Rody Vilchez

Applied ML / AI Systems — Retrieval · Document Intelligence · Evaluation & Robustness

rody.vilchez00@gmail.com | +51 987 082 126 | rosewt.dev | GitHub | LinkedIn

I design applied AI systems for non-ideal conditions: retrieval, document intelligence, and data pipelines over noisy multilingual corpora. Currently at CIP (CGIAR), building document processing and question answering workflows for agricultural research. My focus is evaluation, robustness, and model behavior under real constraints.

Experience

AI / Data Intern — **International Potato Center (CIP, CGIAR)**
Lima, Peru — *Oct 2025 – Present*

- Designed document processing pipelines for an internal GraphRAG workflow over multilingual corpora (Spanish, English, French, Portuguese, Chinese) with noisy OCR, irregular layout and partial classification; ingestion, parsing, chunking, embedding and vector storage
- Implemented LLM-based structured metadata enrichment with schema validation, batching and rate-limit backoff to improve retrieval quality over heterogeneous documents
- Co-built an IT support agent in Copilot Studio deployed in Teams for level-0 resolution over internal technical documentation and escalation to ticketing
- Designed the escalation flow: when the agent cannot solve a case or the user explicitly requests it, it pre-fills the ticket from conversational context via structured output, with human-in-the-loop review through Adaptive Cards before submission

QA Trainee — **Visma LATAM**
Lima, Peru — *Dec 2024 – Oct 2025*

- Built an LLM-based agent that generates automated end-to-end tests from specifications, reducing manual effort in creating and maintaining regression suites
- Developed Cypress regression suites integrated into Jenkins for critical flows that had to remain stable across successive integrations
- Built DOM-aware test generators that extracted selectors and runtime state from live applications, improving maintainability under UI changes

Projects & Research

Imitator — Multimodal Sign Language Translation [GitHub] *Publication: Springer CCIS 2026 · SIMBIG 2025 · WAILAMP 2025*

- Reformulated sign language translation as alignment in an LLM latent space, avoiding gloss as an intermediate representation
- Designed an architecture with latent queries and cross-attention that projects keypoint sequences into token-aligned embeddings, decoupling temporal input length from output length
- Showed that embedding imitation enables robust learning in low-resource settings, with stable alignment (MSE + cosine $\approx$ 8e-4) without retraining the LLM

GENO-MAP — Correspondence-Free Diagnostics for High-Dimensional Data [GitHub] *PCA, UMAP, kNN Graphs*

- Designed a correspondence-free validation framework based on kNN graph invariants to evaluate neighborhood structure in high-dimensional representations
- Showed that neighborhood structure remains robust under severe perturbations, with continuous degradation and no phase transitions
- Showed that PCA preserves structural stability better than autoencoders, and that UMAP changes the visualization rather than the analytical graph
- Presented as a poster at SALA 2026, validating the approach on real-world data without explicit correspondence

ArbitrIA — Legal Retrieval System [Restricted] *LlamaIndex, FastAPI, PostgreSQL, Docker*

- Designed a retrieval system for Peruvian arbitration documents, combining document-level and chunk-level indexing to improve precision on complex queries
- Implemented robust pipelines for heterogeneous PDFs, including multi-column layouts, embedded tables, and inconsistent headers
- Evaluated chunking strategies and showed that finer segmentation improves local precision while hurting global retrieval, motivating dual indexing

Gallstone Risk — ML for Resource-Constrained Screening [Demo] [GitHub] *XGBoost, SHAP, Optuna*

- Reframed gallstone prediction as a decision system under observability constraints, removing dependence on clinical variables unavailable in the field
- Evaluated the trade-off between predictive performance and operational viability, showing controlled degradation as the feature space is reduced
- Designed a human-in-the-loop inspection interface for individual predictions and feature sensitivity analysis with SHAP

Education

B.Sc. Computer Science — **Universidad Peruana de Ciencias Aplicadas (UPC)**
Expected graduation: 2026-2

Skills

ML / AI Systems	PyTorch, scikit-learn, Optuna, model evaluation, multimodal pipelines
Retrieval / Document AI	embeddings, Qdrant, LlamaIndex, chunking, parsing, document processing
Data / Backend	Pandas, FastAPI, Flask, REST APIs, MongoDB, PostgreSQL, ETL
Infrastructure	Docker, Git, Linux, Jenkins, CI/CD

Certifications

Developing Solutions for Microsoft Azure (AZ-204T00, WTC, 2026) · GitHub Foundations (GH-900T00, WTC, 2026) · AI Engineer for Data Scientists (DataCamp, 2025) · Machine Learning Specialization (Google Cloud, 2025) · Google Data Analytics (Google, 2024) · Human-Centered AI (Tec de Monterrey, 2022)

Activities

DataFest - BCP x ESAN (2nd place, 2025) · SALA 2026 - Summit of AI in LatAm (full grant recipient and participant, 2026) · Asociación KP (Volunteer, 95h total, 2022–2023)

Languages

Spanish (native) · English (intermediate)