Gian Paolo Santopaolo

Location: Switzerland
Email: gianpaolo.santopaolo@gmail.com
Phone: +41 (0) 78 340 32 28
Technical Blog: https://genmind.ch/
GitHub personal: https://github.com/gsantopaolo
GitHub CogniX: https://github.com/gen-mind/cognix

Summary

Senior Generative AI Engineer focused on building production-grade AI systems—from custom neural architectures (CNNs, DNNs, Transformers) and LLM/diffusion model fine-tuning to RAG platforms and agentic pipelines. I work daily with Python and PyTorch, delivering AI solutions from research prototypes to scalable, multi-GPU deployments.

- Developed **CogniX**, an enterprise **RAG platform** and multimodal AI stack, and integrated generative AI (LLMs, diffusion, vision, speech, OCR) into **Collaboard**, a real-time collaboration product used worldwide.
- Designed and deployed into production multiple agentic and generative AI systems, including ReforgeAI, Sentinel-AI, CreativeCampaign-Agent, and DeltaE (see Selected Projects).
- Experienced with LLM fine-tuning, diffusion models, distributed inference on multi-GPU/multi-node clusters (vLLM, Hugging Face TGI), and multimodal vector search at scale.
- Strong background refactoring large, research-driven code into modular, testable Python libraries, architecting microservices and GPU-accelerated platforms on Kubernetes/Docker (bare metal and cloud).
- Build model-serving APIs and interactive tools using FastAPI and Streamlit (easily adapted to Gradio) to enable internal researchers and downstream users to experiment with models.
- Recognized as Microsoft MVP (2012–2022) and Microsoft Regional Director (2018–2020) for technical leadership and community impact.
- Own and evolve open-source and internal **foundation model** codebases, with a strong focus on Hydra-based configuration systems, reproducibility, and clean abstraction boundaries.

CORE TECHNICAL SKILLS

Languages & Core

- **Python (expert)**, .NET Core/C# (expert), Go.
- Strong in **modular design**, abstraction boundaries, and collaborative codebase evolution

Deep Learning, Generative AI & Multimodal

- **PyTorch**, Hugging Face Transformers (BERT, GPT, RoBERTa), Diffusers
- **LLM fine-tuning** (e.g., Pixtral multimodal text+vision), **Flux diffusion**, Stable Diffusion (incl. fine-tuning), style transfer
- Embedding model fine-tuning
- **RAG systems**, knowledge graphs, and semantic search
- **Distributed inference** on **multi-GPU/multi-node** setups with **vLLM** and **Hugging Face TGI**
- **DeepSeek-OCR, SmolDocling, Tesseract** for OCR (tables, forms, handwriting)
- **YOLO** (custom-trained) for real-time object detection, **Whisper** for speech.
- **Agentic AI frameworks**: production use of CrewAI, plus familiarity with Microsoft AutoGen, Microsoft Semantic Kernel, Langflow, and Google Antigravity.

Serving, Tooling & Evaluation

- FastAPI, REST APIs, WebSockets; interactive model UIs with **Streamlit and Gradio** for internal researchers and downstream users.
- Design of **evaluation pipelines**, metrics, and experiment tracking for generative/RAG/vision models
- Design of **evaluation pipelines**, metrics, and experiment tracking for generative/RAG/vision models/fine-tuning using Inspect AI, LangTrace, Weights & Biases, and TensorBoard.
- Logging, metrics, and observability for AI services
- Logging, metrics, and observability for AI services with Grafana, Prometheus, Loki, Azure Application Insights, AWS CloudWatch, Google Cloud Monitorin, and Langtrace
- **LLM and agent orchestration** with LangChain, LangGraph, and agentic frameworks (CrewAI, Microsoft AutoGen, Microsoft Semantic Kernel, Langflow, Google Antigravity) for RAG, tools, and multi-step workflows.

AI Platforms, Cloud & GPU

- AI platforms on **bare-metal Kubernetes + Docker** with **NVIDIA Container Toolkit** and GPU drivers (Linux)
- Cloud: **Microsoft Azure** (App Service, Functions, Cosmos DB, SignalR, Service Fabric), plus exposure to **AWS** and **GCP**
- CI/CD with Azure DevOps; highly scalable clusters serving large user bases
- End-to-end observability for AI platforms using Grafana/Prometheus/Loki and cloud-native monitoring (Application Insights, CloudWatch, Google Cloud Monitoring).

