JUAN MANOEL

Senior AI Engineer & AI Solutions Architect | Over 11 years architecting and engineering production-grade AI systems, Multi-Agent GenAI platforms, Advanced RAG architectures, and scalable MLOps/LLMOps pipelines.

SUMMARY

Hands-on AI Engineer and Architect with a proven track record of bridging the gap between cutting-edge Generative AI research and high-scale enterprise production. Expert in designing autonomous Multi-Agent Systems, Advanced RAG pipelines (hybrid search, re-ranking, vector databases), and robust MLOps/LLMOps governance frameworks. Extensive experience with **Python (FastAPI, PyTorch)**, **LangChain / LangGraph / ADK (Agent Developer Kit)**, **Vector DBs (Qdrant, Azure AI Search, Pinecone, pgvector)**, **MLflow, Databricks, and Cloud AI Platforms (Azure AI, AWS, GCP)**.

FEATURED EXPERIENCE

xertica.ai

Senior AI Engineer & Solutions Architect • Feb 2026 - Present

Architecting enterprise-grade Generative AI and Multi-Agent systems for complex corporate workflows. Designing autonomous agentic networks using LangGraph and ADK (Agent Developer Kit) with tool-calling capabilities, multi-LLM orchestration, prompt routing, and automated evaluation guardrails (TruLens/Ragas). Driving scalable LLMOps pipelines to monitor agent drift, latency, and cost efficiency in enterprise cloud environments.

GenAI | Multi-Agent Networks | ADK | LangGraph | LLMOps | Agentic Workflows

Hapvida NotreDame Intermédica

Senior AI Engineer / MLOps Lead • Feb 2025 - Feb 2026

Engineered healthcare AI solutions and enterprise MLOps architectures. Developed clinical RAG systems and predictive ML models (anomaly detection, patient churn, and show/no-show forecasting), boosting operational efficiency by 18%. Architected automated end-to-end MLOps & LLMOps pipelines using Azure ML Studio, MLflow, and Azure DevOps, reducing model deployment cycles from days to under 2 hours with automated testing and model registry governance.

Healthcare AI | Clinical RAG | MLOps / LLMOps | Azure ML Studio | MLflow

Fundação Getulio Vargas (FGV)

Senior AI & Data Solutions Architect • Jan 2025 - Jul 2025

Designed scalable AI-ready data architectures and knowledge base structures to power analytical engines and conversational AI. Modeled enterprise data schemas optimized for Vector Search and RAG integration, establishing clean architecture, SOLID principles, and data pipelines that accelerated strategic decision-making and automated analytical insight generation.

AI-Ready Data Architecture | Vector Knowledge Bases | Analytical AI | SOLID / Clean Arch

NTConsult

Senior Machine Learning & GenAI Engineer • Jul 2024 - Feb 2025

Built custom Generative AI Agents and Advanced RAG (Retrieval-Augmented Generation) frameworks powered by proprietary LLMs. Engineered multi-stage retrieval pipelines combining hybrid search (dense & sparse embeddings), re-ranking algorithms, and vector stores (Azure AI Search / Qdrant) to enable long-term agent memory, semantic context retrieval, and low-hallucination responses for enterprise operations.

GenAI Agents | Advanced RAG | Hybrid Search & Re-ranking | Vector DBs | Long-Term Memory

Ipiranga

Senior Machine Learning & MLOps Engineer • Jan 2023 - Jul 2024

Led an agile MLOps & ML engineering squad building production infrastructure for large-scale model deployment, automated retraining, drift monitoring, and feature stores. Established robust CI/CD/CT pipelines, automated testing (TDD), and model governance, cutting model development-to-production time by 40% while achieving 92% test coverage across ML pipelines.

Enterprise MLOps | CI/CD/CT Pipelines | Model Monitoring & Drift | Feature Store | TDD & Governance

Neoway

Senior MLOps & Software Engineer • Mar 2022 - Feb 2023

Architected and standardized automated MLOps workflows and containerized model deployments using Apache Airflow, Docker, and Kubernetes (K8s). Implemented continuous integration, automated model validation, data quality checks, and monitoring pipelines for ML systems, cutting end-to-end workflow execution time by 35%.

Kubernetes (K8s) | Apache Airflow | ML Pipeline Automation | Containerization | Data Quality & Monitoring

Via Varejo SA

Machine Learning & MLOps Engineer • Mar 2021 - Apr 2022

Orchestrated end-to-end MLOps lifecycle for high-traffic recommendation and forecasting models. Built distributed data and inference pipelines with Azure Data Factory, Apache Spark (PySpark), and MLflow, ensuring high data consistency, low-latency prediction serving, and rock-solid scalability during mega-retail events like Black Friday.

High-Scale MLOps | Apache Spark / PySpark | MLflow | Azure Data Factory | Real-Time Inference

Rede Globo

Software & Data Pipeline Engineer • Nov 2019 - Oct 2020

Re-architected legacy batch data processing systems into a real-time streaming MLOps-ready data pipeline using Apache Kafka, RabbitMQ, and microservices. Reduced data processing latency from 8 hours to under 20 minutes (~96% reduction), empowering downstream ML models and real-time analytics engines with high-availability streaming data.

Real-Time Data Streaming | Apache Kafka | Event-Driven Architecture | High-Availability Pipelines

Academic Background

Stanford University

Postgraduate Degree, Computer Science • Sep 2025 - Jan 2026

Federal University of Technology - Paraná (UTFPR)

Postgraduate Degree, Artificial Intelligence • Jan 2024 - Jun 2025

Pontifical Catholic University of Rio de Janeiro (PUC-Rio)

Postgraduate Degree, Software Engineering • Dec 2022 - May 2024

Pontifical Catholic University of Paraná (PUC-PR)

Big Data and Analytical Intelligence • 2020 - Jul 2022