Jorge Domingues Nunes

Senior Product Engineer · Agentic Systems · MCP & RAG tooling

Pombal, Portugal · +351 914 764 120 · stamina.nunes@gmail.com · linkedin.com/in/stamina · github.com/staminna

SUMMARY

Senior product engineer, language and cloud agnostic, with 7+ years shipping production systems and 3+ years building agentic LLM applications end-to-end. I design and ship multi-agent workflows, MCP tooling, and RAG pipelines that connect models to live business data and run reliably in production. I work natively with AI coding agents (Cursor, Claude Code) as a daily part of how I build. Comfortable owning a product from prototype to deploy: backend, orchestration, data, infra, and the iteration loop with users — in a fast-moving SaaS environment.

FOCUS AREAS

AI agents & agentic workflows: LangGraph / LangChain multi-agent orchestration, planner/executor patterns, tool use, memory, and stateful workflows shipped to real users.

MCP: Designed, built, and published an MCP server (npm) exposing 29+ tools that let LLM agents drive a Directus backend in natural language.

RAG & LLM tooling: Hybrid search, vector stores (ChromaDB, Qdrant, Pinecone), per-user memory, prompt engineering, and evals — wired into production product surfaces.

AI coding agents: Daily driver in Cursor and Claude Code; build with agents, design MCP tools for them, and ship the workflows around them.

Data & PostgreSQL: PostgreSQL / PostGIS, schema design, async query layers, Redis caching — the data foundation under every product I've shipped.

Product engineering in SaaS: Independent end-to-end delivery in fast-moving SaaS environments: backend, frontend, mobile, infra, and the user iteration loop.

Polyglot & adaptable: Daily: Python, TypeScript, Dart. Comfortable picking up new languages and stacks (incl. Go); I optimise for the system, not the language.

EXPERIENCE

Founder & AI Solutions Architect — Exponential

Nov 2025 – Present

Building an agentic AI marketplace and orchestration platform end-to-end — backend, agents, mobile, and infra.

- **Agentic orchestration:** Built a multi-agent runtime on LangChain/LangGraph with planner/executor patterns, MCP-style tool integrations (Terraform, Docker, UI generation), and Ollama-powered local inference.
- **RAG & memory:** WebRAG over ChromaDB with hybrid search, per-user conversation memory, and prompt/response caching — ~90% LLM cost reduction on repeat workloads.

- **Multi-model routing:** Ollama (Mistral 7B, Llama 3.2, Qwen 2.5-Coder, DeepSeek-Coder) primary with Anthropic Claude fallback; token accounting and per-user wallets.
- **Product surface:** FastAPI backend (~200 endpoints), React dashboard, Flutter iOS/Android, embeddable chatbot widget (session persistence, rate limiting, domain allowlisting).
- **Data:** PostgreSQL for app data, ChromaDB for vector store, Redis for sessions and caching.
- **Infra:** Kubernetes with custom CRDs for agent pods, Helm, Kustomize, Terraform across Docker/AWS/GCP/Azure, Ansible for post-deploy config.
- **Built with AI:** Day-to-day pair-programming with Cursor and Claude Code across backend, infra, and mobile.

Principal AI Engineer — Keyprog

Jan 2024 – Oct 2025

LLM-powered e-commerce SaaS & MCP tooling

Remote

Shipped a production LLM-powered e-commerce platform with agent-driven content and ops workflows.

- **MCP server (published on npm):** Designed and shipped @staminna/directus-mcp-server — 29+ tools so AI coding agents and product agents can query, mutate, and reason over a Directus backend via natural language.
- **RAG & prompt engineering:** Retrieval pipelines over product and content data, structured prompting to reduce hallucinations on catalog tasks, MLflow-based evals.
- **Data:** PostgreSQL as the source of truth, Redis for caching and queues, schema and access design via Directus.
- **Delivery:** Independent end-to-end shipping — FastAPI services, Docker, GitHub Actions CI/CD; mentored engineers on LLM integration patterns and prompt/eval conventions.

Senior AI/ML Engineer — Co2Offset.ai

Apr 2023 – Dec 2024

Forest carbon-capture SaaS

Pombal, Portugal

Co2offset is a SaaS that measures carbon captured in forests from satellite and ground data, rewarding landowners for protection.

- **ML pipeline:** Computer-vision models in production behind FastAPI; partnered with ESA on Python microservices for geospatial data.
- **Outcome:** Led pipeline audit with Bureau Veritas, securing scientific reliability certification that unlocked €3.5M of business.
- **Data:** PostgreSQL with PostGIS for geospatial queries, MongoDB for unstructured layers, Redis for caching.
- **Platform:** Kubernetes-deployed Python services with auto-scaling and health probes, Terraform multi-env, Next.js/React frontend, Stripe payments.

Technical Consultant — European Central Bank Mar 2020 – Mar 2023

Web platform & multilingual search at scale *Frankfurt, Germany*

- Main Website ECB design team.
- Drove a multilingual website redesign serving millions of articles across 24 languages.
- Built an Elasticsearch-based search engine indexing internal CMS and public site content in 24 languages, serving millions of pages.
- Owned release engineering, end-to-end testing, and security hardening (CSP, SSL, Apache Server Config).

Senior Drupal Developer — European Commission Apr 2018 – Jan 2020

Public-facing platforms *Brussels / Luxembourg*

- Main-site redesign team member
- Designed the relational data model for tens of thousands of users;
- CI/CD with Jenkins, automated tests with Jest.
- Led several Drupal migrations
- Built Github actions CI/CD Pipelines.

SELECTED PROJECTS

Word-Wide Events (Go and Flutter) - A two-part app for browsing upcoming events from across the world, pulled from license-free sources only.

eventscraper — Go backend. Scrapes Eventbrite, Songkick and Luma, caches everything in SQLite with TTL-based stale-while-revalidate, exposes a JSON HTTP API, and includes an image-proxy that fixes CORS / hotlinking / missing Content-Type problems on the public CDNs.

eventscraper_app — Flutter client (iOS / Android / Web). Responsive 1 / 2 / 3-column grid, persistent search, filter sheet (city, category, source, date range), interactive OpenStreetMap view of every geocoded event, and a polished event detail screen.

RAG-PDF (LangGraph + LangChain) - Retrieval-augmented PDF Q&A built as a LangGraph state machine over LangChain retrievers.

Sentiment Analyzer (Go) - A command-line tool for analyzing the sentiment of text inputs using the Hugging Face API.

SELECTED STACK

Agents & LLM tooling: LangChain, LangGraph, MCP, OpenAI, Anthropic, Gemini, Ollama (Mistral / Llama 3 / Qwen / DeepSeek) Deepagen

AI coding agents: Cursor, Claude Code — daily driver across backend, infra, and mobile work

Retrieval: ChromaDB, Qdrant, Pinecone, hybrid search, embeddings

Languages: Python, TypeScript, Dart, Node, Go

Backend & data: FastAPI, async, PostgreSQL / PostGIS, Redis, MongoDB

Frontend / mobile: React, Next.js, Atmos, Flutter / Dart

Infra & ops: Kubernetes, Helm, Terraform, Docker, GitHub Actions, AWS / GCP / Azure

LLMOps: Langfuse, MLflow, prompt evals, observability

CONTINUOUS LEARNING

- LLM fine-tuning, RAG architectures, prompt engineering, vector databases.
- Agentic workflow patterns (planner/executor, tool use, multi-agent coordination, skills).
- Terraform & Kubernetes operations; Flutter cross-platform mobile architecture.

- CEO and Chairman of Co2Offset.ai recommendation letter.