Smarter technology for all

Scaling AI Responsibly

David Ellison | August 2024

Responsible AI Pillars

1

Diversity and Inclusion

2

Privacy and Security

3

Accountability and Reliability

4

Explainability

5

Transparency

6

Environmental and Social Impact

Use Cases

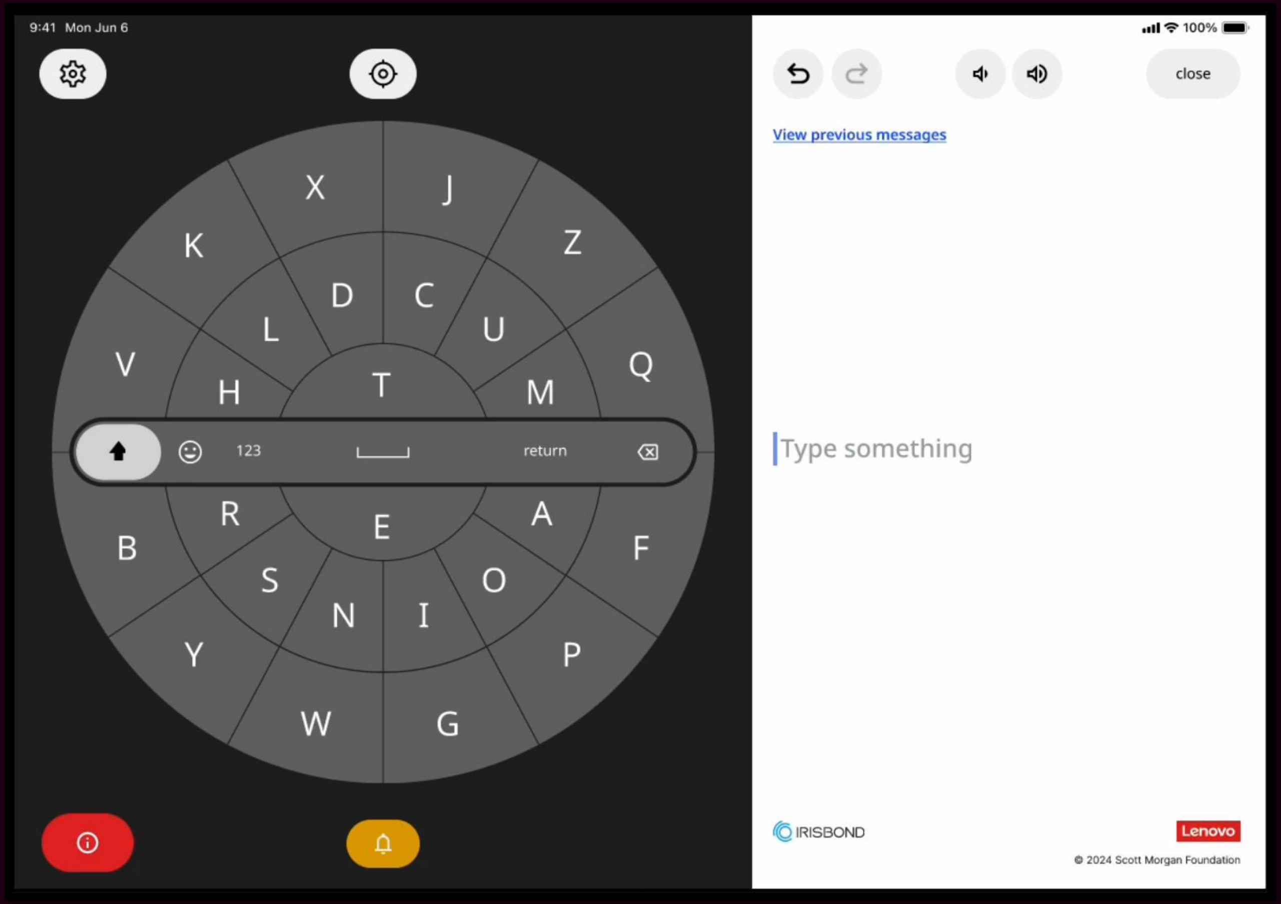
Look at exciting use cases Lenovo developed and assisted to scale AI responsibly

Examples

- Assistive Text Prediction with Personalized LLMs – D&I, E&SI
- Island Conservation – Sustainability with scaling at edge
- NASCAR SmartPitbox – Reliability to deliver in all scenarios
- ElephasCare AI Patient Activity Recognition – Privacy protected data

Lenovo & the Scott-Morgan Foundation are building assistive technology for people with severe disabilities.

