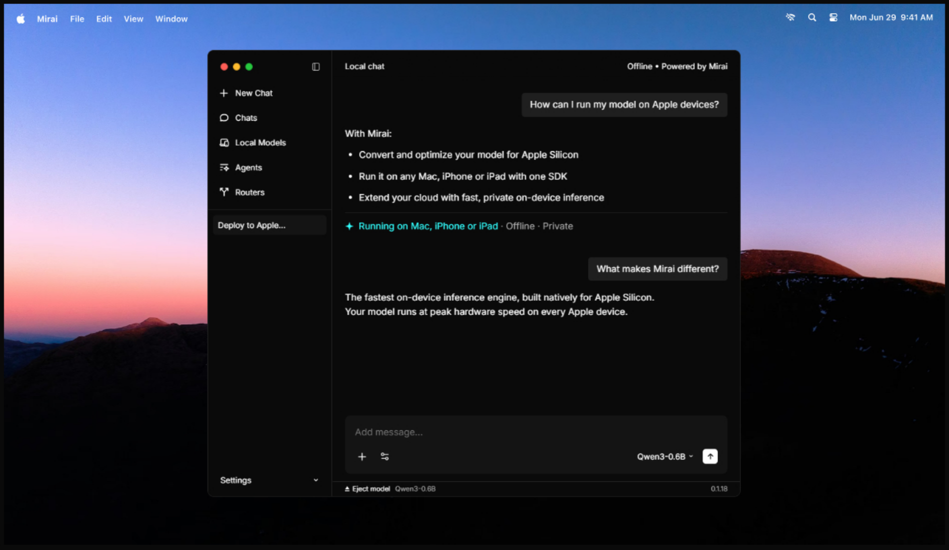
Your models. Every Apple device.

The fastest inference engine for Apple Silicon.

Talk to us →



How can I run my model on Apple devices?

With Mirai:

- Convert and optimize your model for Apple Silicon
- Run it on any Mac, iPhone or iPad with one SDK
- Extend your cloud with fast, private on-device inference

Running on Mac, iPhone or iPad · Offline · Private

What makes Mirai different?

The fastest on-device inference engine, built natively for Apple Silicon.
Your model runs at peak hardware speed on every Apple device.

Convert, optimize, distribute & run your models on Apple devices.

```
user %
```

Convert.

Convert and optimize your model for edge devices.

- ✓ One line model conversion.
- ✓ Quantize with outstanding quality.
- ✓ A lot of supported architectures out of the box.

Benchmark.

Distribute.

Run.

What Apple Silicon delivers today with Mirai.

Benchmarks

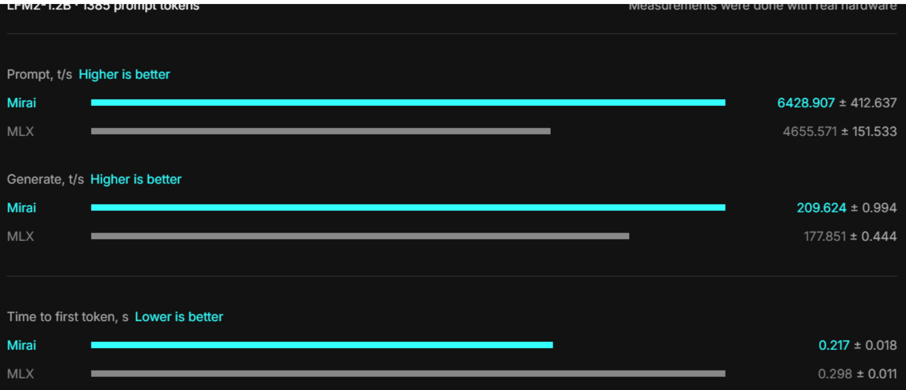
LFM2-1.2B ⌄ M1 Ultra ⌄

LFM2-1.2B · 1285 prompt tokens

Measurements were done with real hardware

LFM2-1.2B · 1385 prompt tokens

Measurements were done with real hardware



47 %

Real-world AI queries

Can be served locally on consumer hardware.

Stanford IPW ↗

20 TOPS

Neural Engine on M4

Mirai squeezes everything out of it.

