

Javier Pontón González

AI Engineer / LLM Engineer

I build AI systems that survive an audit, not just a demo.

Email: javierpontongonzalez@gmail.com
Phone: +34 623 920 307
Location: Asturias, Spain
Time zone: Europe/Madrid
Remote: Fully remote (EU, UK, US)
Availability: Available immediately

Engagement: Independent B2B contractor, invoiced from Spain
Work authorisation: EU citizen. Invoices as a Spanish B2B contractor; no visa sponsorship required for EU work.
LinkedIn: <https://www.linkedin.com/in/javierpontongonzalez>
GitHub: <https://github.com/sktjpg>
Website: <https://javierpontongonzalez.com>

PROFESSIONAL SUMMARY

AI Engineer and LLM Engineer with 7+ years of production backend engineering for Iberia, KPN, Mercadona, Inditex, Openbank and Idealista, and 2+ years shipping Generative AI to production. I design, build and operate LLM systems end to end: AI agents and multi-agent systems (LangChain, LangGraph, MCP, function calling), Retrieval-Augmented Generation (RAG) with embeddings, hybrid search and cross-encoder reranking, structured extraction with vision-language models, and self-hosted inference including on-premise and air-gapped deployments.

What separates my work from prototypes: every system ships with an evaluation harness (golden datasets, field-level precision and recall, regression testing), OpenTelemetry observability, source-level traceability and human-in-the-loop review. Core stack: Python, Java, Kotlin, FastAPI, Spring Boot, PostgreSQL and pgvector, Kafka, AWS, Docker. Independent B2B contractor since 2023 with no gaps between engagements, and founder-engineer of four live products.

KEY ACHIEVEMENTS

- Architected an air-gapped, on-premise multi-agent AI system that automates Arabic public-tender response end to end, from heterogeneous vendor quotations to costing model and drafted technical proposal, with source-level traceability on every extracted figure.
- Built a production Model Context Protocol (MCP) server and a RAG service over engineering documentation at Iberia, giving the team governed function calling and natural-language access to the flight pricing platform, instrumented with OpenTelemetry and Dynatrace.
- Led a zero-downtime Azure to AWS migration of a global logistics platform at Inditex and a Java to Kotlin migration at Openbank inside a 100+ engineer organisation; shipped a consumer app (ZORRO) solo to iOS and Android, reaching 400+ users with 50% D1 retention in three weeks.

TECHNICAL SKILLS

AI & LLM	Large Language Models (LLMs), Generative AI, Applied AI, Machine Learning, Deep Learning, Transformers, Vision-Language Models (VLMs), NLP, Computer Vision, Prompt Engineering, Structured Outputs, JSON Schema, Function Calling, Tool Use, Hallucination Mitigation, Grounding
AI Agents & RAG	Retrieval-Augmented Generation (RAG), AI Agents, Agentic AI, Agentic Workflows, Multi-Agent Systems, LangChain, LangGraph, Model Context Protocol (MCP), Embeddings, Semantic Search, Hybrid Search, Cross-Encoder Reranking, Vector Databases, pgvector, HNSW
LLM Evaluation & LLMOps	LLM Evaluation, Golden Datasets, Regression Testing, Precision, Recall, F1, LLM-as-a-Judge, Human Evaluation, RAG Evaluation, Hit Rate, MRR, A/B Testing, Shadow Evaluation, RAGAS, Promptfoo, LLMOps, OpenTelemetry, Prometheus, Grafana, Dynatrace, Cost Optimization, Latency Optimization
Model Serving & AI Infrastructure	vLLM, Ollama, mistral.rs, OpenAI-Compatible APIs, PyTorch, HuggingFace, Quantization, GGUF, AWQ, LoRA, QLoRA, GPU/VRAM Optimization, Self-Hosted Inference, On-Premise AI, Air-Gapped Deployment, Qwen3, Qwen2.5-VL, DeepSeek, Claude API, OpenAI API
Backend & Data	Python, Java, Kotlin, Elixir, TypeScript, SQL, Spring Boot, FastAPI, Spring Batch, REST, OpenAPI, Microservices, PostgreSQL, pgvector, MongoDB, Elasticsearch, Redis, DynamoDB, Apache Kafka, AWS SQS
Cloud & DevOps	AWS, Azure, GCP, Docker, Kubernetes, GitHub Actions, Jenkins, Git, CI/CD
Architecture & Engineering	Hexagonal Architecture, Domain-Driven Design (DDD), CQRS, Event Sourcing, SOLID, TDD, API-First, Technical Leadership, Mentoring

PROFESSIONAL EXPERIENCE

Independent Backend & AI Engineer – Freelance / B2B contractor Nov 2023 – Present · Remote, Europe and Middle East
Back-to-back enterprise engagements, no gaps between clients.

