| JUAN GABRIEL GARASSINO

GenAI/LLM & Machine Learning Engineer | +6 years exp

📍 | Berlin, Germany
📧 | juan.garassino@gmail.com
📞 | +49 152 24024 860
💼 | linkedin.com/in/juan-garassino
💻 | github.com/juan-garassino
🌐 | artificial-artifacts.com

Generative AI & Machine Learning Engineer specializing in LLM systems, multi-agent orchestration, and reinforcement learning. Designs agentic infrastructures and multi-agent, multi-tool systems built from scratch as well as with LlamaIndex, LangChain, and LangGraph, integrating advanced fine-tuning and optimization techniques such as GRPO, PPO, DPO, LoRA, QLoRA, quantization, and distillation. Bridges rigorous machine learning engineering with creative AI research in generative design and neural spatial systems for architecture, supported by a strong foundation in deployment, containerization, and CI/CD automation.

EXPERIENCE

Forward Deployed AI/ML Engineering Consultant | Engenious Inc

March 2026 – Present

Hands-on technical lead driving customer engagements end-to-end — architecting solutions, building core AI services, managing clients, and aligning technical delivery with business objectives. Architected an enterprise Document AI service processing high-volume multi-format documents (PDFs, images, spreadsheets) with hybrid OCR and multi-modal LLMs into structured Pydantic schemas. Deployed scalable multi-agent pipelines as Dockerised SQS workers on AWS ECS. Implemented Model Context Protocol (MCP) integrations provisioned via Terraform and automated CI/CD via GitHub Actions.

Tech Stack: Python, PyTorch, OpenAI API, Anthropic API, Google Gemini, Model Context Protocol (MCP), FastAPI, Docker, AWS (ECS, SQS, S3), Terraform, GitHub Actions, Pydantic, LlamaIndex, LangGraph.

Lead Data Science & AI Instructor | Le Wagon · neuefische

(Berlin / Tokyo / Amsterdam / Kuala Lumpur / London / Dubai / Munich / Barcelona / Riyadh) | 2021 – Present

Taught ML, DL, LLM Engineering, MLOps, and Data Engineering to 300+ students and professionals across global campuses including Berlin, Tokyo, Amsterdam, Kuala Lumpur, London, Dubai, Munich, Barcelona, and Riyadh. Delivered complete end-to-end training and real-world technical solutions.

Tech Stack: Python, PyTorch, TensorFlow, Scikit-learn, NumPy, Pandas, LlamaIndex, LangChain, FastAPI, Streamlit, Docker, MLflow, Vertex AI, Prefect.

Head of Generative AI Engineering | Caidera GmbH

Berlin, Germany | Sep 2024 – Jun 2025

Designed, fine-tuned, and deployed multi-agent systems with tool calling, memory, context engineering, automated evaluations, and monitoring. Architected end-to-end Generative AI products from agent intelligence to production APIs using Docker, Azure, and FastAPI.

Tech Stack: Python, PyTorch, Transformers, LlamaIndex, LangChain, LangGraph, Hugging Face, MLflow, Docker, Azure, FastAPI.

Design Manager | European Stone Masons Ltd

Auckland, New Zealand | 2016 – 2019

Directed CAD/BIM workflows, 3D modeling, and on-site coordination for civic and commercial projects. Integrated parametric design and digital fabrication using Rhino + Grasshopper.

Stack: AutoCAD, Revit, Rhino + Grasshopper, SketchUp, Adobe Suite, MS Project.

TECHNICAL SKILLS

Languages & Frameworks

Python • PyTorch • TensorFlow • Transformers • Pydantic • LlamaIndex • LangChain • LangGraph • Model Context Protocol (MCP) • HuggingFace • FastAPI • Gradio • Streamlit • Prefect • MLflow • deepEval • OpenAI / Anthropic / Gemini APIs

Architectures & Algorithms

Transformers • Diffusion • GANs • VAEs • LoRA • QLoRA • MoE • Mamba • RLHF • GRPO • PPO • DPO • PEFT • Unsloth • FlashAttention-2 • Quantization • Distillation • YOLO • U-Net

Data & Infrastructure

Docker • Kubernetes • Terraform • GitHub Actions • AWS (ECS, SQS, S3, Lambda, SageMaker) • GCP (Vertex AI, Cloud Run, BigQuery, GKE) • Azure ML • Postgres • Redis • ChromaDB • Neo4j • Pinecone • FAISS • DuckDB • ONNX

