LANDSCAPE · APRIL 2026

The Edge & On-Device AI Landscape

A snapshot of the companies building AI that runs outside the cloud — from silicon to applications.

READ THIS FIRST

Edge ≠ On-Device

EDGE

The broader term. **Anywhere AI runs outside a centralised cloud data centre** — factory-floor servers, on-prem appliances, base stations, and physical devices. Edge computing is a deployment location, not a device class.

ON-DEVICE

A stricter subset. **AI that runs on the end device itself** — the phone in your hand, the camera on the wall, the sensor on the machine, the chip in the robot. No network round-trip, no external inference server.

The Landscape

01 EDGE HARDWARE — 7 players

Silicon, accelerators, and servers. The layer that defines what's physically possible to run at the edge.

- **EdgeScale AI** — Edge servers · on-prem
- **Hailo** — 26 TOPS @ 2.5W
- **Axelera AI** — Europa AIPU · 629 T…
- **SiMa.ai** — 50+ TOPS vision SoC
- **Ambarella** — Automotive · CCTV S…
- **NVIDIA Jetson** — Thor · robotics default
- **Qualcomm Drag…** — Mobile + robotics NPU

02 ON-DEVICE LLMS — 8 players

Three kinds of player: independent model labs, open-source model families from frontier labs, and the runtimes that optimise and ship them to devices.

MODEL PROVIDERS

Independent labs training their own edge-sized foundation models — a scarce category.

- **Liquid AI** — LFM / LFM2 · open w…

OPEN-SOURCE MODELS

Model families released by frontier labs — the raw material most on-device runtimes serve.

- **Llama (Meta)** — 3.2 1B/3B · default ba…
- **Gemma (Google)** — Nano · mobile-tuned
- **Qwen (Alibaba)** — Small VLMs · multilin…
- **Phi (Microsoft)** — Reasoning per param…
- **Mistral** — Ministral · EU open-w…

OPTIMISATION / RUNTIME

Engines that take open-source models and make them run fast on phones, NPUs, embedded hardware.

- **Cactus** — Mobile inference engi…
- **Nexa AI** — NPU-aware runtime

03 ON-DEVICE INFRASTRUCTURE — 6 players

The plumbing enterprises embed to ship AI into apps and devices. Two lanes — free platform-vendor toolkits vs. paid cross-platform applications.

PLATFORM-VENDOR TOOLKITS

Shipped by Apple / Google / Microsoft with their platforms. Free, generic, single-platform, subsidised for ecosystem lock-in.

- **Core ML** — Apple · iOS/macOS
- **ML Kit** — Google · drop-in APIs
- **MediaPipe** — Google · perception
- **LiteRT** — Google · TFLite succe…

CROSS-PLATFORM APPLICATIONS

Commercial, cross-platform, mobile-native enterprise AI SDKs — split by modality. The paid production-grade alternatives.

- **Captur** — Vision · mobile-native
- **NimbleEdge** — Voice + text · privacy-…

04 EDGE AI APPLICATIONS — 4 players

End products where edge AI *is* the product. Robotics, autonomy, embodied AI.

- **Figure AI** — Humanoids · BMW pro…
- **Boston Dynamics** — Atlas + Spot
- **Wayve** — L4 embodied driving
- **Nuro** — Autonomous delivery

05 ADJACENT INFRASTRUCTURE — 4 players

Well-funded infra companies in neighbouring slices — different buyer, form-factor, or workflow. Often conflated with the main buckets but worth showing separately.

- **Edge Impulse** — TinyML · sensors
- **Roboflow** — Vision training · cloud-…
- **Latent AI** — Drones · robotics · defe…
- **Wallaroo.AI** — Enterprise MLOps · sp…

AUTHOR'S NOTE

Full disclosure: I founded **Captur**. You'll find us in Bucket 3 under Cross-Platform Applications. I've tried to place us honestly and to make the map useful for anyone thinking about the space — not as a pitch. Disagreements, missing players, better framings welcome. Drop them in the comments.

METHODOLOGY

Curated April 2026 from public sources — company sites, recent funding, CES 2026 launches, analyst coverage. 29 players chosen as recognisable representatives of each bucket; not exhaustive. Companies placed in their *primary* bucket.

CAVEATS

- Device-edge, not CDN-edge.
- Open-source runtimes (llama.cpp, ONNX, ExecuTorch, MLX, TVM) are ecosystem infrastructure, not shown as player tiles.
- Platform vendors (Apple, Google, NVIDIA, Qualcomm) span multiple buckets in practice.