Simone Montali

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

Summary

Highly skilled software engineer with a strong foundation in machine learning and a passion for building impactful products. Proven track record of delivering **complex ML systems** and driving significant business growth.

Experience

Meta (July 2025 - Present)

July 2025 - Present — **Senior ML Software Engineer**, *Meta Reality Labs*

- Lead development of **new data engine** supporting Reality Labs team working on **foundation models for VR**
- Build robust **ML infrastructure** and design efficient **data pipelines** for large-scale model training
- Implement **state-of-the-art vision-language models (VLMs)** spanning computer vision and VR applications
- Create extensive **automation systems** to streamline ML workflows and model deployment processes
- Drive **significant business impact** by enabling researchers to iterate faster while improving model quality and performance

Google (May 2022 - end of June 2025)

May 2024 - Present — **ML Software Engineer**, *YouTube Shopping Recommendations Quality*

- Improved worldwide **product CTR by around 3%** 6 months after joining the team
- Leading cross-organization, cross-geo project on **LLM embeddings** for Shopping videos
- Currently the **first PR author** (by delta and PR number) in a team of 9 people, 3rd in the organization.

May 2022 - May 2024 — **ML Software Engineer**, *TuneLab Studio*

- Founding engineer for TuneLab Studio, a **unified ML/LLM platform** now responsible for training over 50% of models at Google.
- Developed a user-friendly platform for **LLM fine-tuning, inference, and prompt engineering**, empowering teams across Google to leverage LLMs in their products.
- Owned the TuneLab Studio **analytical and quality stack** (TLS Vault), a critical infrastructure for data-driven decision-making, leading multiple designs and refactors.
- Reduced **88% of the latency** on the analytical platform AutoTFX Explorer, used by the team for quality experiments (hundreds of model trainings running every day), by leading a redesign of

its APIs.
- Spearheaded the integration of TuneLab Studio with M2P (Model-to-production), enabling seamless one-click **deployment of large language models to Servomatic**, Google's high-performance model serving platform.
- Recognized for outstanding contributions, ranking among the **top 22% highest performers** at Google in 2022 and earning a promotion in 2023.
- Enabled a **$100M revenue uplift** for Google's sales team by launching the first fully automated and checked-in ML pipeline for the global publisher intelligence team.

Sinfonia Media (July 2017 - May 2022)

July 2017 - May 2022 **Consultant/developer**

- Delivered **full-stack solutions and ML projects** as a consultant and developer, working directly with stakeholders and tech teams to achieve measurable impact for medium-sized Italian companies.
- Implemented projects using TensorFlow, React, JavaScript, PHP, WordPress, Node.js, Python, and Scikit-learn, showcasing a diverse technical skillset.
- Managed a CentOS 8 **VPS** with almost-continuous uptime, ensuring high availability and reliability for client applications.
- Proactively identified and integrated bleeding-edge technologies to meet client requirements and enhance project outcomes.

Skills

- **Machine Learning:** Strong foundation in machine learning, deep learning, and reinforcement learning (RL), with experience in both classic ML and cutting-edge research.
- **Large Language Models (LLMs):** Expertise in LLM infrastructure, fine-tuning, inference, prompt engineering, and deploying LLM-based products.
- **Full-stack Development:** Proficient in full-stack development with experience in Python, C++, Java, SQL, TensorFlow, Keras, and Angular.
- **DevOps and Infrastructure:** Managed cloud infrastructure (CentOS 8 VPS) with near-continuous uptime and experience deploying and maintaining ML models and web applications in production environments.
- **Collaboration and Communication:** Excellent communication and collaboration skills, demonstrated through knowledge sharing, teamwork, and clear communication of technical concepts.

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: Machine Learning, Deep Learning, Reinforcement Learning, Natural Language Processing.

Artificial Intelligence M.Sc. (Exchange Student), *Vrije Universiteit, Amsterdam, Netherlands*
Research focus: Reinforcement Learning and Deep Learning. Successfully completed a C2 English exam.

Computer Engineering B.Sc., *Università di Parma, Parma, Italy*
Graduated **cum laude with 110/110**. Maintained a GPA of 4.0/4.0 over the 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
- "His clear communication keeps **everyone informed and ensures efficient use** of meeting time for focused discussions on next steps and troubleshooting." - A.F. (L5 TPM)
- "...went above and beyond to deeply understand the problem space to think of **solutions beyond core coding!**" - N.A. (L5 SWE)