Cloud & Platforms

- **Microsoft Azure (strong experience):** Azure Kubernetes Service (**AKS**), App Service, Functions, Azure Container Apps, **Azure Cosmos DB**, Azure SQL, Azure Storage, Azure Service Bus / Event Hubs, Azure Key Vault; AI services including **Azure OpenAI Service, Azure AI Studio, Azure Machine Learning, Azure Cognitive Services (Vision, Speech, Language), Azure AI Search**.
- **Amazon Web Services (AWS):** EC2, **EKS/ECS**, Lambda, S3, RDS, DynamoDB, API Gateway, SQS/SNS, CloudWatch; AI services including **Amazon Bedrock, Amazon SageMaker, Amazon Comprehend, Amazon Rekognition, Amazon Transcribe, Amazon Translate, Amazon Lex**.
- **Google Cloud Platform (GCP):** GKE, Cloud Run, App Engine, **BigQuery**, Cloud Storage, Cloud Functions, Pub/Sub, Firestore; AI services including **Vertex AI (Generative AI Studio), Vision AI, Natural Language AI, Translation AI, Dialogflow**.
- **Lamda Labs/Cloud**, **RunPod**
- Cross-cloud monitoring and logging with Azure Application Insights, AWS CloudWatch, and Google Cloud Monitoring (Operations Suite).

Configuration, Packaging & Code Quality

- YAML / environment-scoped config systems; **strong focus on reproducibility and override logic,** including **Hydra-based** hierarchical configs with reusable templates and environment scoping.
- Experience in **packaging and releasing Python modules** using modern tooling such as **uv, just, and pydantic**, managing environment consistency, and shipping libraries as internal and OSS artifacts
- API design, directory structure, dependency isolation, unit & integration tests, **docstrings**, exception hygiene, code review.

Data, Storage & Infra

- Vector DBs: **Qdrant, Milvus**
- SQL and NoSQL (Cosmos DB, Cassandra, SQL Server, Mongo, time-series DBs)

- Familiar with data engineering patterns, batch/stream processing, Apache Spark and Databricks

---

PROFESSIONAL EXPERIENCE

IBV Solutions, Switzerland
CTO & Principal Generative AI Engineer (hands-on)
December 2019 – August 2025

AI Platform & Foundation Model Engineering

- **Owned and evolved** the core codebase of **CogniX**, an open-source **Retrieval-Augmented Generation (RAG) platform**, transforming it from research scripts into a **modular, reusable Python library** consumed across teams.
- Fine-tuned and operated **LLMs and diffusion models** (Pixtral multimodal text+vision, Flux, Stable Diffusion, HF Transformers) for enterprise use cases, deploying them via **multi-GPU / multi-node distributed inference** (vLLM, TGI).
- Integrated **off-the-shelf models** (YOLO, Whisper, OCR engines, Stable Diffusion, style transfer) as configurable **pre- and post-processors** inside unified pipelines, with robust checkpoint loading and model wrappers.
- Implemented shared Python modules for RAG orchestration, feature extraction, and evaluation and distributed them as internal packages to standardize AI development across projects.
- Designed and maintained **Hydra-style hierarchical configuration systems** for training, inference, and evaluation pipelines, ensuring reproducibility, safe overrides, and clean separation between environments (dev/staging/production).
- Introduced LangChain and LangGraph for modular RAG and agent orchestration, with LangTrace and Inspect AI used to trace, debug, and evaluate complex multi-step flows.

Productized Generative & Multimodal Features (Collaboard)

- Transformed **Collaboard** into a **multimodal, generative AI-driven platform** by shipping features such as **background removal**, **sticky-note detection**, **AI search**, **knowledge graph-powered retrieval**, **summarization**, and **text/image generation**.
- Built **FastAPI-based model-serving endpoints** for internal and customer-facing use, enabling low-latency inference and easy experimentation with new models and configurations.
- Developed interactive evaluation and demo tools using **Streamlit** to let product managers, researchers, and customers experiment with generative and RAG capabilities; patterns are directly transferable to **Gradio-based** interfaces.
- Designed and maintained **evaluation pipelines** for RAG quality, vision tasks, and transcription using Inspect AI, LangTrace, TensorBoard, and Weights & Biases, with metrics and dashboards that guided model selection, regression detection, and release decisions.
- Instrumented Collaboard's AI features with **end-to-end observability** via Grafana, Prometheus, Loki, and cloud-native monitoring (Application Insights, CloudWatch, Google Cloud Monitoring).

Cloud, GPU & Microservices Architecture

- Led the **migration from monolithic architecture to microservices** using **Kubernetes** and **Docker**, improving scalability, reliability, and deployment frequency.
- Architected an **AI platform on bare-metal Kubernetes** with **NVIDIA Container Toolkit** (Linux) for GPU-accelerated workloads, supporting both on-prem and cloud deployments (Azure, with experience in AWS/GCP).
- Employed **Infrastructure-as-Code** principles for highly available, regulated systems, enabling **zero-downtime updates** and reproducible environments.
- Standardized **logging, tracing, and metrics** across microservices using Grafana, Prometheus, Loki and cloud services such as Azure Application Insights, AWS CloudWatch, and GCP Monitoring.

