Agustín *Vivancos*

Senior AI / Data Engineer — LLM Systems & Data Pipelines in Production · Python / FastAPI · AWS / GCP

- 17 YEARS SHIPPING WEB
- AI-FIRST · 11 LLM AGENTS IN PRODUCTION
- PYTHON · FASTAPI · DATA PIPELINES
- AWS · GCP · DOCKER · K8S · CI/CD

agusvc@gmail.com
Salamanca, Spain · open to relocation (Madrid)

Available immediately

Freelance 12m+ ⟶ permanent · on-site / hybrid · Madrid · EU

Madrid · EU

ENGAGEMENT

Contract	Freelance 12m+ ⟶ perm
Mode	On-site / hybrid
Base	Madrid · EU
Home	Salamanca, ES
Start	Immediately

STACK IN PRODUCTION

AI / LLM	LangChain · **MCP servers** · agent-proxy orchestration · OpenAI · Anthropic
PYTHON	FastAPI · async services · Pydantic · pytest
LLMOPS	prompt eng · behavior tuning · **cost / latency / traceability** · evals · open-source models
DATA	ETL / ELT · streaming · data quality · post-call analytics · **RAG / vector search**
STORES	PostgreSQL · Redis · Qdrant / Pinecone · WebSockets
INFRA	AWS · GCP · **Docker** · **Kubernetes** · CI/CD · multi-tenant · audit trail
JS / TS	TypeScript · Node.js · Next.js · React · React Native (Expo)
INTEGR.	REST · **OpenAPI** · webhooks (HMAC) · Stripe · Meta CAPI

METHODOLOGY

- **Spec-driven**: every feature starts with the spec.

SUMMARY

Senior engineer (**17 years**) building **AI-driven, scalable data solutions** in production. I design and operate **LLM systems** and the **data pipelines** around them in **Python / FastAPI** on **AWS / GCP** — RAG, agent-proxy orchestration, streaming and post-call analytics — with full observability (cost, latency, traceability), Docker / Kubernetes, CI/CD and microservices. I run **11 LLM agents live in production**. I work *AI-first and spec-driven* (TDD, contract testing, ADR per decision) and I like getting my hands dirty in the code. Obsessed with measurable efficiency: voice-LLM at **$0.04/min**, latency **−35%**, compute **−30%**, Meta Ads ROAS **4% → 9%**. Based in Spain, available immediately for an on-site Madrid / EU contract that transitions to permanent.

SELECTED PROJECTS

pilis.app LIVE LLM + DATA

in production

Production LLM agent API + data & analytics layer · Python / FastAPI · built end-to-end

- **Versioned public API** (OpenAPI / Swagger) for chat, query, case-analysis, training-feedback and call-summarization — consumed by a Next.js / TS web and a React Native (Expo) app.
- **Data & analytics layer**: post-call analysis pipelines, model management, internal/external dashboards — including Meta Ads performance feeding the CRM and reporting.
- Contract testing + TDD over every critical flow; CI-gated on each push.

Migro LIVE LEGAL-TECH

2022 — 2025

11 LLM agents · voice-LLM · data pipelines · web, mobile & API · live at migro.es

- **11 LLM agents in production** (MCP + LangChain): legal research, case mgmt, paid acquisition, voice analysis, compliance, document processing, governance.
- **Voice-LLM call-center** (ElevenLabs → Telnyx) → structured CRM records and analytics; cost brought to $0.04/min by routing to open-source LLMs.
- **Meta Ads MCP agent** operating paid acquisition (ROAS 4% → 9%, CPL −24%); clean API layering for −35% latency and −30% compute spend.

SELECTED EXPERIENCE

- **TDD + contract testing** on every critical flow.
- ADR per decision; documented with diagrams.
- Full observability — cost, latency, traceability — and prompt-injection hardening.

HIGHLIGHTS

- 11 LLM/MCP agents + voice-LLM call-center in production — $0.04/min via open-source routing
- −35% latency and −30% compute spend on critical flows
- Meta Ads MCP agent: ROAS 4% → 9%, CPL −24%
- 17 years shipping production software; led teams of up to 15

EDUCATION

BSc in Software Engineering
Universidad de Salamanca · 2006 — 2010

LANGUAGES

Spanish	Native
English	Bilingual
Portuguese	Professional

Full-Stack / AI & Data Engineer · *Independent* NOW

2023 — present

B2B via own LLC · AI-first, spec-driven + TDD end-to-end

- Ship Python / FastAPI microservices and JS / TS clients on versioned API contracts with TDD; LLM agents and data pipelines in production (OpenAI · Anthropic · MCP · AWS / GCP).
- Technical leadership: migrations, safe refactors under coverage, observability, ramp-up of junior teams.

Founder & Full-Stack Engineer · *Migro*

2022 — 2025

Legal-tech for immigration processes in Spain · Madrid · Remote

- Owned product & architecture end-to-end — 11 LLM agents, voice-LLM and data pipelines (see Selected projects above).

Expansion & Driver-Acquisition Manager · *Uber*

2015 — 2018

Contractor · Lima, Peru · on-site · 3 years 5 months

- Drove driver-acquisition strategy, onboarding flows and data-led growth for city launches across Peru, Colombia and Ecuador; cross-functional work across ops, marketing and expansion.
- **Lead-acquisition sharing agreement** with a top advertising agency (2016) — ~$2M saved in acquisition spend.

Founder · *HoyRed*

2014 — 2020

Salamanca, Spain · tourism web network + hotel & apartments · full P&L for 6 years

- Web properties driving qualified SEO / content traffic to offline assets; booking funnels, pricing, channel management.
- Kept an active portfolio of technical clients throughout — shipping on Python and Next.js stacks.

Founder · Impulsa Consultores

2010 — 2013

First company, founded at 22 · digital consultancy

- **Led a team of 15** across engineering, design and accounts; enterprise clients BBVA · Banco Santander · Inditex + 90 SMEs.

CTO · *enterbio*

2011 — 2012

Organic brand moving to D2C · Madrid

- Online store, order pipeline and operational integrations: catalog, fulfilment, billing.