On-Device AI Computing Platforms

for Robotics and In-Cabin Applications

Multimodal Perception and Real-Time Inference



Dongchao Xu

Why Do Robots Still Fail on the Factory Floor?

For the same job all day, a standard industrial arm can be more reliable.

Grip	Can it hold the part without slipping?
Alignment	Can it fit the part in the right place?
Cycle time	Can it keep up with the line all shift?
Recovery	Can it restart after a part gets stuck?
Safety and integration	Can it work safely with other machines?

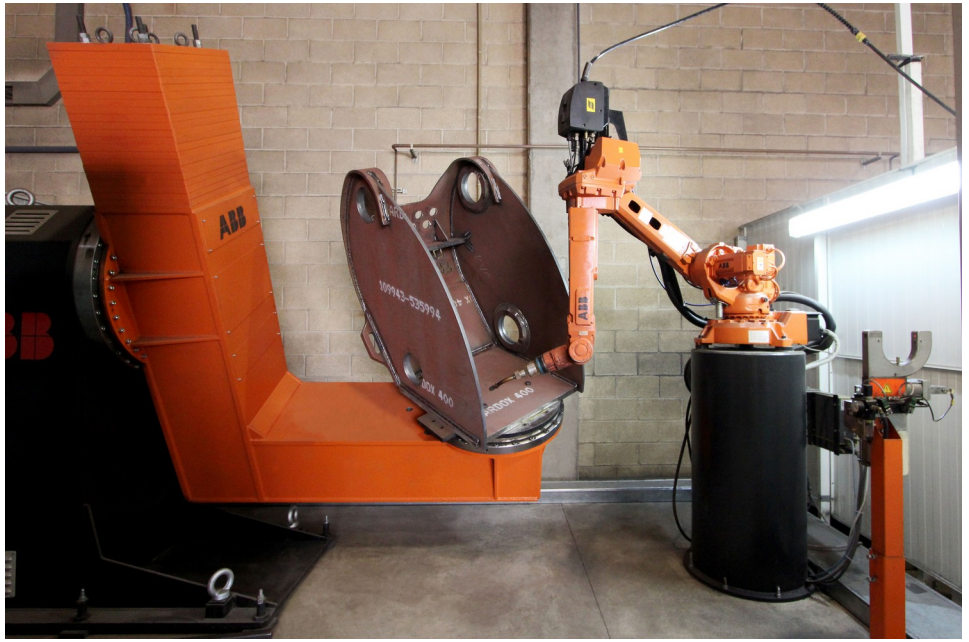


ABB welding cell · Ana 2016 / Wikimedia Commons / CC BY-SA 4.0

A 99.9999% success target means about one failure in a million tries.

Example target for a job that can stop the line. The real target depends on the cost of a failure.

If Driving Took Years, What About Robots?

Self-driving took years, even with roads, lanes and traffic rules.

Driving on a road

Stay on the road and avoid other vehicles and obstacles.

No vehicle finished the 2004 DARPA challenge. In 2007, cars had to drive through a test town with traffic.

Handling objects in 3D

Move the arm and gripper together. Adjust the force as the robot touches the part.

The part, its position and the setup can all change how the robot needs to move.

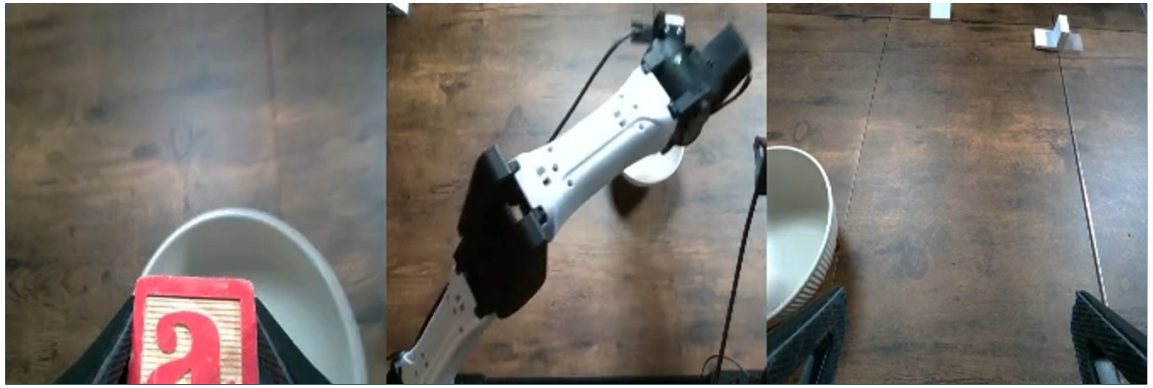
The long road to self-driving made many engineers doubt robots could handle unfamiliar objects.

