

Rody Vilchez

Applied ML Engineer — Retrieval · Agent Systems · Evaluation · Urban ML

rody@rosewt.dev | +51 987 082 126 | rosewt.dev | [GitHub](#) | [LinkedIn](#)

Ingeniero de ML aplicado en la intersección de sistemas agénticos, retrieval y evaluación bajo datos imperfectos. En CIP-CGIAR construyo plataformas multiagente en producción, GraphRAG multilingüe e infraestructura de evaluación; en investigación, estudio fallos de medición en urban ML y soy coautor de una arquitectura multimodal publicada en Springer CCIS.

Experiencia

AI Intern

Lima, Perú

International Potato Center (CIP-CGIAR)

Oct 2025 – Presente

- **Plataforma multiagente en Microsoft Teams** — Arquitecté y llevé a producción sobre Copilot Studio un sistema institucional para soporte TI y servicios corporativos, con delegación entre dominios y escalamiento a tickets prellenados desde el contexto conversacional, sujetos a revisión humana.
- **Suite de evaluación agent-based** — Construí una suite donde agentes evaluadores ejecutan escenarios derivados de registros operativos; combina regresión determinista, LLM-as-judge agnóstico de proveedor y simulación por perfiles para atribuir fallos de ruteo, delegación y calidad de respuesta.
- **GraphRAG multilingüe** — Construí end-to-end la capa de document intelligence que alimenta un GraphRAG agrícola en cinco idiomas sobre corpus con OCR ruidoso y layouts irregulares: parsing, chunking, extracción estructurada de metadatos, embeddings y almacenamiento vectorial.

QA Trainee

Lima, Perú

Visma LATAM

Dic 2024 – Oct 2025

- **Síntesis de pruebas con LLM** — Construí un agente que transforma especificaciones funcionales en escenarios E2E ejecutables.
- **Automatización DOM-aware** — Desarrollé extracción de selectores y estado del DOM, integrando suites Cypress generadas en Jenkins como cobertura de regresión para flujos críticos.

Investigación y sistemas seleccionados

Infelix — Riesgo delictivo latente bajo sesgo de medición

Investigación de tesis · UPC · 2025–2026

- **Modelado del delito no denunciado** — Construí un modelo bayesiano sobre encuestas de victimización y un panel H3 de 11 fuentes geospaciales (Sentinel-2, VIIRS, OSM, WorldPop y Street View) para estimar riesgo delictivo en 43 distritos de Lima-Callao; los datos censales fueron la única fuente que mejoró consistentemente las predicciones dentro de cada distrito.
- **Transferencia entre ciudades** — Las features aprendidas en Lima igualaron modelos con un año de datos locales en Arequipa y se mantuvieron competitivas en Cusco y Piura; las pruebas mostraron que la calidad del registro, no solo la arquitectura, determina cuándo un sistema puede evaluarse y transferirse.

Multimodal Sign Language Model

Springer CCIS 2895, 2026

- Coautor de una arquitectura multimodal publicada, sin glosas, que usa queries latentes y cross-attention para proyectar keypoints 2D al espacio de embeddings de un modelo de lenguaje preentrenado, evaluada en corpus peruanos y argentinos. DOI: [10.1007/978-3-032-20322-9_23](https://doi.org/10.1007/978-3-032-20322-9_23)

Wachi — Ruteo peatonal sensible al riesgo

El Comercio Lab

- Desarrollé un sistema geoespacial que reordena rutas peatonales por riesgo espaciotemporal en Lima mediante superficies H3, decaimiento temporal y seis franjas horarias para una experiencia interactiva de periodismo de datos.

Contribuciones open source

- **Microsoft Copilot Studio** — Diagnostiqué fallos de sincronización de knowledge sources en subagentes y validé correcciones de prerelease en la extensión oficial de VS Code.
- **Gemini CLI** — Corregí un esquema de UI y añadí cobertura de regresión al CLI agéntico open source de Google.
- **scikit-learn** — Mejoré la documentación técnica de HDBSCAN.
- **beads (bd)** — Implementé prefix-routed write-through y el test de regresión que desbloqueó la consolidación de la implementación upstream.

Educación y reconocimientos

Ciencias de la Computación

Universidad Peruana de Ciencias Aplicadas (UPC)

Graduación prevista: fines de 2026

Segundo puesto, DataFest BCP × ESAN 2025 · Beca completa SALA 2026

Foco técnico

Retrieval y agentes: LlamaIndex, Qdrant, Copilot Studio, Power Automate, orquestación multiagente

Evaluación e investigación: Python, PyTorch, scikit-learn, LLM-as-judge, diseño experimental, robustez e incertidumbre

Datos y backend: FastAPI, H3, DuckDB, PostgreSQL, ETL, APIs REST, Docker, GitHub Actions, Linux

Idiomas: Español (nativo) · Inglés (intermedio)