Agustín *Vivancos*

Senior AI / Data Engineer — Sistemas LLM e Data Pipelines em Produção · Python / FastAPI · AWS / GCP

- 17 ANOS CONSTRUINDO WEB
- AI-FIRST · 11 AGENTES LLM EM PRODUÇÃO
- PYTHON · FASTAPI · DATA PIPELINES
- AWS · GCP · DOCKER · K8S · CI/CD

agusvc@gmail.com
Salamanca, Espanha · aberto a realocação (Madrid)

Disponibilidade imediata

Freelance 12m+ ⟶ permanente · on-site / híbrido · Madrid · EU

Madrid · EU

ENGAGEMENT

Contrato	Freelance 12m+ ⟶ fixo
Modo	On-site / híbrido
Base	Madrid · EU
Origem	Salamanca, ES
Início	Imediato

STACK EM PRODUÇÃO

AI / LLM	LangChain · **MCP servers** · agent-proxy orchestration · OpenAI · Anthropic
PYTHON	FastAPI · serviços async · Pydantic · pytest
LLMOPS	prompt eng · ajuste de comportamento · **custo / latência / rastreabilidade** · evals · modelos open-source
DATA	ETL / ELT · streaming · qualidade de dados · analítica pós-chamada · **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

METODOLOGIA

- **Spec-driven**: cada feature começa pela spec.

RESUMO

Engenheiro sénior (**17 anos**) construindo **soluções de dados escaláveis com IA** em produção. Desenho e opero **sistemas LLM** e os **data pipelines** à sua volta em **Python / FastAPI** sobre **AWS / GCP** — RAG, orquestração agent-proxy, streaming e analítica pós-chamada — com observabilidade completa (custo, latência, rastreabilidade), Docker / Kubernetes, CI/CD e microsserviços. Tenho **11 agentes LLM em produção**. Trabalho *AI-first e spec-driven* (TDD, contract testing, ADR por decisão) e gosto de meter as mãos no código. Obcecado por eficiência mensurável: voice-LLM a **$0.04/min**, latência **−35%**, compute **−30%**, ROAS de Meta Ads **4% ⟶ 9%**. Baseado em Espanha, disponível imediatamente para um contrato on-site Madrid / EU que transite para permanente.

PROJETOS EM DESTAQUE

pilis.app LIVE LLM + DATA

em produção

API de agentes LLM em produção + camada de dados e analítica · Python / FastAPI · end-to-end

- **API pública versionada** (OpenAPI / Swagger) para chat, query, análise de casos, training-feedback e resumo de chamadas — consumida por uma web Next.js / TS e uma app React Native (Expo).
- **Camada de dados e analítica**: pipelines de análise pós-chamada, model management, dashboards internos/externos — incluindo o performance do Meta Ads alimentando o CRM e os relatórios.
- Contract testing + TDD em cada fluxo crítico; gated por CI em cada push.

Migro LIVE LEGAL-TECH

2022 — 2025

11 agentes LLM · voice-LLM · data pipelines · web, mobile e API · em migro.es

- **11 agentes LLM em produção** (MCP + LangChain): legal research, gestão de casos, paid acquisition, análise de voz, compliance, processamento documental, governance.
- **Call-center voice-LLM** (ElevenLabs ⟶ Telnyx) ⟶ registos CRM estruturados e analítica; custo reduzido a $0.04/min encaminhando para LLMs open-source.
- **Agente MCP de Meta Ads** operando paid acquisition (ROAS 4% ⟶ 9%, CPL −24%); API limpa para −35% latência e −30% compute.

EXPERIÊNCIA SELECIONADA

- **TDD + contract testing** em cada fluxo crítico.
- ADR por decisão; documentado com diagramas.
- Observabilidade completa — custo, latência, rastreabilidade — e hardening anti prompt-injection.

HIGHLIGHTS

- 11 agentes LLM/MCP + call-center voice-LLM em produção — $0.04/min via routing open-source
- −35% latência e −30% de gasto de compute em fluxos críticos
- Agente MCP de Meta Ads: ROAS 4% → 9%, CPL −24%
- 17 anos entregando software em produção; equipes de até 15 pessoas

EDUCAÇÃO

Eng. de Software (BSc)
Universidad de Salamanca · 2006 — 2010

IDIOMAS

Espanhol	Nativo
Inglês	Bilíngue
Português	Profissional

Full-Stack / AI & Data Engineer · *Independente* AGORA — 2023 — presente

B2B via sociedade própria · AI-first, spec-driven + TDD end-to-end

- Microsserviços Python / FastAPI e clientes JS / TS sobre contratos de API versionados com TDD; agentes LLM e data pipelines em produção (OpenAI · Anthropic · MCP · AWS / GCP).
- Liderança técnica: migrações, refactors seguros sob cobertura, observabilidade, ramp-up de equipes júnior.

Founder & Full-Stack Engineer · *Migro* — 2022 — 2025

Legal-tech para processos de imigração em Espanha · Madrid · Remoto

- Produto e arquitetura end-to-end — 11 agentes LLM, voice-LLM e data pipelines (ver Projetos em destaque acima).

Expansion & Driver-Acquisition Manager · *Uber* — 2015 — 2018

Contractor · Lima, Peru · presencial · 3 anos 5 meses

- Estratégia de driver-acquisition, onboarding flows e crescimento data-led em launches de cidades no Peru, Colômbia e Equador; trabalho cross-functional com ops, marketing e expansão.
- **Acordo de lead-acquisition partilhado** com uma agência de publicidade top (2016) — ~$2M poupados em gasto de aquisição.

Founder · *HoyRed* — 2014 — 2020

Salamanca, Espanha · rede de sites de turismo + hotel e apartamentos · P&L completo 6 anos

- Web properties com SEO / conteúdo direcionando tráfego qualificado a ativos offline; booking funnels, pricing, channel management.
- Carteira ativa de clientes técnicos em paralelo — shipping sobre stacks Python e Next.js.

Founder · Impulsa Consultores — 2010 — 2013

Primeira empresa, fundada aos 22 · consultoria digital

- **Equipe de 15** entre engineering, design e account management; clientes enterprise BBVA · Banco Santander · Inditex + 90 SMEs.

CTO · *enterbio* — 2011 — 2012

Brand de orgânicos movendo-se para D2C · Madrid

- Online store, order pipeline e integrações operacionais: catálogo, fulfilment, billing.