IT systems integrator, Saudi Arabia (client under NDA) – AI Engineer Jul 2026 – Present · Remote
Air-gapped multi-agent AI platform for public-tender (RFP) response.

- **Multi-agent AI system.** Architected a two-agent solution for tender response. A financial agent performs structured extraction from vendor quotations in heterogeneous formats (PDF, Word, Excel, email, supplier portals) into a costing model; a technical agent applies RAG and Arabic OCR over RFPs of hundreds of pages, extracts scope of work and bill of quantities, and drafts the technical proposal from a fixed corporate template.
- **Human-in-the-loop and traceability.** Engineered source-level traceability from every extracted figure back to its document and page, mandatory human review gates, and visual flagging of AI-generated content with no vendor source behind it, as required by tender confidentiality.
- **LLM evaluation.** Defined the evaluation harness before the build: a golden dataset of historical tenders scored on field-level precision and recall, regression testing on every prompt or model change, and OpenTelemetry traces per agent step for latency, token cost and failure analysis.
- **On-premise AI infrastructure.** Designed a model-agnostic architecture against OpenAI-compatible APIs, benchmarked Qwen3, DeepSeek, Falcon-H1 Arabic and ALLaM, sized GPU and VRAM for vLLM and Ollama serving, and met Saudi PDPL and NCA requirements in a fully isolated network.

Technologies: Python, LangChain, LangGraph, vLLM, Ollama, Qwen3, DeepSeek, RAG, pgvector, Arabic OCR, OpenTelemetry, Docker

Iberia, S.A. – AI Engineer Feb 2026 – Jul 2026 · Remote, Spain
Flight pricing platform. Technical owner of the fare pricing domain.

- **MCP server.** Built and deployed an internal Model Context Protocol server exposing pricing platform tooling to LLM clients, enabling governed function calling and agentic queries over fares, providers and configuration instead of ad-hoc scripts.
- **RAG service.** Delivered retrieval over internal specs and domain documentation with embeddings, hybrid search and cross-encoder reranking, cutting the time engineers spent locating pricing and NDC domain answers, instrumented with OpenTelemetry and Dynatrace like any other

production service.

- **AI-assisted software engineering.** Introduced Claude Code for service scaffolding, Karate and JUnit test generation and large refactors, driven by project-specific SKILL.md and CLAUDE.md conventions I defined and rolled out to the team.
- **Fare pricing services.** Engineered the services that calculate flight prices by orchestrating calls to external fare-calculation providers, normalising and aggregating heterogeneous responses under IATA NDC for internal channels and distribution partners; optimised latency and cut provider calls with Redis caching on the highest-traffic pricing flows.
- **Event-driven integration.** Delivered Kafka and PostgreSQL integration on AWS (ECS, SQS, S3) in hexagonal architecture with DDD, owning resilience of the pricing path (timeouts, retries and fallbacks) against third-party provider degradation.

Technologies: Python, Java, Kotlin, FastAPI, Spring Boot, MCP, Redis, Kafka, PostgreSQL, AWS, Karate, Claude Code

KPN – AI Engineer

Mar 2025 – Feb 2026 · Remote, Netherlands

Fibre installation tracking platform, ODF International Team.

- **Generative AI in the delivery workflow.** Introduced LLM-assisted development across the international team (test generation, legacy refactors, PR review) with shared prompt and review conventions, so AI output was always human-verified before merge.
- **Retrieval assistant.** Built semantic search with embeddings over runbooks and ServiceNow incident history, cutting the time to find the right precedent when triaging field-operations incidents.
- **Platform delivery.** Delivered a large-scale fibre installation tracking platform running in three European countries under international SLAs, building microservices and field-operations features and owning production incident resolution.
- **Data migration.** Migrated historical production records from PostgreSQL to DynamoDB with chunk-based storage, offloading the operational database and keeping the hot read path fast as volume grew.

