

**seeed studio**  
the AI hardware partner

Eric Pan

2024. 5

# TinyML devices enabling physical GPT

# About us

## the AI hardware partner

Seed Studio has been a leading Open Hardware company since 2008, empowering half a million direct users to create real-world digital solutions. Through relentless efforts and earned trust, our ever-growing product lines now form around emerging AI scenarios:

- Sensor networks to fetch extensive real-time data
- Edge computing to push intelligence to new frontiers

We provide industrial-ready modules and devices, and open up the capability of prototype, produce, and promote as Fusion service. Innovators from different vertical domains co-create with us to make their creations widely available for diversified markets.

By embracing open source, community building and integrated software suites like SenseCraft, we are proactively lowering the tech barriers and including users with diverse expertise for globalized matters.



**From possibilities to  
productivities**



# From Technology to Industry

## Technologies

- Open Source Hardware
- Machine Learning
- Advanced Sensors
- Home Assistant
- Wireless. DePIN
- LLM

## Emerging Technologies



## Applications

- Asset Tracking
- Smart Building
- Industrial space
- Smart City
- Smart Agriculture
- Smart Energy
- Tech for good
- Open Science

## Digital Economy



## Traditional Industries

- Digital Infrastructure
- Smart Environment
- Smart Agriculture
- Smart Energy

- ARDUINO
- Arduino
- Beaglebone
- Home Assistant
- NVIDIA
- NVIDIA@ Jetson
- Raspberry Pi
- Raspberry Pi
- SenseCAP
- TINY ML
- TinyML

# Adding AI to almost anything

## multi-modal

vision  
sound  
speech  
sensors

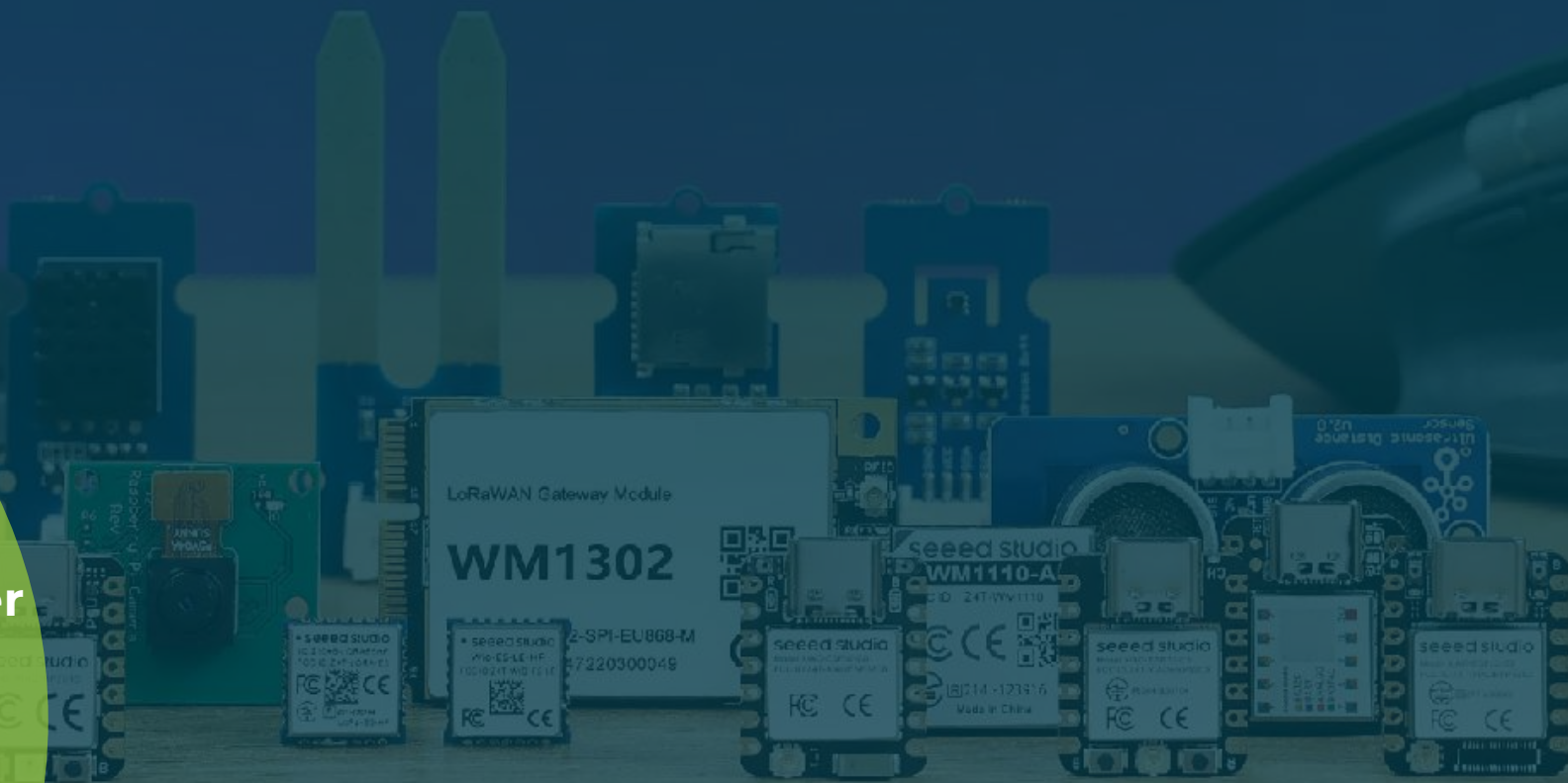
**Faster**  
new architecture  
(Cortex-M55)

**low power**  
<1 w

# TinyML

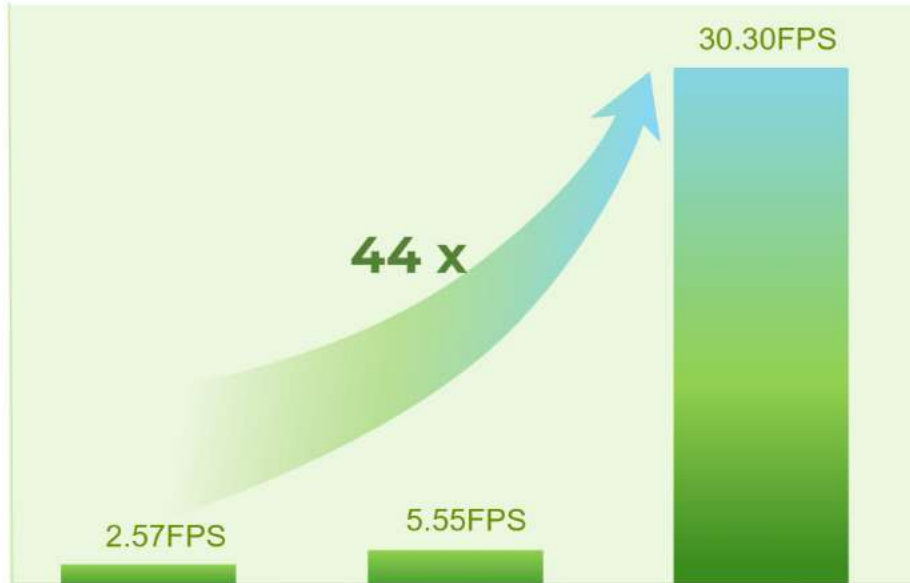
**Cheaper**  
<10 \$

**Easier**  
few shot  
training  
no-code  
web server



# Faster and Lower power

## Refresh Rate



Grove Vision AI v1

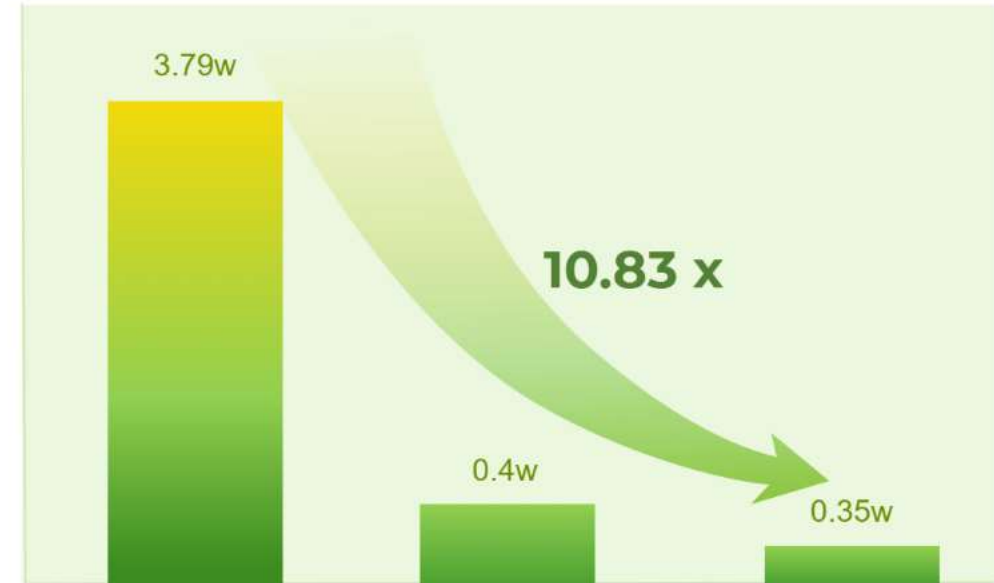


XIAO ESP32S3



Grove Vision AI v2  
(M55 + U55 AI addon)

## Energy Efficiency



Raspberry Pi 4B



Grove Vision AI v1



Grove Vision AI v2  
(M55 + U55 AI addon)

