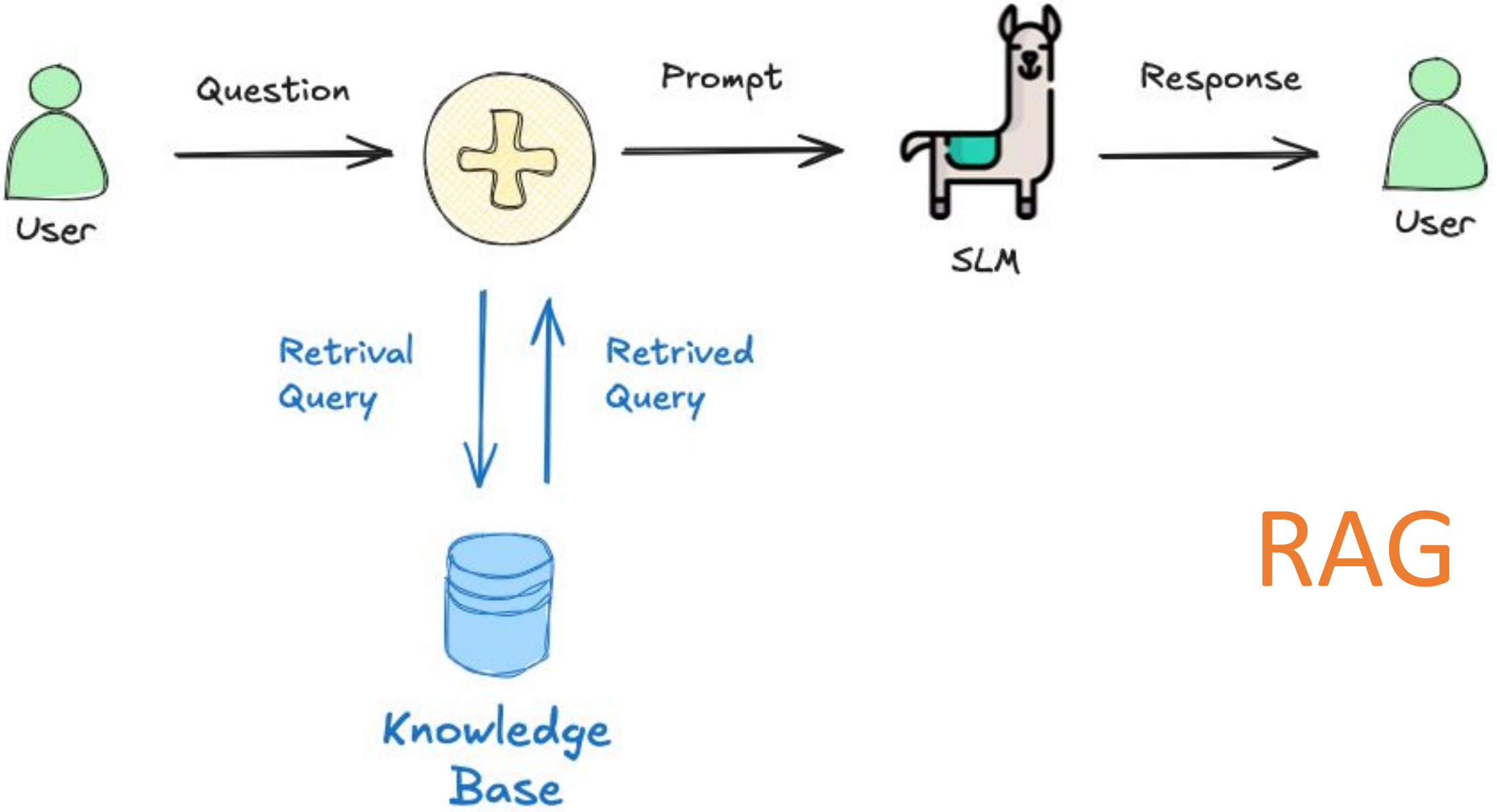


Agents


How much is 123456 multiplied by 123456?
I need to select a tool! - A calculator!
User
Agent
Tools
Tool
1
2
3
4
User Request
Think & Plan
Select a tool
Act using the Tool

Function Calling

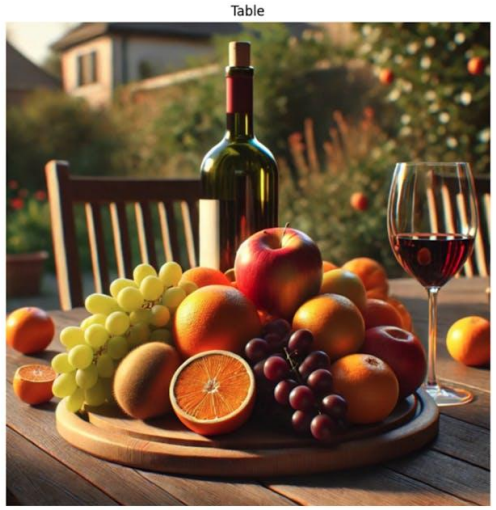



User
Question
Prompt
SLM
Response
User
Retrival Query
Retrived Query
Knowledge Base