AI for Assistive Typing



View previous messages

Type something

© 2024 Scott Morgan Foundation

The circular keyboard, optimized for inputs through eye-gaze tracking, accelerates typing for people with severe disabilities.

It uses LLMs to suggest characters, words, and full responses.

Responsible AI at scale:

- Environmental Impact: reduce energy consumption through model compression and optimized inference
- Diversity & Inclusion: Train and deploy many personalized text suggestion LLMs

Environmental Impact

Model compression can reduce memory consumption by 2-4x

Continuous batching maximizes GPU capacity for varying length sequences

Paged attention minimizes cache waste, achieving 2-4x throughput

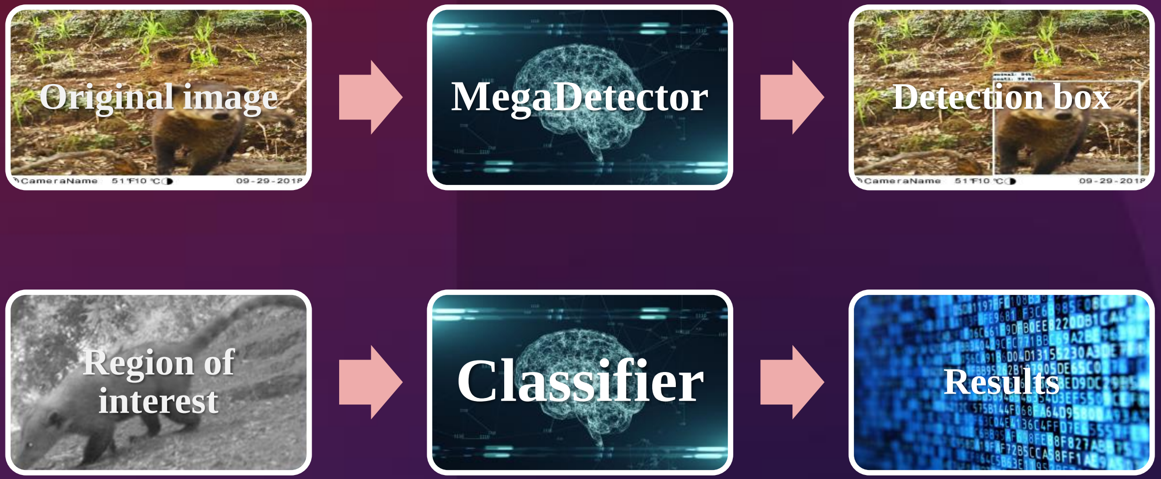
Diversity & Inclusion

Parameter-efficient fine-tuning (PEFT) enables cheap, customized LLMs

LLM personalization at scale by dynamically switching many fine-tuned models on top of a single foundation model

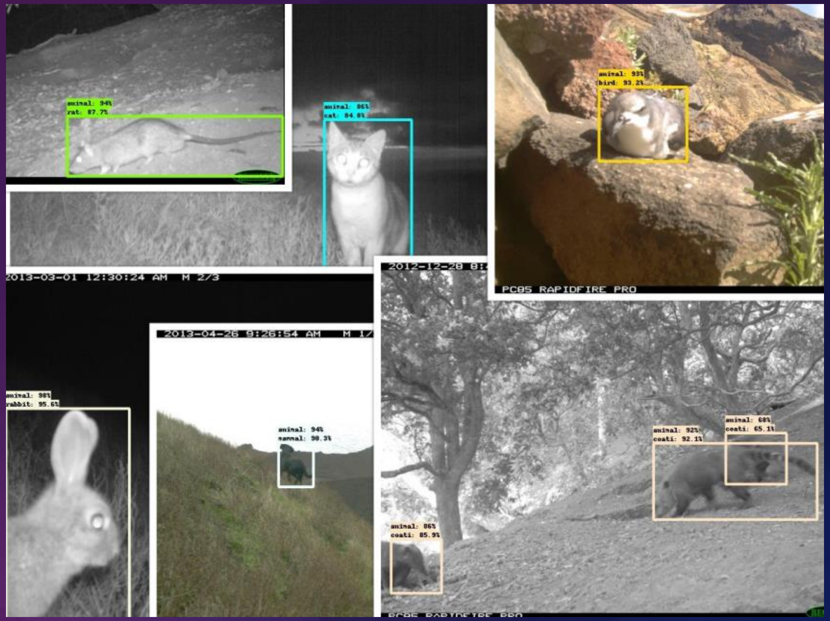
Island Conservation

- AI model to detect invasive species
- Scaled project to multiple islands
- Sustainable computation at the edge
- No need to transfer data through helicopters
- Reduced carbon emissions and quick action