# Faster

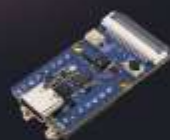
## 2024 MCU AI Vision Boards: Performance Comparison



**Power Consumption: 0.40W**  
**Inference Time: 389.0ms**  
**Frame Rate: 2.57FPS**  
**Ease of Use: 8.0**  
**Price: \$25.99**

### Grove Vision AI

Himax HM6327-A (ASIC: 518K DSR) 400 MHz



**Power Consumption: 0.35W**   
**Inference Time: 33.0ms**  
**Frame Rate: 30.30FPS**  
**Ease of Use: 9.0**  
**Price: \$23.89**

### Grove Vision AI V2

Himax WiseEye2 HM6530 (Cortex-M55 + Ethos-U55) 400 MHz + 300MHz (M55) + 400MHz (U55)



**Power Consumption: 0.46W**  
**Inference Time: 180.0ms**  
**Frame Rate: 5.55FPS**  
**Ease of Use: 4.0**  
**Price: \$45.00**

### ESP32-EYE

ESP32S3 (Dual-Core Tensilica LX6) 240 MHz



**Power Consumption: 0.45W**  
**Inference Time: 180.0ms**  
**Frame Rate: 5.55FPS**  
**Ease of Use: 9.0**  
**Price: \$13.99**

### XIAO ESP32S3 Sense

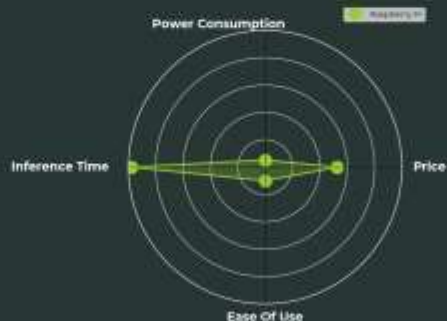
ESP32S3 (Dual-Core Tensilica LX6) 240 MHz



(Based on OpenMV)  
**Power Consumption: 0.59W**  
**Inference Time: 178.89ms**  
**Frame Rate: 5.59FPS**  
**Ease of Use: 6.0**  
**Price: \$115.00**

### Arduino Nicla Vision

STM32H7A3IIE (ARMv8 Cortex-M7/M4) 480MHz (M7) + 340MHz (M4)



**Power Consumption: 3.79W**  
**Inference Time: 8.83ms**  
**Frame Rate: 113.21FPS**  
**Ease of Use: 1.0**  
**Price: \$55**

### Raspberry Pi 4B

Broadcom BCM2711 (Cortex-A72 ARM V8) 1.5GHz

## How I conduct the test

1. flash the same test model on board - Swift-YOLO Tiny 96x96
2. feed the camera with the same human face picture under the same condition
3. record their performance

### Notes:

1. Raspberry Pi 4B is Included for perspective on CPU vs. MCU performance, despite its difference from MCU boards
2. Nicla Vision: Due to compatibility issues with the test model, it is tested with an alternative method; results are not directly comparable with other boards.
3. The inclusion of these 2 boards is intended to offer a broader perspective on processing capabilities across different hardware platforms. Their results are color-highlighted for clear distinction.



Scan the code to read the article

\*The larger the colour block area, the better the combination.

# Smaller and cheaper

## XIAO - tinyML MCUs

### Add AI to Almost Everything

The Seeed Studio XIAO Series is a collection of thumb-sized, powerful microcontroller units (MCUs) tailor-made for space-conscious projects requiring high performance and wireless connectivity. Embodying the essence of popular hardware platforms such as ESP32, RP2040, nRF52840, and SAMD21, the Arduino compatible XIAo series is the perfect toolset for you to embrace tiny machine learning (tinyML) on the Edge. Trusted by 500,000 developers globally!



- **Module and Development Board Hybrid**

Enabling Rapidly Prototype While Easily Integrate, Significantly Streamline Product Development Process



- **Invest for Your Future**

Unified Form Factor Enables You to Seamlessly Upgrade or Downgrade Your Product at the Lowest Cost



