Rocco Fortuna

AI ENGINEER
STARTUP FOUNDER

AI Engineer & Full-Stack Developer skilled in turning research models into production systems. Experienced in LLM pipelines, evaluation frameworks, and scalable cloud deployment that deliver measurable product impact.

Contact

☎ +39 327 55 06 955

✉ roccofortuna98@gmail.com

in linkedin.com/in/roccofortuna/

Highlights

- Strong leadership
- Ownership & Initiative
- Time Management

Work Experience (1/2)

Apr 2025 - Present

AI Engineer

Indigo.ai - Customer Support Chatbots

- Built **LLM-powered analytics pipelines** to extract subscription intent, churn reasons, and customer satisfaction drivers from chat data, **enabling retention and upsell** strategies for enterprise clients.
- Performed competitor analysis of evaluation frameworks (LangChain, Guardrails.ai, UpTrain, DeepEval) and **developed LLM evaluation and guardrail services** to assess factuality, reliability, and policy compliance.
- **Improved speech-to-text latency by ~33%** by benchmarking 20+ STT providers and integrating a low-latency EU endpoint, validated in production.
- **Designed** and shared a **serverless job orchestration toolkit using Terraform and Google Cloud Run**, enabling one-click setup of scheduled extraction and evaluation pipelines used across the team.
- **Enhanced CI/CD workflows** by implementing automated linting, merge checks, and consistent package management, reducing integration errors and improving release stability and overall Developer Experience.
- Supported internal AI knowledge-sharing programs by preparing and presenting sessions on LLM evaluation, speech benchmarking, and emerging AI trends.

Jun 2025 - Present

Lead Engineer | part-time

Qdrant - Vector Search

- **Led development** of a hosted conversational AI platform extending **open-source project by Microsoft** into a SaaS.
- **Designed the system architecture** for scalable, retrieval-augmented conversational interfaces supporting multiple LLM, embedding, and retrieval providers.
- Built the backend for **automated content ingestion and retrieval**, allowing websites to integrate conversational access to their structured and unstructured data.
- **Presented** the live MVP platform *nlweb.one* **at Qdrant Vector Space Day 2025 in Berlin**, directly following Microsoft's announcement, creating the first distribution channel for managed agents through Microsoft NLWeb.

Skills

Machine Learning & AI
Deep Learning, NLP, Generative AI, Probabilistic Modeling, Recommendation Systems, Retrieval-Augmented Generation (RAG), LLM Fine-tuning, Evaluators & Guardrails.

ML Frameworks & Libraries
PyTorch, TensorFlow, Scikit-learn, NumPy, Pandas, Hugging Face Transformers/Datasets, LangChain, Qdrant, ONNX Runtime.

MLOps & Optimization
Model Quantization, Data Pipelines, Experiment Tracking, Model Evaluation, Continuous Training (CT), Asynchronous Inference, Parallelization, Batch Processing, Prompt Engineering, API Design for LLM Systems.

Infrastructure & Deployment
Docker, Kubernetes, Redis, Terraform, Google Cloud Platform, IBM Cloud, Azure, CDKTF, Pulumi, CI/CD Automations.

Backend & Data Engineering
Python (FastAPI, Sanic), SQL, RESTful APIs, Async I/O, Data Ingestion, Message Queues, Caching, WebSockets, Vector Databases.

Fullstack & Product Integration
Next.js, React, TypeScript, HTML/CSS, Supabase, Looker Studio, Stripe API, HubSpot API, Printful API, OpenAI Embedding Integration, Analytics & Tracking Tools.

Programming & Data Formats
Python, Go, TypeScript, Bash, R, C; Protobuf, JSON, YAML, SQL, Markdown.

Work Experience (2/2)

Apr 2023 - Apr 2025

Co-Founder, CEO & CTO

Dobles - AI Apparel Startup

- Acquired 1,000+ contacts, facilitated 25,000+ AI-generated graphics, and **generated €4,000+ revenue** validating early market demand.
- **Built and deployed an AI-powered platform** from the ground up, **leading a team of 3 developers** to deliver **30+ production-grade API endpoints**, handling hundreds of design requests monthly.
- **Integrated 10+ APIs**, including HubSpot (CRM & automation), Printful (fulfillment), Stripe (payments), OpenAI (text-to-image), and Google Cloud services.
- **Provisioned and managed infrastructure with Terraform**, implementing **CI/CD pipelines** for one-click deployments across **10+ cloud services** (GKE, Cloud SQL, Storage, Redis) in **dev/staging/prod**, ensuring reproducible environments and **cutting infrastructure costs by 50%**.

Mar 2022 - Apr 2023

Machine Learning Engineer

Pieces for Developers - Developer Productivity Tool

- **Expanded** code snippets **dataset by 25%**, increasing coverage from 37 to 43 programming languages, and **trained a Programming Language Classification model**, improving from 78% to 85% accuracy.
- Integrated classification, smart suggestions and **vector search** for code snippets for users.
- Introduced **DVC** and **CML** practices, streamlining data preprocessing, model training, and deployment, reducing iteration time by 50%.

Education

Sept 2019 - Mar Sept 2020

MSc Artificial Intelligence

The University of Edinburgh - Edinburgh, UK

- Graduated with distinction.
- Thesis **in collaboration with Amazon** – Researched the Conditional Variational Autoencoder (**CVAE**) architecture to enhance recommendation diversity.

Sept 2019 - Mar 2022

BSc Software Science & Engineering

Eindhoven University of Technology - Eindhoven, NL

- Bachelor End Project: Developed a PoC for the Netherlands' Healthcare Information System with **enhanced privacy**.