

Jimi Vaubien

Lead AI Engineer & Consultant

Senior ML Engineer with 7+ years building production AI systems end-to-end. Led GenAI platform serving 4 enterprise clients at Prometee, managing a team of 3 engineers. Deep expertise in LLM agents, RAG systems, human-in-the-loop workflows, model training, and data pipelines. Creator of open-source tools with 1,300+ GitHub stars.

jimi.vaubien@protonmail.com ■ jimivaubien.me ■ github.com/jimzer ■ linkedin.com/in/jimi-vaubien ■ youtube.com/@bitswired



01 Experience

Prometee

Founding + Lead AI Engineer

2024 - Present

Lausanne, CH

Enterprise GenAI platform for regulated industries (finance, accounting, trading) with client-isolated Azure deployments.

YourGPT Platform

GenAI platform deployed by Terraform into each client's own Azure subscription: private networking inside it, microservices backend, config-driven SPA assembly.

- ▶ 4 production clients (10-500 users each)
- ▶ One deployment per client subscription, private networking inside

Trade Settlement Agent

Stateful email workflow automation with multi-turn conversation management and intelligent escalation.

- ▶ 10h/week back-office savings
- ▶ Hundreds of counterparty interactions processed daily

Customer Support LLM Service

Llama 3 8B on 2x GPU nodes for HA with vLLM serving, intent routing, 68% autonomous resolution, 32% to human agents.

- ▶ 20 QPS peak, p99 1.8s, 80% peak GPU utilization
- ▶ Response time 1h → 45s
- ▶ 3+ hours saved per human agent per day

Financial Audit Auto-Tagging

AI agent automating email monitoring, document extraction, and audit software integration via Microsoft Graph API.

- ▶ 2h saved per agent per day (25% of workday returned)
- ▶ Continuous processing: plain text, PDF, images, scanned docs

Advanced RAG Systems

Hybrid retrieval (BM25 + vector + reranking) with SQL agent and multi-turn agentic coordination.

- ▶ 90% recall vs 30% SQL-only baseline
- ▶ 100s GB corpora at 2 clients (dozens to hundreds of employees each)

AI Safety & Guardrails Framework

Prompt injection classifier (DeBERTa), PII detection, and a red teaming pipeline.

- ▶ 99.2% injection detection, 0 PII leaks, 2.1% false positive rate
- ▶ 150ms avg overhead, passed internal audit

Proxinea

Senior Data Engineer

Dec 2023 - Mar 2024

Zurich, CH

Modernized data infrastructure supporting pricing optimization and churn prediction models.

Data Platform Migration

Transformed manual SQL pipelines into automated DBT/Dagster architecture with semantic versioning, data quality testing, and unified orchestration of SQL and Python ML nodes.

- ▶ Eliminated manual pipeline failures and data corruption
- ▶ 10s to 100s GB pipelines migrated to fully automated scheduled execution
- ▶ CI/CD with semantic versioning enabling safe deployments

Visium Technologies SA

Senior Machine Learning Engineer / Tech Lead

2018 - 2023

Lausanne, CH

End-to-end AI solutions for enterprise clients across banking, pharma, and e-commerce.

Unsupervised Fraud Detection (Banking)

Anomaly detection system with explainable PDF reports: isolation forest + ML ensemble, POC to production.

- ▶ 30% precision (30x lift), 1000s transactions/day
- ▶ 2+ years in production, passed security audits

Real-Time Sound Event Detection

Edge (MobileNet INT8 on Raspberry Pi) + cloud (ResNet) for gunshot/accident detection.

- ▶ Client won a Swiss startup competition with this PoC

Pharma Data Engineering

AWS data lake with Spark, DBT, and fine-tuned BERT for Real World Evidence. PII removal, NER, sentiment analysis.

- ▶ 100s GB to TB data, PII removal, NER, sentiment analysis

B2B Recommender System

Two-tower neural network empowering B2B sales reps with offline batch ranking engine.

- ▶ +15% AOV uplift
- ▶ 10K+ products per client

B2C Recommender System

Neural matrix factorization replacing broadcast notifications with personalized delivery.

