Simone Montali

Senior Machine Learning Engineer

Zurich, CH
☎ *+41 0762279718*
✉ *sim.montali@gmail.com*
in *linkedin.com/in/simonemontali*

Summary

Senior ML Engineer and Tech Lead with deep expertise in **Large Language Models**, **Machine Learning**, and **Platform Infrastructure**. Proven track record at **Meta** and **Google** leading 0-to-1 product launches, scaling inference to millions of users, and architecting cost-saving data pipelines. Specialist in bridging the gap between research prototypes and production reliability.

Experience

July 2025 – Present

Meta Reality Labs, *Senior ML Software Engineer (Tech Lead)*, Zurich

- **Tech Lead for a v-team** driving the platformization of LLM inference for wearable devices (Ray-Ban Meta). Engineered the transition from a prototype supporting tens of users to a robust production environment serving **millions of users**.
- **Architected a cost-efficient data generation pipeline** for foundation models using public data sources. Reduced data collection costs by an estimated **$4M** while cutting data rejection rates by **8x** (40% → 1%) via semantic sampling.
- **Designed, launched and tech-lead DRE**, a scalable platform that reduced the build time for custom data engines from months to weeks. Successfully secured adoption across multiple internal teams.
- **Automated the end-to-end ML training stack (CI/CD) of a CV model for VR**, transitioning from manual canary training to a fully automated pipeline with daily retraining. Reduced lead time for new use cases from **months to minutes**.
- **Spearheaded the organizational ”AI for Productivity” initiative**, driving engineering adoption of AI tooling from **64% to 83%** in three months. Organized and implemented an AI-assisted fixathon with 100+ participants, automating code quality changes.
- **Co-authored a paper submitted to CVPR** regarding novel data methodologies, while managing complex cross-functional legal and privacy compliance for large-scale data usage.

May 2024 – June 2025

Google, *ML Software Engineer (YouTube Shopping)*, Zurich

- **Served as Technical Point of Contact (POC)** and primary on-call engineer for critical YouTube Shopping events (Black Friday, Cyber Monday), ensuring system stability and zero downtime during peak global traffic.
- **Boosted worldwide product CTR by ~3%** within 6 months by optimizing recommendation algorithms.
- **Led a cross-org initiative** on LLM embeddings for Shopping videos, aligning stakeholders across multiple geographies to unify video understanding signals.

May 2022 – May 2024 **Google**, *ML Software Engineer (Core ML / TuneLab)*, Zurich

- **Founding Engineer of TuneLab Studio**, a unified ML/LLM platform now responsible for training **over 50% of all models** at Google.
- **Enabled a $100M revenue uplift** for the sales team by launching the first fully automated ML pipeline for the global publisher intelligence team.
- **Reduced latency by 88%** on the AutoTFX Explorer platform by re-architecting the API layer, enabling real-time visualization for large-scale quality experiments.
- **Designed the "TLS Vault" analytical stack**, a critical infrastructure for data-driven decision-making used by hundreds of engineers daily.
- **Spearheaded the integration** with Google's high-performance serving platform (Servomatic), enabling seamless one-click deployment of Large Language Models.

July 2017 – May 2022 **Sinfonia Media**, *Consultant / Full Stack Developer*, Italy

- Delivered **full-stack solutions and ML projects** working directly with stakeholders to achieve measurable impact for medium-sized enterprises.
- Implemented production systems using TensorFlow, React, Node.js, and Python.
- Managed cloud infrastructure (VPS) ensuring high availability and reliability for client applications.

Skills

- **Machine Learning:** Strong foundation in Deep Learning, Reinforcement Learning (RL), and Computer Vision. Experience bridging cutting-edge research with production constraints.
- **Large Language Models (LLMs):** Expertise in LLM infrastructure, fine-tuning, inference optimization, prompt engineering, and deploying agentic workflows.
- **Full-stack Development:** Proficient in Python, C++, Java, SQL, TensorFlow, Keras, and modern web frameworks (React/Angular).
- **DevOps & Infrastructure:** Experience managing cloud infrastructure, CI/CD pipelines, and deploying ML models at scale.
- **Leadership:** Technical leadership, cross-functional collaboration, mentorship, and driving organizational AI adoption.

Education

Artificial Intelligence M.Sc., *Alma Mater Studiorum, Bologna, Italy*
Graduated with **110/110 cum laude**. Maintained a **GPA of 4.0/4.0** while working full time. Research focus: ML, Deep Learning, RL, NLP.

Artificial Intelligence M.Sc. (Exchange Student), *Vrije Universiteit, Amsterdam, Netherlands*
Research focus: Reinforcement Learning and Deep Learning.

Computer Engineering B.Sc., *Università di Parma, Parma, Italy*
Graduated **cum laude with 110/110**. Maintained a GPA of 4.0/4.0 over 3 years while working part-time.

Feedback

- "**It's been years** since I've seen an engineer who, given a problem, proposes varied / creative solutions. Many stay in their comfort zone or ignore scaling problems." - M.M., YT Shopping
- "...not only did he accomplish the work required to operate the [NTK], but also **automated and improved the ops by a margin of magnitude**." - D.A., Direct Manager