Rody Vilchez

IA Aplicada / ML Systems — Retrieval · Inteligencia Documental · Evaluación y Robustez
| rody.vilchez00@gmail.com +51 987 082 126 | rosewt.dev | GitHub | LinkedIn

Diseño sistemas de IA aplicada para condiciones no ideales: retrieval, inteligencia documental y pipelines de datos sobre corpus ruidosos y multilingües. Actualmente en CIP (CGIAR) construyendo pipelines de procesamiento documental y question answering sobre investigación agrícola. Publicación aceptada en Springer CCIS 2026 sobre traducción multimodal de lengua de señas. Mi foco está en la evaluación, la robustez y el comportamiento de modelos bajo restricciones reales.

Experiencia

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

- Diseñé pipelines de procesamiento documental para alimentar un GraphRAG interno de CIP/CGIAR, sobre corpus multilingüe (español, inglés, francés, portugués, chino) con OCR ruidoso, layout irregular y clasificación parcial; ingesta, parsing, chunking, embedding y almacenamiento vectorial
- Implementé enriquecimiento de metadata con structured output basado en LLM (validación de esquema, batching, backoff frente a rate limits) para mejorar recuperación sobre documentos heterogéneos
- Diseñé y construí en conjunto un agente de soporte TI en Copilot Studio desplegado en Teams, cubriendo resolución de consultas nivel 0 (resolución sobre documentación técnica interna) y escalamiento a ticketing
- Diseñé el flujo de escalamiento: cuando el agente detecta que no puede resolver o el usuario lo pide explícitamente, genera un prellenado del ticket (structured output) a partir del contexto conversacional, con revisión human-in-the-loop vía Adaptive Cards antes del envío

QA Trainee — **Visma LATAM**
Lima, Perú — *Dic 2024 – Oct 2025*

- Construí un agente basado en LLM que genera tests automatizados end-to-end a partir de especificaciones, reduciendo el esfuerzo manual en la creación y mantenimiento de suites de regresión
- Desarrollé suites de regresión automatizadas con Cypress integradas en Jenkins, cubriendo flujos críticos que debían permanecer estables a través de integraciones sucesivas
- Construí generadores de pruebas DOM-aware para extraer selectores y estado desde aplicaciones en ejecución, mejorando mantenibilidad del testing frente a cambios de UI

Proyectos e Investigación

Imitator — Traducción Multimodal de Lenguaje de Señas [GitHub] *Publication: Springer CCIS 2026 · SIMBIG 2025 · WAILAMP 2025*

- Reformulé el problema de traducción de señas como alineación en el espacio latente de un LLM, evitando el uso de gloss como representación intermedia
- Diseñé una arquitectura con queries latentes y cross-attention que proyecta secuencias de keypoints a embeddings token-aligned, desacoplando longitud temporal de la salida
- Demostré que la imitación de embeddings permite aprendizaje robusto en escenarios low-resource, con alineación estable (MSE + cosine ≈ 8e-4) y sin necesidad de retraining del LLM

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

- Diseñé un framework de validación sin correspondencia basado en invariantes de grafos kNN para evaluar la estructura de vecindad en representaciones de alta dimensionalidad
- Mostré que la estructura de vecindad es robusta bajo perturbaciones severas, con degradación continua y sin cambios de fase
- Evidencié que PCA preserva mejor la estabilidad estructural que autoencoders y que UMAP no altera el grafo analítico, solo su visualización
- Presentado como poster en SALA 2026, validando el enfoque en un entorno de datos reales sin correspondencia explícita

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

- Diseñé un sistema de retrieval para documentos de arbitraje peruano, combinando indexación a nivel documento y chunk para mejorar precisión en consultas complejas
- Implementé pipelines robustos para PDFs heterogéneos (multi-columna, tablas, encabezados inconsistentes)
- Evalué estrategias de chunking mostrando que segmentación fina mejora precisión local pero degrada recuperación global, motivando indexación dual

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

- Reformulé el problema de predicción de colelitiasis como un sistema de decisión bajo restricciones de observabilidad, eliminando dependencia de variables clínicas no disponibles en campo
- Evalué el trade-off entre desempeño predictivo y viabilidad operativa, mostrando degradación controlada al reducir el espacio de features
- Diseñé una interfaz de inspección human-in-the-loop para analizar predicciones individuales y sensibilidad de variables mediante SHAP

Educación

B.Sc. Ciencias de la Computación — **Universidad Peruana de Ciencias Aplicadas (UPC)**
Graduación estimada: 2026-2

Reconocimientos y actividades

DataFest - BCPxESAN (2do lugar, 2025) · SALA 2026 - Summit of AI in LatAm (full grant y participante, 2026) · Asociación KP (voluntariado, 95h, 2022–2023)

Habilidades

ML / AI Systems	PyTorch, scikit-learn, Optuna, evaluación de modelos, pipelines multimodales
Retrieval / Document AI	embeddings, Qdrant, LlamaIndex, chunking, parsing, procesamiento documental
Data / Backend	Pandas, FastAPI, Flask, REST APIs, MongoDB, PostgreSQL, ETL
Infraestructura	Docker, Git, Linux, Jenkins, CI/CD

Certificaciones

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) · IA centrada en el ser humano (Tec de Monterrey, 2022)

Idiomas

Español (nativo) · Inglés (intermedio)