Sources DARPA Grand Challenge history and Autonomous Robotic Manipulation program

GPT-6 Astra on Two Robot Tasks

Real robot arms. Same controls. 20 tries per task.

Place a block in a bowl

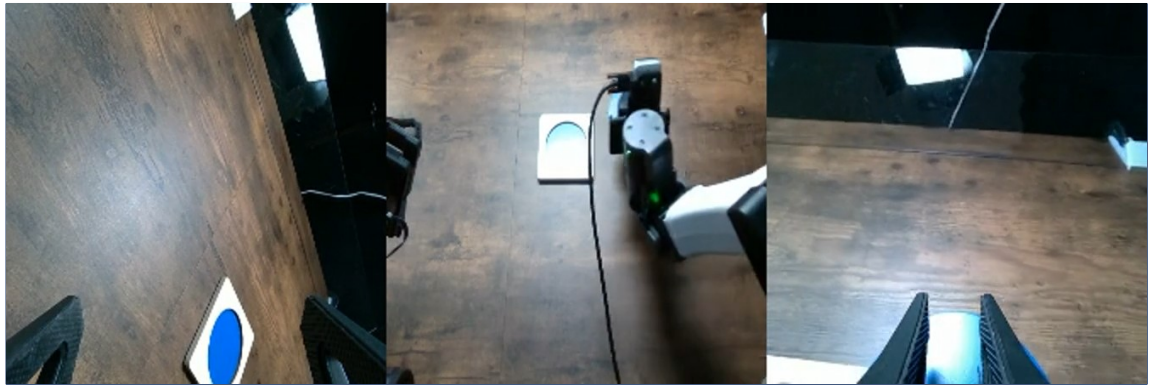


Frame from a recorded trial

19 / 20

Successful trials

Fit a puzzle piece into a groove



Frame from a recorded trial

2 / 20

Successful trials

GPT-6 showed that a general AI model can turn what it sees and reads into real robot actions.

Source Robocurve, GPT-6 Astra on robotic manipulation
https://openai.robocurve.org/gpt-6-astra/