Apple specs ↗

64 GB/s

Unified memory bandwidth on M4

Your model loads once, runs everywhere on chip.

Apple specs ↗

105 t/s

Qwen3-0.6B on M4 Max

Fast real-time generation on device.

Benchmarks ↑

Use cases that benefit from local inference.

Text

Summarization & extraction
Documents, emails, meeting notes

Classification
Intent detection, content tagging

Routing
Easily route complex requests to cloud model

Translation
With no internet connection

Voice coming soon

Speech-to-text
Real-time transcription on device

Text-to-speech
Easily narration of the text

Speech-to-speech
Translation and voice assistants

Voice commands
Predictable outputs

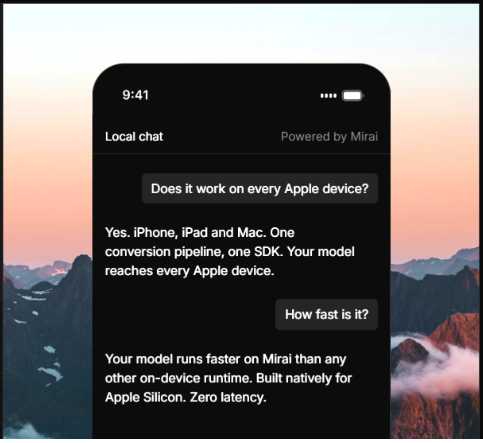
We built on-device native inference layer for Apple Silicon.

Seamless distribution.

✓ **Reach every Apple device.**
Your model works on every Apple device.

✓ **Performance without compromise.**
Your model runs faster on Mirai than any other on-device runtime.

✓ **One conversion pipeline.**
Convert from Hugging Face. Quantize, optimize, distribute.



Does it work on every Apple device?

Yes. iPhone, iPad and Mac. One conversion pipeline, one SDK. Your model reaches every Apple device.

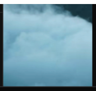
How fast is it?

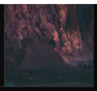
Your model runs faster on Mirai than any other on-device runtime. Built natively for Apple Silicon. Zero latency.

Offline by default.

$ Zero inference cost.

🔒 Data stays on device. ⌄





Supported models

LFM LiquidAI Partner	GPT-OSS OpenAI	Qwen 3 Alibaba
Gemma-3 Google	Llama-3.2 Meta	SmolLM2 Hugging Face
DeepSeek-R1 DeepSeek	Llamba Cartesia	+ Your model can be next Talk to us →

Explore all models →

Route models between device and cloud.

Run compact models locally on Apple Silicon. Route larger workloads to cloud infrastructure when necessary.

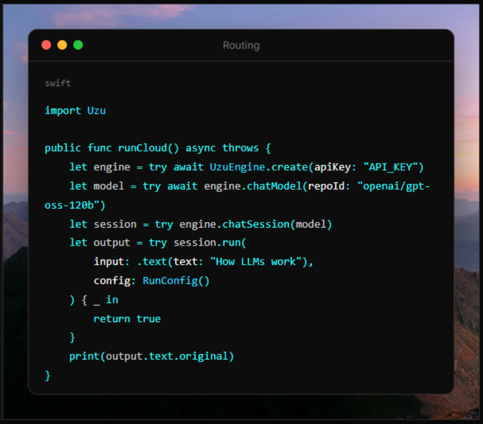
✓ Reduce cloud cost.

✓ Maintain full control.

✓ Keep sensitive data on user device.

Learn more ↗

baseten × mirai



```swift
swift

import Uzu

public func runCloud() async throws {
    let engine = try await UzuEngine.create(apiKey: "API_KEY")
    let model = try await engine.chatModel(repoId: "openai/gpt-
oss-120b")
    let session = try engine.chatSession(model)
    let output = try session.run(
        input: .text(text: "How LLMs work"),
        config: RunConfig()
    ) { _ in
        return true
    }
    print(output.text.original)
}
```

Common questions:

How does model support work? +

What architectures are supported? +

How does Mirai compare to other inference engines? +

What is the maximum supported model size? +

How can I run benchmarks myself? +

How can we discuss a specific use case? +

Backed by leading AI builders and investors

builders and investors.

