ENRICO ZANETTI

AI Engineer

ezan@enricozanetti.dev | EnricoZanetti | Enrico-Zanetti | enricozanetti.dev

WORK EXPERIENCE

AI Engineer at HPA | High Performance Analytics, a division of Terranova Software — **Trento**, Italy

Python | FastAPI | Django | LangGraph | HuggingFace | Celery | RabbitMQ | Docker | React | Linux — June 2024 – Present

- **Developed and maintained the backend** of an enterprise LLM/RAG chatbot serving **5,000+ daily queries** across **10+ client organizations**, including Django REST API design, task orchestration, PostgreSQL management, and Linux server administration. System maintained **99.9% uptime**.
- Designed and implemented a secure enterprise SSO system (OAuth2.0, OIDC, JWT, Microsoft Entra ID) with fine-grained tenant-level access control, **eliminating unauthorized access vectors** present in standard OIDC flows and enabling onboarding of enterprise clients with strict compliance requirements.
- Re-architected the async task processing system from a shared 2-queue setup to a per-tenant isolated queue pool with round-robin worker scheduling, reducing worst-case task wait time by ∼**80% under peak load** and guaranteeing fair resource allocation across all clients regardless of individual burst traffic.
- Developed LangGraph AI Agents with tool-using capabilities (web search, retrieval, custom tools), automating complex enterprise workflows and reducing manual processing time by ∼**60% on targeted tasks**, directly contributing to client ROI justification.
- Conducted **LLM evaluation research** across open-source and proprietary models (GPT, Claude, Mistral, LLaMA), producing structured trade-off analyses on latency, cost, and hallucination rate that directly informed model selection for **5+ production deployments**.
- Designed and built an end-to-end Analytics dashboard (React + Django) for client-facing usage insights, query volume, response quality trends, model performance over time, and per-user interaction breakdowns; enabling **data-driven ROI reporting** across all enterprise accounts.
- Mentored a University of Trento student team through a tech challenge, guiding AI engineering best practices; the team delivered a professional-level consultancy report for an enterprise client, with 2 team members subsequently hired by the company.
- Represented the company at industry events including a talk on AI Agents for enterprise automation at **Confindustria Trento** (100+ professionals).

EDUCATION

Natural Language Processing with Deep Learning - XCS224N — **Stanford University**

Successfully Completed — Sept 2024 - Nov 2024

Core topics: neural NLP, question answering, machine translation, PyTorch model design, word representations (Word2Vec, GloVe), RNN language models, Transformer pre-training and fine-tuning.

Master's Degree in Data Science — **University of Trento**

Grade: 108/110, GPA: 3.9/4.0 — Sept 2022 - Sept 2024

- **Relevant Coursework:** Advanced Deep Learning & ML, Statistical Learning, Data Visualization, Big Data Technologies, IoT, Linear Algebra.
- **Erasmus Exchange Program at Utrecht University (Aug 2023 - Feb 2024):** Advanced Machine Learning, Pattern Recognition, Deep Learning.
- **Thesis:** "RAG-based Personalization and LLMs Evaluation for AI Chatbots", implementation and evaluation of retrieval-augmented generation for chatbot personalization.

Bachelor's Degree in International Studies — **University of Trento**

Grade: 101/110, GPA: 3.2/4.0 — Sept 2017 - Feb 2021

PROJECTS

MoodStream - Real-Time Facial Emotion Detection — **IoT**

Python | PyTorch | TFLite | OpenCV | MQTT | InfluxDB | Grafana | Docker — github.com/EnricoZanetti/moodstream

- Designed and trained a compact CNN (3 conv blocks) on FER-2013 (~30,000 images, 6 emotion classes) with PyTorch; quantised and exported to TFLite (float16) for efficient CPU inference, retaining a legacy deployment path on OpenMV Cam H7.
- Built a real-time detection loop using OpenCV for face localisation and TFLite for per-frame emotion classification, publishing classified events over MQTT on each detection interval.
- Orchestrated a fully containerised telemetry stack (Docker Compose): MQTT (Mosquitto) → Node-RED → InfluxDB → Grafana, delivering a live emotion-stream dashboard with sub-second end-to-end latency.

GAIA AI Agent — **AI Agent**

LangGraph | Python | HuggingFace | RAG | Web Search — EnricoZanetti/gaia-ai-agent