Light-O1 Learns by Watching People

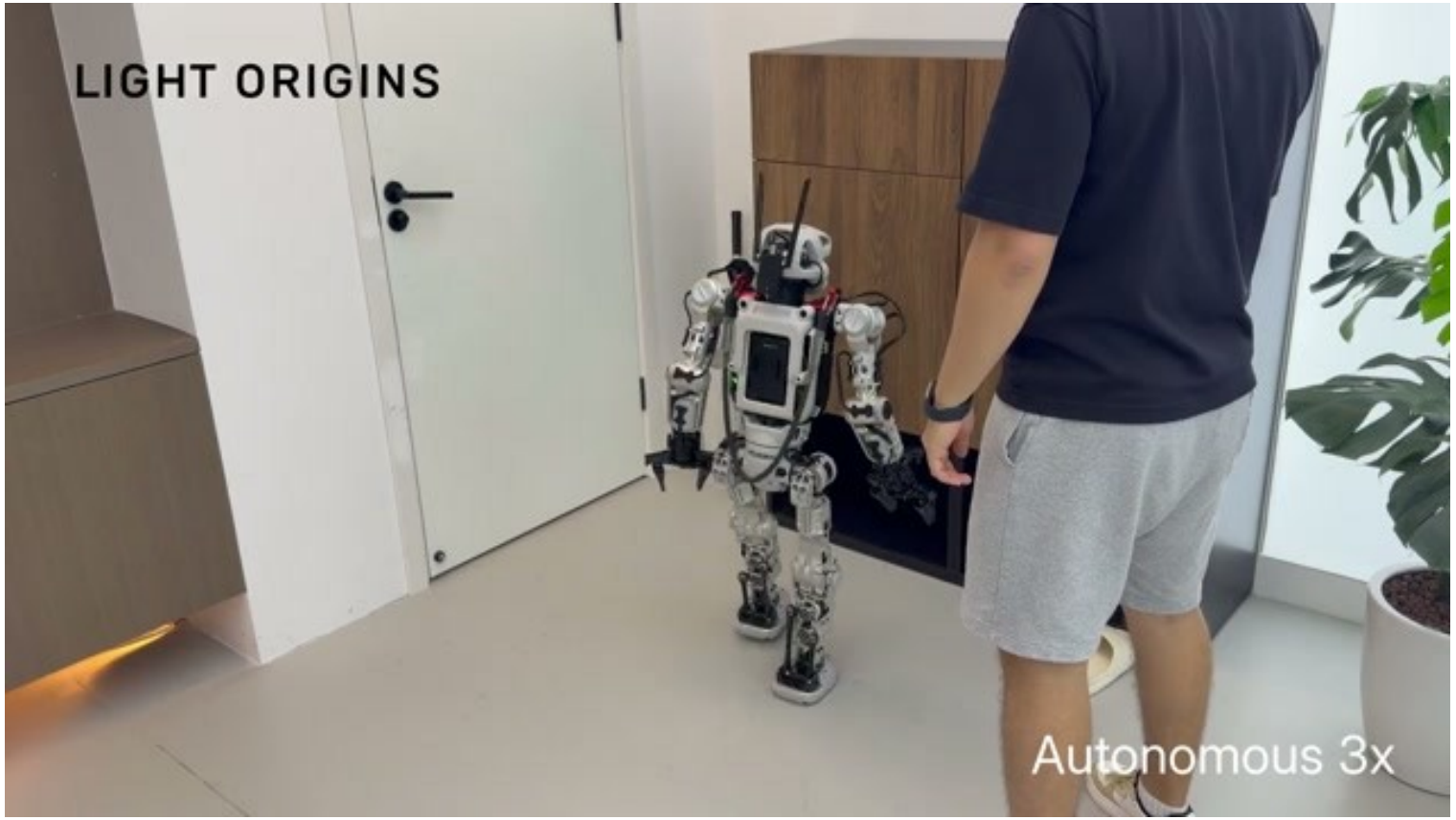
LightBot demo. Open the cabinet and put away the slippers.

100,000

hours of human movement

The team first used videos from the internet, then used data from real robots.

When the door closes, LightBot opens it again and keeps putting away the slippers.



Frame from the LightBot shoe cabinet demo

Source Light Origins, Light-O1 technical report, Figure 06
https://www.lightorigins.cn/blog/light-o1

Robot tests show that AI can turn what it sees into action.

In a car, action starts with knowing who spoke and what they need.

Why Does In-Cabin AI Get Simple Requests Wrong?

“Intelligent cockpits have gone end-to-end on language but not on perception.”

It understands the words

One speech model can hear a request, understand it and reply.

“sensor data stays in silos”

Cameras, radar and vehicle systems each see part of what is happening. Their data often stays separate.

The car needs to know who is inside and what is happening around them.

Defining Cabin Context: A Perception Framework for End-to-End Intelligent Cockpits
https://donfrontera.com/articles/defining-cabin-context/

Four Ways In-Cabin AI Gets It Wrong

PAIN POINT	SCENARIO	MISSING CONTEXT
“Listens but does not see”	A passenger is shaking from the cold. The car waits for her to ask for heat.	How the person feels physically
Wrong person	A passenger in the back asks for music. The car plays the owner’s playlist.	Who spoke and where they sit
Bad timing	The assistant starts a long reply just as the driver joins heavy traffic.	What the driver and car are doing
“Ignores the audience”	The car reads out a private job interview reminder while the user’s boss is in the car.	Who is in the car together

The car needs to know who to help, when to speak and what it can share.