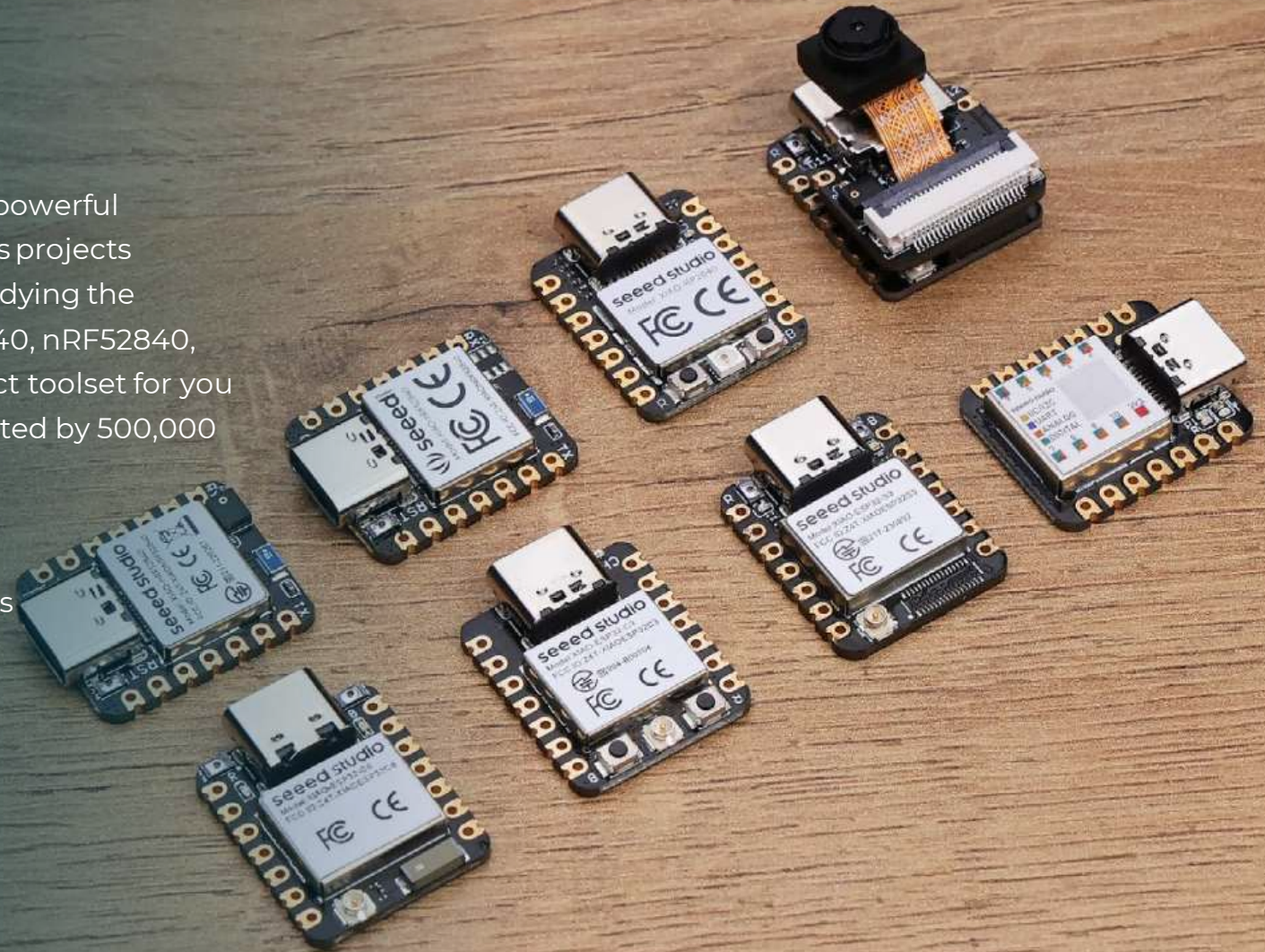
- **Single-Sided Surface Mount Design**

Effortlessly Incorporate XIAO into Other Boards for Large-Scale Manufacturing



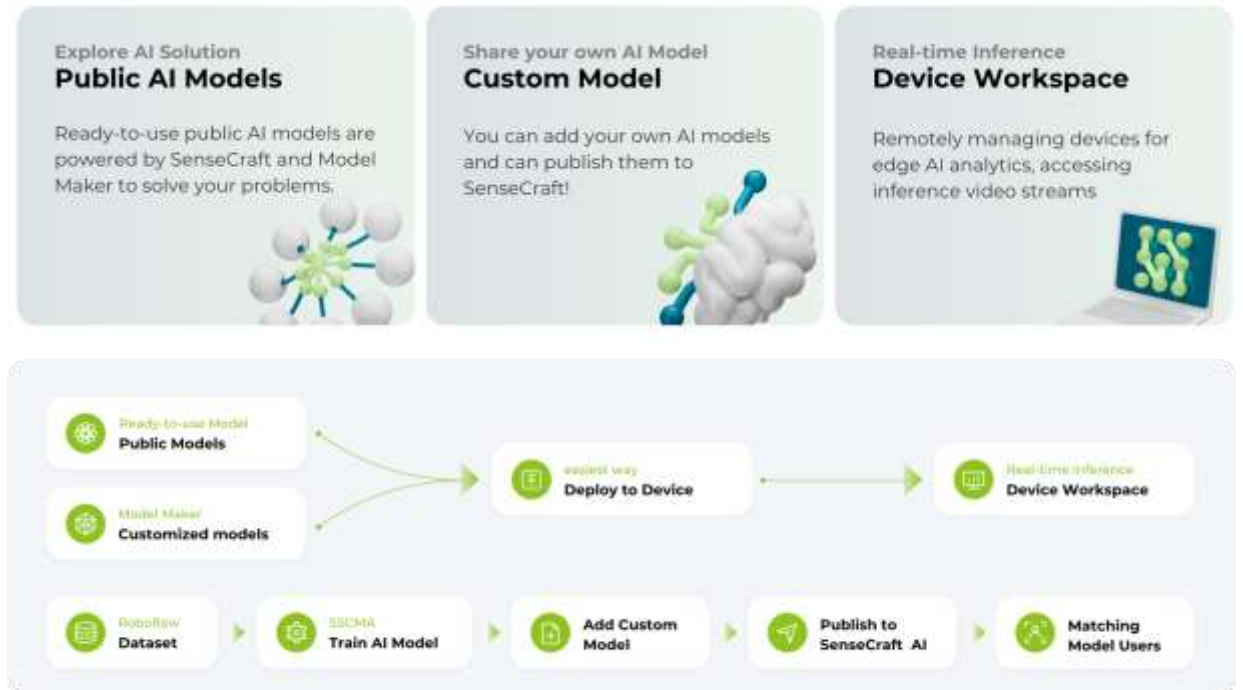
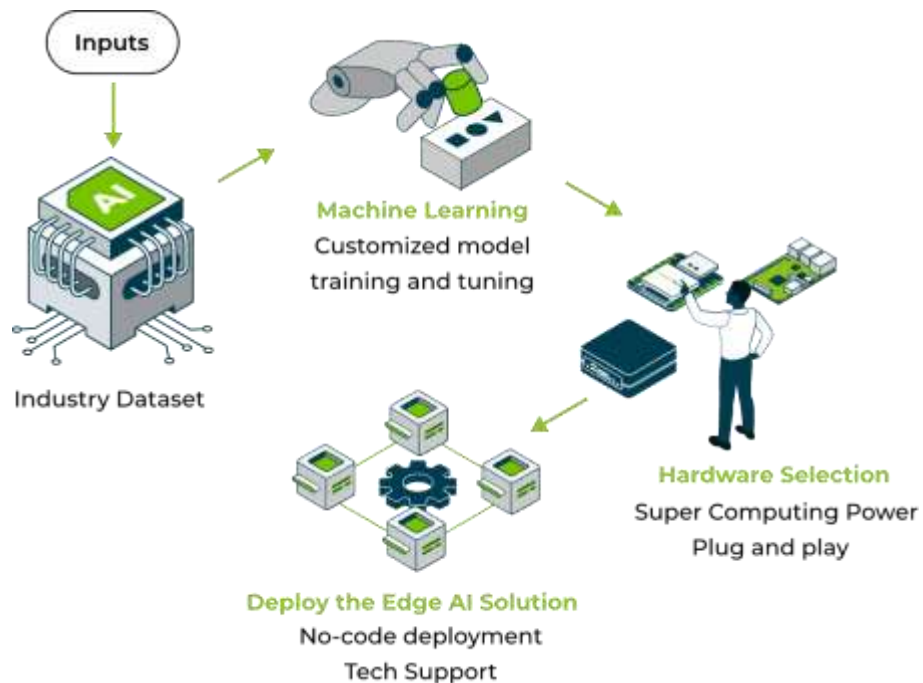
- **Strong Ecosystem Support**

Extensive Software Compatibility, Abundant Community Resources, and Dedicated Technical Assistance



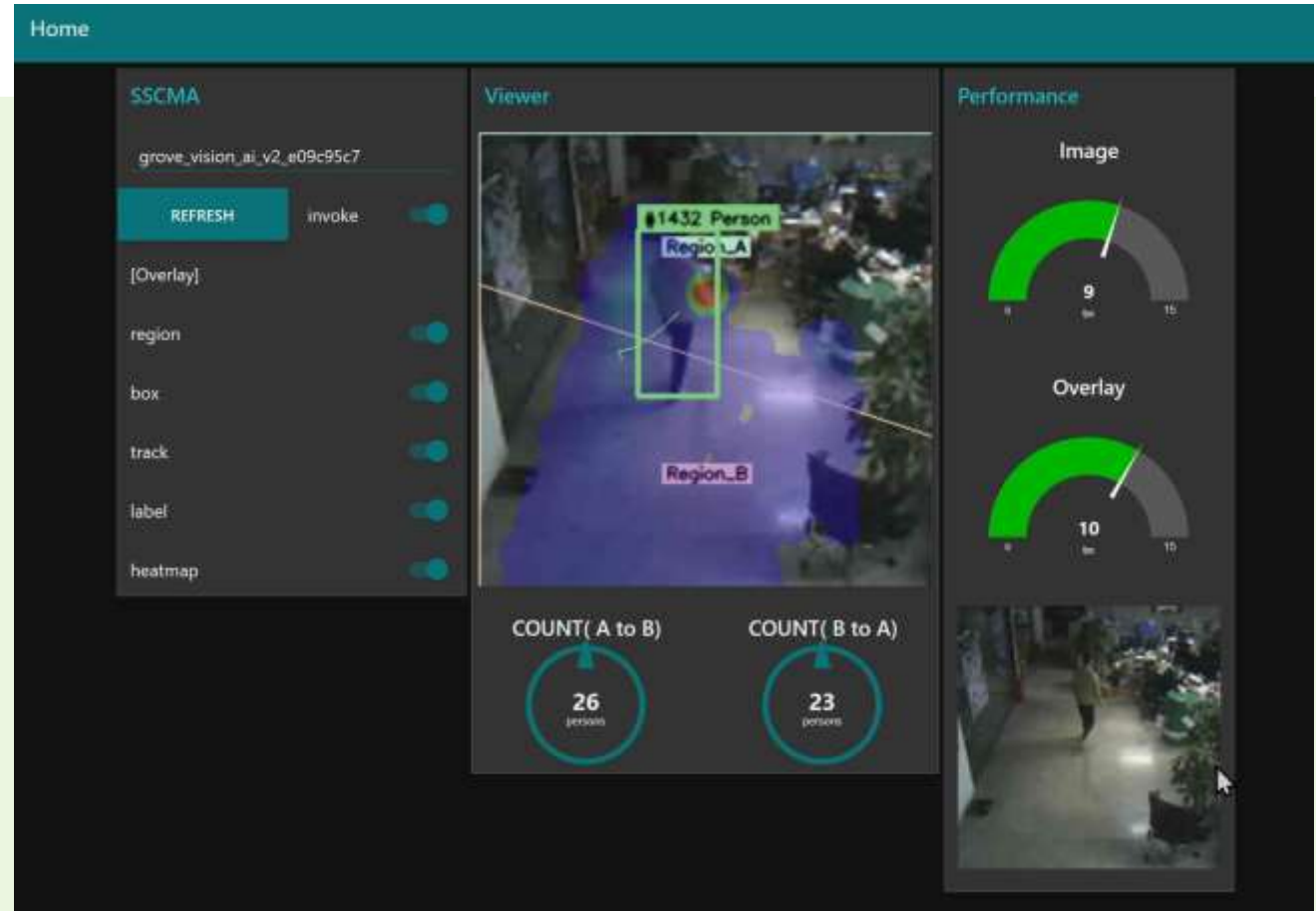
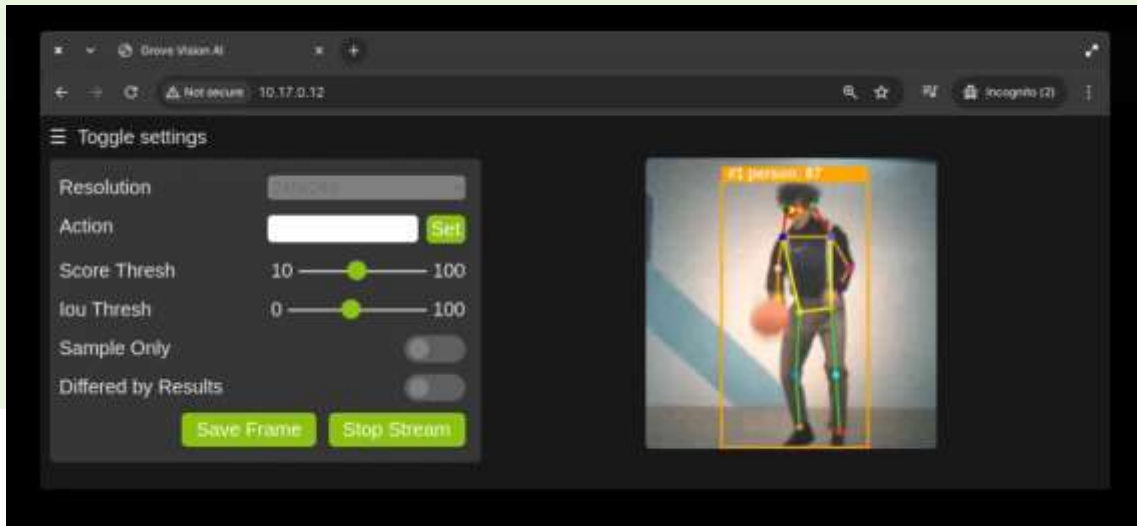
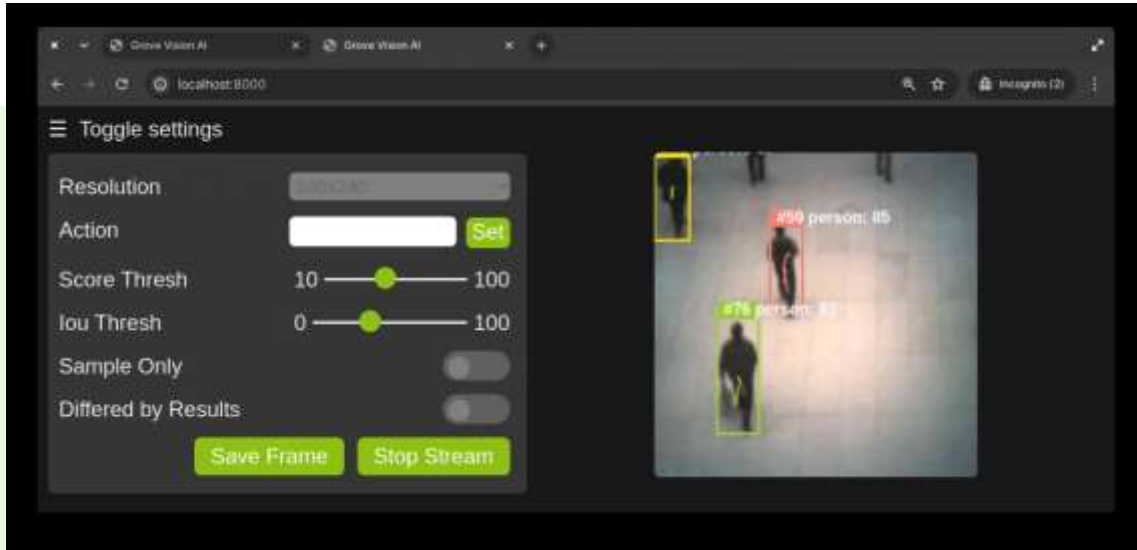
# Easier - No code model MLops

