Xybrid: On-Device AI, At Scale.

The economics behind cloud AI is about to break.

Compute is scarce: even OpenAI and Anthropic are rationing

The companies with the deepest GPU allocations and the most capital are already cutting products, capping usage, and rewriting pricing to control demand.

- **Mar '26:** OpenAI shuts down Sora — $15M/day in inference against $2.1M lifetime revenue.
- **Apr '26:** Anthropic cuts third-party agents from Claude subscriptions; $200 plans were powering thousands of dollars of agent compute per user.
- **Apr '26:** Anthropic restructures enterprise billing — flat seats replaced by pre-committed spend.

If the frontier labs can't make the math work, no one downstream can.

Cloud inference punishes growth

The better your product does, the harder cloud math hits.

- **The math gets worse as you grow** — 1M MAU × 100 LLM = **$3M/month in spend**, at today's *subsidized* rates.
- **Agentic features multiply the bill 5–30×** — the more capable your product, the more tokens per task.
- **Premium UX needs sub-150ms latency** — cloud round-trips can't deliver it.

On-device is the answer — but on-device alone won't ship

- **Fragmented runtimes** — every platform speaks a different dialect.
- **Hardware-specific code paths** — each chip family wants different optimization.
- **Model rotation outpaces integration** — new SOTA every month, every model × platform × SDK is a fresh integration surface.
- **No graceful fallback** — devices fail, models OOM, batteries die. Without a cloud safety net, on-device means unreliable.

The Solution

Xybrid is the operating layer for on-device AI. Developers integrate once, Xybrid handles the rest.

Integration becomes an implementation detail.

One runtime and one stop API to run models locally on any device

- Run any model: LLM, Voice, Transcription, Vision
- Seamlessly route between device and cloud **cut inference costs by 80%**
- Supports new SOTA models on day one
- Eliminate engineering overhead

A Platform to operate at scale.

Observe, deploy, rollout, update in one click all from a fully-featured control-plane.

- **OTA model updates** — ship new models without touching code; progressive rollout built in
- **Observability** — telemetry, model analytics, device-level insights across your fleet
- **Evals & prompt libraries** — model-aware harness across runtimes
- **Prompt optimization** — per-model tuning = the same intent runs cheaper on every chip

- **Cross-device benchmarks** — large-scale evals for AI labs and OEMs

Traction - within < 5 months build

- 5 SDKs + Platform with alpha testers (telemetry, model analytics, device-level insights).
- +5k organic downloads across SDKs
- Early design partners
- Active partnerships with AI Labs (Stellon Labs, Neuphonic, Liquid)
- Model-agnostic runtime supporting open ecosystems

Why Now

The closed API moat is cracking, the silicon is in people's pockets, and the rules are catching up.

Small models cleared the utility bar

- **Mar '26:** Mistral releases Voxtral TTS, outperforming ElevenLabs.
- **Mar '26:** LFM2.5-350M proves 350M params deliver real utility on any device.
- **Apr '26:** Bonsai-8B shrinks 8B intelligence to 1.28GB.
- **Apr '26:** Google's Gemma 4 drops optimized 2B/9B variants built for phones and edge.

AI silicon is reaching critical mass in consumer devices

- AI-capable smartphones = **16% of shipments in 2024, 54% by 2028**
- AI PCs = **31% in 2025, on track for 94% by 2028.**
- Billions of devices already shipping with Apple Neural Engine, NPUs, and commodity GPUs; **and less than 5% of that silicon is ever utilized.**

The regulatory picture is converging globally

- **EU AI Act** enforcement begins August 2026.
- **APAC** countries firming up guardrails through 2026.

Why We Win

Breadth

- Every major on-device runtime and every major hardware target under one API
- When a new SOTA model drops, we support it on day one
- When cloud inference is the right call, we route there transparently

Proprietary data

- Every deployment generates cross-device execution data across thousands of real devices running real workloads.
- **Every new deployment makes the dataset harder to replicate.**
- The longer Xybrid runs, the more this dataset defines us.

The loop

- Data collected sharpens the routing. The routing sells the runtime and the runtime generates more data.
- Every deployment makes every future deployment faster, cheaper, and more reliable.

GTM

Initial target: Companies with 100K+ users across Europe and the US (consumer applications, game studios, no-code app engines)

Phase 2: Expansion into the APAC region, where device diversity and cost sensitivity amplify the value of a unified runtime.

Business Model

1. **Open-source adoption:** Free runtime drives developer gravity.
2. **Cloud routing, usage-based pricing:** Intelligent fallback to cloud when the device can't keep up.
3. **Platform (enterprise):** compliance, observability, control
4. Insights layer: model labs and enterprises pay for cross-device benchmarks and pre-deployment simulation.

Team

Glenn Sonna (l.in) — 10+ years building software and scaling products to millions of users

- ex-Entrepreneur First, BlaBlaCar
- **Scaled YOLO to 50M+ users** - Funded by T**hrive Capital, General Catalyst**

Sami Moustachir (l.in) — 10+ years in data/ML systems engineering

- ex-Entrepreneur First, ML at Thales
- Built large-scale payment infrastructure at **Xendit** (YC15) - **$70 billion/annually** in the APAC region

The Ask

$1.5M Pre-Seed

- Expand runtime capabilities
- Grow developer adoption (DevRel hires, hackathons)
- Convert design partners into enterprise deployments

xybrid.ai