IA on edge, une romance qui s'affirme

Fun exploration of AI on NVIDIA edge board

Most of this work was done for Gemma3n Kaggle hackathon

Frédéric Collonval

Full-stack developer freelance @WebScIT

Open-source contributor

JupyterLab SSC rep 2023 – Jupyter distinguish contributor 2021

linked.com/in/fcollonval

fcollonval

Motivation for Edge AI and Problem Statement

Challenges of Running Large AI Models

High Computational Power

Large AI models demand extensive processing power, which often requires specialized computing hardware.

Deployment Limitations

Deploying large AI models on mobile or embedded devices is challenging due to resource constraints.

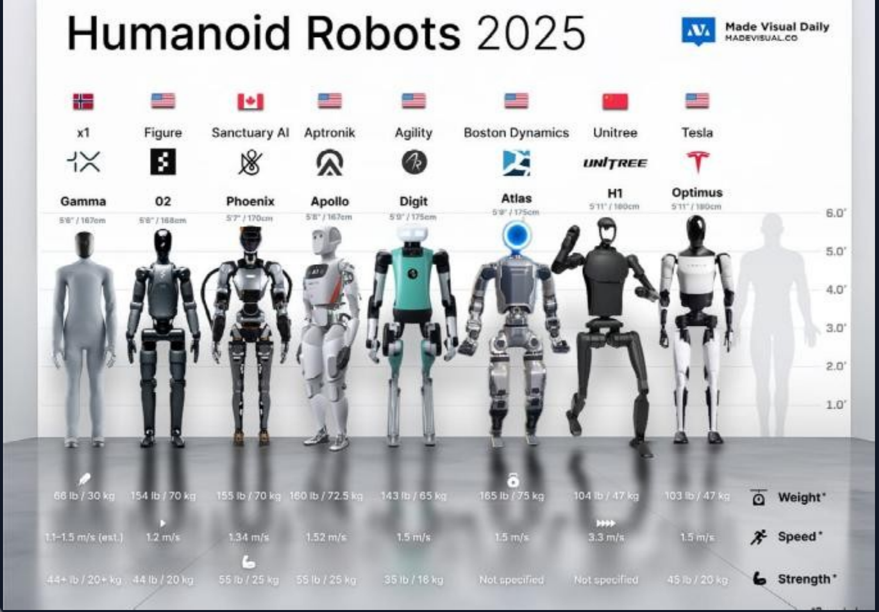
Accessibility Issues

High hardware costs and power consumption restrict accessibility to advanced AI technologies.

Local-first for privacy and fun 😈

High hardware costs and power consumption restrict accessibility to advanced AI technologies.

Growing Need for On-Device (Edge) AI for Mobile, Robotics, Cars



April 9, 2025 - https://www.madevisual.co/the-current-generation-of-humanoid-robots-2025/

Real-Time Processing

On-device AI enables instant data processing for immediate decision-making without delays.

Low Latency Benefits

Edge AI reduces communication delays by processing data locally rather than relying on the cloud.

Enhanced Privacy

Processing AI data on-device minimizes data exposure and protects user privacy effectively.

Autonomous Mobility

Edge AI empowers cars, drones, mobile and robotic systems to operate independently without cloud reliance.

Noticeable facts

Volvo will use 2x Nvidia AGX Orin 64Gb

Nvidia AGX Orin card found in Russian drones

Project Overview: Curious Frame

Concept: AI Tutor for Children

Idea

Offer kids an AI-powered device that can help them fulfill their curiosity.