SenseCraft Model Assistant is an AI platform dedicated to simplifying the training, distribution, and deployment of AI models. With just a few clicks, you can easily deploy models and say goodbye to tedious configuration and coding. It supports users to upload and share self-trained models, build a shared model library, and promote collaboration and innovation among AI enthusiasts. Currently supports computer vision algorithms (such as target detection, image classification, image segmentation, and pose ) and LLM, making it possible to realize high-speed and accurate inference on low-cost hardware, unlocking the powerful potential of AI in edge devices.





# Easier - No code model MLops



[https://github.com/Seed-Studio/Seed\\_Arduino\\_SSCMA/tree/main/examples/camera\\_web\\_server](https://github.com/Seed-Studio/Seed_Arduino_SSCMA/tree/main/examples/camera_web_server)

# Easier - chat with TinyML

The image shows a Telegram chat interface with a bot named "Vision AI V2". The chat history displays a series of eight identical messages: "Alert: Human presence detected. Check the area." Each message is timestamped at 12:28 PM or 12:29 PM. The messages are grouped under "Today" and "Unread Messages".

The "Bot Info" panel on the right shows the bot's profile picture, which is a hand holding a small electronic device (likely a Raspberry Pi Zero) with a camera module attached. The bot's name is "Vision AI V2" and its username is "@RoomCamBot". The "Notifications" toggle is turned on. The "Media" section is currently empty, showing "No media files yet".

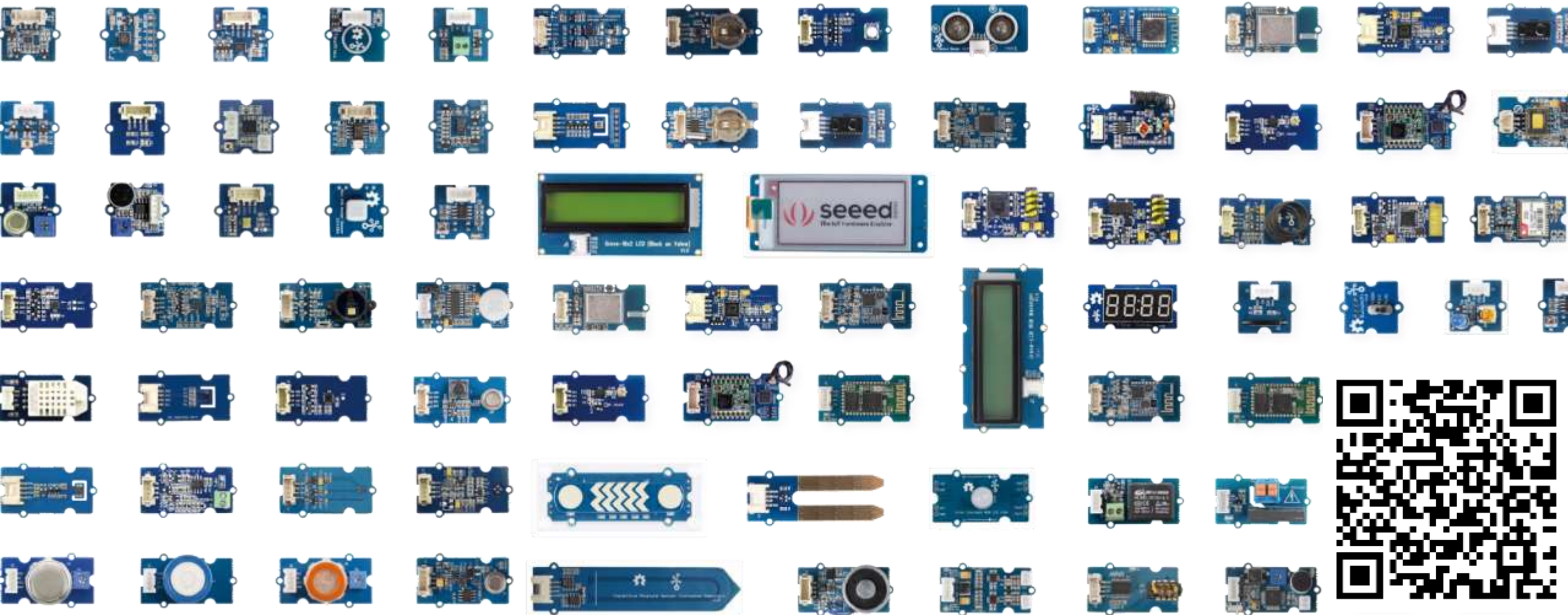
At the bottom of the chat, there is a text input field with the placeholder "Message" and icons for emojis, attachments, and voice recording.