RAG

Caption

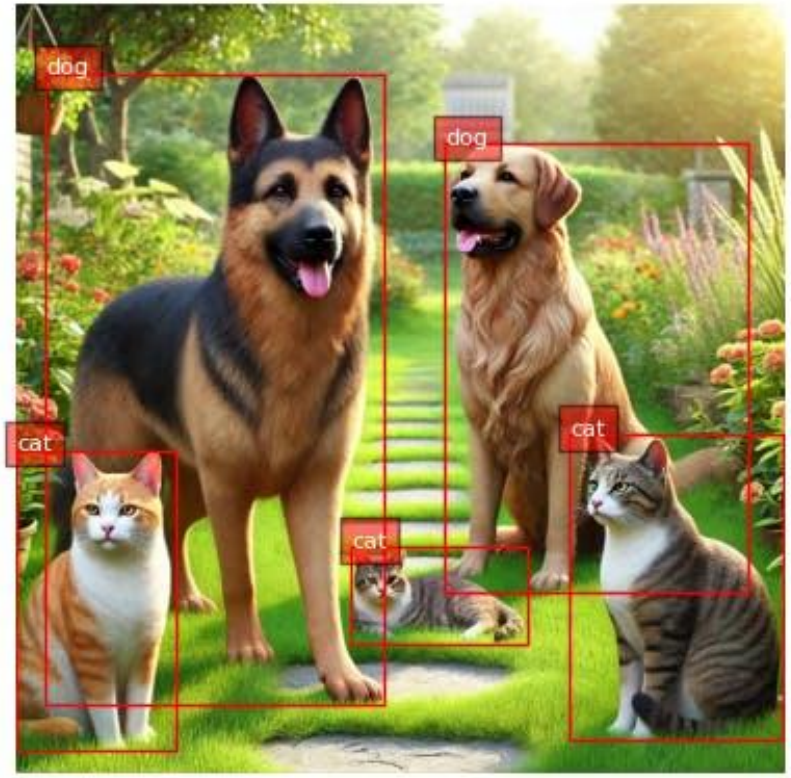
{'<MORE_DETAILED_CAPTION>': 'The image shows a wooden table with a wooden tray on it. On the tray, there are various fruits such as grapes, oranges, apples, and grapes. There is also a bottle of red wine on the table. The background shows a garden with trees and a house. The overall mood of the image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



{'<CAPTION>': 'A group of dogs and cats sitting in a garden.'}

{'<DETAILED_CAPTION>': 'The image shows a street with cars and people walking down it, surrounded by buildings with windows, railings, and balconies. There is a tree in the foreground and a clock tower in the background. The sky is filled with clouds and there is a watermark on the image.'}

Object Detection



Segmentation

Caption to Phrase Grounding

```
task_prompt = '<CAPTION_TO_PHRASE_GROUNDING>'
results = run_example(task_prompt, text_input="a church clock tower",image=city)
plot_bbox(table, results['<CAPTION_TO_PHRASE_GROUNDING>'])
```

```
[INFO] ==> Florence-2-base (<CAPTION_TO_PHRASE_GROUNDING>), took 12.7 seconds to execute.
```

A church clock tower

```
task_prompt = '<CAPTION_TO_PHRASE_GROUNDING>'
results = run_example(task_prompt, text_input="a person dressed in white",image=city)
plot_bbox(table, results['<CAPTION_TO_PHRASE_GROUNDING>'])
```

```
[INFO] ==> Florence-2-base (<CAPTION_TO_PHRASE_GROUNDING>), took 12.3 seconds to execute.
```

