



Imanol Gomez

Applied AI Engineer | LLM Agents, RAG & Evaluation | Customer-Facing Technical Lead

✉ imanol@steinergomez.com

☎ +49 15773950505

📍 Berlin, Germany

🌐 linkedin.com/in/imanolgo

🔗 github.com/ImanolGo

🚩 Spanish

Profile

Applied AI engineer and technical lead with 15+ years of building production software end to end, in Python and C++. I design and build LLM-powered agentic workflows with RAG, vector databases, tool calling, structured outputs and evaluation frameworks that measure accuracy, safety, latency and cost. As Head of AI Deployment at CARIAD (Volkswagen Group), I take AI from prototype to production under strict reliability and compliance requirements, and work directly with enterprise stakeholders (Volkswagen, Audi) and technology partners (Qualcomm, Samsung, NVIDIA). Background in regulated industries: healthcare (IEC 60601) and automotive (ISO 26262, ASPICE).

Skills

LLM & Agentic



- LLM-powered agents, agentic workflows, multi-step agents with tool use, RAG, embeddings, vector databases (pgvector), LangGraph (agent orchestration framework), MCP (Model Context Protocol), structured outputs (Pydantic), prompt engineering, LiteLLM, OpenRouter, frontier models

Software Engineering



- Python (production), C++, Rust, FastAPI, PostgreSQL, REST APIs, Docker, Docker Compose, GitHub Actions, CI/CD, testing, HTMX/Jinja (frontend)

Customer & Delivery



- Enterprise stakeholder management, translating business needs into technical specifications, end-to-end use-case ownership, technical workshops, talks

Evaluation & Observability



- Evaluation frameworks, LLM-as-judge, automated eval gates in CI, accuracy/safety/latency/cost metrics, OpenTelemetry, Arize Phoenix, tracing, cost and token accounting

Machine Learning



- PyTorch, TensorFlow, fine-tuning (DistilBERT), ONNX, TensorRT, model optimization and deployment

Regulated Environments:



- Security, compliance and auditability, healthcare (IEC 60601), automotive safety (ISO 26262, ASPICE)

Professional Experience

10/2025 – Present
Berlin, Germany

CARIAD



Head of AI Deployment

- Lead the team that takes AI models from research prototypes to production on a volume-production ADAS platform (CARIAD–Bosch Automated Driving Alliance).
- Own production reliability, performance and auditability under ISO 26262 and ASPICE: latency, memory and power targets, release quality and traceability.
- Work directly with enterprise stakeholders at Volkswagen Group brands (Volkswagen, Audi) to turn requirements into technical plans and deliveries.
- Lead technical discussions with technology partners Qualcomm, Samsung and NVIDIA on integration and performance.
- Drive shared tooling and automation patterns so model teams deliver consistently across multiple platforms.

04/2023 – 10/2025
Berlin, Germany

CARIAD

Senior ML Software Engineer

- Technical Product Owner for the AI Runtime Environment: built a Python SDK for model conversion, quantization and optimization, plus a runtime abstraction layer used by engineering teams across the organization.
- Supported internal customer teams end to end, from scoping their use cases to production integration.
- Developed AI perception models (sensor fusion, 3D occupancy, behaviour prediction) and deployed them to production hardware under real-time constraints.

08/2020 – 04/2023
Berlin, Germany

Nela Carepath Technologies

AI Software Engineer → Senior IoT Engineer

- Built NELA Cough, an AI-based tele-monitoring product for chronic-disease patients in a regulated healthcare environment (IEC 60601).
- Owned the system architecture end to end, from requirements to production, including cloud IoT back end (Azure IoT) and device firmware.
- Worked directly with customers, partners and medical hardware suppliers on integration and deployment issues; built demos and technical documentation.

06/2014 – 08/2020
Berlin & International

Self-employed

Creative Technologist (Freelance)

- Scoped and delivered client projects end to end, from ambiguous briefs to production: interactive installations, light and video projection, live performance software; clients included ProtoPixel, ElektroCouture and Moritz Waldemeyer.
- Technical evangelist at ProtoPixel; gave talks and ran workshops and hackathons in Porto, Barcelona, Berlin, Bangkok and Accra.

10/2011 – 07/2014
Linz, Austria

Ars Electronica Linz GmbH

Creative Engineer

- Built interactive computer-vision installations (C++, Python, OpenGL, OpenCV); lead developer of GeoPulse, a spatial-data visualization platform.

Projects

- **Alibi – Agentic LLM application with RAG and evaluation framework** · <http://github.com/ImanolGo/alibi>  | Python, LangGraph, pgvector, MCP, FastAPI, LiteLLM, OpenTelemetry
- Multi-agent LLM system: each agent is a LangGraph state machine (retrieve → think → act → remember) with RAG memory in PostgreSQL/pgvector and code-enforced access control over sensitive information.
- Exposed the application as an MCP server so any MCP client can call its tools; FastAPI + HTMX web interface; deployed with Docker Compose.
- Built an evaluation framework with an automated tool-using agent and LLM judge: 20-run eval with CI quality gates on accuracy, safety (contradiction rate 2.9%), cost (\$0.03 per run) and 90% agreement with a labeller.
- Fine-tuned a DistilBERT classifier exported to ONNX that beats an LLM-judge baseline on F1 (0.33 vs 0.18) at 40 ms instead of 7.7 s per prediction.
- OpenTelemetry tracing of every agent step and LLM call in Arize Phoenix, with token, cost and latency accounting and a daily spend cap.
- **Open source:** contributor to openFrameworks (C++); 80+ public repositories on GitHub.

Education

2009 – 2011
Barcelona, Spain

Master's Degree, Sound and Music Computing

Universitat Pompeu Fabra 

Focus: Computational modeling, perception, cognition, and interactive systems

2003 – 2009
Madrid, Spain

Master's Degree, Telecommunication Engineering

Universidad Autónoma de Madrid

- Specialization: Multimedia Signal Processing, Electronic Systems
- Erasmus year at TU Berlin (2007–2008)

- Research student at Fraunhofer HHI, Image Processing Department (2008–2009), researching forward error correction for video compression.

Languages

Spanish

Native Speaker



English

Professional proficiency



German

Professional proficiency

