

Jesús Enrique Roncal Huatta

Senior Software Engineer | AI & Backend

Milán, Italia jesusroncal94@gmail.com linkedin.com/in/jesús-enrique-roncal-huatta-453a15109
github.com/jesusroncal94

RESUMEN

Ingeniero de software senior con más de 8 años en sistemas backend, APIs e infraestructura cloud; los últimos 4+ construyendo y operando aplicaciones en producción basadas en LLM. Recientemente Technical Lead de dos productos de IA, y actualmente AI Engineer en un tercero, diseñando sistemas multiagente, pipelines RAG y frameworks de evaluación para usuarios reales a escala. Trabaja a diario con agentes de programación con IA (Claude Code) y con las APIs de Claude, Gemini y GPT. Desarrolla y opera servicios en AWS y GCP con Python, Go, Java y TypeScript, con experiencia previa en producción con PHP (Laravel, CodeIgniter). Colabora de forma transversal con equipos distribuidos en EE. UU., la UE, LATAM y Asia-Pacífico.

EXPERIENCIA

AI Engineer — InterCorp Management (NGR, Innova LATAM), Perú - remoto sept 2024 - hoy

- Diseño y opero el sistema LLM en producción detrás de Nauta, la plataforma docente de Innova Schools, sobre Gemini 2.5 Pro y 3.x (Vertex AI), que genera planes de clase alineados al currículo para Perú, Colombia y México: ~25.000 peticiones, más de 23.000 sesiones y 362 docentes activos en cuatro semanas, a ~\$0,09 por plan.
- Propuse el paso de n8n a un servicio en Spring Boot / Spring AI (Java 21) y llegué a ser su segundo mayor contribuidor; corre en AWS ECS (DEV/STG/PRD) con prompts versionados con SemVer, monitorización del coste por llamada y un dashboard de Grafana de 19 paneles.
- Entregué el motor de generación de unidades (v3): generación concurrente por lotes de unidades completas con tolerancia a fallos por sesión y guardrails deterministas (sin llamadas extra al LLM), con 14.900 sesiones pregeneradas.
- Lideré el paso de 15 tareas de IA a Gemini 3.x con un arnés de mediciones pareadas sobre casos de producción (coste, latencia, calidad con LLM-as-judge), con un +2,9 % de coste total; construí una POC de agente conversacional de edición sobre Google ADK (gate, planificador, ejecutor en paralelo).
- Soy responsable de la gobernanza del coste de IA: informes mensuales al Comité Ejecutivo conciliados con la factura cloud, un modelo de costes por país, una banda de presupuesto aprobada y alertas de presupuesto en Slack.
- Antes, para NGR (a través del programa Salto de La Victoria Lab), construí un asistente de ventas por WhatsApp para Bembo sobre Azure OpenAI GPT-4o con detección de intención mediante function calling.

Technical Lead - AI Products — MindFortress Inc (via Revelo), EE. UU. - remoto oct 2025 - sept 2026

- Fui Technical Lead de dos productos de IA (Freya, un compañero de escritura con IA en producción; Reeve, una plataforma de IA para el crecimiento de comercios en fase MVP) en un equipo repartido entre husos horarios de EE. UU., Europa y Asia-Pacífico: fui responsable de la revisión de código y el control de releases, coordiné las decisiones técnicas con el equipo y apoyé a mis compañeros con revisiones y guías de incorporación.
- Diseñé y operé el sistema de entrega asistida por IA del equipo sobre Claude Code, empaquetado como mindfortress-claude y adoptado por el equipo: agentes por rol y comandos para cada etapa del SDLC, herramientas de sesión que cruzan Linear, GitHub y Slack, gates de aprobación humana para los despliegues y los estándares que codifica (desarrollo guiado por especificaciones, estrategia de ramas).
- Hice de gate de releases de Reeve (más de 1.900 PRs fusionados, la mayoría de compañeros del equipo); en la campaña de merge del MVP, vacié una cola de más de 110 PRs en 4 repositorios en ~30 h orquestando agentes de Claude Code en paralelo (worktrees aislados, cadenas de PRs apiladas) tras merge gates seriales, y corregí hallazgos de seguridad por el camino.
- Co-diseñé como Technical Lead la capa de IA de Freya y fui autor de su modelo en "reloj de arena" (muchos canales de entrada alimentan una única base de conocimiento solo con hechos que guía la generación de prosa, con subagentes de entrada y de salida), y la preparé para producción: enrutamiento unificado de modelos, aislamiento por proyecto, seguimiento real de tokens y costes y llamadas de IA no bloqueantes. La plataforma abarca 7 agentes con RAG, un motor de consistencia basado en un grafo de conocimiento y edición con Yjs/CRDT sobre FastAPI (más de 900 endpoints).
- Devolví al CI la suite completa de tests del backend de Freya (pytest y upgrades de Alembic sobre un Postgres real) y llevé la base de fallos heredada de 64 a 0, encontrando la causa en una memoización obsoleta de un comparador de SQLAlchemy.
- Encontré la causa de una fuga de gasto de IA no atribuido de ~\$16/día: un worker de staging obsoleto que trabajaba contra la base de datos de producción; entregué una protección en el código y separé las credenciales de producción y staging.