A person dressed in white

OCR


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

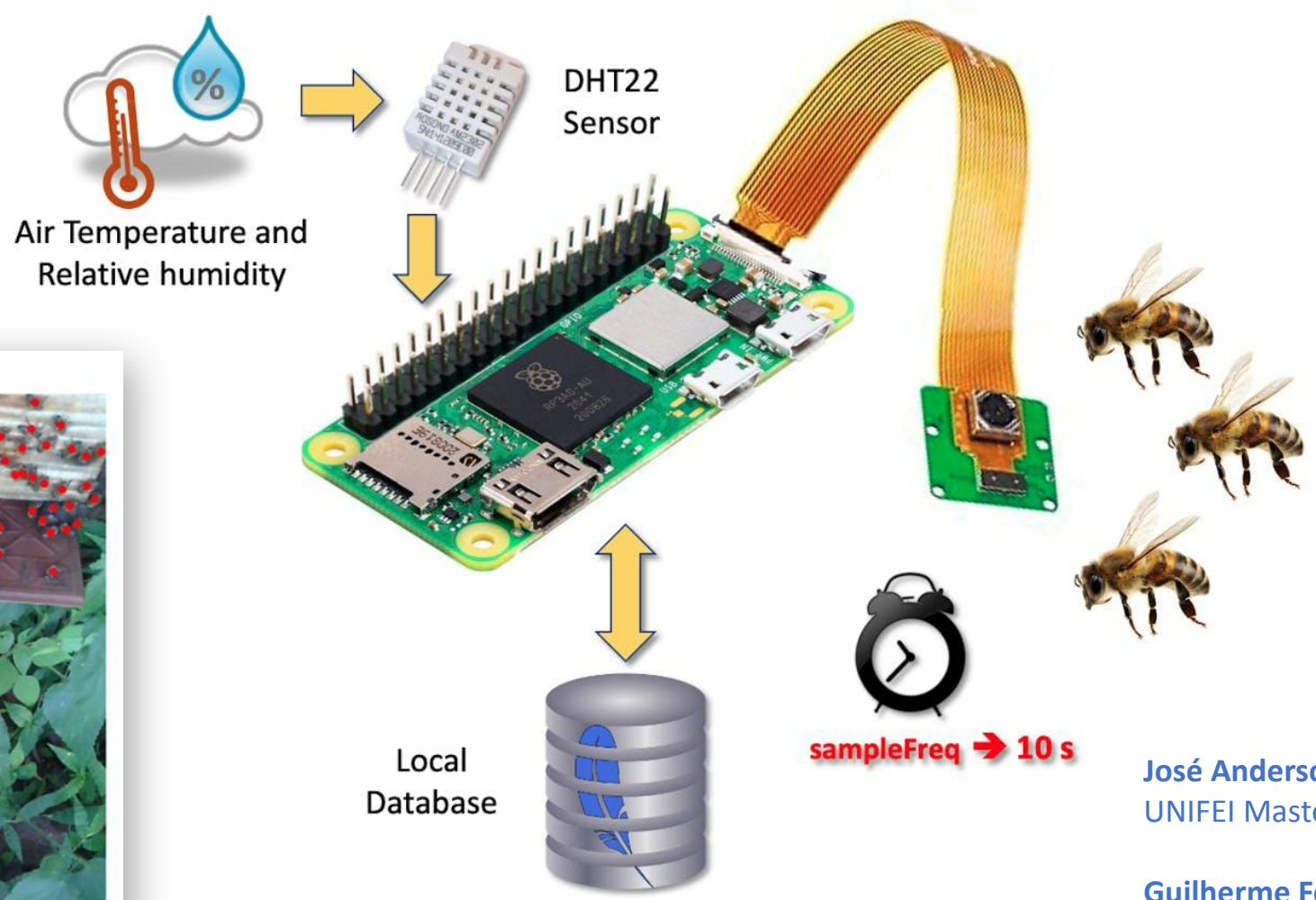
```
['</s>Machine Learning',
 'Café',
 'com',
 'Embarcado',
 'Embarcados',
 'Democratizando a Inteligência',
 'Artificial para Paises em',
 '25 de Setembro ás 17h',
 'Desenvolvimento',
 'Toda quarta-feira',
 'Marcelo Roval',
 'Professor na UNIFIEI e',
 'Transmissão via',
 'in',
 'Co-Director do TinyML4D']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

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], [35.36000061035156, 36.31999969482422, 104.15999603271484, 112.15999603271484], [0.1599999964237213, 0.47999998927116394, 319.5199890136719, 319.5199890136719]], "labels": ["wheel", "wheel", "box", "box", "box", "box", "wheel", "box"]}}

Real-World Applications

Bee Counting



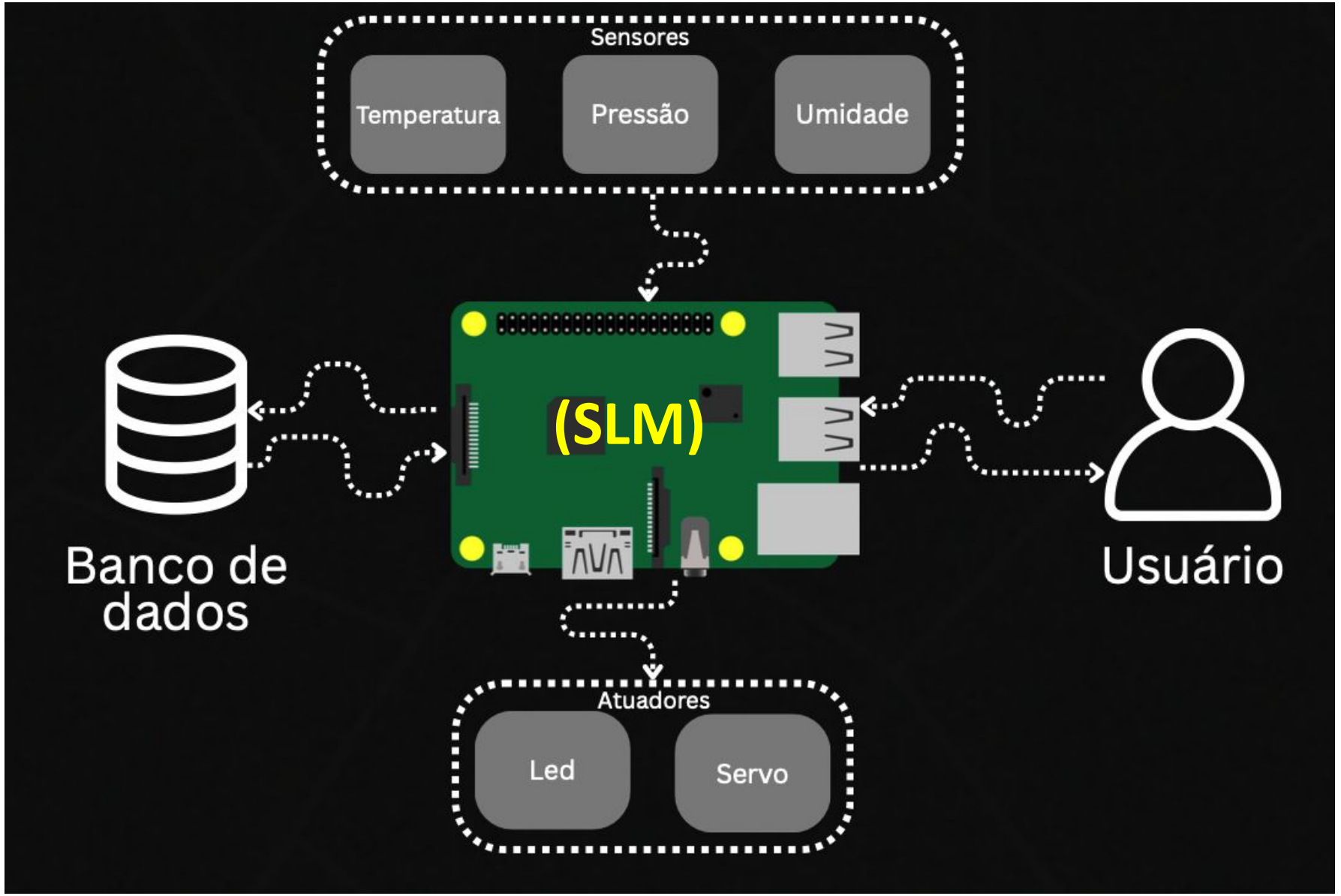
José Anderson Reis
UNIFEI Master's Student

Guilherme Fernandes
UNIFEI Grad Student

Ant Detection

FURG UTEC UNRaf

Juan Camilo Abedala
PRIA



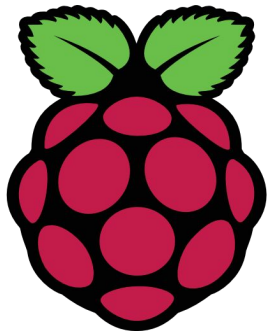
João Pedro Fonseca
UNIFEI Grad Student (IESTI)

Matheus Souza