Software Engineering, Refactoring & Code Health

- Systematically refactored legacy, research-heavy services into **clean, testable, maintainable Python and .NET libraries**, eliminating anti-patterns and reducing technical debt.
- Defined standards for **directory structure, API boundaries, unit/integration testing, logging and exception handling, and dependency isolation** across AI and non-AI services.
- Championed automated testing and CI/CD for AI features, increasing reliability of deployments and reducing production incidents.
- Ensured reproducible experiments and model comparisons through structured experiment tracking and visualization with Weights & Biases and TensorBoard.

Leadership & Mentorship

- Led and mentored **cross-functional teams** (frontend, backend, DevOps, SQA) delivering AI projects from concept to deployment, aligning technical work with business objectives.
- Bridged **R&D, product, and customer teams**, translating research prototypes into **robust, production-ready modules** and coaching stakeholders on realistic AI capabilities and constraints.

Senior Software Architect & Team Lead – Various Companies (Banking, Finance, Tech)
2001 – 2019

- Led teams designing and implementing **large-scale, secure enterprise systems** with high availability across banking, finance, and technology sectors.
- Migrated **monolithic systems** to **service-oriented and microservices architectures**, driving adoption of cloud and modern DevOps practices.
- Pioneered **real-time collaboration tools and natural user interfaces (NUI)**, including an interactive whiteboard platform that influenced products such as Miro and Mural, later commercialized as **Collaboard** by a Swiss company.
- Developed an innovative **bank teller solution** on Samsung SUR40 with Microsoft PixelSense and a biometric signature pad, which earned the customer **ABI (Italian Banking Association) Lab 2012** innovation recognition.
- Over 15 years of experience leading teams in developing large-scale enterprise software across banking, finance, and technology industries.
- Delivered scalable, secure, and efficient systems with a focus on cloud technologies, service-oriented architecture (SOA), and AI integration.
- Raised overall engineering quality in a research setting by turning exploratory notebooks and scripts into production-grade Python modules, adding tests, observability, and standardized wrappers around complex generative model codebases.

SELECTED AI & GENERATIVE PROJECTS (PRODUCTION DEPLOYMENTS)

CogniX – Enterprise RAG Platform (Production, Open Source)

- Designed and led development of **CogniX** (https://github.com/gen-mind/cognix), an **enterprise-grade Retrieval-Augmented Generation (RAG) platform** for semantic search and question-answering over large organizational knowledge bases.
- Built as a modular, production-ready system with connectors to multiple enterprise data sources and vector databases, enabling **secure, high-relevance retrieval** and chatbot-style interactions over millions of documents, **deployed in production** for real customers.
- Implemented **RAG orchestration** and tool-calling flows using LangChain and LangGraph, with LangTrace for tracing end-to-end conversations and retrieval chains.

ReforgeAI – Agentic AI for Legacy App Modernization (Production)

- Designed and implemented **ReforgeAI** (https://github.com/gsantopaolo/reforge-ai), an **agentic AI system for modernizing legacy Java codebases**.

- The system automatically analyzes existing code, generates documentation and a transformation plan, and executes modernization tasks (dependency updates, refactors, framework migrations, and security hardening) with **human-in-the-loop feedback**.
- **Deployed in production** to accelerate modernization of real customer codebases, replacing large portions of manual analysis and refactoring work.
- Implemented **CrewAI-based agent teams** AI to coordinate research, reasoning, and content-generation workflows, improving execution of complex multi-step tasks.

Sentinel-AI – Real-Time Event & Anomaly Monitoring Platform (Production)

- Built **Sentinel-AI** (https://github.com/gsantopaolo/sentinel-AI), an **event-driven, microservice platform** that ingests, filters, ranks, and detects anomalies in IT-related event streams using **embeddings and LLMs**.
- Designed to run on **Kubernetes** and **scale to millions of users**, with asynchronous services communicating over a message bus.
- **Deployed in production** as a continuous monitoring and alerting system, providing noise-filtered, actionable summaries to dashboards and email.
- Used **CrewAI to orchestrate specialized agents** for log analysis, anomaly triage, and incident summarization, turning Sentinel-AI into an agentic monitoring and response system.
- Integrated **metrics, logs, and anomaly alerts** into Grafana/Prometheus/Loki and cloud monitoring (Application Insights, CloudWatch, Google Cloud Monitoring) for real-time observability.