- Fui responsable de la salud de los entornos y de los despliegues (runbook de despliegue de Reeve, refuerzo del monitor de disponibilidad de Freya, operaciones en AWS EC2 y Railway) y reforcé el gateway de agentes multicanal de Reeve (gate de autenticación, idempotencia, detección de bucles).

Backend Developer (concurrent) — AccelOne (US client), EE. UU. - remoto ago 2023 - ene 2024

- Construí y operé el backend de una plataforma de video en Django / PostgreSQL sobre AWS, a cargo de los pipelines de almacenamiento, subida y actualización de archivos multimedia muy grandes, con procesamiento asíncrono por colas en AWS SQS.
- Modelé los metadatos de los videos para el filtrado y la búsqueda, e implementé la búsqueda con Elasticsearch.
- Integré Stripe para los pagos y la gestión de contenidos de la plataforma.

Backend Developer & AI Engineer — Wordbox, México / Reino Unido - remoto ene 2022 - nov 2024

- Integré LLMs de Google y OpenAI en el pipeline de contenidos para transcripción, traducción y aprendizaje adaptativo automatizados, con más de 500.000 usuarios en más de 10 mercados, aumentando la satisfacción de los usuarios un 35 %.
- Diseñé un procesamiento asíncrono y orientado a eventos con Redis y Google Cloud Pub/Sub para desacoplar los servicios de la plataforma de contenidos.
- Reescribí servicios centrales de Python a Go, reduciendo la latencia de la API un 40 % y soportando un 50 % más de usuarios concurrentes.

Backend Developer - Freelance (concurrent, part-time) — Indra (Peru) | Pretty Technical 2021 - 2022
(Netherlands)

- Construí APIs que sirven modelos predictivos de ML, sobre todo en Python y con servicios en PHP (Laravel), con un 95 % de cobertura de tests; mejoré el CI/CD, acelerando los ciclos de despliegue un 40 % (Indra, Perú).
- Construí sistemas backend y APIs sobre todo en Go, además de servicios en PHP (Laravel), para el procesamiento de datos de videojuegos, finanzas y secuenciación de ADN (Pretty Technical, Países Bajos; cliente: Elimbio, EE. UU.).

Backend Developer — Red Científica Peruana, Lima, Perú jul 2020 - ene 2022

- Internacionalicé Yachay (yachay.lat), una plataforma de hosting y dominios que da servicio a más de 11.000 empresas, impulsando un 25 % de crecimiento de ingresos en nuevos mercados.

Software Engineer / IT Support Analyst L3 — Codrise (Vilbra Group), Lima, Perú dic 2017 - jul 2020

- Pasé del soporte TI L3 a Software Engineer, desarrollando aplicaciones web y APIs REST en PHP (Laravel, CodeIgniter y PHP puro); integré aplicaciones satélite con bases de datos transaccionales y servicios de reportes.

FORMACIÓN

Bachiller en Ingeniería de Sistemas — Universidad Nacional Tecnológica de Lima Sur, Lima, Perú 2019

HABILIDADES

Lenguajes	Python, Go, Java, TypeScript / Node.js, PHP, SQL, Bash
Backend y APIs	FastAPI, Django, Flask, Spring Boot, Laravel, CodeIgniter, REST, GraphQL, gRPC, WebSocket, SQLAlchemy, Alembic
Asincronía y mensajería	AWS SQS, Google Cloud Pub/Sub, Celery, Redis, event-driven workflows, n8n orchestration
Cloud e infraestructura	AWS (ECS, EC2, RDS, S3, SQS, Cognito, IAM/SSO), GCP (Cloud Run, Vertex AI, Cloud SQL, Pub/Sub), Docker, CI/CD (GitHub Actions), Terraform and Kubernetes (working knowledge)
Bases de datos y búsqueda	PostgreSQL, Redis, Elasticsearch, Qdrant, pgvector, Typesense, Neo4j
Ingeniería de IA / LLM	Multi-agent orchestration, RAG, prompt engineering (few-shot, chain-of-thought, eval-driven), LLM evaluation (LLM-as-judge, quality gates), AI cost governance, Anthropic Claude, Google Gemini / Vertex AI, OpenAI / Azure OpenAI, LangChain, LangGraph, LangFuse, Spring AI, MCP, Claude Code
Monitoreo y DevOps	Grafana, Sentry, PostHog, GitHub Actions, CI/CD
Frontend	React, Next.js (App Router), TypeScript, Tailwind / shadcn/ui, Yjs / CRDT

IDIOMAS

Español — Nativo (C2) · Inglés — Competencia profesional (C1) · Italiano — Básico (A2)