# Atlas of Sensors

Nerve endings of the digital world

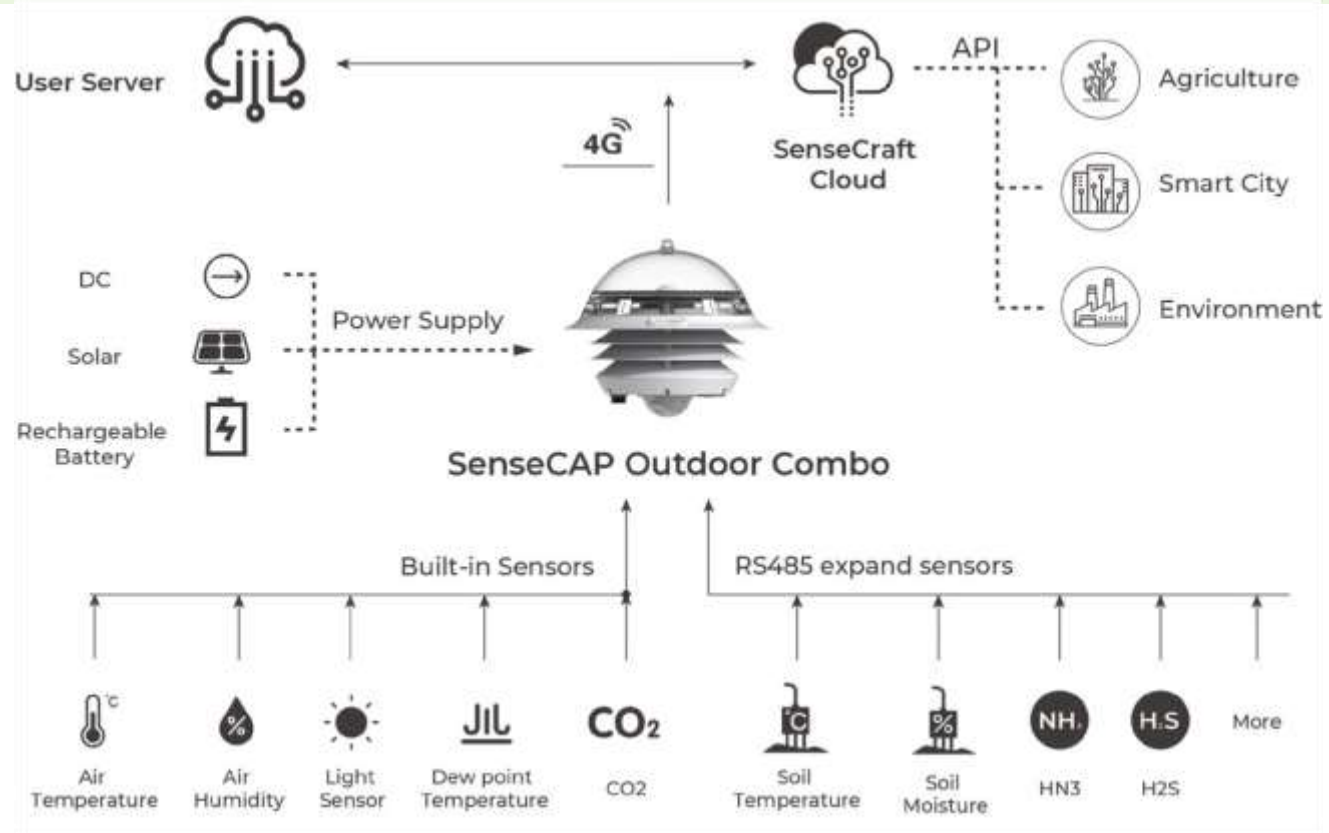
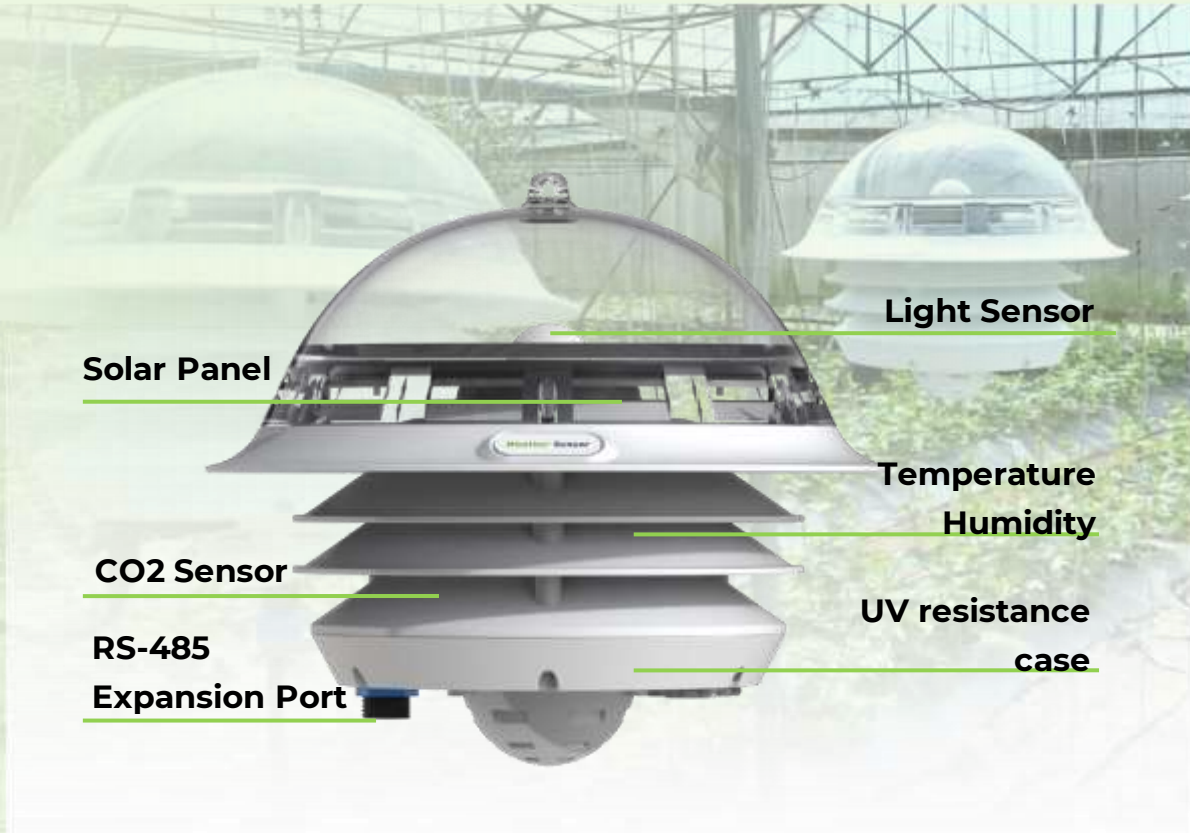


