Valerio Francesco Puglisi

Email: valeriopuglisi.vp@gmail.com | Phone: (+39) 340 005 9665 | Linkedin, Scholar

Summary

ML/DL Engineer specialised in applied AI. Focused with **GenAI** on **LLMs systems** and in particular **multilingual Agentic AI** and **LLM-powered information retrieval systems**. Designed and shipped **production RAG systems** (LangGraph, PGVector, FAISS, Airflow, FastAPI, OpenAI, LLaMA) with **automated evaluation** (faithfulness, relevance, context precision, context recall). Experience also in **Computer Vision**, **Audio** Systems, and shipping ML features end-to-end. Recognised as collaborative, curious, and impact-driven, with strong communication skills gained through both industry and academic experience.

Keywords (ATS): indexing · retrieval · ranking · vector search (PGVector/FAISS) · RAG · RAG evaluation (RAGAS) · LangGraph · Agno · agents/tools · planning · multi-step reasoning · Python · PyTorch · TensorFlow · FastAPI · Airflow · Docker/K8s · multilingual NLP · MT · data quality · evaluation harness

Work Experience

Hlpy — Senior ML Engineer (Remote) | 2024 – Present

- **RAG system (FastAPI, LangGraph, PGVector, Airflow)**
  - Designed & deployed **Graph-RAG (LangGraph, PGVector, FastAPI, Airflow)** for multilingual customer support.
  - Built **PostgreSQL/PGVector schema** for conversations and retrieved context (auditability, reproducibility).
  - Developed **Airflow ingestion pipeline**: SharePoint **Docs sync, Preprocess** the PDF **docs** to enhance format**,** OpenAI **embeddings** (text-embedding-3-large)**, evaluation on testset** with RAGAS **with test knowledge** (faithfulness, answer relevance, context precision)**, promotion to production**.
  - Exposed **FastAPI endpoints** for chat & knowledge hot-reload.
- **Conversational AI platform (ElevenLabs, Quarkus)**
  - Designed & deployed **multi-agent system service** for **multilingual conversational customer support** with ElevenLabs Conversational AI platform that uses ad hoc proxy Quarkus Service to expose APIs as tools for Agents to interact with the rest of the system.

Stack: LangGraph, PGVector, FastAPI, Airflow, OpenAI API, Python, ElevenLabs, Quarkus.

Epicode* - *Online Computer Architecture Lecturer* | *2024: Improved remote communication skills.

***Department of Electrical electronic and computer engineering* (*University of Catania*)**
— Deep Learning Engineer Consultant | 2024
• Developed SOTA detection (**YOLOv8, DINO, DiffusionDet, RTMDet, CO-DETR**) and **tracking** (DeepSORT, ByteTrack) for cattle detection in position estimation, Data Cleaning and Data Augmentation.
• Noise-robust **audio classifiers**;
• Research contract on applied AI with **Ferrari**.
Stack: PyTorch, TorchAudio, TorchVision, Scikitlearn, OpenMMLab, Ultralytics.

Park Smart — Deep Learning Engineer Consultant | 2023
• Dataset creation and processing and Model training and exportation of YOLOv8 in onnx format.
• Real-time CV pipelines with **NVIDIA DeepStream with an inference plugin** (i.e. YOLOv8), accuracy/latency benchmarking.
Stack: Python, Nvidia Deepstream SDK Python Bindings, Ultralytics.

Vicosystems S.r.l. — ML/DL Engineer Consultant | 2021 – 2023
• Designed and developed a Multimodal DL platform for **predictive maintenance** on time-series (EC-funded).
• Multimodal Data generation with **CTGAN** to resolve data quantity in industrial context with multivariate sensors.
• Audio classification frameworks for robust inference.

Earlier (2016 – 2021):
Paycasso/Xydus (ML & Full-Stack), Reply (Software Analyst), WebRTC Streaming Engineer, VSearch (Flask/Celery search engine).

Education

- **PhD, Computer Science (Deep Learning)** — University of Catania (2020–2024)
- **MS, Computer Science (Security & AI)** — University of Catania (2016–2019)
- **BS, Computer Science** — University of Catania (2011–2016)

Skills

Search & Retrieval: indexing, retrieval, ranking, **RAG**, vector DB (**PGVector, FAISS**), evaluation (**RAGAS**).
ML/DL: TensorFlow, PyTorch, Hugging Face.
NLP/Multilingual: embeddings, machine translation, query understanding with LLMs.
Systems: FastAPI, Airflow, Docker, Kubernetes, NVIDIA DeepStream.
Programming/Data: Python, Java, SQL/NoSQL, C.
Languages: Italian (native), English (C1). **Interests:** IT, Music, playing piano jazz, singing, anime.