Technologies: Python, Java, Kotlin, Spring Boot, Kafka, PostgreSQL, AWS (DynamoDB, Lambda), Embeddings, Semantic search, Vue.js, Docker

Mercadona, S.A. – Lead Backend Engineer

Mar 2024 – Mar 2025 · Remote, Spain

Product analytics platform.

- **Ingestion pipeline.** Engineered a Spring Batch ingestion pipeline processing millions of product records per day from heterogeneous sources, with restartable jobs and idempotent writes.
- **API-first services.** Designed API-first REST services (OpenAPI) in hexagonal architecture with DDD; sustained over 85% automated test coverage under real TDD.

Technologies: Java, Kotlin, Spring Boot, PostgreSQL, Kafka, AWS, GCP, Docker, Flyway

Inditex, S.A. – Lead Backend Engineer

Nov 2023 – Mar 2024 · Remote, Spain

Global logistics, garment-sorting event platform. Technical lead for the Azure to AWS migration.

- **High-concurrency platform.** Engineered a serverless platform on Azure and AWS processing thousands of garment-sorting events per minute across the logistics network.
- **Zero-downtime migration.** Led the Azure to AWS migration with zero downtime; built async REST APIs in hexagonal architecture with DDD and introduced Karate integration testing alongside the existing JUnit suite.

Technologies: Java, Kotlin, Spring Boot, MongoDB, PostgreSQL, Kafka, AWS, Azure, Docker, Liquibase

Open Bank, S.A. (Santander Group) – Backend Engineer

Oct 2022 – Nov 2023 · Madrid, Spain

Investment automation platform. Technical lead for the Java to Kotlin migration.

- **Serverless batch at scale.** Implemented serverless batch pipelines on AWS Lambda processing millions of financial transactions per day on top of the T24 core banking system, inside a 100+ backend engineer organisation.
- **Java to Kotlin migration.** Led the migration across services and Lambdas; delivered internal tech talks on DDD and hexagonal architecture to around 50 engineers.

Technologies: Java, Kotlin, Spring Boot, PostgreSQL, AWS (Lambda, S3), Elasticsearch, Docker, MockK, T24

Idealista, S.A. – Backend Engineer

Mar 2021 – Oct 2022 · Madrid, Spain

Digital contract-signing platform.

- **Contract-signing platform.** Built a digital contract-signing platform for Spain, Portugal and Italy: REST APIs plus Kafka event streams for real-time signing, with event sourcing keeping an auditable, replayable history of every contract.

Technologies: Java, Kotlin, Spring Boot, PostgreSQL, Elasticsearch, Kafka, Docker, Event sourcing, DDD, TDD

Empathy.co – Backend Engineer

Jan 2019 – Mar 2021 · Gijón, Spain

Playboard, e-commerce search configuration platform.

- **Search configuration platform.** Built the configuration platform enterprise clients (Kroger, Carrefour, Inditex) use to tune their Elasticsearch-backed search engines; executed zero-downtime migrations: Java 8 to 11, GCP to AWS, monolith to API Gateway.

Technologies: Java, Spring Boot, MongoDB, Elasticsearch, Docker, Kubernetes, GCP, AWS, JUnit

SELECTED PROJECTS

Facturias – Multimodal invoice-processing SaaS

Website

Founder and sole engineer · Live · Web, iOS, Android

Structured extraction from invoices with a self-hosted vision-language model (Qwen2.5-VL) plus RAG over Spanish tax regulation to classify and validate entries, with confidence scoring, a labelled evaluation set tracking field-level precision and recall across model versions, and human review on low-confidence fields. VeriFactu compliant, multi-tenant FastAPI backend on PostgreSQL 17 with row-level security and pgvector.

- **Vision-language extraction.** Self-hosted Qwen2.5-VL reads invoices as documents rather than as OCR text dumps, preserving table structure, line items and stamps that flat OCR loses.
- **RAG over Spanish tax regulation.** Retrieval over the Spanish tax code classifies and validates each entry, so a deduction is justified by a retrievable rule rather than by model intuition.
- **Confidence and human review.** Every field carries a confidence score; anything under threshold is routed to human review instead of being written silently.
- **Measured, not assumed.** A labelled evaluation set tracks field-level precision and recall across model versions, so a model upgrade is a measurable decision.
- **Multi-tenant by construction.** PostgreSQL 17 row-level security isolates tenants at the database, not in application code. pgvector holds the regulation embeddings.