Cloud Platforms

GCP (Vertex AI, Cloud Run, BigQuery, GKE) • AWS (ECS, SQS, S3, Lambda, SageMaker) • Azure (ML, Container Instances)

Tools & Design

Git • Poetry • Pyenv • VS Code • Jupyter • OpenCV • Weights & Biases • Adobe Suite • Revit • Rhino + Grasshopper

CONSULTING EXPERIENCE

AI Consultant – Energy Sector (Confidential Client)

Abu Dhabi & Kuala Lumpur | 2023 – 2026 (3 on-site engagements)

Built **U-Net segmentation pipelines** for satellite image analysis and landform detection across three on-site consulting engagements in the Middle East and Southeast Asia. Delivered **Deep Learning and Computer Vision solutions** for energy applications, implementing PyTorch, TensorFlow, OpenCV, and FastAPI training–inference pipelines deployed on GCP Vertex AI.

Stack: Python, PyTorch, TensorFlow, U-Net, OpenCV, FastAPI, Docker, GCP Vertex AI.

AI Consultant – Pharmaceutical & Diagnostics (Confidential Client)

Heidelberg, Germany | Jan 2025 (1-week sprint)

Provided **pre-development consulting and technical feasibility analysis** in an intensive one-week sprint, designing internal AI tools for document retrieval and workflow automation. Advised on the architecture and development of LLM assistants built with Python and API-based integrations for research and process optimization.

Stack: Python, PyTorch, Transformers, LlamaIndex, LangChain, LangGraph, FastAPI, Docker, Azure ML.

Note: Certain client identities and project specifics are anonymized due to non-disclosure agreements. Detailed descriptions available upon request.

EDUCATION

Le Wagon | Data Science & Engineering Certification

Berlin, Germany | 2020

Completed an intensive international bootcamp focused on applied machine learning, deep learning, and data engineering, gaining strong proficiency in Python, SQL, and cloud deployment. Developed hands-on projects in model orchestration, MLOps, and data visualization, combining technical rigor with modern software-engineering practices.

Universidad de Buenos Aires (UBA), Argentina

Diploma in Architecture & Urban Design | 2007 – 2016

Earned a professional architecture degree from one of Latin America's top technical universities. Specialized in architecture and urbanism, with focus on parametric design, smart urban systems, construction, and large-scale project management, blending spatial reasoning, structural logic, and digital fabrication.

École nationale supérieure d'architecture (ENSA) – Beaux-Arts, France

Exchange Program in Parametric Design | 2012 – 2013

Completed an academic exchange within the Beaux-Arts network focused on parametric design and computational architecture. Explored advanced algorithmic modeling, generative form-finding, and digital fabrication, bridging architectural creativity with early principles of procedural and data-driven design.

LANGUAGES & SOFT SKILLS

Languages

Spanish (Native) • English (C1 Cambridge) • German (B2 Berufssprache, TELC) • Italian (A2) • French (A2)

Soft Skills

Leadership • Adaptability • Independence • Proactivity • Creativity • Communication • Problem Solving

AI RESEARCH & PROJECTS

Artificial Artifacts

Personal research lab and open-source platform exploring creative AI, neural spatial generation, autonomous agent workflows, and physical-digital machine interactions.

PromptPlot

Real-time **LLM→pen-plotter interface** enabling **machine-to-machine artistic expression** through live token streams. Translates language model outputs directly into **G-code trajectories**, exploring the intersection of physical motion and generative language.

DeepSculpt

3D **GAN-based framework** for architectural form exploration through **latent-space walks and neural geometry synthesis**. Combines **StyleGAN, diffusion, and VAE hybrids** to generate complex spatial topologies and volumetric abstractions.

Scaffold

Open-source framework for building, testing, and evaluating multi-agent orchestration architectures. Supports dynamic tool discovery, memory management, structured output enforcement, and real-time observability across heterogeneous model providers.

Also built: Neural Architecture Atlas (50 models from scratch) • Tiny LLM Foundry (GPT, MoE, Mamba, GRPO) • NYC Taxi-Fare MLOps (GKE, ONNX INT8) • YAML Agent Harness • MCP Server Suite.