Defining Cabin Context: A Perception Framework for End-to-End Intelligent Cockpits
https://donfrontera.com/articles/defining-cabin-context/

Cabin Context in Three Layers and Twelve Dimensions

The three layers help the car decide what to do, who to help and how to reply.

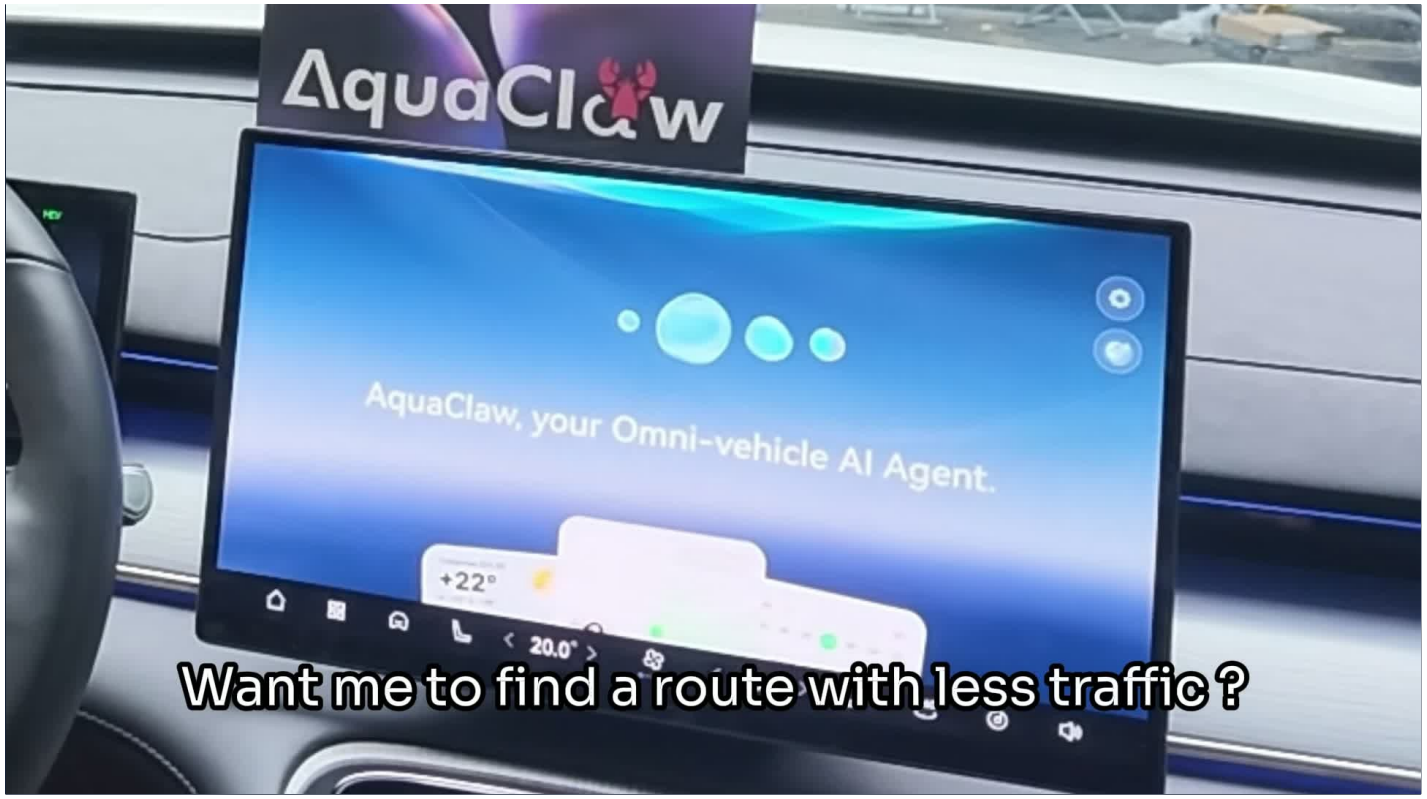
Occupant D1–D7	Who they are · Seat and role · Physical state · Mental state Mood · Activity · Past behavior and preferences
Relationship D8	How people know and interact with each other Who is talking to whom and who should get the reply
Shared D9–D12	Car and driving status · What is happening and where we are in the trip Surroundings · Trip and task

For each estimate, track how we know it, how sure we are and when to check again.

