

Alessandro Flaborea

AI generativa e computer vision | RAG, pipeline agentiche, VLM, produzione su AWS

📍 Venezia, Italia ✉ flaborea.alessandro@gmail.com ☎ +39 346 720 2432 🌐 aleflabo.github.io

📄 alessandro-flaborea 🔄 aleflabo 🎓 Google Scholar

Sintesi

Costruisco sistemi di AI generativa e li porto in produzione. Come co-fondatore e CTO di Proceso ho gestito lo stack LLM e vision e guidato il team tecnico, portando il prodotto a cinque contratti enterprise fra petfood, carta e macchinari medicali, dove i clienti hanno riportato fino al 70% di tempo in meno per documentare le procedure. Prima di questo, un dottorato alla Sapienza, dieci pubblicazioni a CVPR, ICCV, ICLR e IROS con un best paper award a un workshop CVPR, e quasi 500 citazioni.

Esperienza

Proceso, Advisor

Supporto la direzione tecnologica, le scelte di architettura su computer vision e AI e la valutazione delle pipeline in produzione.

Treviso, Italia

Lug 2026 – presente
2 mesi

Proceso, Co-Founder & CTO

Startup pre-seed di AI per il manifatturiero, cinque persone in tutto. Il prodotto trasforma i video di reparto in conoscenza produttiva strutturata: la documentazione, il materiale di formazione e le analisi di processo che i team industriali costruiscono normalmente a mano in settimane. Cinque contratti enterprise fra petfood, carta e macchinari medicali, con i clienti che hanno riportato fino al 70% di tempo in meno per documentare le procedure.

Treviso, Italia

Lug 2024 – Lug 2026
2 anni 1 mese

- Ho costruito la pipeline multi-agente che trasforma ore di video di reparto in procedure strutturate. Agenti vision-language analizzano audio e frame in una timeline allineata nel tempo, uno stage mixture-of-experts esegue tre esperti di estrazione in parallelo e un consensus judge grounded sceglie e completa l'estrazione vincente. L'orchestrazione è codice nostro, senza framework per agenti.
- Ho salvato prompt, response schema e modello di ogni agente come record a database invece che nel codice, così il routing dei modelli per cliente e le modifiche ai prompt vanno in produzione senza deploy. Ogni agente ricade su un percorso deterministico non-LLM, quindi una risposta malformata non può interrompere un'esecuzione. In produzione giravano i modelli Gemini tramite Google Vertex AI, quindi cambiare provider è configurazione e non un refactor.
- Ho progettato lo stack di retrieval-augmented generation dietro il question answering grounded: embedding in pgvector con ricerca HNSW cosine, fusi con il ranking full-text di Postgres da un retriever ibrido, più visibility gating per chunk e risposte cordate di citazioni.
- Ho costruito il layer di valutazione che governa le modifiche a modelli e prompt, con quality scoring multi-dimensionale per stage, un consensus judge, confronto diretto fra estrazioni e logging completo delle esecuzioni.
- Ho rilasciato un assistente conversazionale agentico con tool calling su retrieval, Q&A su procedure e allegati, redazione e modifica. Le risposte sono in streaming e un feedback loop alimenta una dashboard di revisione. Il team ha scritto il frontend oltre al backend.
- Ho progettato e gestito la piattaforma su AWS: CloudFront, WAF e ALB al perimetro, ECS Fargate per API e worker Celery, RDS PostgreSQL Multi-AZ criptato ed ElastiCache in subnet private, Lambda e AWS Batch per l'elaborazione media. Isolamento multi-tenant su tutta la piattaforma: database e storage per tenant, controllo degli accessi basato su ruoli, policy MFA, accessi di supporto tracciati, catene di approvazione con hash per la ISO 27001 e cancellazione dei dati per il GDPR, con accesso ai modelli solo via API.
- Ho guidato il team tecnico di tre persone: stand-up giornaliero, retrospettiva ogni due settimane e una pipeline CI/CD su GitHub che verifica backend, frontend, unit test e test di migrazione, con code review automatizzata tramite Claude. Ho portato lo sviluppatore full-time da nessuna esperienza professionale alla piena autonomia.
- Proceso è stata selezionata dall'acceleratore B4i (Bocconi for Innovation).

Università di Amsterdam, Visiting Researcher

- Ho lavorato con il Prof. Pascal Mettes su representation learning iperbolico. Dalla collaborazione sono nati i lavori a ICLR 2025 (compositional entailment learning per modelli vision-language iperbolici) e IROS 2024 (navigazione iperbolica tra la folla).

Amsterdam, Paesi Bassi

Dic 2023 – Mar 2024

4 mesi

HomeSweatHome Training, Co-Founder

Un personal trainer virtuale basato su AI, due co-fondatori, portato fino a un MVP funzionante: un coach in tempo reale che ti osserva mentre ti alleni attraverso la webcam di un normale portatile, legge la postura e corregge l'esecuzione mentre la fai. Nessun wearable, nessun round-trip verso il cloud.

Roma, Italia

Gen 2021 – Dic 2023

3 anni

- Ho architettato una pipeline di analisi di postura e movimento che gira interamente on-device e in tempo reale, con modelli di computer vision abbastanza leggeri per un portatile di consumo ma capaci di restituire un feedback immediato e azionabile.
- Ho costruito il modello di action recognition on-device che stabilisce se un esercizio viene eseguito correttamente, chiudendo il ciclo fra il movimento dell'utente e l'indicazione del coach.

Sapienza Università di Roma, Assistente alla didattica

- Corsi magistrali *Algorithmic Methods of Data Mining*, *Fundamentals of Data Science* e *Advanced Machine Learning*.

Roma, Italia

Set 2019 – Dic 2023

4 anni 4 mesi

Pubblicazioni selezionate

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 · codice pubblicato, 8 stelle su GitHub)

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 · codice pubblicato, 34 stelle su GitHub)

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 · codice pubblicato, 93 stelle su GitHub)

Formazione

Dottorato

Sapienza Università di Roma, Informatica

- Conseguito *cum laude*.

Roma, Italia

Mar 2021 – Mar 2024

Magistrale

Sapienza Università di Roma, Data Science (LM-91)

- 110/110 e lode.

Roma, Italia

Set 2018 – Ott 2020

Triennale

Università degli Studi di Udine, Informatica (L-31)

Udine, Italia

Set 2014 – Ott 2017

Competenze

AI generativa: Retrieval-augmented generation, retrieval ibrido, prompt engineering, output strutturato e response schema, pipeline agentiche e tool calling, modelli vision-language, orchestrazione multi-agente, valutazione di LLM e VLM

Modelli e framework: PyTorch, Gemini tramite Google Vertex AI, orchestrazione di agenti sviluppata internamente (senza framework), modelli di diffusione, reti neurali iperboliche, inferenza on-device

Computer vision: Video understanding, action recognition, anomaly detection, comprensione procedurale, pose estimation

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

AWS e produzione: ECS Fargate, Lambda, AWS Batch, RDS, S3, ElastiCache, CloudFront, WAF, ALB, isolamento multi-tenant, RBAC, audit logging, CI/CD, observability

Leadership: Guida e mentoring di team tecnici, assunzioni, gestione della roadmap, clienti enterprise, processo di sviluppo

Lingue: Italiano (madrelingua), Inglese (professionale completo)

Disponibilità: Italia, oppure da remoto per aziende estere