Osman Vilchez

+51 943 610 349 osman.vilchez@utec.edu.pe linkedin.com/in/osman-vilchez-aguirre-163708167 github.com/osoman2

Summary

AI/ML Engineer with strong experience designing and deploying machine learning systems, generative AI solutions, and multi-agent architectures in cloud-native environments. Proven track record building production-ready ML pipelines across model training, evaluation, deployment, and monitoring, as well as LLM orchestration and computer vision workflows. Experienced in Azure, AWS, and GCP, with hands-on work integrating vector databases, cloud storage, and scalable inference services. Strong background in Python and applied machine learning, focused on translating business requirements into robust, measurable AI-driven systems.

Technical Competencies

Languages Python, Go, C++
ML & GenAI LLM Orchestration, LangChain, LangGraph, Prompt Engineering, RAG, Computer Vision
Libraries TensorFlow, Pandas, Dask, FastAPI, Streamlit, React
Cloud Azure, AWS, GCP
Data & Storage Vector Databases, MongoDB, Redis
DevOps & MLOps CI/CD, Model Deployment, Monitoring
Certificates Professional ML Engineer (GCP), Neo4j Professional

Work Experience

Senior AI Engineer — **Plain Concepts** — **January 2026 – Present**

- Design and implementation of AI agents for object identification and classification using Azure AI Foundry.
- Development of multi-agent workflows for document understanding and automated decision pipelines.
- Integration of cloud-native storage and data services to support scalable AI inference and processing.
- Collaboration with cross-functional teams to deliver production-grade AI solutions for enterprise use cases.

Machine Learning Engineer — **Zazmic** — **July 2024 – January 2026**

- Design and deployment of GenAI and multi-agent systems for enterprise AI solutions on GCP.
- Development of computer vision pipelines and LLM-powered workflows for production environments.
- ML benchmarking and architecture evaluation to select optimal cloud and AI stacks.

AI Developer — **Quanta Solutions** — **August 2023 – Present**

- Development of LLM-based applications using LangChain, vector databases, and cloud APIs.
- Implementation of TTS, voice cloning, object detection, and generative image/video pipelines.
- Design of multi-agent architectures for automation and decision support systems.

Senior Artificial Intelligence Developer — **Arroyo Consulting** — **August 2022 – April 2025**

- Design and implementation of serverless data pipelines and AI services in cloud environments.
- Led internal AI transformation initiatives, introducing GenAI tools and ML workflows.
- Development of multi-agent orchestration frameworks for internal AI platforms.
- **Achievements:** Accelerated AI adoption across technical and non-technical teams through reusable POCs and internal AI tooling.

LLM Consultant — **Securitec and YaVendio** — **November 2023 – August 2024**

- Optimization of LLM inference pipelines using caching and parallel processing.
- Implementation of vector-based retrieval systems for semantic search and summarization.

Data Science Analyst — **MiBolsillo App** — **March 2022 – December 2022**

- Development of clustering and predictive models for user behavior analysis.
- Statistical analysis and data extraction for product analytics.

Education

Bachelor of Science in Computer Science — **Universidad de Ingeniería y Tecnología (UTEC)**

Languages

- Spanish: Native
- English: B2
- Portuguese: B2

Extra Experience - Research

Real-time Potato Classification **CIP - UTEC** **August 2022**

- Open Dataset Publications: Image recognition for native potatoes and Mobile prototype for the recognition of images of native potatoes.
- Support in labeling platform development and dataset curation.
- Training and optimization of YOLO and CNN models for crowded scenes.

Master AI Advisor **UTEC** **May 2024**

- Design of lab sessions focused on applied ML, fine-tuning strategies, and real-world AI use cases for Master students.