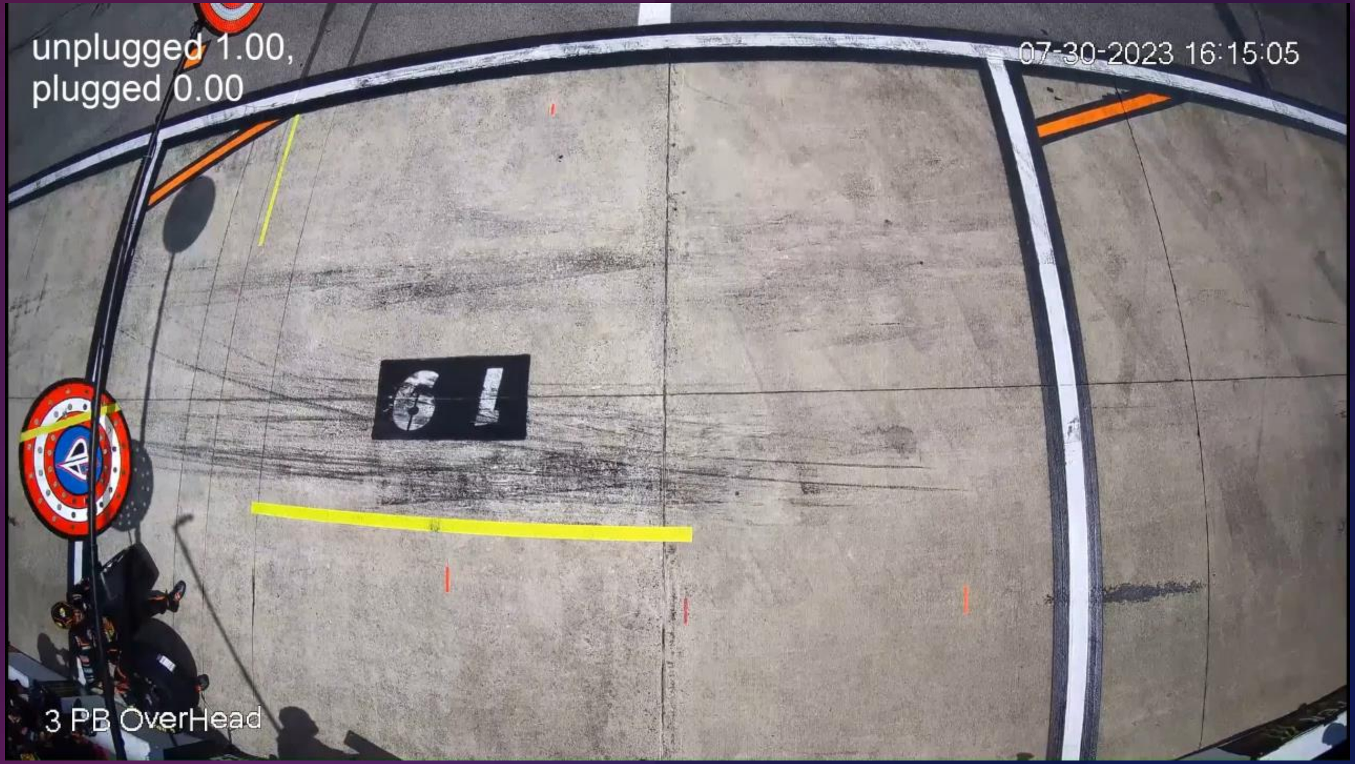
Original image → MegaDetector → Detection box