UNIFEI Grad Student (IMC)

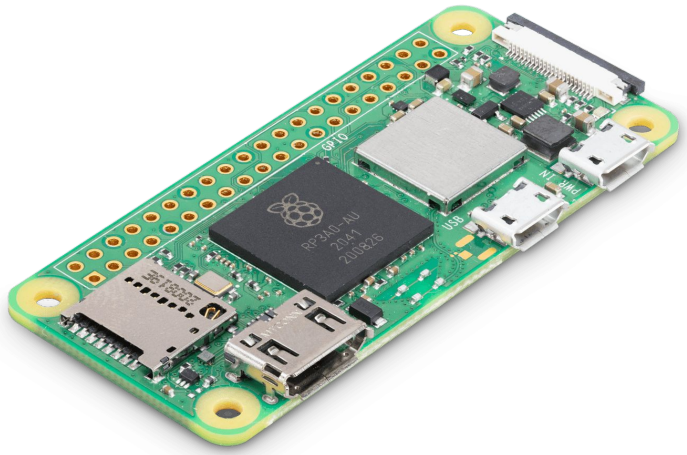
About the Course
& Syllabus

Hands-on Learning

- **Software / Tools**
  - Linux
  - Python
  - TensorFlow and PyTorch
  - Google Colab or Jupyter Notebook
  - Edge Impulse Studio
  - Ultralytics and RoboFlow
  - LangChain
  - And more…
- **Hardware**
  - Raspberry Pi Zero W2
  - Raspberry Pi 5
  - Miscellaneous items such as LEDs, Sensors, etc.



Part 1: Fixed Function AI (Reactive)

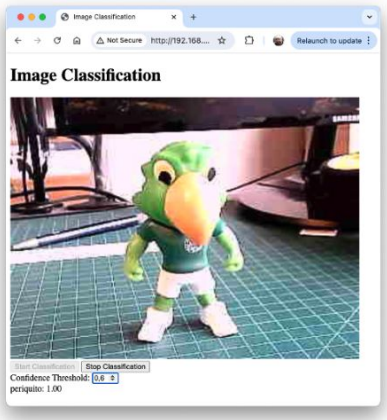


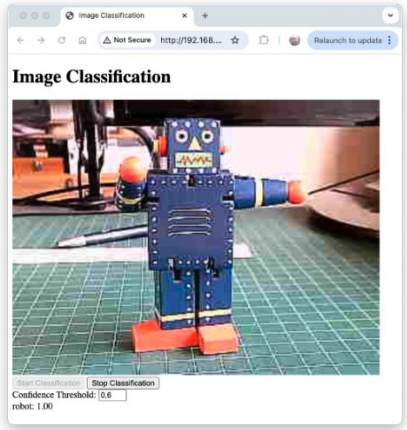
Raspberry Pi Zero W2

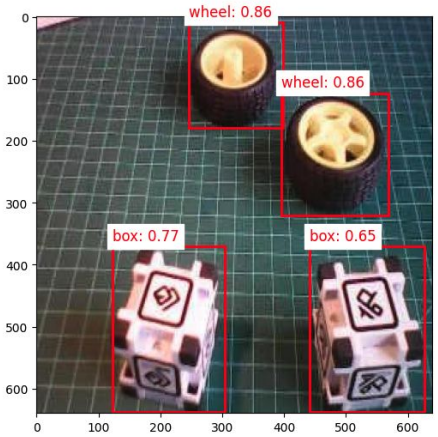
Part 2: Generative AI (Proactive)

Raspberry Pi 5

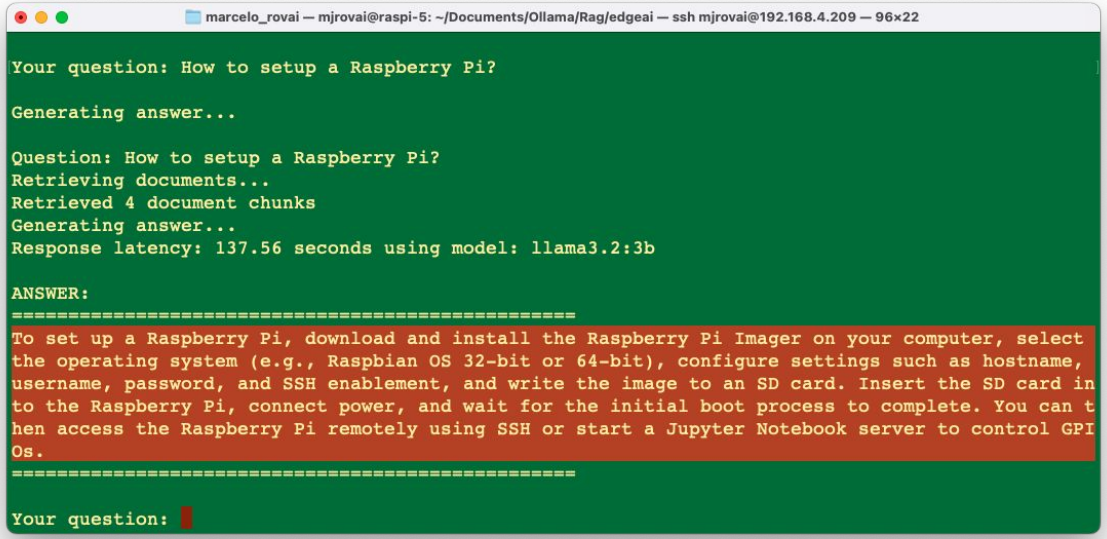
Part 1: Fixed Function AI (Reactive)

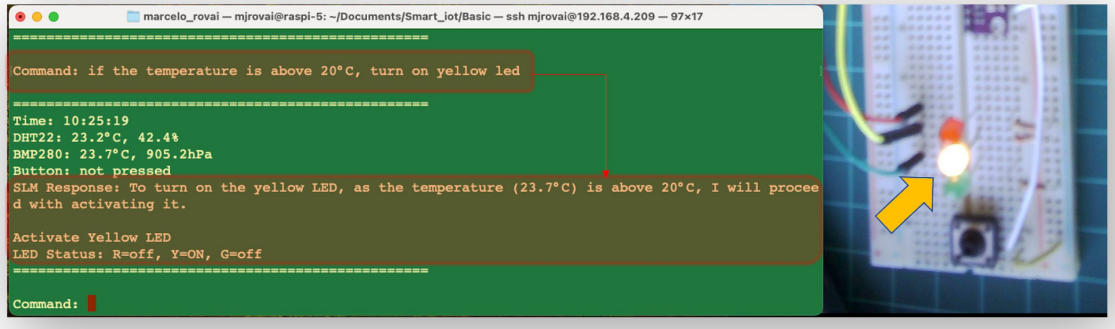






Part 2: Generative AI (Proactive)





Course Structure

The course is divided into two main parts (based on the textbook: Edge AI Engineering by Marcelo Rovai):

- **Part 1 (Weeks 1-7): Fixed Function AI** - Focus on image classification and object detection
- **Part 2 (Weeks 8-15): Generative AI** - Focus on Small Language Models (SLMs) and RAG systems