Defining Cabin Context: A Perception Framework for End-to-End Intelligent Cockpits
https://donfrontera.com/articles/defining-cabin-context/

ThunderSoft AquaClaw in Action

How tone of voice, memory and the current trip shape a navigation response



AquaClaw offers to find a route with less traffic.

Occupant · D5 and D7

Tone hints at stress (D5). Memory recalls a past event and the user's tastes (D7).

Relationship · D8

The people in the car talk to each other and to AquaClaw.

Shared · D10 and D12

Live traffic (D10) and the search for a place to go (D12).

Judgment and action

It suggests two places that fit the request and the traffic. It sets the route after the user chooses.

Demo ThunderSoft AquaClaw, Agentic Travel in Action

From GPT-6 to Perception, Language and Action

The Astra robot tests give us a way to think about end-to-end AI in robots and cars.

	PERCEPTION	→ LANGUAGE REASONING	→ ACTION
Robotics 	Vision, depth, joints and force	Find the part and decide how to pick it up	Move, pick up the part and check where it lands
In-cabin AI 	Speech, cabin cameras, occupied seats and vehicle data	Work out who spoke, what they need and when to reply	Reply or make an allowed change in the cabin

Connect what the system senses to what it does. Check safety and confirm the result.

Robots and cars have to keep sensing while they decide.

Extra reasoning helps only if the answer arrives in time to act.

Inference-Time Scaling on the Device

Spend additional compute where it can improve a decision before its deadline.

Routine decision	Use fresh context and a lightweight model for the common path.
Ambiguous decision	Acquire another observation, run a stronger model or compare candidate actions.
Time-critical response	Protect the feedback loop. Use a bounded action or fallback if inference runs out of time.

Measure task success and deadline misses at each inference budget.

On-Device Hardware Requirements

Handle sensor streams and peak AI workloads together, within power and response limits.

Sensor throughput	Measure ingest, synchronization, frame drops and the age of observations at decision time.
Memory and acceleration	Measure data movement, working-set size and sustained multi-model inference.
Power and thermals	Run the workload to thermal steady state. Measure how much performance the device sustains.
Response deadlines	Measure P99 sensor-to-action latency and protect reserved capacity for critical work.

Run the representative product workload before committing to a platform.

On-Device Software Requirements

Keep sensor data fresh, critical work on time and actions within safety limits.

Zero-copy data path	Keep perception current by sharing large buffers where supported. Track timestamps, ownership and buffer lifetime.
Deterministic scheduling	Reserve resources for time-critical tasks. Bound queues and isolate bursts of model inference.
Independent safety gate	Check permissions, device state and limits before executing an action. Apply a bounded fallback on timeout.
End-to-end observability	Trace sensor age, inference queues, safety decisions and verified outcomes across the loop.

The runtime must keep context fresh, meet critical deadlines and verify what the device did.

Thor and IQ10 for On-Device AI

Specifications for two robotics configurations

Metric	NVIDIA Jetson AGX Thor (T5000)	Qualcomm Dragonwing IQ10 Robotics Reference Design (IQ-10800)
Peak AI compute (INT8 TOPS)	Up to 1,035 (sparse, GPU)	Up to 700 (sparse, NPU)
Memory capacity and type	128 GB LPDDR5X	64 GB LPDDR5X
Peak memory bandwidth	273 GB/s	273 GB/s
Camera inputs	Up to 20 via Holoscan Sensor Bridge	Up to 12 via GMSL2

Measure sustained inference and latency using the same model, precision and sensor load.

https://developer.nvidia.com/blog/introducing-nvidia-jetson-thor-the-ultimate-platform-for-physical-ai
https://www.qualcomm.com/internet-of-things/products/iq10-series/iq-10800
https://docs.qualcomm.com/doc/87-A0789-1/87-A0789-1_REV_A_Qualcomm_Dragonwing_IQ10_Robotics_Reference_Design_Product_Brief.pdf

Li Auto MACH M100 vs NVIDIA Thor-U

Both benchmark configurations list 273 GB/s of peak DDR bandwidth.

