Emanuele Czofei

AI Product Engineer

Production LLM systems · RAG · Agents · LLM evaluation · Python + TypeScript
Gipuzkoa, Spain (CET) · Open to Remote (EU / UK / US-overlap timezones) · EU work authorization
www.emanuelecz.dev · github.com/emanuelecz · linkedin.com/in/emanueleczofei · emanuelecz06@gmail.com

SUMMARY

AI Product Engineer specializing in production LLM systems: RAG pipelines, agent orchestration, and the evaluation harnesses that keep them reliable. Shipped a Spanish regulatory RAG assistant with retrieval validated at **Hit@8 0.98 / MRR 0.955** against a human-labeled golden set, plus a meta-evaluated LLM-as-judge pipeline that runs as a CI quality gate. Comfortable across the full stack with Python/FastAPI and TypeScript/Next.js.

TECHNICAL SKILLS

AI / LLM: RAG pipelines, LangChain, LangGraph (agent orchestration, state machines), Claude API (Anthropic), OpenAI API, prompt engineering, structured outputs, LLM evaluation (golden sets, LLM-as-judge, judge meta-evaluation, TPR/TNR validation), VoyageAI embeddings & reranking, cross-encoder reranking, ChromaDB, pgvector, hybrid retrieval (BM25 + dense)

Backend: Python, FastAPI, Node.js, Express, PostgreSQL, MongoDB, SQLAlchemy, Redis, Celery, REST API design, async subprocess execution, Zod validation

Frontend: TypeScript, Next.js, React, Tailwind CSS, streaming UI (server-sent events), internationalization (ES/EN)

Infra / DevOps: Docker, Docker Compose, GitHub Actions (CI/CD with automated eval gates), Vercel, AWS S3, Git

Languages: Italian (native), Romanian (native), Spanish (intermediate), English (professional working proficiency)

SELECTED PROJECTS

Spanish Occupational-Safety RAG Assistant (INSHT/NTP corpus)

Live demo & case study: emanuelecz.dev

Python · FastAPI · LangChain · ChromaDB · VoyageAI voyage-3 + rerank-2.5 · Claude API · Next.js · Tailwind

- Built an end-to-end RAG system over the Spanish INSHT/NTP occupational-safety corpus: hybrid BM25 + dense retrieval (EnsembleRetriever), cross-encoder reranking, and grounded answer generation with source citations.
- Validated retrieval against a 50-item human-labeled golden set: **Hit@8 0.98, MRR 0.955, Precision@8 0.69**.
- Designed a three-layer evaluation harness: retrieval metrics, an LLM-as-judge scoring faithfulness/relevance/completeness, and meta-evaluation of the judge itself against human labels.
- Diagnosed a structurally broken judge (TNR 0.00 from threshold miscalibration and missing reference answers) and rebuilt it with reference-aware prompting and fault enumeration, revalidating to faithfulness **TPR/TNR 1.00/1.00** and completeness **1.00/0.875**.
- Evaluations run in GitHub Actions CI and fail the build on retrieval- or answer-quality regressions.

Triage (RepoDoctor): Automated Code-Diagnosis Agent

Python · LangGraph · OpenAI API · Next.js · React · TypeScript · Docker Compose

- Built a full-stack agent that clones a target repo, auto-detects its stack (Python/Node + package manager), installs dependencies, and runs its own test/build/lint suites to identify and explain failures.
- Implemented the agent as a LangGraph state machine (clone → detect → install → diagnose → report) orchestrating an LLM with custom tools and a grounding-focused prompt that bases every claim on real command output.
- Engineered safe execution: each repo runs in an isolated sandbox with async subprocess execution and per-command timeouts; a Next.js/TypeScript frontend streams progress live via server-sent events.

E-Commerce Backend API

Node.js · Express · MongoDB · Stripe · Docker

- RESTful API for authentication, product catalog, and order management with JWT and role-based access control; Stripe payments with webhook-driven order lifecycle, containerized with Docker Compose.

Async Document Conversion Service

React · Celery · Redis · AWS S3

- File-conversion tool with asynchronous Celery/Redis task processing, S3-backed storage, and client-side job-status polling.

EXPERIENCE

AI & Full-Stack Engineer Intern, Gaddr

Sep 2025 – Dec 2025

- Shipped authentication pages for an AI-driven web application (Next.js / TypeScript).
- Implemented Zod request/response validation for an external API integration.

EDUCATION

Self-Taught Software Engineering

2024 – Present

- Structured, intentional curriculum in production LLM systems: RAG architecture, evaluation discipline (Hamel Husain methodology), FastAPI/PostgreSQL backend engineering, and agent orchestration. Primary text: Chip Huyen, *AI Engineering*.