CO<sub>2</sub>

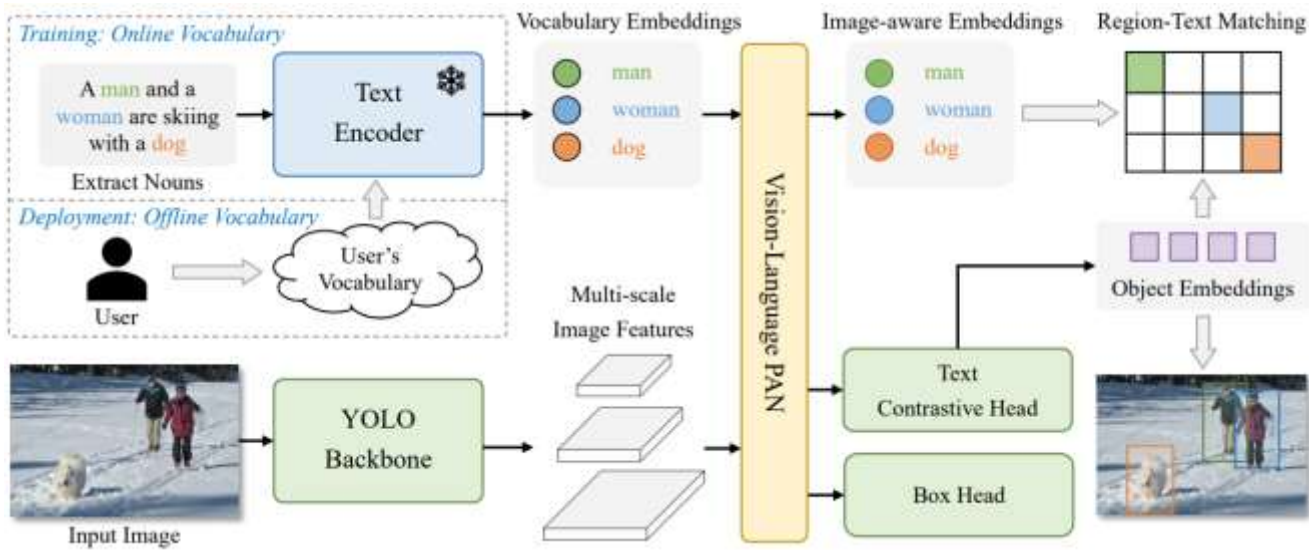


Scan code to download

# Multi-modal



# what's next: few shot training?



# Generative AI expand to real world

**multi-modal**

image  
Speech  
Sensors

**Stronger**

> 40T

**Distributed**

Locally deployed

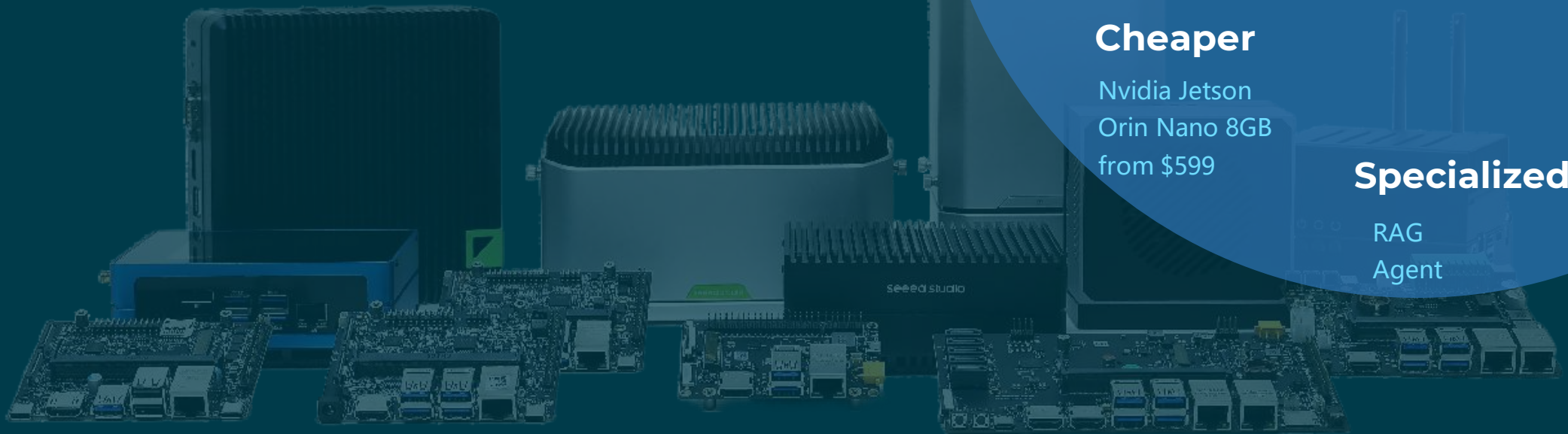
# Generative AI

**Cheaper**

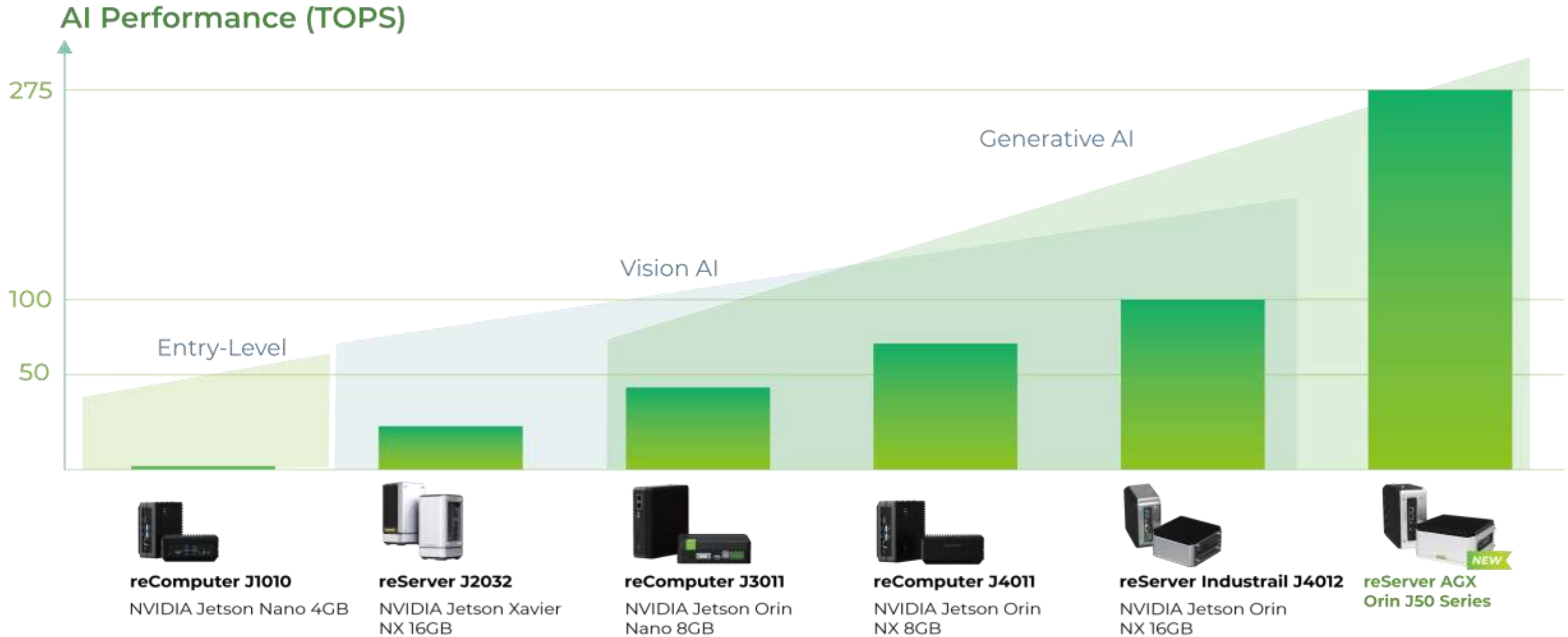
Nvidia Jetson  
Orin Nano 8GB  
from \$599

**Specialized**

RAG  
Agent



# Evolving edge AI capability



**reComputer J1010**  
NVIDIA Jetson Nano 4GB



**reServer J2032**  
NVIDIA Jetson Xavier NX 16GB



**reComputer J3011**  
NVIDIA Jetson Orin Nano 8GB

**From \$599**



**reComputer J4011**  
NVIDIA Jetson Orin NX 8GB



**reServer Industrial J4012**  
NVIDIA Jetson Orin NX 16GB



**reServer AGX Orin J50 Series**

**NEW**

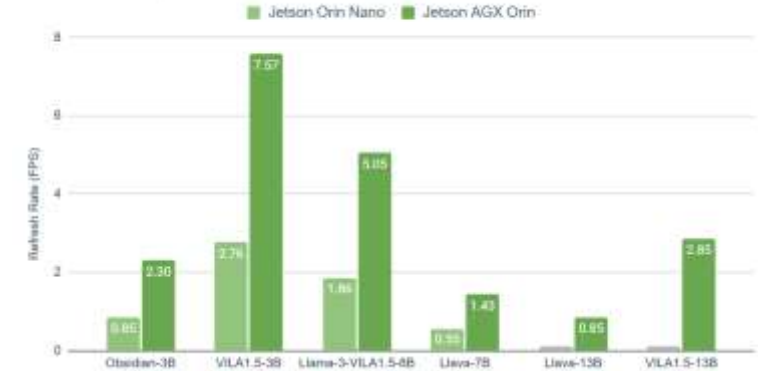
# Generative AI at the Edge

Describe the image concisely.



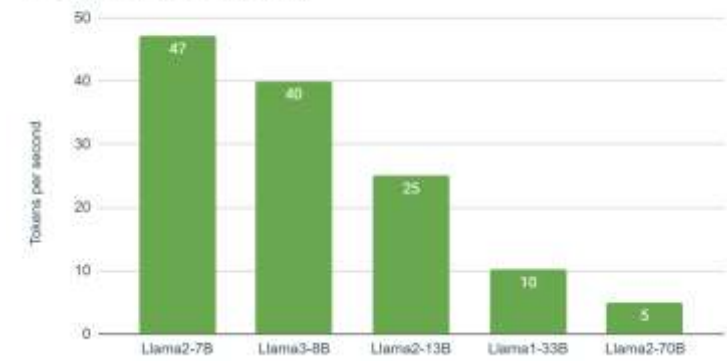
Multimodal Streaming Rate

Vision Encoder + Projector + VLM



LLM Text Generation Rate

Jetson AGX Orin, 4-bit Quantization

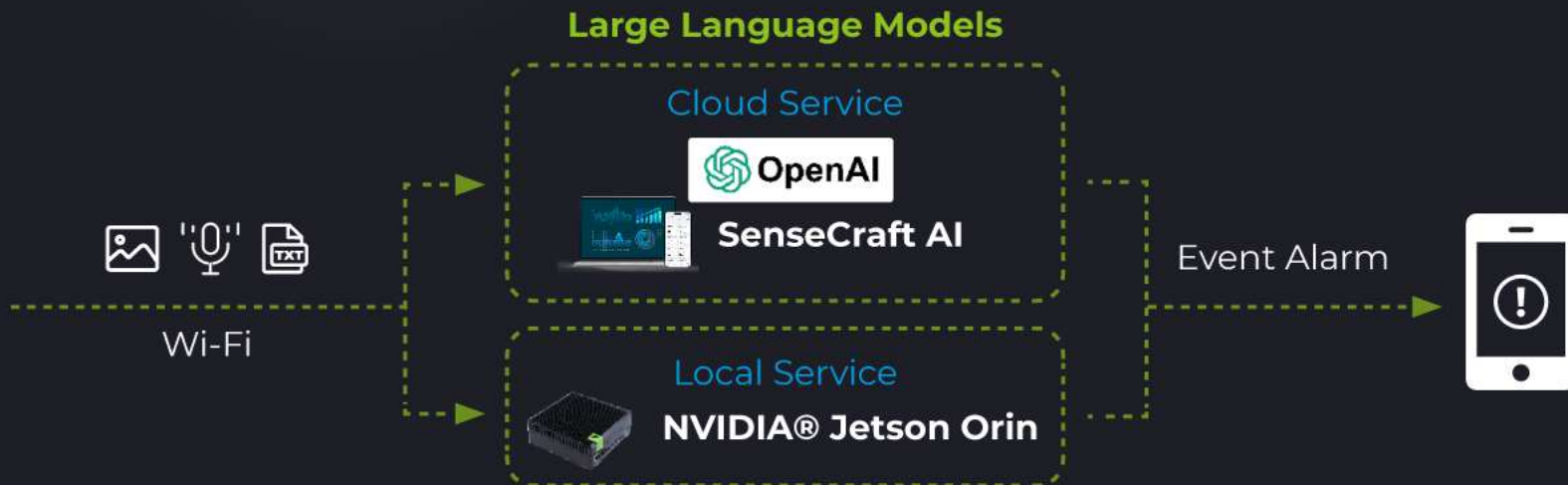




# Cost comparison of Open AI vs Local LLM

## # 5 Years Usage of Solution

	Times/day	Times for 5years	Jetson Solution Cost	OpenAI Solution Cost/year	Number of months to break-even
(24h) 1 request per day	1	1825	899	1.825	5993.33
(24h) 1 request per hour	24	43800	899	43.8	249.72
(24h) 1 request per 30min	48	87600	899	87.6	124.86
(24h) 1 request per 15min	99	180675	899	<b>903.4</b>	60.83
(24h) 1 request per 60s	1440	2628000	899	2628	4.16
(24h) 1 request per 10s	14400	26280000	899	26280	0.42