Workload	Inference phase	MACH M100	Thor-U	M100 speedup
LLaMA2-7B	Prefill (W8A8)	79 ms	154 ms	**1.95×**
LLaMA2-7B	Decode (W4A16)	21.34 ms	20 ms	**0.94×**
MindVLA language component	Prefill	0.84 ms	1.74 ms	**2.1×**
MindVLA language component	Decode	0.10 ms	0.30 ms	**3.0×**

LLaMA2-7B used a 1,024-token input. MindVLA figures cover the language component only.

Prefill processes the input. Decode generates one token at a time.

Benchmark prefill and decode separately. The speedup depends on the workload.

M100: An Orchestrated Dataflow Architecture Powering General AI Computing
https://arxiv.org/abs/2604.17862

Edge Compute Products



ThunderX AIBOX-N1 · Official product image

ThunderX for vehicles

AIBOX-N1 uses DRIVE AGX Orin. Published specs are 200 TOPS and 205 GB/s.

AIBOX N2 · Thor-based · Upcoming release

Thundercomm for robots

TurboX IRB10 is an IQ10-based robotics development and reference platform.

Size the product with its sustained workload and response deadlines.

Sources ThunderX AIBOX-N1 and ThunderSoft TurboX IRB10 product announcements

The chip gives the device room to run several models at once.

Software keeps critical work on time and checks each action.

One On-Device Software Architecture

Robotics and in-cabin AI share the same loop, with different sensors and device interfaces.

Domain inputs Robotics · Vision, depth, joints and force In-cabin · Speech, vision, occupancy and vehicle state

Perception

Fuse time-aligned sensor data. Build scene or cabin context.

Language reasoning

Use local models and the agent runtime to interpret the goal and propose an action.

Governed action

An independent safety gate checks the request. The device controller or service executes it.

Device interfaces Robotics · Motion, navigation and gripper control In-cabin · Climate, navigation and cabin services

Outcome feedback updates context and informs the next action

Shared runtime services Zero-copy data path Deterministic scheduling for critical tasks End-to-end observability

Governed Action and Outcome Feedback

The runtime checks each proposed action and verifies the result using device feedback.

	Propose	Check and execute	Verify
	The model and agent runtime request a task-level action.	An independent gate checks permission and current state. The controller or service enforces limits.	Compare device feedback with the intended outcome. Update context or recover.
Robotics	Pick up the tray at Station B.	Check workspace and force limits, then call the grasp controller.	Verify the lift from pose and gripper feedback.
In-cabin AI	Raise the rear zone’s setpoint.	Check the passenger’s zone and permissions, then call the climate service.	Read back the zone’s new setpoint.

Critical control keeps its deadlines even when model inference is delayed.

AquaDrive AIOS for Robotics and In-Cabin AI

ThunderSoft’s on-device software platform for both domains

Shared AquaDrive AIOS capabilities

Perception

Multimodal data and context

Language reasoning

Local model and agent runtime

Governed action

Skill invocation, policy and feedback

Robotics interfaces

Vision, depth, joint and force signals

Navigation, grasping and motion control

In-cabin interfaces

Speech, cabin vision, occupancy and vehicle state

Navigation, climate, media and vehicle services

Each product integrates its own sensors, skills, control interfaces and safety limits.

On-Device Physical AI

1. Size compute for continuous sensing and peak inference
2. Keep context current and actions within their deadlines
3. Verify the outcome and recover when the scene changes

Test the full loop under realistic sensor load and response deadlines.

Personal Website

Research and reflections

www.donfrontera.com

Thank You