Uncork
CAPITAL

stripe
David Singleton
/dev/agents
x-Stripe, Google

Stanford · Y
Francois Chaubard
Stanford AI Lab,
YC Partner

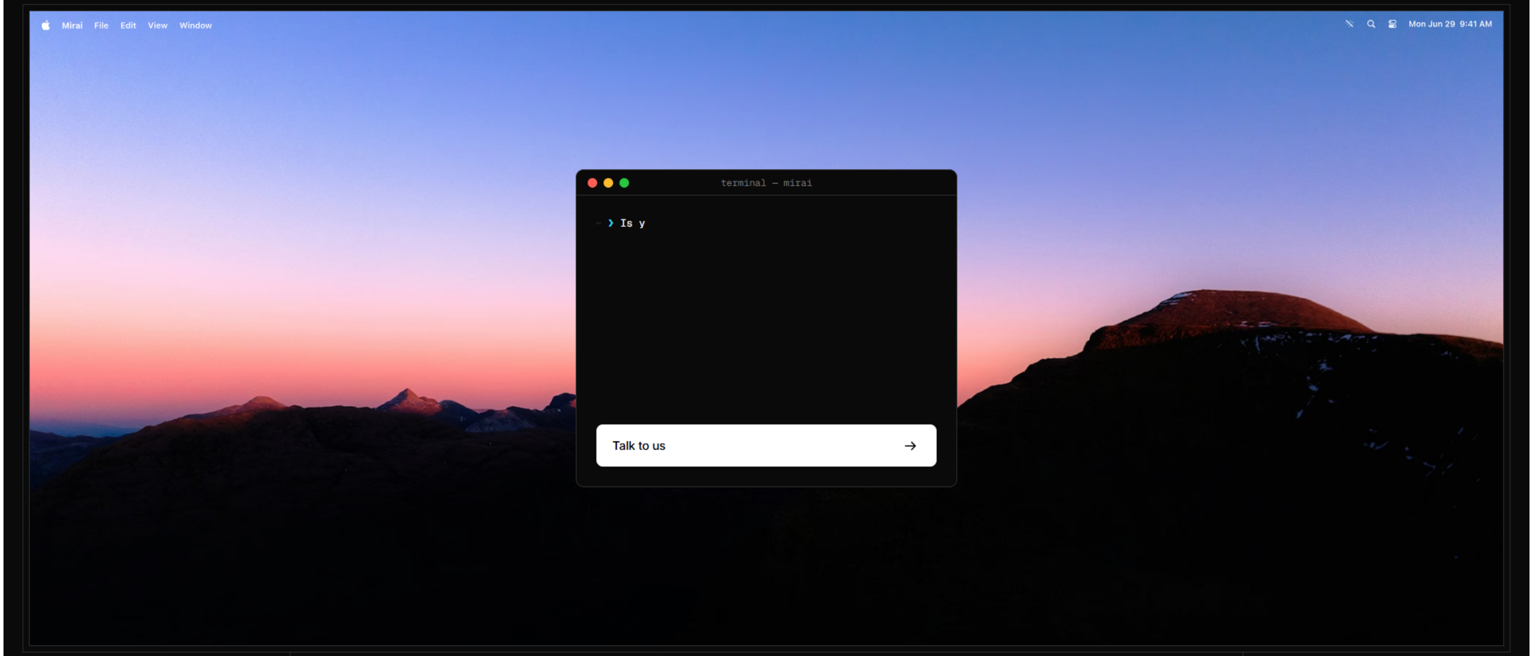
THEORY FORGE · moltbook
Ben Parr
TheoryForge VC,
Moltbook

IIElevenLabs
Mati Staniszewski
Co-founder, ElevenLabs

snowflake
Marcin Żukowski
Co-founder,
Snowflake

Google · coinbase
Gokul Rajaram
Google, Coinbase,
Trade Desk

groq · scale
Scooter Braun
Groq, Scale AI,
Justin Bieber



```
~ › Is y
```

mirai

Main

Platform / SDK

Models Library

MacOS App

Blog

Docs

Company

About us

Careers

Contact Us

Privacy Policy

Terms of Use

Links

X (Twitter)

Github

LinkedIn

Discord