

JAVIER PONTÓN GONZÁLEZ

Forward Deployed Engineer (FDE)

Customer-Embedded AI Delivery | LLM Systems | RAG | AI Agents | On-Premise & Air-Gapped GenAI | Python · Java · Kotlin · TypeScript
javierpontongonzalez@gmail.com | +34 623 920 307 | Asturias, Spain | Remote (CET; comfortable with EU and US overlap)
EU citizen, no sponsorship required for EU work | Independent B2B contractor | Available immediately
linkedin.com/in/javierpontongonzalez | github.com/sktjpg | javierpontongonzalez.com

PROFESSIONAL SUMMARY

Forward Deployed Engineer with 7+ years of production backend engineering for Iberia, KPN, Mercadona, Inditex, Openbank and Idealista, and 2+ years shipping Generative AI (GenAI) systems inside customer environments. I embed with the client team, scope the use case against their real data and constraints, build on their infrastructure, and stay through production and handover. Five back-to-back enterprise engagements since 2023, contracted directly with no agency layer, across airline pricing, telecoms field operations, retail logistics and Middle East public tendering. Technically: AI agents and multi-agent systems (LangChain, LangGraph, MCP), Retrieval-Augmented Generation (RAG) with hybrid search and cross-encoder reranking, structured extraction with vision-language models, and self-hosted inference on-premise and in air-gapped environments. Every system ships with an evaluation harness (golden datasets, field-level precision and recall, regression testing), OpenTelemetry observability, guardrails and human-in-the-loop review.

- Deployed into a Saudi customer's fully isolated on-premise environment to architect an air-gapped multi-agent AI platform automating Arabic public-tender response over RFPs of hundreds of pages, with source-level traceability on every extracted figure, under Saudi PDPL and NCA requirements.
- Embedded in Iberia's pricing engineering team as technical owner of the fare pricing domain, building the Model Context Protocol (MCP) server and RAG service for the flight pricing platform: governed function calling and natural-language access to fares and domain documentation, instrumented with OpenTelemetry and Dynatrace.
- Comfortable owning whatever the customer needs shipped: front end to infrastructure, zero-downtime Azure-to-AWS and Java-to-Kotlin migrations inside 100+ engineer organisations, and a consumer app shipped solo to iOS and Android reaching 700+ users with 50% D1 retention in three weeks.

TECHNICAL SKILLS

Forward-Deployed Delivery: Customer-embedded engagements, use-case scoping against real customer data, build-versus-buy assessment, solution architecture, production operation and handover, on-premise and air-gapped deployment, regulated environments (Saudi PDPL and NCA, GDPR, IATA NDC), human-in-the-loop review workflows, direct technical communication with customer stakeholders in English and Spanish

AI & LLM Engineering: Large Language Models (LLMs), Generative AI (GenAI), RAG (Retrieval-Augmented Generation), AI Agents, Agent Orchestration, Multi-Agent Systems, LangChain, LangGraph, LlamaIndex, MCP (Model Context Protocol), Function Calling, Tool Use, Prompt Engineering, Structured Outputs (JSON Schema), Embeddings, Semantic Search, Hybrid Search, Cross-Encoder Reranking, Vector Databases (pgvector), Vision-Language Models (Qwen2.5-VL)

LLM Evaluation, LLMops & Observability: LLM Evaluation, Evals, RAGAS, LangSmith, Golden Datasets, Field-Level Precision and Recall, Regression Testing, Guardrails, Human-in-the-Loop Review, LLMops, OpenTelemetry, Prometheus, Grafana, Dynatrace, Inference Latency and Cost Optimization

Model Serving, AI Platforms & Self-Hosted AI: vLLM, Ollama, mistral.rs, Hugging Face, OpenAI-Compatible APIs, Model Deployment, Quantization (GGUF), On-Premise and Air-Gapped Deployment, Qwen3, DeepSeek, Claude API, OpenAI API, AWS Bedrock, Google Vertex AI, managed-versus-self-hosted trade-off analysis (latency, cost, data privacy, scalability, model choice)

Backend & Data: Python (async), FastAPI, Java, Kotlin, Spring Boot, TypeScript, React, React Native, Vue.js, SQL, REST, OpenAPI, Streaming APIs, WebSockets, Microservices, Apache Kafka, PostgreSQL, Redis, DynamoDB, Elasticsearch, MongoDB

Cloud & DevOps: AWS (ECS, Lambda, SQS, S3), Azure, GCP, Docker, Kubernetes, GitHub Actions, CI/CD

Architecture: Hexagonal Architecture, Domain-Driven Design (DDD), Event Sourcing, TDD, API-First

PROFESSIONAL EXPERIENCE

Independent Consultant (Freelance / B2B Contractor) — Forward Deployed Engineer | Nov 2023 – Present

Remote, Europe and Middle East. Five back-to-back enterprise engagements contracted directly, no agency layer. End-to-end ownership per engagement: use-case scoping, build-versus-buy assessment, architecture, delivery, production operation and handover.

Confidential IT systems integrator, Saudi Arabia (under NDA) — Forward Deployed Engineer | Jul 2026 – Aug 2026

Sole engineer on a fixed-scope architecture-and-build engagement, delivered inside the customer's isolated on-premise environment: air-gapped multi-agent AI platform for public-tender (RFP) response.

