

Giolivo Garcia Santarelli

PRODUCT & APPLIED AI ENGINEER | FINTECH SYSTEMS

Argentina (UTC-3) | Argentina + EU citizen | Remote employment or global contractor | maat.work@giolivosantarelli.com | [GitHub](#) | [LinkedIn](#) | hola@maat.work

SUMMARY

Product engineer with hands-on work across production fintech, local LLM infrastructure, native macOS audio and agent operations. Turns ambiguous business workflows into shipped systems, then validates them with tests, benchmarks and observable outcomes. Licensed financial advisor with direct domain experience in regulated advisory operations. English C2; Spanish native.

CORE SKILLS

Product: workflow modeling, requirements, systems design, production operations, client communication | **Engineering:** TypeScript, Python, Next.js, React, Node.js, PostgreSQL, Prisma, REST, MCP, Playwright | **AI / Systems:** LLM APIs, MLX, Metal, agent orchestration, local inference, observability, failure handling

SELECTED ENGINEERING PROJECTS

Financial Advisory CRM | Product & Full-Stack Engineering

Private production system | Next.js, TypeScript, PostgreSQL, Prisma, Playwright

- Built and operated a CRM used by two advisory teams, with a 56-table domain model spanning clients, portfolios, positions, risk profiles and compliance.
- Implemented production workflows, API surfaces and automated verification across a repository containing 700+ test/spec files.
- Modeled the product from real regulated-advisory work rather than a sample domain; architecture walkthrough available.

Local LLM Inference & Gateway | Systems Engineering

Python, MLX, Metal, SSE, GPU scheduling

- Profiled a 27B model on an M2 Max, isolated a quantized matrix-vector bottleneck and replaced the kernel in Metal.
- Increased decode throughput from 13.3 to 27.8 tok/s (2.1x) while preserving hash-identical output at every optimization step.
- Built an OpenAI-compatible local gateway with streaming, tier routing, GPU turn-taking, wake-on-demand and abort-on-disconnect; served 7,400 requests in a measured 24-hour window.

MeetCapture | Native macOS Product

Open source | Swift, SwiftUI, Core Audio, sherpa-onnx, SQLite

- Built private local meeting capture without a meeting bot or cloud audio service.
- Combined Core Audio process taps, live transcription, swappable ASR engines, speaker diarization and atomic summary handoff.
- Source: github.com/Gigisanta/MeetCapture

QuotaMax Router | Open-Source AI Reliability

Public source | Python, FastAPI, Redis, Prometheus, GitHub Actions

- Built an OpenAI-compatible router with provider fallback, per-model quota accounting, budget controls and circuit-breaker behavior.
- Added Prometheus metrics, structured call logs, health probes, failure alerts and an incident runbook; public CI is green with 81.89% measured line coverage.
- Source: github.com/Gigisanta/hermes-quota-max-router

DOMAIN EXPERIENCE

Licensed Financial Advisor | Argentina | Jun 2024 - Present

Advises a personal book of 40 clients and applies that operational experience to fintech product design, portfolio data, reporting and compliance workflows.

EDUCATION & CREDENTIALS

Quantitative Finance Certificate - UCEMA | **Idoneo CNV** - Argentina securities-agent qualification | **High School Diploma** - ESRN N 35