Technologies: Python, FastAPI, SQLAlchemy (async), PostgreSQL 17, Row-level security, pgvector, Qwen2.5-VL, Docker · **Keywords:** invoice OCR, VeriFactu, vision language model, RAG, multi-tenant SaaS, Spanish tax

ZORRO – Consumer dating app for the gay and queer community, iOS and Android Website · App Store · Google Play
Founder and sole engineer · Live · iOS, Android

Shipped solo to Google Play and the App Store in June 2026. Semantic matchmaking with embeddings and pgvector plus cross-encoder reranking, and self-hosted vision-language model moderation of user photos handling GDPR special-category data with no third-party processors. Polyglot production backend in Kotlin/Spring Boot, Elixir and Python.

- **Semantic matchmaking.** Profiles are embedded and retrieved with pgvector, then reordered by a cross-encoder reranker: the same retrieve-then-rerank architecture as a production RAG pipeline, applied to people instead of documents.
- **Self-hosted photo moderation.** A vision-language model screens user photos on my own infrastructure. Sexual orientation data is GDPR special-category data, so it never reaches a third-party processor.
- **Polyglot backend.** Kotlin and Spring Boot for the domain, Elixir for realtime chat and presence, Python for the AI services.
- **Traction.** 400+ users with 50% D1 retention within three weeks of launch.

Technologies: Kotlin, Spring Boot, Elixir, Python, PostgreSQL, pgvector, Cross-encoder reranking, AWS SES, Hetzner, Docker · **Keywords:** semantic matching, embeddings, pgvector, content moderation, GDPR, mobile app

Apunta – Multi-tenant SaaS for shooting clubs, from custom PCB to mobile app Website · App Store
Founder and sole engineer · Live · iOS, Android, Web, Hardware

Multi-tenant SaaS for shooting clubs owned end to end: custom NFC PCB (ESP32-C6, PN532, PoE) and C/ESP-IDF firmware, Spring Boot backend and React Native app. Range access, training sessions, scoring and club administration on one stack.

- **Custom hardware.** Designed the NFC access PCB around an ESP32-C6 and a PN532 reader with Power over Ethernet, and wrote the firmware in C on ESP-IDF.
- **Backend and multi-tenancy.** Spring Boot services with per-club tenancy, membership, range booking, training sessions and scoring.
- **Mobile.** React Native app shipped to the App Store and Google Play for shooters and club administrators.

Technologies: Kotlin, Spring Boot, React Native, PostgreSQL, C, ESP-IDF, ESP32-C6, PN532, Docker · **Keywords:** IoT, NFC, embedded firmware, multi-tenant SaaS, React Native

Grabia – Self-hosted meeting intelligence, fully local AI
Founder and sole engineer · Private inference node · Self-hosted

Local pipeline with WhisperX transcription and pyannote diarisation feeding a LangChain and LangGraph agent that produces structured summaries, decisions and action items, with RAG over the meeting archive. Served from a private node (Ryzen AI MAX+ 395, 128 GB unified memory, ROCm) running Qwen3-30B under Ollama and mistral.rs; no audio or transcript leaves the machine.

Technologies: Python, LangChain, LangGraph, WhisperX, pyannote, Ollama, mistral.rs, ROCm, Qwen3 · **Keywords:** local LLM, speech to text, diarisation, agentic summarisation

icekar – Distributed scraping and search over 100,000+ car listings nightly Website
Founder and sole engineer · Live · Web

Distributed scraping and Elasticsearch search over 100,000+ car listings nightly, with an agentic LLM layer that rewrites scraper extraction rules when target sites change markup.

Technologies: Python, Elasticsearch, LLM agents, Docker · **Keywords:** web scraping, self-healing scrapers, Elasticsearch

EDUCATION

M.Sc. in Artificial Intelligence Research (official) Sep 2026 – Sep 2027
UIMP / AEPIA

Specialisation in Machine Learning and Data Science. Incoming, starting September 2026.

B.Sc. in Computer Science and Software Engineering
University of Oviedo
Oviedo, Spain.

LANGUAGES

- **Spanish.** Native (CEFR C2)
- **English.** Professional working proficiency (CEFR C1)