

Gabriele Desimini

dabi.ai.eng@gmail.com linkedin.com/in/gabriele-desimini dabih2.github.io

Work Authorization: Spouse of U.S. Citizen. Authorized to work for any U.S. employer upon arrival. No sponsorship required.

Profile

AI/ML Software Engineer with 8+ years shipping production systems across backend, AI, and full stack work, including 4+ years building LLM and agentic applications with LangChain, LangGraph, and PyTorch. Deep experience with microservices, REST APIs, Kubernetes, distributed pipelines processing billions of records, RAG platforms, and evaluation harnesses under strict reliability SLOs. Track record of high ownership in fast moving environments, from founding engineer roles at AI startups to safety critical decision support at NASA.

Skills

Backend & APIs: Python (deep, 8+ yrs), FastAPI, Node.js, REST API design, OpenAPI, authentication and authorization patterns

Distributed Systems: Kubernetes (4+ yrs production), high throughput pipelines over billions of records, batch and streaming processing, fault tolerance, partitioning, performance tuning

AI & Agents: LangChain, LangGraph, RAG, vector databases (Pinecone, Weaviate), tool calling, multi step reasoning, evaluation harnesses, PyTorch, model serving (TorchServe, vLLM)

Integration: Service decomposition, asynchronous messaging, event driven patterns, API gateways, middleware, contract testing

Full Stack: React, TypeScript, component driven UI, operator facing dashboards

CI/CD & Reliability: GitHub Actions, GitLab CI, GitOps patterns, blue/green and canary rollouts, automated rollback, structured logging, tracing, metrics, SLO driven on call

Cloud & Infra: AWS (ECS, EKS, S3, IAM), GCP, Azure, Terraform, Docker

Other Languages: SQL, Bash

Experience

AI Engineer (Research)

11/2024

NASA / San Jose State University

- Led agent architecture and technical direction for transformer based decision support services, staying hands-on across design, production code, and deployment on Kubernetes under strict latency and reliability SLOs.
- Designed and shipped Python microservices exposing REST APIs consumed by operator applications and downstream systems; orchestrated multi step reasoning agents over heterogeneous aerospace datasets.
- Built CI/CD pipelines with automated testing, blue/green deployments, and automated rollback for safety critical AI services; instrumented end to end observability with structured logs, metrics, and tracing.
- Diagnosed and resolved production failures spanning API integrations, data ingestion, inference, agent orchestration, and downstream consumers, reducing mean time to detection and recovery.

Freelance Machine Learning Engineer

11/2023 – 09/2024

BagsID

- Built and operated distributed Vision Transformer inference services on Docker and Kubernetes for live customer traffic, owning capacity planning, latency, and cost optimization.
- Designed batching, caching, and resource aware deployment patterns; improved throughput by 40% while cutting infrastructure cost.
- Built CI/CD pipelines covering training, evaluation, and deployment; resolved production incidents across the data, model, and infrastructure layers.

Senior Quantitative Analyst

10/2022 – 11/2023

Atradius

- Built and supported high throughput ML pipelines in regulated financial environments, processing billions of records on Linux based distributed systems.
- Redesigned data architecture to improve scalability, fault tolerance, and recoverability under partial system failure.
- Implemented automated monitoring, validation, and drift detection workflows; investigated production failures through structured root cause analysis.

ML Infrastructure Engineer

01/2022 – 10/2022

Supplai

- Architected Kubernetes based ML deployment stacks with Terraform and GitHub Actions; built containerized microservices with clear service boundaries, contract testing, and performance budgets.
- Exposed REST APIs for real time predictions on multimodal inputs; built automated retraining and evaluation loops integrated with upstream data sources and downstream serving systems.

AI Consultant

01/2020 – 12/2021

ESA / OHB / ATG Europe

- Built telemetry pipelines and anomaly detection systems for mission critical satellite operations on Linux infrastructure with strict reliability SLAs.
- Operated within structured engineering processes covering change management, validation, and continuous testing.

AIT Engineer

06/2018 – 01/2020

SITAEL

- Built Linux based automation and orchestration tooling supporting distributed hardware integration; executed system level validation and structured failure analysis.

Selected AI Platform Work

GrantSpider (Founding Engineer, Production Multi Agent RAG Platform)

- Owned technical direction for a multi agent LLM platform with RAG pipelines, retrieval optimization, and evaluation feedback loops; built operator facing React dashboards for review and human in the loop intervention.
- Implemented guardrails against hallucination and prompt injection, plus a structured evaluation harness driving iteration in production.

Text4Quote (Co-founder & CTO, LangGraph Agent)

- Built an end to end LangGraph agent with multi turn flow, tool calling, persistent memory, and structured outputs, backed by FastAPI and persistent state; owned architecture, production code, and integrations end to end.

Sphinx (LLM Compliance Extraction)

- Hybrid rule plus LLM extraction pipeline with automated CI/CD validation, structured evaluation, and end to end test coverage.

Education

M.Sc. Artificial Intelligence

University of Amsterdam

B.E. Computer Science & Automation

Politecnico di Bari (110/110 cum laude)