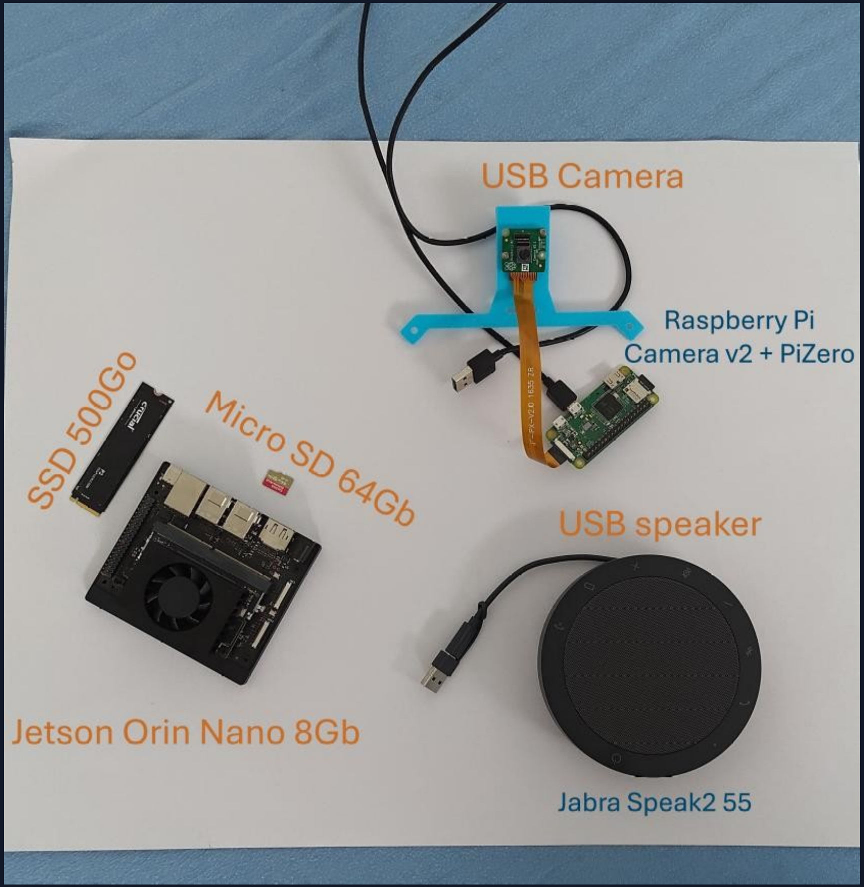
Explaining Objects Simply

The AI explains objects in an understandable manner that children can easily grasp.

Constraints

Children between 2 and 7 years old can't read and have limited speech

- No screen
- Collect only visual information
- Provide only vocal feedback in the kid language



System Components: RPi Camera, NVIDIA Jetson Orin Nano, Gemma3n Model

Image capture with a RaspberryPi camera v2

USB-connected camera

Edge Computing Platform

Nvidia Jetson Orin Nano – ARM board with 8Gb shared VRAM
SSD for better performance (but OS on micro SD)
No desktop to reduce default RAM resource