- ▶ 4x CTR uplift, 100K+ users

Cyclist Detection (Sports)

MaskRCNN instance segmentation for automated finish-line detection in dense peloton scenarios.

- ▶ 80% standard / 55% hard scenarios
- ▶ Architecture reused for gymnastics detection

Université de Lausanne

Main Teaching Assistant & Full Stack Developer

2018 - 2025

Lausanne, CH

Main TA for the Python course (part-time, 2018 to 2019), then freelance development for the faculty research projects until 2025.

Job Seeker Platform

Full-stack recommendation system with A/B testing framework deployed on GCP, built for academic research on job seeker behavior.

- ▶ Adopted by official Swiss state administrations
- ▶ Cohort-based A/B testing framework built from scratch

NEC Labs

Research Intern, Master Thesis

2019

Princeton, USA

Unsupervised deep learning research for manufacturing defect detection.

Unsupervised Defect Detection

One-Class SVM, Autoencoders, and GANs for defect recognition with no positive training examples available.

- ▶ 6/6 Master thesis grade
- ▶ Supervised by Eric Cosatto (NEC Labs)

Schlumberger Doll Research

AI Research Intern

2017

Boston, USA

Geological simulation acceleration using deep generative models.

Geological Simulation Acceleration

Conditional GANs with chunk-based positional conditioning replacing slow physics-based simulation for soil composition generation.

- ▶ 10x speedup vs physics-based simulation
- ▶ Predictions integrated into downstream geological modules

Sea Level Time Series Prediction

RNN/LSTM for reconstructing historical sea levels from soil composition measurements at varying depths.

TasteHit

Full Stack Developer

2016 - 2017

Lausanne, CH

eCommerce recommendation engine startup, MassChallenge finalist, subsequently acquired.

Embeddable Real-Time Search Widget

React widget included via script tag with recommendation-powered live search ranking.

- Deployed to 10+ production eCommerce clients
- Company became MassChallenge finalist and was acquired

02 Toolbox

Languages Python 7y · TypeScript 5y · SQL 6y · Rust 3y

Cloud Azure · GCP · AWS · Cloudflare

Data Engineering DBT · Dagster · Apache Beam

Web FastAPI · React · Node.js / Bun · HTMX · Axum

ML / AI PyTorch · TensorFlow / Keras · Hugging Face · LLMs · RAG Systems · Agents / Multi-Agent · Vector Databases

DevOps Docker · Terraform · Kubernetes

Databases PostgreSQL · Redis · MongoDB

03 Open source

RustGPT 699 ★

ChatGPT UI in Rust + HTMX. Zero JavaScript framework.

Website-to-Knowledge-Base 79 ★

RAG-powered Q&A system converting websites into searchable knowledge bases with source citations.

Kiru 8 ★

High-performance text chunking for RAG. Rust core, Python bindings. 4,370 MB/s (4,000x faster than LangChain).

LazyCodr 45 ★

AI-powered CLI automating PR descriptions, commit messages, and documentation.

FuseAI 293 ★

Self-hosted OpenAI web app with multi-user Docker deployment.

Demos 62 ★

Source code for my YouTube tutorials: Effect, Python, MCP, AI agents.

FoldCMS 23 ★

Type-safe CMS with Effect framework. SQLite-backed, multiple format support.

04 Education & the rest

Education Master in Data Science, EPFL 2019 (GPA 5.16/6) · Bachelor in Communication Systems, EPFL 2016

Certifications AZ-104 Azure Administrator · AZ-900 Azure Fundamentals · MQ Academy Strategic Thinking · Coursera Build Better GANs

Competitions 1st place (Cisco Challenge), LauzHack 2017 · 1st / 63, Machine Learning Competition 2017 · 15th / 160, Data Visualization 2017 · 1st place (Zühlke Robochallenge), Start Hack 2016 · 135th / 1900, IEEE Extreme 2015

Teaching Main TA, Python (UNIL, 2018 - 2019) · TA, C++ (EPFL, 2018) · TA, Computer Networks (EPFL, 2016 - 2018) · TA, ICC (EPFL, 2016) · 61 tutorials (YouTube)

Talks AI Tinkerers Lausanne 2024 · Webmardi 2023 · LauzHack Workshop 2024

Details French, English · French · Swiss C Permit