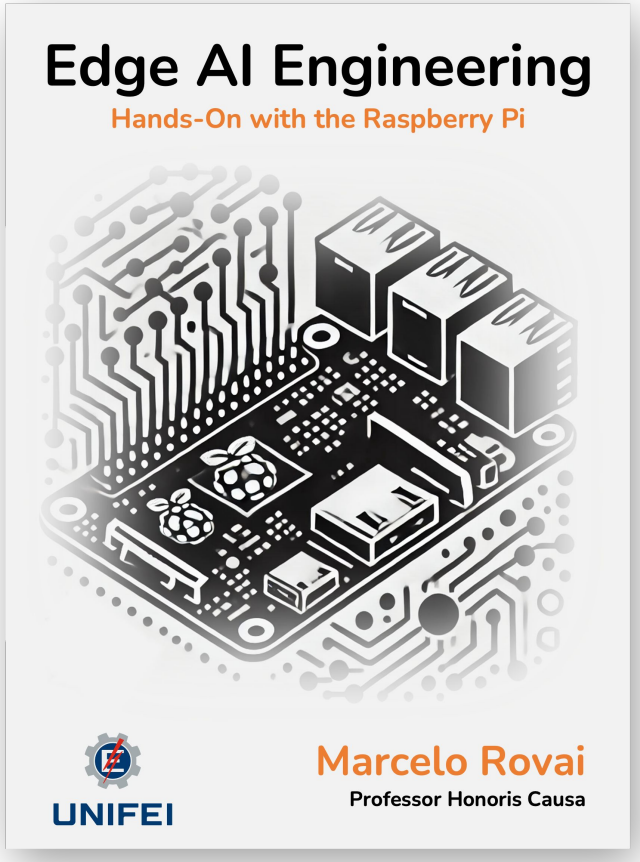
Key Learning Areas

- Raspberry Pi setup, configuration, and optimization
- Computer vision with OpenCV and TensorFlow Lite
- Image classification and object detection implementation
- Small Language Model deployment and integration
- Retrieval-Augmented Generation (RAG) systems
- Physical computing integration with sensors and actuators

Assessment Structure

- **40%** - Weekly hands-on labs, quizzes, and surveys
- **20%** - Midterm project (Fixed Function AI system)
- **30%** - Final project (Generative AI application)
- **10%** - Participation and documentation

EdgeAI Engineering



Editar Tópico: IESTI05 - Sum

https://moodle.unifei.edu.br/course/section.php?id=82357

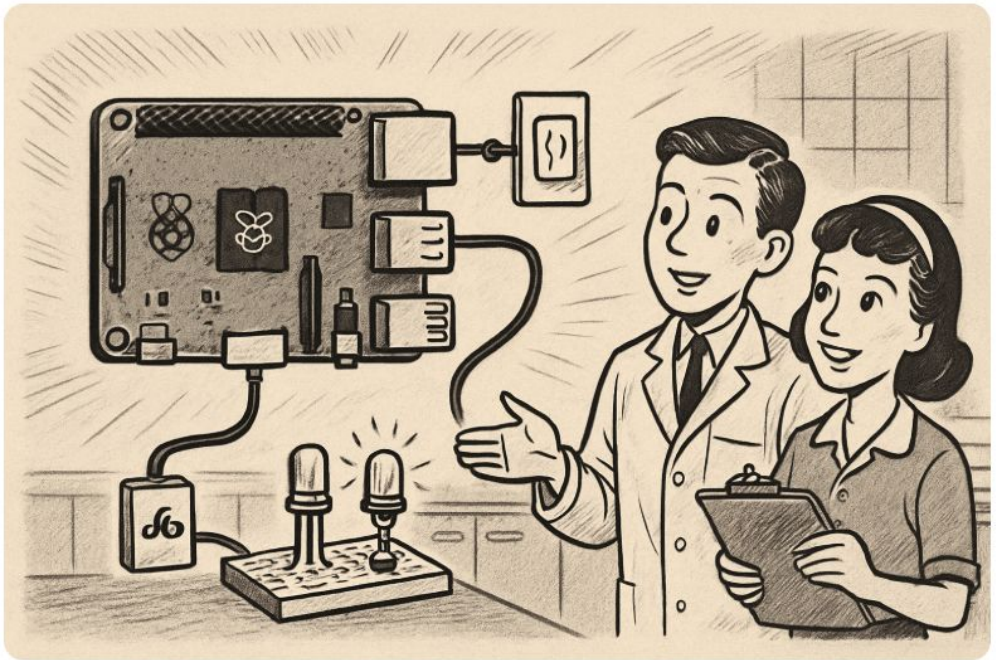
Página inicial Painel Meus cursos

Modo de edição

- IESTI05 - Summary
  - Avisos
- Pre-Course Activities
  - Suggested Review
  - Pre-Course Survey
- Week 1
  - About the Course and Syll...
  - Introduction to Edge AI
    - Lab 1a: Raspberry Pi Config...
    - Lab 1b: Development Enviro...
  - Optional Readings

IESTI05 - Summary

Ações em lote