Region of interest → Classifier → Results



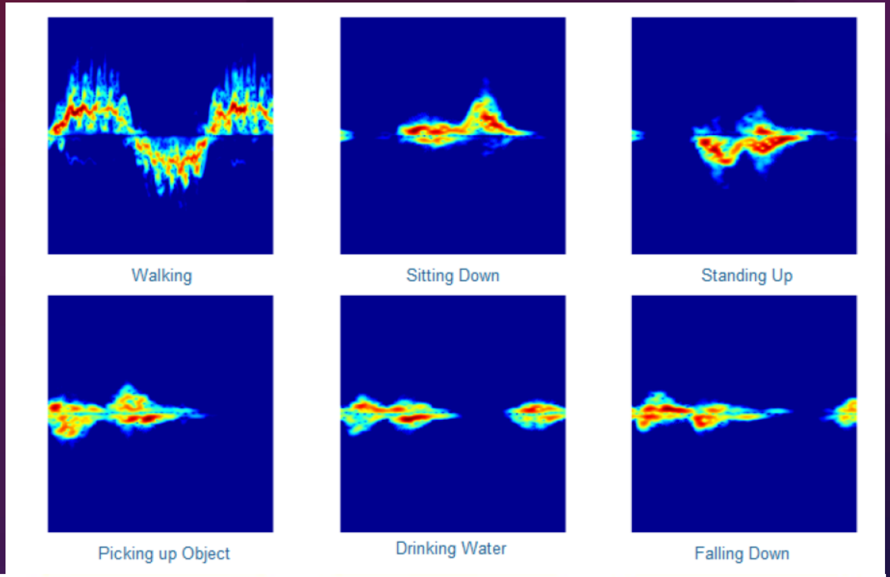
NASCAR SmartPitbox

- AI model to predict fuel level using plug duration
- Scalable solution to 48 Chevy member cars
- Reliable computation at the edge & consistent results
- Resource efficient, only runs when needed
- Explainable through heatmaps



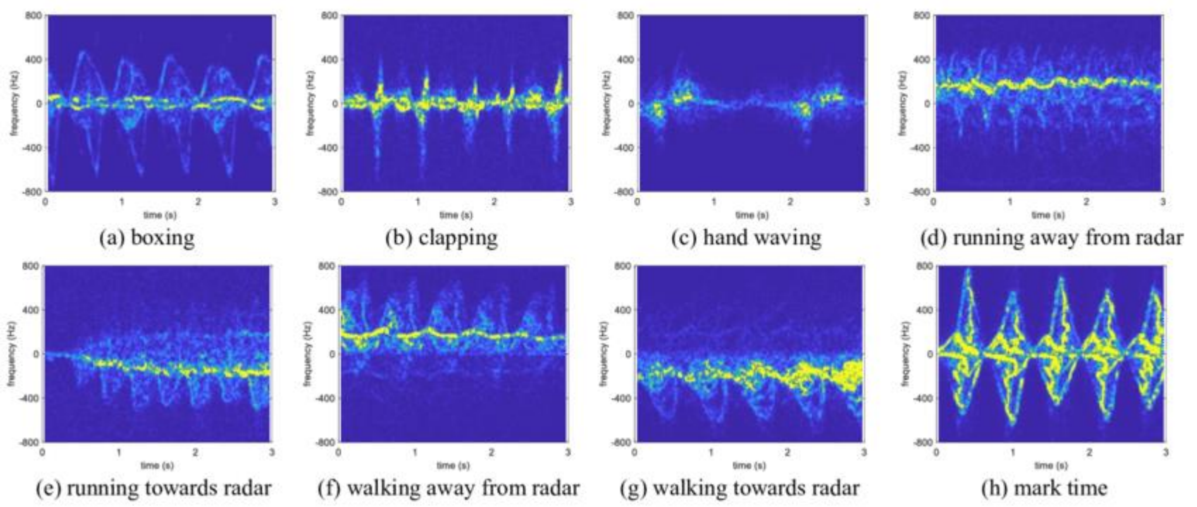
ElephasCare – Patient Activity

- RADAR based patient activity recognition
- Scaled to 150 rooms in a single facility. Used across multiple facilities
- Private tracking, no cameras / videos used
- Resource efficient running 24/7 – 450 sensors per device



Walking · Sitting Down · Standing Up

Picking up Object · Drinking Water · Falling Down



(a) boxing · (b) clapping · (c) hand waving · (d) running away from radar

(e) running towards radar · (f) walking away from radar · (g) walking towards radar · (h) mark time

Smarter
technology
for all

Lenovo

thanks.