

Joaquin Bonifacino

Montevideo, Uruguay • +59898734604 • joaquinbonifacino7@gmail.com • linkedin.com/in/joaquin-bonifacino • jabr7.github.io

AI/ML Engineer (Agentic Systems, RAG, LLM Infrastructure)

- AI/ML Engineer specializing in production-grade agentic LLM and RAG systems, owning the full path from architecture and prototyping to deployment and evaluation.
- Shipped multi-agent and RAG systems for clients across the US, the EU, and Latin America, some serving 5,000+ monthly users in production.
- Led an applied R&D lab turning research into production prototypes, and delivered 15+ AI workshops while teaching Applied AI at ORT University.

WORK EXPERIENCE

Zyte • Ireland (Remote) • 02/2026 - Present

AI/ML Engineer (Agentic Systems, Data Quality) • Full-time

- Technical lead for LLMNL, an AI-powered data-QA engine, and Evolution, the platform on top, on a 3-person team. Drove the full rewrite to a scalable Cloud Run, Celery, and LangGraph architecture and shipped the major features to production, now processing real audit workloads.
- Cut QA turnaround from hours to 30 to 60 minutes for the team. The production launch tripled the user base in two weeks, and a follow-up demo tripled it again.
- Own the system end to end: LangGraph deep agents with per-role model routing, sandboxed execution, a deterministic streaming profiler (HyperLogLog, reservoir sampling), FastAPI and Next.js services, and Langfuse observability. Validates thousands of items in parallel at roughly \$0.15 to \$0.80 per run, serving 20B+ tokens per week
- Exposed the full QA surface over a native MCP server so any developer's agent can call it as an automated quality gate, and act as the product's point of contact for continuous delivery and support.

Promptior • 03/2024 - 02/2026

Applied AI Engineer | Lead, Augmented Behaviors Lab (ABL) • Full-time

- Led Promptior's Augmented Behaviors Lab (ABL), an applied R&D team building POCs in agent memory, real-time voice interfaces, multimodal interaction, and local fine-tuning, turning research into prototypes that shaped product and strategy decisions.
- Cut a financial institution's operations-manual synthesis from ~6 months and \$15K+ of expert labor to ~4 minutes at cents per run, by architecting a production RAG pipeline (Azure Document Intelligence + AI Search + GPT-4o) with strict JSON-schema validation and automated DOCX assembly.
- Delivered, as sole engineer, a legal-research orchestrator over 600K+ Ecuadorian laws: a multi-node LangGraph system with a dual-model (GPT-4o + SLM) retrieval pipeline, multi-stage relevance validation, and a custom citation engine that eliminated hallucinated references in legal advice.
- Architected a multi-agent financial companion for a US client, leading a team of 10, a Supervisor-orchestrated LangGraph system with state-isolated agents, three-tier memory (episodic/semantic/procedural), and per-task model routing (Cerebras + AWS Bedrock) for low-latency intent handling.
- Drove a new AI product from concept to 5,000+ monthly users and built the MLOps behind it, Dockerized deployments, CI/CD for AI models, Langfuse observability, and internal evaluation libraries that kept multi-agent performance consistent at scale.

ORT University • 07/2025 - Present

Course Instructor – Applied Technologies II: Tools for Artificial Intelligence • Full-time

- Design and teach a university AI course on production LLM systems: RAG pipelines, vector databases, and agentic systems with tool calling.

SKILLS

Languages: C#, Java, Python, TypeScript

AI & Agentic Systems:

LangChain, LangGraph, Multi-Agent Systems, Multimodal AI, Real-Time Voice Interfaces, Retrieval-Augmented Generation (RAG)

Frameworks & Platforms:

Angular, ASP.NET Core, .NET, Next.js

Cloud & Infrastructure:

Amazon Web Services (AWS), Azure, CI/CD Pipelines, Docker, Terraform

Databases: DocumentDB, MongoDB, PostgreSQL, SQL Server, Vector Databases

Development Practices: API Design, Git, Scalable Systems

EDUCATION

Bachelor's Degree in Systems Engineering

ORT University
Montevideo, URY
GPA: 3.8
02/2020 - 12/2024

CERTIFICATIONS

AI Fluency: Frameworks and Foundations

03/2025 - 03/2025
Anthropic

Long-Term Agentic Memory With LangGraph

03/2025 - 03/2025
DeepLearning.AI