Course Summary

Edge AI Engineering with Raspberry Pi is a 15-week undergraduate course designed to teach students how to implement AI systems on edge devices, specifically using Raspberry Pi platforms.

The course is based on the e-book: "Edge AI Engineering" by Prof. Marcelo Rovai, UNIFEI 2025

Course Structure

The course is divided into two main parts:

- **Part 1 (Weeks 1-7): Fixed Function AI** - Focus on image classification and object detection
- **Part 2 (Weeks 8-15): Generative AI** - Focus on Small Language Models (SLMs) and RAG systems

Tópico: Week 1 | IESTI05_202

https://moodle.unifei.edu.br/course/section.php?id=82817

Página inicial Painel Meus cursos

Modo de edição

- IESTI05 - Summary
  - Avisos
- Pre-Course Activities
  - Suggested Review
  - Pre-Course Survey
- Week 1
  - About the Course and Syll...
  - Introduction to Edge AI
    - Lab 1a: Raspberry Pi Config...
    - Lab 1b: Development Enviro...
  - Optional Readings

Introduction to Edge AI

For this week's lab, please review the Book Chapter: Setup.

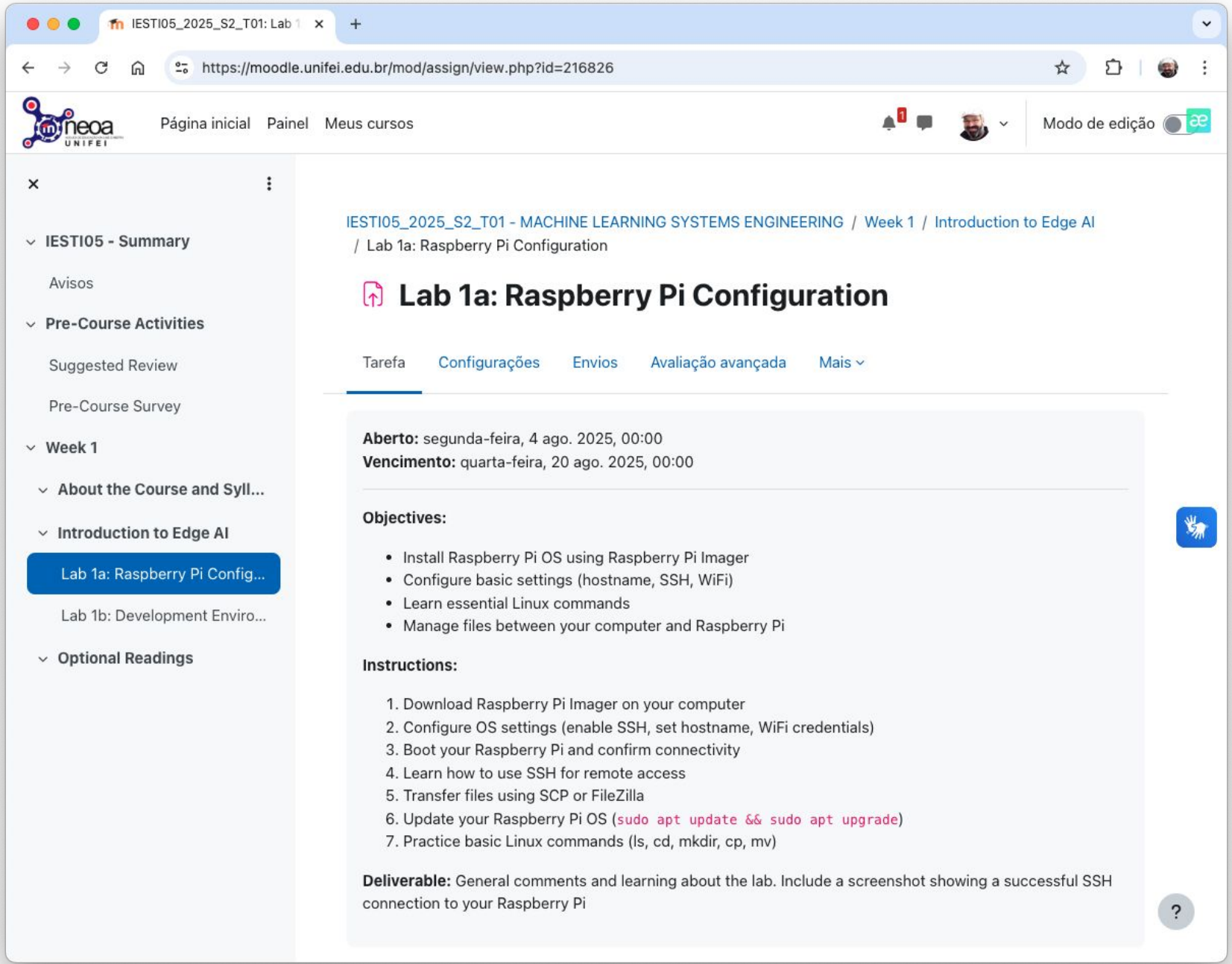