- Architected a two-agent platform: a financial agent performing structured extraction from vendor quotations in heterogeneous formats (PDF, Word, Excel, email, supplier portals) into a costing model, and a technical agent applying RAG and Arabic OCR over RFPs of hundreds of pages to extract scope of work and bill of quantities and draft the technical proposal from a corporate template.
- Defined the LLM 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; extraction cleared 90%+ field-level precision and recall on the golden dataset before go-live.
- Engineered guardrails for tender confidentiality: source-level traceability from every extracted figure to its source document and page, a full-stack review application (web front end and backend API) enforcing mandatory human review gates, and visual flagging of unsourced AI content; met Saudi PDPL and NCA requirements on fully isolated on-premise infrastructure, benchmarking Qwen3, DeepSeek, Falcon-H1 Arabic and ALLaM under vLLM and Ollama.

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

Iberia, S.A. (Contract) — Forward Deployed Engineer (FDE) | Feb 2026 – Present

Remote, Spain. Flight pricing platform. Technical owner of the fare pricing domain.

- Built and deployed an internal MCP (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.
- Delivered a RAG service over internal specs and NDC domain documentation with embeddings, hybrid search and cross-encoder reranking, turning manual searches across specs into sub-minute, source-grounded answers for the engineering team; instrumented with OpenTelemetry and Dynatrace as a production service.
- Engineered the fare pricing services orchestrating external fare-calculation providers under IATA NDC on Kafka, PostgreSQL and AWS (ECS, SQS, S3) in hexagonal architecture with DDD; reduced latency and provider calls with Redis caching on the highest-traffic pricing flows and owned resilience (timeouts, retries, fallbacks) against provider degradation.
- Deployed managed model serving on AWS Bedrock and Google Vertex AI to production alongside self-hosted inference, selecting per use case on latency, cost, data privacy, scalability and model flexibility, and integrated it with the MCP server and RAG service under OpenTelemetry and Dynatrace observability.
- Introduced Claude Code for service scaffolding, Karate and JUnit test generation and large refactors, driven by SKILL.md and CLAUDE.md conventions I defined and rolled out to the team.

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

KPN (Contract) — Senior Full Stack Engineer | Mar 2025 – Feb 2026

Remote, Netherlands. Fibre installation tracking platform, ODF International Team.

- Delivered a fibre installation tracking platform running in three European countries under international SLAs, owning features end to end from Spring Boot microservices and Kafka event flows to Vue.js and TypeScript screens for planners and field engineers, including production incident resolution.
- Migrated ~30 million 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.
- Built a semantic search assistant with embeddings over runbooks and ServiceNow incident history, cutting time-to-precedent when triaging field-operations incidents; introduced LLM-assisted development conventions (test generation, refactors, PR review) with human verification before merge.

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

Mercadona, S.A. (Contract) — Lead Backend Engineer | Mar 2024 – Mar 2025

Remote, Spain. Product analytics platform.

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

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

Inditex, S.A. (Contract) — Lead Backend Engineer | Nov 2023 – Mar 2024

Remote, Spain. Global logistics, garment-sorting event platform.

- Technical lead of the zero-downtime Azure-to-AWS migration of a serverless platform processing thousands of garment-sorting events per minute across the logistics network.
- 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

Earlier roles (employee positions):

Openbank (Open Bank, S.A., Santander Group) — Senior Backend Engineer | Oct 2022 – Nov 2023 • Madrid, Spain

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

Idealista, S.A. — Backend Engineer | Mar 2021 – Oct 2022 • Madrid, Spain

- 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; contributed features to the platform's React front end.

Empathy.co — Backend Engineer | Jan 2019 – Mar 2021 • Gijón, Spain

- 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, and monolith to API Gateway.

SELECTED PROJECTS

Facturias — Invoice-processing SaaS, founder-built and live in production | facturias.es

- Structured extraction from invoices with a self-hosted vision-language model (Qwen2.5-VL) plus RAG over Spanish tax regulation; confidence scoring with human review on low-confidence fields, and a labelled evaluation set tracking field-level precision and recall across model versions, at 93% field-level extraction precision in production. VeriFactu-compliant, multi-tenant FastAPI backend on PostgreSQL 17 with row-level security and pgvector.

ZORRO — Consumer app shipped solo to iOS and Android | Jun 2026

- 700+ users with 50% D1 retention within three weeks of launch. React Native apps on both stores; semantic matchmaking with embeddings, pgvector retrieval and cross-encoder reranking; self-hosted vision-language photo moderation handling GDPR special-category data with no third-party processors; realtime chat over WebSockets.

EDUCATION & LANGUAGES

M.Sc. in Artificial Intelligence Research (official), UIMP / AEPIA — Specialisation in Machine Learning and Data Science | Sep 2026 – 2027, part-time alongside consulting work.

B.Sc. in Computer Science and Software Engineering, University of Oviedo — Oviedo, Spain | 2017 – 2021.

Languages: Spanish (native) · English (professional working proficiency; daily working language on international engagements).