

Alessandro Flaborea

Generative AI and Computer Vision | RAG, agentic pipelines, VLMs, production on AWS

📍 Venice, Italy ✉ flaborea.alessandro@gmail.com ☎ +39 346 720 2432 🔗 aleflabo.github.io

📄 alessandro-flaborea 🔄 aleflabo 🎓 Google Scholar

Summary

I build generative AI systems and put them in production. As co-founder and CTO of Procedo I owned the LLM and vision stack and led the technical team, taking the product to five enterprise accounts across petfood, paper and medical machinery, where customers reported up to 70% less time spent documenting procedures. Before that, a PhD at Sapienza, ten publications at CVPR, ICCV, ICLR and IROS with a best paper award at a CVPR workshop, and close to 500 citations.

Experience

Procedo, Advisor

Advising on technical direction, on computer vision and AI architecture choices, and on the evaluation of the production pipelines.

Treviso, Italy

July 2026 – present
2 months

Procedo, Co-Founder & CTO

Pre-seed AI startup for manufacturing, five people in total. The product turns raw shop-floor video into structured manufacturing knowledge: the documentation, training material and process insight that industrial teams normally assemble by hand over weeks. Five enterprise accounts signed across petfood, paper and medical machinery, with customers reporting up to 70% less time spent documenting procedures.

Treviso, Italy

July 2024 – July 2026
2 years 1 month

- Built the multi-agent pipeline that turns hours of shop-floor video into structured procedures. Vision-language agents analyse audio and visual frames into a time-aligned timeline, a mixture-of-experts stage runs three step-extraction experts in parallel, and a grounded consensus judge picks and completes the winning extraction. The orchestration is our own code, with no agent framework.
- Stored every agent's prompt, response schema and model as a database record instead of code, so per-customer model routing and prompt changes ship without a deploy. Each agent falls back to a deterministic non-LLM path, so a malformed model response cannot abort a run. Production ran on Gemini through Google Vertex AI, so changing provider is configuration rather than a refactor.
- Designed the retrieval-augmented generation stack behind grounded question answering: embeddings in pgvector with HNSW cosine search, fused with Postgres full-text ranking by a hybrid retriever, plus per-chunk visibility gating and citation-backed answers.
- Built the evaluation layer that gates model and prompt changes, covering multi-dimensional quality scoring per stage, a consensus judge, head-to-head extraction comparison and full execution logging.
- Shipped an agentic chat assistant with tool calling over retrieval, procedure and attachment Q&A, drafting and editing. Responses stream, and a thumbs-up/down feedback loop feeds an admin review dashboard. The team wrote the frontend as well as the backend.
- Designed and ran the AWS platform: CloudFront, WAF and ALB at the edge, ECS Fargate for the API and Celery workers, Multi-AZ encrypted RDS PostgreSQL and ElastiCache in private subnets, Lambda and AWS Batch for media processing. Multi-tenant isolation throughout: database and object storage per tenant, RBAC, MFA policies, audit-logged support access, hash-chained approval trails for ISO 27001 and GDPR retention cleanup, with model access kept API-only under the same controls.
- Led the three-person technical team: daily stand-ups, two-week retrospectives and a GitHub CI/CD pipeline gating backend, frontend, unit and migration tests, with automated code review by Claude. Took the full-time engineer from no professional experience to working autonomously.
- Selected for Bocconi's B4i (Bocconi for Innovation) accelerator.

University of Amsterdam, Visiting Researcher

Amsterdam, Netherlands

- Worked with Prof. Pascal Mettes on hyperbolic representation learning. The collaboration produced papers at ICLR 2025 (compositional entailment learning for hyperbolic vision-language models) and IROS 2024 (hyperbolic crowd navigation).

Dec 2023 – Mar 2024
4 months

HomeSweatHome Training, Co-Founder

An AI virtual trainer, two co-founders, taken as far as a working MVP: a real-time coach that watches you exercise through a standard laptop camera, reads your posture and corrects your form as you go. No wearables, no cloud round-trip.

Rome, Italy
Jan 2021 – Dec 2023
3 years

- Architected a real-time posture and movement analysis pipeline that runs entirely on-device, using computer vision models small enough for a consumer laptop while still giving immediate, actionable feedback.
- Built the on-device action recognition model that decides whether an exercise is being performed correctly, closing the loop between user motion and coaching output.

Sapienza University of Rome, Teaching Assistant

- MSc courses *Algorithmic Methods of Data Mining*, *Fundamentals of Data Science* and *Advanced Machine Learning*.

Rome, Italy
Sept 2019 – Dec 2023
4 years 4 months

Selected Publications

Compositional Entailment Learning for Hyperbolic Vision-Language Models

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, *Alessandro Flaborea*, Fabio Galasso, Pascal Mettes

arxiv.org/abs/2410.06912 (International Conference on Learning Representations (ICLR), 2025)

Contracting Skeletal Kinematics for Human-Related Video Anomaly Detection

Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Marco Aurelio Sterpa, Alessio Sampieri, Fabio Galasso

(Pattern Recognition, vol. 156, 2024 · code released, 8 GitHub stars)

PREGO: Online Mistake Detection in Procedural Egocentric Videos

Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Galasso

arxiv.org/abs/2404.01933 (IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024 · code released, 34 GitHub stars)

Multimodal Motion Conditioned Diffusion Model for Skeleton-Based Video Anomaly Detection

Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, Fabio Galasso

arxiv.org/abs/2301.09489 (IEEE/CVF International Conference on Computer Vision (ICCV), 2023 · code released, 93 GitHub stars)

Education

PhD Sapienza University of Rome, Computer Science

- Awarded *cum laude*.

Rome, Italy
Mar 2021 – Mar 2024

MSc Sapienza University of Rome, Data Science (LM-91)

- 110/110 *cum laude*.

Rome, Italy
Sept 2018 – Oct 2020

BSc University of Udine, Computer Science (L-31)

Udine, Italy
Sept 2014 – Oct 2017

Skills

Generative AI: Retrieval-augmented generation, hybrid retrieval, prompt engineering, structured output and response schemas, agentic pipelines and tool calling, vision-language models, multi-agent orchestration, LLM and VLM evaluation

Models & Frameworks: PyTorch, Gemini via Google Vertex AI, in-house agent orchestration (no framework), diffusion models, hyperbolic neural networks, on-device inference

Computer Vision: Video understanding, action recognition, anomaly detection, procedural understanding, pose estimation

Engineering: Python, Flask, Celery, PostgreSQL, pgvector, Redis, Docker

AWS & Production: ECS Fargate, Lambda, AWS Batch, RDS, S3, ElastiCache, CloudFront, WAF, ALB, multi-tenant isolation, RBAC, audit logging, CI/CD, observability

Leadership: Team leadership and mentoring, hiring, roadmap ownership, enterprise customers, delivery process

Spoken: Italian (native), English (full professional)

Availability: Italy, or remote for companies abroad