Sound

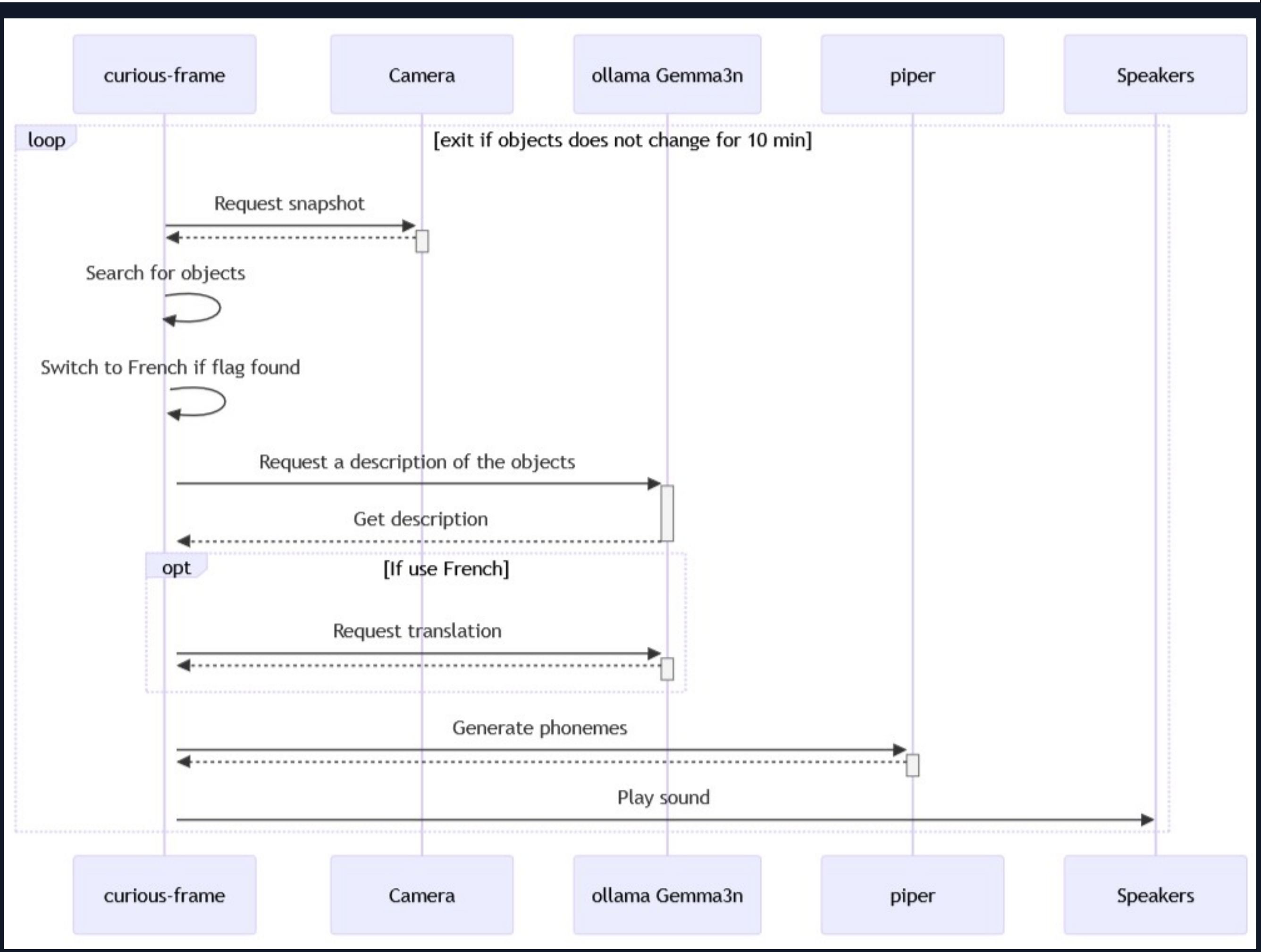
Jabra Speak2 55 connected through USB

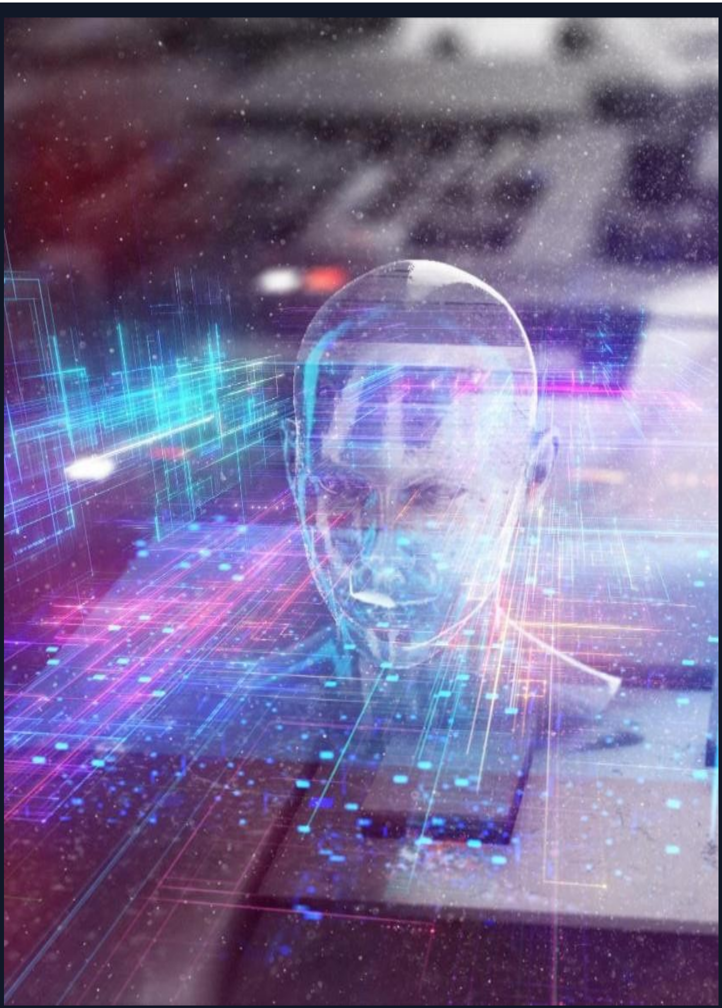
Cardboard frame

To point to an object to describe

Technical Implementation and Workflow

Image Capture and Processing Pipeline





Integration of Vision Language Model for Object Recognition

Vision
Gemma3n is executed on Ollama with no support for image input.
So moondream2 model is used to analyze the images.