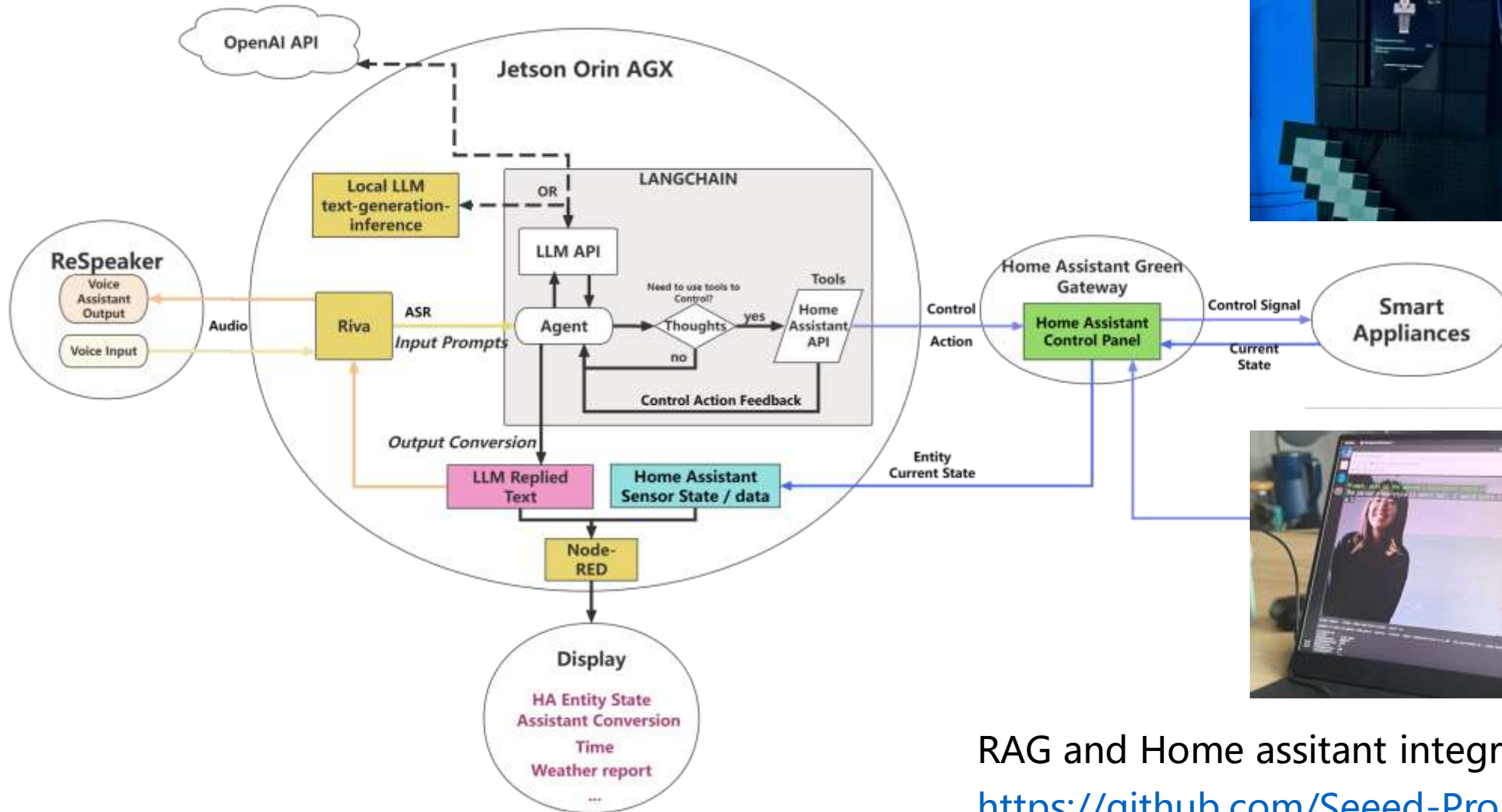
## Cons:

- Privacy
- High demand, lower cost with local LLM
- Local LLM speed is network-independent

## Pros:

- Low demand, lower cost with OpenAI
- OpenAI offers high accuracy

# Next generation human machine interface to complex system



RAG and Home assistant integration

<https://github.com/Seed-Projects/LocalJARVIS>

# Convergence of TinyML and Generative AI

## multi-modal

vision  
sound  
speech  
sensors

## Faster

new architecture  
(Cortex-M55)

