Balazs Thomay

+31643589332 • balazs.thomay@gmail.com • balazs-thomay
balazsthomay

Summary

I've spent the last year building ML workflows, computer vision prototypes, small LLM finetunes, agentic systems, and some streaming and distributed components. Most of my work uses PyTorch, scikit-learn, LangGraph, or the OpenAI Agents SDK.

Technical Skills

ML/AI: PyTorch, scikit-learn, YOLO, LoRA/QLoRA, PEFT, MediaPipe, time-series forecasting

LLM/Agents: finetuning workflows, LangChain/LangGraph, OpenAI Agents SDK, MCPs, vector DBs

Systems: Python, FastAPI, Kafka, Redis Streams, PostgreSQL, Docker, AWS S3, GitHub Actions, WebSockets

Key Projects

LLM Agents & Automation

Multi-Agent Code Review System (code)

- Built an automated PR reviewer using a multi-agent (5) LLM pipeline with RAG (82 curated patterns) to detect bugs, security issues, and style violations.
- Achieved high bug/security detection on BugsInPy and real Python CVEs using location-based and semantic evaluation.
- Integrated as a GitHub Action that blocks merges on critical findings and runs across multiple repositories.

Cloud ML Infrastructure

Fraud Detection MLOps Pipeline (code)

- Implemented a full training → packaging → deployment workflow for fraud detection (Random Forest).
- Automated model comparison by F1 score, saved artifacts to S3, and triggered API reload when a better model appeared.
- Containerized inference server with FastAPI and other 2 services; CI/CD built with GitHub Actions.

Distributed Systems

Real-Time Event Streaming Platform (code)

- Built a multi-service event pipeline: Kafka (durable input log) → Redis Streams (per-user buffering) → Postgres (filter storage) with a WebSocket broadcaster.
- Containerized with Docker for local multi-service orchestration. Prometheus metrics, logging, and Locust for load testing

Computer Vision

Knife Safety Monitor (code)

- Developed a real-time CV prototype that combines MediaPipe hand tracking, a fine-tuned YOLO knife detector, and basic CoreML depth analysis.
- Designed as a research-level safety system; no mobile or edge deployment yet.

NLP / Finetuning

LLM Fine-tuning (code)

- Fine-tuned a LLaMA model using QLoRA (bitsandbytes + PEFT) for controlled-style text generation.
- Wrote training and inference scripts; small proof-of-concept demonstrating data preprocessing, adaptor training, and quantized inference.

Time series Forecasting

Neuron Voltage Forecasting (code)

- Implemented a CNN-LSTM-based forecasting model for neuron voltage traces.
- Conducted exploratory error analysis and built reproducible experiments.

Education

International Business Administration — **Erasmus University Rotterdam**
Bachelor of Science, Rotterdam, Netherlands — *2020–2023*