CONTACT

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

HIGHLIGHTS

1333 GitHub Stars
7 Years In Industry
4 Published Packages
3 Competition Wins
2x Founding Engineer

EDUCATION

Master – Data Science
EPFL
2019 | GPA: 5.16/6

Bachelor – Communication Systems
EPFL
2016

CERTIFICATIONS

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

MISC

Nationality: French
Swiss Residency: **C Permit**
Languages: French, English
Available for consulting

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.

PROFESSIONAL EXPERIENCE

Founding + Lead AI Engineer

2023 - Present · Lausanne, CH

Prometee

Enterprise GenAI platform for regulated industries (finance, accounting, trading) with client-isolated Azure deployments. Team: 3 engineers + 2 interns.

KEY PROJECTS

YourGPT Platform
Multi-tenant GenAI platform on Azure with Terraform IaC, private networking, microservices backend, and config-driven SPA assembly.
- 4 production clients (10-500 users each)
- Zero data crossing tenant boundaries via private networking

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

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

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 production clients

High-Throughput Customer Support LLM Service
Llama 3 8B on 3x A100 GPUs with vLLM serving and intent routing – 68% autonomous resolution, 32% to human agents.
- 60 QPS, p99 1.8s, 84% GPU utilization
- Response time 4h → 45s
- 3+ hours saved per human agent per day

AI Safety & Guardrails Framework
Prompt injection classifier (DeBERTa), PII detection (Presidio), and automated red teaming pipeline.
- 99.2% injection detection, 0 PII leaks, 2.1% false positive rate
- 150ms avg overhead – passed internal audit

SKILLS

Languages
Python (7y) · TypeScript (5y) · SQL (6y) · Rust (2y)

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

Cloud
Azure [AZ-104] · GCP · AWS · Cloudflare

DevOps
Docker · Terraform · Kubernetes

Data Engineering
DBT · Dagster · Apache Beam

Databases
PostgreSQL · Redis · MongoDB

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

COMPETITIONS

LauzHack 2017 – 1st place (Cisco Challenge)
Machine Learning Competition 2017 – 1st / 63
Data Visualization 2017 – 15th / 160
Start Hack 2016 – 1st place (Zühlke Robochallenge)
IEEE Extreme 2015 – 136th / 2000 (5th Switzerland)

SPEAKING

AI Tinkerers Lausanne 2024
Webmardi 2023
LauzHack Workshop 2023

Senior Data Engineer

2023 · Zurich, CH

Proxinea

Modernized data infrastructure supporting pricing optimization and churn prediction models. Team: 1-2 engineers.

KEY PROJECTS

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

Senior Machine Learning Engineer / Tech Lead

2018 - 2023 · Lausanne, CH

Visium Technologies SA

End-to-end AI solutions for enterprise clients across banking, pharma, and e-commerce. Junior ML Engineer (part-time 2018) → Senior / Tech Lead (2022). Team: 1-4 engineers per project.

KEY PROJECTS

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

B2B Recommender System
Two-tower neural network empowering B2B sales reps with offline batch ranking engine.
- +20% AOV uplift (top of industry benchmark 15-22%)
- 10-90K products, 100K-1M orders per client

Real-Time Sound Event Detection
Edge (MobileNet INT8 on Raspberry Pi) + cloud (ResNet) for gunshot/accident detection.
- 40-80% precision vs ShotSpotter 10-11%
- Sub-10s offline on Raspberry Pi, client won competition

B2C Recommender System
Neural matrix factorization replacing broadcast notifications with personalized delivery.
- 4x CTR uplift, 100K+ users
- PoC accepted into client production

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
- Warehouse design, life-critical quality requirements met

Cyclist Detection (Sports)
MaskRCNN instance segmentation for automated finish-line detection in dense peloton scenarios.
- 95% standard / 55% hard scenarios
- Architecture reused for gymnastics detection

OPEN SOURCE

Packages

Kiru 8 ★
High-performance text chunking for RAG — Rust core, Python bindings. 4,370 MB/s (4,000x faster than LangChain).

FoldCMS 23 ★
Type-safe CMS with Effect framework — SQLite-backed, multiple format support.

LazyCodr 45 ★
AI-powered CLI automating PR descriptions, commit messages, and documentation.

Projects

RustGPT 699 ★
ChatGPT UI in Rust + HTMX — zero JavaScript framework.

FuseAI 293 ★
Self-hosted OpenAI web app with multi-user Docker deployment.

Website-to-Knowledge-Base 79 ★
RAG-powered Q&A system converting websites into searchable knowledge bases with source citations.

Demos 62 ★
Source code for Bitswired YouTube tutorials — Effect, Python, MCP, AI agents.

Main Teaching Assistant & Full Stack Developer

2018 - 2025 · Lausanne, CH

Université de Lausanne

Principal TA for Python curriculum and freelance full-stack development for research projects. Team: Led 1 additional developer.

KEY PROJECTS

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
- Hundreds to thousands of active users
- Cohort-based A/B testing framework built from scratch

Research Intern — Master Thesis

2019 · Princeton, USA

NEC Labs

Unsupervised deep learning research for manufacturing defect detection.

KEY PROJECTS

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)

AI Research Intern

2017 · Boston, USA

Schlumberger Doll Research

Geological simulation acceleration using deep generative models.

KEY PROJECTS

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.

Full Stack Developer

2016 - 2017 · Lausanne, CH

TasteHit

eCommerce recommendation engine startup — MassChallenge finalist, subsequently acquired.

KEY PROJECTS

Embeddable Real-Time Search Widget
React widget included via script tag with recommendation-powered live search ranking, deployed to 2-3 production eCommerce clients.
- Deployed to 2-3 production clients
- Company became MassChallenge finalist and was acquired