CreativeCampaign-Agent – Agentic Creative Automation for Social Ads (Production)

- Created **CreativeCampaign-Agent** (https://github.com/gsantopaolo/CreativeCampaign-Agent), an **agentic AI system for social advertising creatives**: generates brand-safe hero images and layouts, adds logos intelligently, localizes copy, and exports multiple formats.
- Uses **OpenAI DALL·E 3 + GPT-4o-mini**, an **event-driven microservice architecture**, and production-grade components (S3, MongoDB, NATS) with health checks, retries, and observability.
- **Deployed in production** to automate creative generation and localization workflows, replacing manual design steps at scale.

DeltaE – Automated Color Correction for Fashion AI (Production)

- Developed **DeltaE** (https://github.com/gsantopaolo/DeltaE), a **production-ready color fidelity correction pipeline** for AI-generated fashion imagery, ensuring ΔE2000-accurate garment colors with texture preservation.
- Combines advanced segmentation and computer vision to correct garment colors while preserving material appearance and leaving skin/background untouched.
- **Deployed in production** for fashion e-commerce imagery, enabling scalable AI-generated on-model photos with color fidelity suitable for commercial use.
- Enables scalable **synthetic on-model fashion imagery** in production e-commerce pipelines while preserving strict color fidelity requirements.

Multilingual, Multimodal Vector Search (10M+ Users)

- Architected a **scalable multilingual and multimodal vector search system** using **Qdrant** and advanced embedding models, enabling rich semantic search experiences across languages and content types for a user base exceeding **100 million**.

AI Code Generator for Developer Productivity

- Built an **AI-powered code generation tool** that analyzes existing codebases to generate repetitive code blocks and boilerplate, integrated into development workflows to significantly improve developer velocity.

AI Agents & Bots (Agentic Systems)

- Created **AI agents and bots** for industries such as real estate and travel, including a **WhatsApp bot** for appointment scheduling and an **AI-driven trading bot**, orchestrating model inference with external APIs and business rules.

Fine-tuning & Distributed AI R&D

- Conducted R&D on **fine-tuning LLM**s and **diffusion models** for diverse business applications, focusing on **distributed inference** across multi-GPU setups (vLLM, TGI), with experiment tracking in Weights & Biases/TensorBoard and evaluation pipelines using Inspect AI and LangTrace.

AWARDS & RECOGNITIONS

- **Microsoft Regional Director (2018–2020)**
  Selected among ~140 experts worldwide for technical excellence, business acumen, and community leadership; provided strategic feedback to senior Microsoft leadership under NDA.

- **Microsoft Most Valuable Professional (MVP, 2012–2022)**
  Recognized for deep technical expertise and community contributions through speaking, writing, and mentoring.

- **Innovation Recognition – Real-Time Whiteboard Technology**
  Inventor of real-time whiteboard technology that influenced platforms like **Miro** and **Mural**, today commercialized as **Collaboard** by a Swiss company.

EDUCATION & FORMAL TRAINING

Stanford Artificial Intelligence Professional Program (three exams)

- CS224N: Natural Language Processing with Deep Learning
- CS234: Reinforcement Learning
- CS236: Deep Generative Models

Machine Learning Specialization (DeepLearning.AI and Stanford University, three exams):

- Supervised Machine Learning: Regression and Classification
- Advanced Learning Algorithms
- Unsupervised Learning, Recommenders, Reinforcement Learning

Mathematics for Machine Learning Specialization (Imperial College London, three exams):

- Linear Algebra
- Multivariate Calculus
- Principal Component Analysis - PCA

Liceo Scientifico "E. Fermi", 1994

Continuous Professional Development

- Ongoing self-education in AI, ML, and software development through hands-on projects, research, and industry engagement.

SELECTED SPEAKING ENGAGEMENTS

- **Stanford AI Professional Certificate Program - Show and Tell Session, October 2025** - CV-Pilot and Fine-Tuning (200+ attendees)

- **Stanford AI Professional Certificate Program - Show and Tell Session, September 2025** - Reforge AI and Sentinel AI (200+ attendees)
- **API World, San Jose (2023)** – “Our Journey from Monolithic to Microservice with Kubernetes” (300+ developers)
- **Future Tech, Amsterdam (2019)** – Multiple AI/ML sessions for .NET developers (audiences up to 500+)
- **DevSum, Stockholm (2018)** – “Machine Learning for Developers”
- **.NET Conference (2018)** – “AI for Every Developer” (3,000+ online attendees)
- **Insider Dev Tour, Zurich & Milan (2018)** – Keynote speaker
- **ESPC – European SharePoint, Office 365 and Azure Conference, Copenhagen (2018)** – “Machine Learning for Developers”

ADDITIONAL INFORMATION

- **Languages:** English (fluent), Italian (fluent)
- **Availability:** Immediate
- **References:** Available upon request