For a quick review, please take a look at the video below, which covers the most essential Linux terminal commands, from navigating the file system to managing files, processes, and permissions directly from the command line. The video has an audio track in Portuguese. We recommend that you see it before starting work in the labs.

Lab 1a: Raspberry Pi Configuration
Aberto: segunda-feira, 4 ago. 2025, 00:00 **Vencimento:** quarta-feira, 20 ago. 2025, 00:00

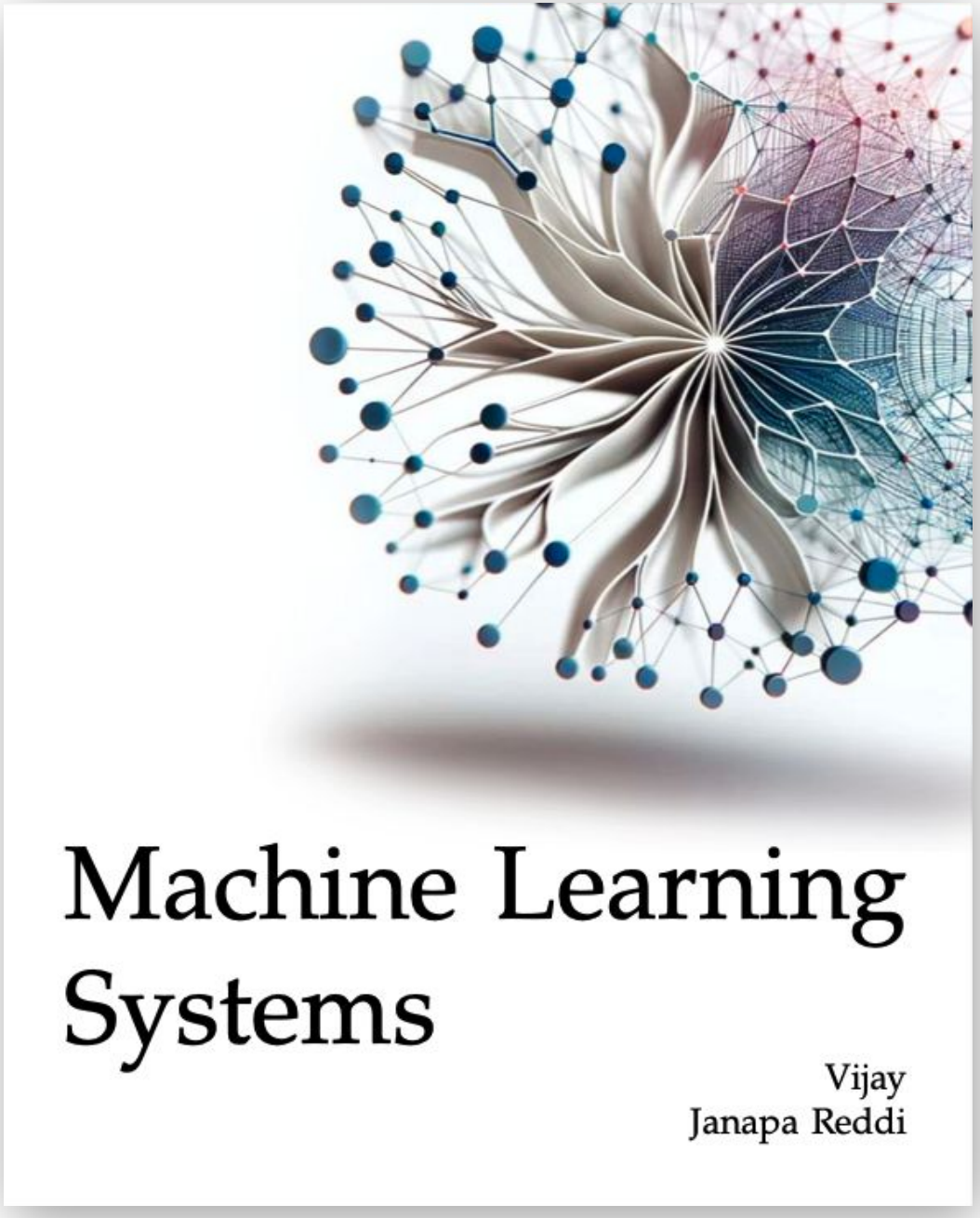
Lab 1b: Development Environment Setup
Aberto: segunda-feira, 4 ago. 2025, 00:00 **Vencimento:** quarta-feira, 20 ago. 2025, 00:00

Optional Readings

- Cutting AI down to size, Science Magazine
- Generative AI at the Edge: Challenges and Opportunities, Vijay Janapa Reddi, ASM


IESTI05_2025_S2_T01: Lab 1
https://moodle.unifei.edu.br/mod/assign/view.php?id=216826
Página inicial
Painel
Meus cursos
Modo de edição
IESTI05 - Summary
Avisos
Pre-Course Activities
Suggested Review
Pre-Course Survey
Week 1
About the Course and Syll...
Introduction to Edge AI
Lab 1a: Raspberry Pi Config...
Lab 1b: Development Enviro...
Optional Readings