## low power

<1 w

## Easier

few shot  
training  
no-code  
web server

# TinyML

## Cheaper

<10 \$

## Mixed Reality

## next gen HMI

# Embodied AI

## Droids

## Autonomous Machines

## multi-modal

imagine  
Speech  
Sensors

## Stronger

> 40T

## Distributed

Locally  
deployed

## Cheaper

Nvidia Jetson  
Orin Nano 8GB  
from \$599

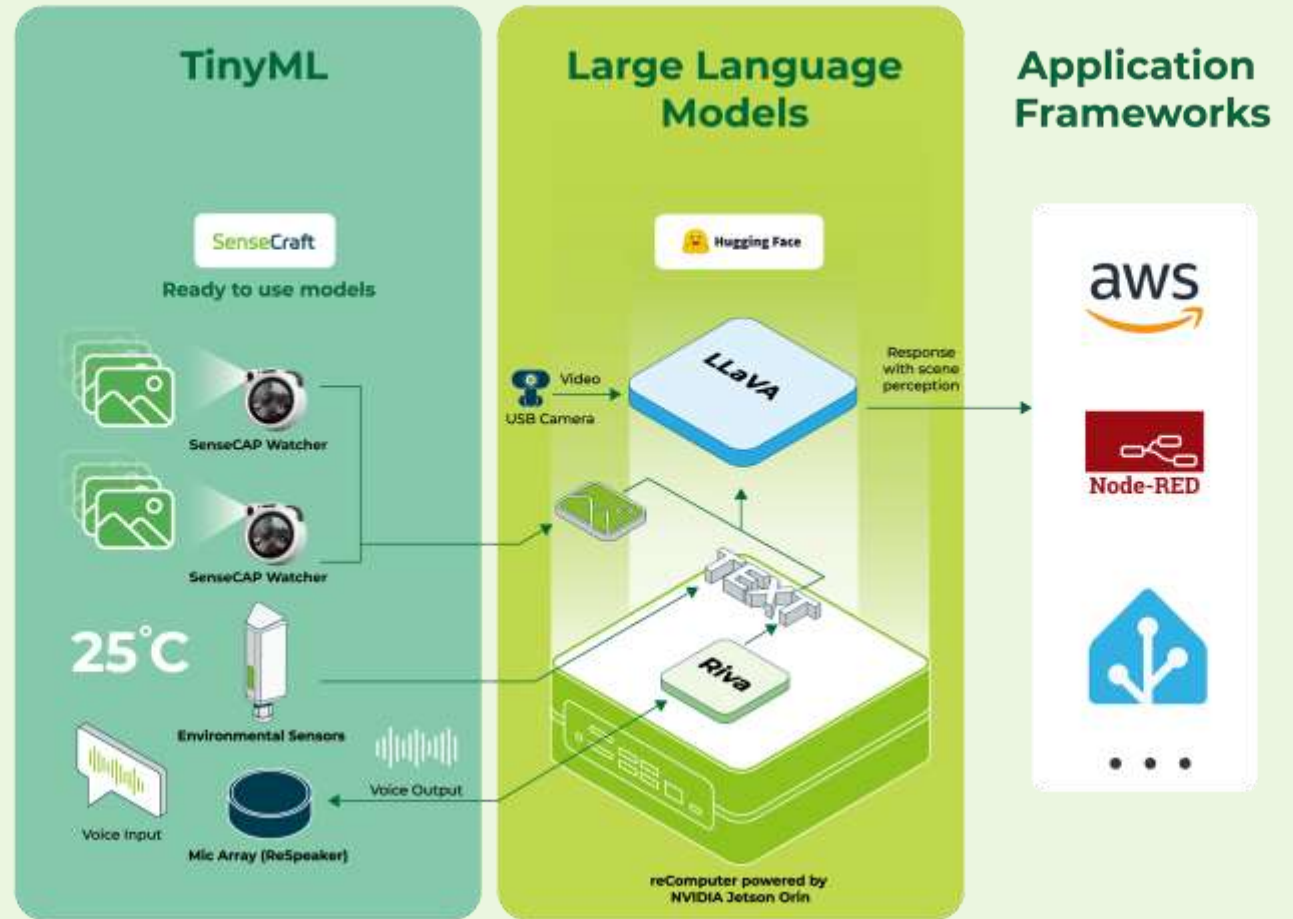
## Specialized

RAG  
Agent

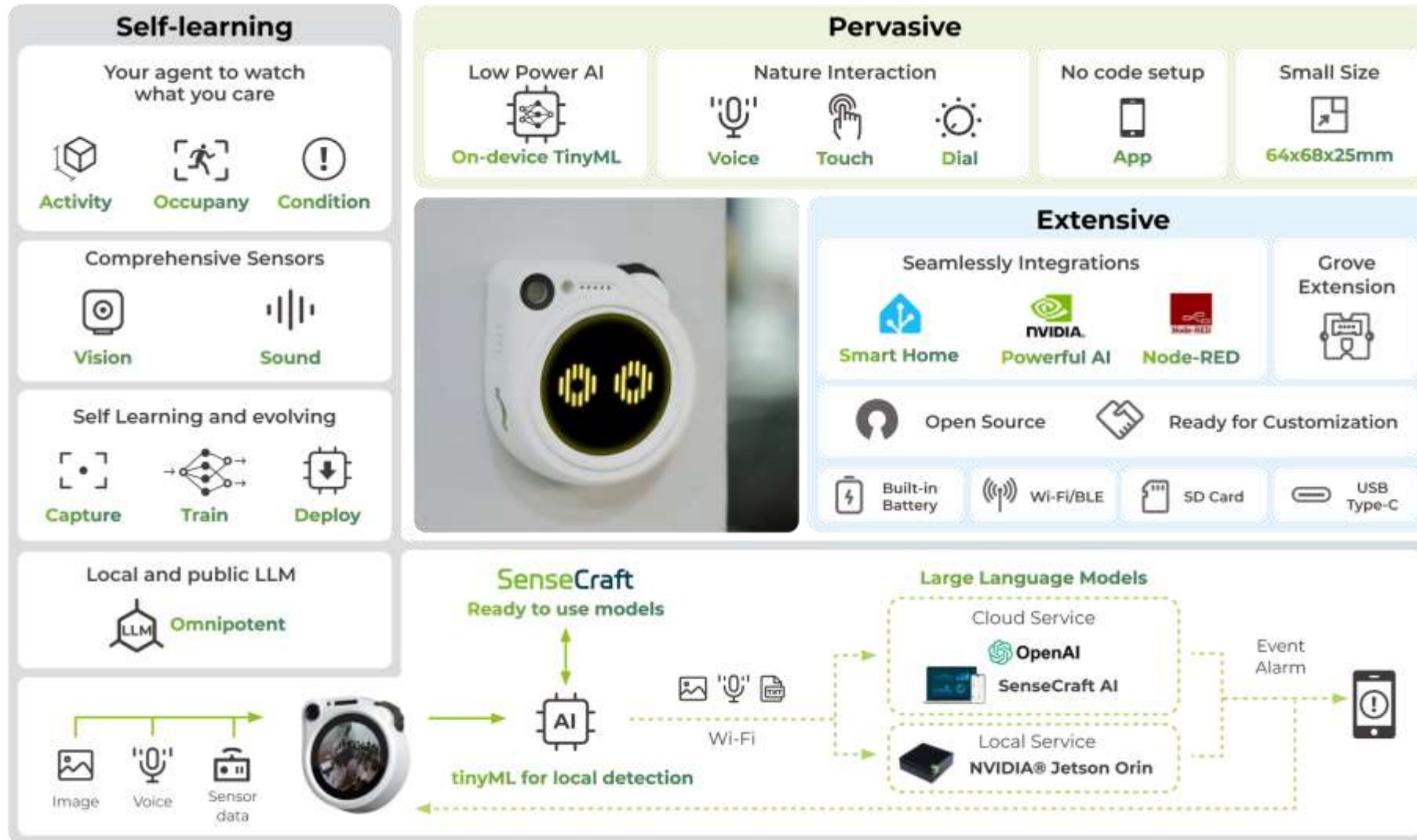
# Generative AI

# Convergence of TinyML and Generative AI

rhapsody of TinyML  
and local LLM

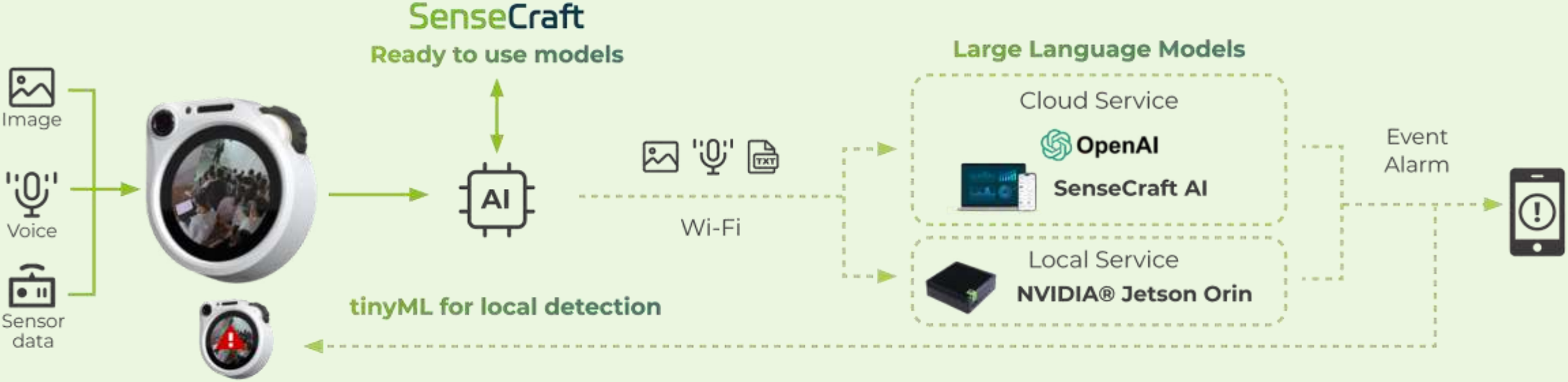


# Physical AI agent for smarter facility and machines

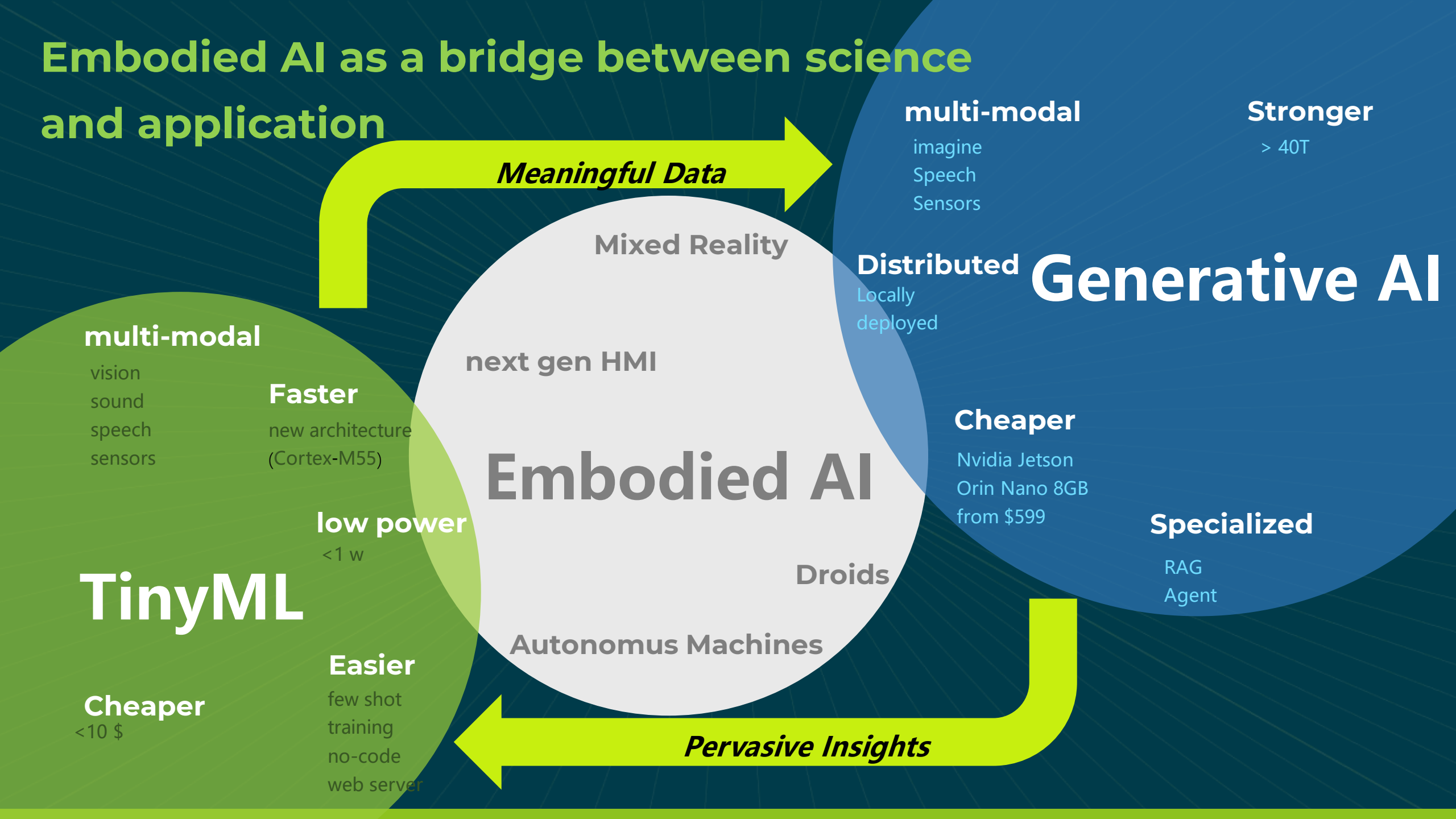




# Physical AI agent for smarter facility and machines



# Embodied AI as a bridge between science and application



## multi-modal

vision  
sound  
speech  
sensors

## Faster

new architecture  
(Cortex-M55)

## low power

<1 w

# TinyML

## Cheaper

<10 \$

## Easier

few shot  
training  
no-code  
web server

*Meaningful Data*

Mixed Reality

next gen HMI

# Embodied AI

Droids

Autonomus Machines

## multi-modal

imagine  
Speech  
Sensors

## Stronger

> 40T

## Distributed

Locally  
deployed

# Generative AI

## Cheaper

Nvidia Jetson  
Orin Nano 8GB  
from \$599

## Specialized

RAG  
Agent

*Pervasive Insights*



# Embodied AI as a bridge between science and application

## IoT2wild Contest

Website: <https://www.hackster.io/contests/iotinthewild>

Winner announced at Hackster Impact Summit on October 11, 2022

seed studio | hackster.io | Impact Summit

### Winner Announcement: IoT Into the Wild Contest for Sustainable Planet 2022

October 11th, 2022  
14:15P.M.-14:30P.M.  
Pacific Standard Time

**REGISTER NOW**

**Early Detection  
of Harmful  
Algae Bloom**

**Early flash flood warn system**

**GATE KEEPER**  
An IOT BASED ELEPHANT DETECTION SYSTEM

**Wild Animal Tracker**

**MonChan**  
Mountain Chain

Danger notification  
to save people!

**TO CHECK PLASTIC  
BOTTLE DUMPERS  
TO THE LAKES**

**BIOLIGHT**  
Harmful  
Algal Bloom  
Prediction and  
Monitoring

**NOMOS**  
NOMOS

**Landscape Pollution Alert**

**Monitoring Illegal Sand Mining**

**FIGHT FIRE**  
PREDICT MILD FIRE. RESPOND QUICKLY & SAVE NATURE

**Livestock / Wildlife Counting from  
Drone with FOMO algorithm**

**Black Soldier Fly Farming**

# Embodied AI as a bridge between science and application

## Co-Invent Solutions

Based on various digital transformation scenarios, we continue to develop smart devices that integrate the latest technologies, and work closely with developers and industry experts to provide software and hardware solutions for vertical industries at multiple levels.





**seeed studio**

The AI Hardware Partner

**Let's talk!**

**ep@seeed.cc**



LinkedIn  
@Seed Studio



Twitter  
@seeedstudio



Discord  
discord.seeed.cc



YouTube  
@Seed Studio



Project Hub  
hackster.io/seeed