Antonio Puerto Borreguero

AI Engineer · End-to-end AI systems — from the LLM to the firmware

Seville, Spain · Hybrid / Remote · info@antoniopuerto.com · (+34) 665 694 029
github.com/PuertOcho · linkedin.com/in/antonio-puerto-borreguero · antoniopuerto.com

PROFILE

Computer engineer who takes AI to production end-to-end: from the language model and agents, to the microservices that serve them, down to the hardware they run on. Working with AI since 2020, learning alongside the field as a self-taught practitioner: from academic foundations (genetic algorithms, MCTS, decision trees) to **three generations of personal assistants** — the latest, **tony**, an agentic platform with 1,000+ commits and 36 releases — whose techniques (agents, MCP, observability) I apply directly in my professional work.

My edge: I come from embedded systems (Renesas, STM32, FreeRTOS, CAN), so I don't stop at the cloud prototype — I understand the physical device end to end. Today I design and ship AI improvements into industrial workflows and build agents, **MCP** integrations and microservice architectures that run in production.

TECH STACK

AI & Agents	LLMs · Generative AI · Agents · RAG · MCP · function calling · LiteLLM · Langfuse · LangChain/LangGraph/LangSmith · NLU · STT/TTS · OCR (Tesseract) · prompt engineering · classic AI (genetic, MCTS, ID3/C4.5)
Backend & Cloud	Java · Spring Boot · C#/.NET (WPF) · microservices · Eureka · REST · hexagonal architecture · PostgreSQL · MariaDB · Redis · Kafka · Maven · JUnit · Jenkins
Embedded & HW	C · C++ · Renesas RX · STM32 · ESP32 · FreeRTOS · CAN · UART · SPI · I2C · industrial protocols
DevOps	Docker · Docker Compose · Nginx · Grafana · Prometheus · Linux · self-hosting · Git
Frontend	Vue.js / Quasar · TypeScript · Svelte · Ionic

EXPERIENCE

Firmware & AI Engineer · R&D

Apr 2023 – present · Seville, Spain

COBER (Control of Biomedical Embedded Robotics)

R&D in biomedical robotics and embedded systems, with growing ownership of applying AI to the product and the engineering workflow — backed by the agent engineering from my personal R&D.

- Designed and built the **internal platform for hardware-board and contract lifecycle management**: Spring Boot 3 + MariaDB + JWT, WPF (.NET) desktop client, barcode-scanner integration, Docker and Jenkins CI/CD — 7 modules, 40+ REST endpoints (310 commits).
- Added **automatic document and contract classification with OCR and AI** (Python/Tesseract): field extraction, per-manufacturer classification with confidence scoring and cross-validation.
- **Collaboration with MP Lifts (since Sep 2024)**: full migration of the proprietary CAN-based protocol (v2 to v3) across the lift controller and its diagnostic web UI — **9/9 modules verified on real hardware**, 12 monitoring services, 200+ commits across firmware (C/C++, FreeRTOS, Renesas RX) and frontend (Vue.js/Quasar, i18n in 6 languages).
- Cut the event-request response time **from 12 s to 0.7 s (-94%)** by optimizing timeouts and retries in the communications stack.
- Automated end-to-end verification with a test suite (PowerShell + Playwright) covering **130+ HTTP endpoints** against real hardware.
- Built **MCP servers** connecting LLM agents to real engineering tools: e2 Studio IDE (10 tools; extension **published on the Visual Studio Marketplace**) and an industrial CAN bus (17 tools, LLM chat with function calling).

Full Stack Developer

Sep 2021 – Apr 2023

Insinno España · extracurricular internship Mar–Sep 2021

- Built the **chemical-product management application for BASF** (the world's largest chemical company): Java/Spring Boot backend with microservices, REST APIs, PostgreSQL/Redis and Kafka messaging.
- Scalable product held to a multinational client's quality standards: hexagonal architecture, JUnit testing and Maven-based CI.

Sales Associate (Mountain team)

Sep 2020 – Apr 2021

Decathlon · Camas, Seville

- Customer service and technical advice, alongside university studies.

SELECTED PROJECTS (PERSONAL AI R&D)

tony — 3rd-generation agentic platform (2025–present)

Multi-provider (OpenAI, Anthropic, Google via LiteLLM, local Ollama fallback) with a custom MoE: parallel model voting with consensus and debate. Bio-inspired brain on LangGraph, 20+ MCP services, desktop client (Tauri + Svelte), continuous semantic evaluation (122 cases, LLM judge), observability (Langfuse, Grafana, Prometheus) and its own hardware (ESP32-S3). 1,000+ commits, 36 releases, 6,300+ tests.

puertocho-assistant — Voice AI assistant (2nd generation, 2025)

Spring Boot + Eureka microservices with an E2E voice layer: local Whisper, LLM-RAG intent manager with MoE voting, dynamic task decomposition and a TTS chain (Azure, Kokoro, XTTS-v2, F5/E2). Deployed on Raspberry Pi.

nuka — Assistant (1st generation, 2023–2024)

Where the series began, in the first wave of generative AI: GPT + Whisper + Azure TTS + DALL-E + RAG over Obsidian. Ionic/Angular app + Spring Boot server. 485 commits.

Custom MCP servers

e2studio-mcp (Renesas firmware build/debug; extension on the Visual Studio Marketplace), mp-can-mcp (industrial CAN bus + LLM chat), taiga-mcp-ms (project management).

EDUCATION

B.Sc. in Computer Engineering — University of Seville · 2016 – 2021 · AI foundations (2020): genetic algorithms (TSP), MCTS, decision trees (ID3/C4.5)

Artificial Intelligence Applied to Business — EDUCATIC GAP PUE · Nov 2022 – Feb 2023

Big Data Architecture — Fundación CONFEMETA · Jun 2021 – Sep 2022

LANGUAGES

Spanish (native) · English (B2 — professional working proficiency)