IESTI05_2025_S2_T01 - MACHINE LEARNING SYSTEMS ENGINEERING / Week 1 / Introduction to Edge AI
/ Lab 1a: Raspberry Pi Configuration
Lab 1a: Raspberry Pi Configuration
Tarefa
Configurações
Envios
Avaliação avançada
Mais
Aberto: segunda-feira, 4 ago. 2025, 00:00
Vencimento: quarta-feira, 20 ago. 2025, 00:00
Objectives:
Install Raspberry Pi OS using Raspberry Pi Imager
Configure basic settings (hostname, SSH, WiFi)
Learn essential Linux commands
Manage files between your computer and Raspberry Pi
Instructions:
1. Download Raspberry Pi Imager on your computer
2. Configure OS settings (enable SSH, set hostname, WiFi credentials)
3. Boot your Raspberry Pi and confirm connectivity
4. Learn how to use SSH for remote access
5. Transfer files using SCP or FileZilla
6. Update your Raspberry Pi OS (sudo apt update && sudo apt upgrade)
7. Practice basic Linux commands (ls, cd, mkdir, cp, mv)
Deliverable: General comments and learning about the lab. Include a screenshot showing a successful SSH connection to your Raspberry Pi

Learning Deeper



https://mlsysbook.ai/

Questions?

Prof. Marcelo J. Rovai

rovai@unifei.edu.br

UNIFEI