Explanatory Output Generation

Gemma3n generates relevant textual explanations to describe recognized objects based on visual input.

Translation

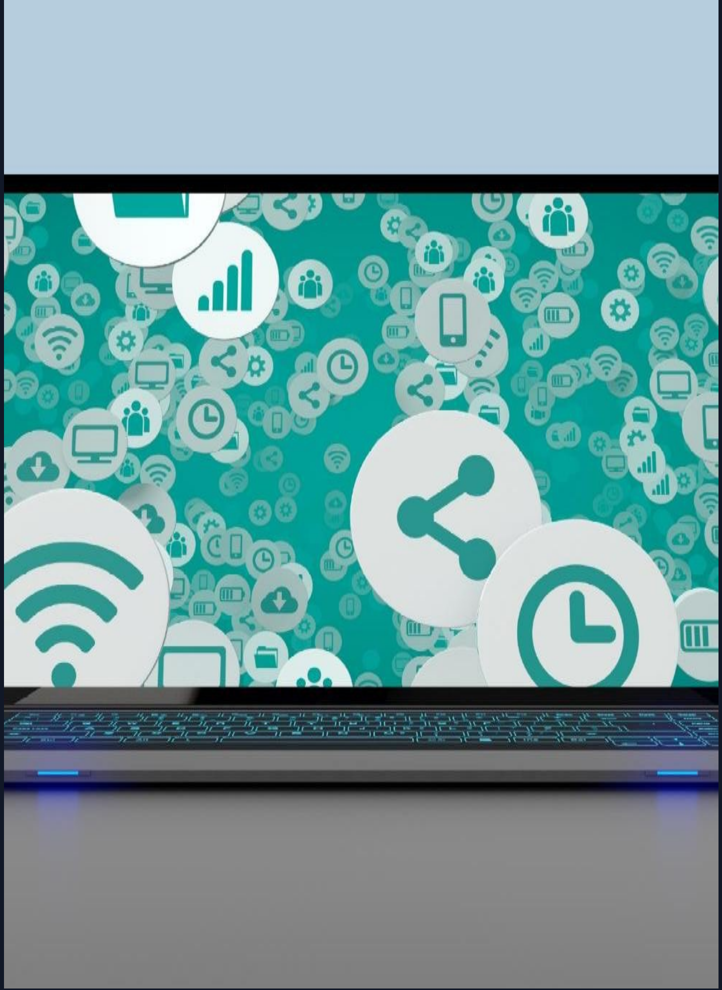
Gemma3n is used a second time to translate the description if the language is French.

Text-To-Speech (TTS)

Piper is used to transform text to phonemes.

Demonstration

Challenges, Lessons Learned, and Follow-ups



Edge Constraints: Latency, Memory, and Tooling

Latency Challenges

Minimizing latency is crucial for a smooth user experience on edge devices with limited processing capabilities.

Memory Limitations

Edge devices have limited memory capacity, requiring efficient data storage and processing approaches.

Tooling

Challenge to get tools stack working on the Jetson Orin Nano.

Potential Improvements

Vision and Language Fusion

Use a single model that can take text and image as inputs.

✅ using ministral-3:3b

TTS

Piper has a known issue with dropping the first phonemes. An alternative would be interesting

Next Step

Integration with Reachy mini – Santa should bring it in some weeks.

Conclusion

Edge AI Integration

Curious Frame integrates cutting-edge edge AI devices to enable advanced offline educational experiences.

Offline Learning Empowerment

Providing engaging educational content without internet connectivity enhances accessibility and usability.

Future Innovation Potential

This technology paves the way for innovations in learning methods and AI deployment in education.

References

First iteration done for Kaggle hackathon: https://kaggle.com/competitions/google-gemma-3n-hackathon/writeups/the-curious-frame-an-offline-ai-based-tutor-for-cu

Youtube video: https://youtu.be/yx0OXfG8UnQ?si=vukv0psQrXsM_f51

Code link: https://github.com/webscit/curious-frame