Jesús Enrique Roncal Huatta

Senior Software Engineer | AI & Backend

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

SUMMARY

Senior software engineer with 8+ years across backend systems, APIs, and cloud infrastructure, the last 4+ building and operating production LLM-powered applications. Most recently Technical Lead on two AI products, and currently AI Engineer on a third, designing multi-agent systems, RAG pipelines, and evaluation frameworks for real users at scale. Daily hands-on with AI coding agents (Claude Code) and with Claude, Gemini, and GPT APIs. Ship and operate services on AWS and GCP in Python, Go, Java, and TypeScript, with earlier production experience in PHP (Laravel, CodeIgniter). Strong cross-functional collaborator across distributed US, EU, LATAM, and Asia-Pacific teams.

EXPERIENCE

AI Engineer — Intercorp Management (NGR, Innova LATAM), Peru - Remote Sep 2024 - now

- Design and operate the production LLM system behind Nauta, Innova Schools' teacher platform, on Gemini 2.5 Pro and 3.x (Vertex AI), generating curriculum-aligned lesson plans for Peru, Colombia and Mexico: ~25K requests, 23K+ sessions and 362 active teachers in four weeks, at ~$0.09 per plan.
- Proposed the move from n8n to a Spring Boot / Spring AI service (Java 21) and became its second-largest contributor; it runs on AWS ECS (DEV/STG/PRD) with SemVer-versioned prompts, per-call cost monitoring and a 19-panel Grafana dashboard.
- Shipped the unit-generation engine (v3): concurrent batch generation of whole units with per-session fault tolerance and deterministic guardrails (zero extra LLM calls), powering 14.9K pregenerated sessions.
- Led the move of 15 AI tasks to Gemini 3.x with a paired-measurement harness on production cases (cost, latency, LLM-as-judge quality), at +2.9% total cost; built a conversational editing-agent POC on Google ADK (gate, planner, parallel executor).
- Own AI cost governance: monthly Executive Committee reporting reconciled against the cloud invoice, a per-country cost model, an approved budget band and Slack budget alerts.
- Earlier, for NGR (via La Victoria Lab's Salto program), built a WhatsApp sales assistant for Bembos on Azure OpenAI GPT-4o with function-calling intent detection.

Technical Lead - AI Products — MindFortress Inc (via Revelo), USA - Remote Oct 2025 - Sep 2026

- Technical Lead for two AI products (Freya, an AI writing partner in production; Reeve, an AI commerce-growth platform at MVP) in a team spread across US, European and Asia-Pacific time zones: owned code review and release gating, coordinated technical decisions with the team and supported teammates through reviews and onboarding guides.
- Designed and ran the team's AI-assisted delivery system on Claude Code, packaged as mindfortress-claude and adopted by the team: role agents and commands for each SDLC stage, session tooling that cross-references Linear, GitHub and Slack, human approval gates for deploys, and the standards it encodes (spec-driven development, branching strategy).
- Release gate for Reeve (1,900+ PRs merged, mostly teammates'); in the MVP merge campaign, cleared 110+ queued PRs across 4 repos in ~30h by orchestrating parallel Claude Code agents (isolated worktrees, stacked PR chains) behind serial merge gates, fixing security findings along the way.
- Co-architected Freya's AI layer as Technical Lead, authoring its "hourglass" model (many input channels feed one facts-only knowledge base that drives prose generation, with inbound and outbound subagents), and hardened it for production: unified model routing, per-project isolation, real token/cost tracking and non-blocking AI calls. The platform spans 7 RAG-grounded agents, a knowledge-graph consistency engine and Yjs/CRDT editing on FastAPI (900+ endpoints).
- Restored Freya's full backend test suite to CI (pytest and Alembic upgrades on a real Postgres) and drove the inherited failing baseline from 64 to 0, root-causing a stale SQLAlchemy comparator memoization.
- Root-caused a ~$16/day unattributed AI-spend leak to a stale staging worker running against the production database; shipped a code guard and split prod/staging credentials.
- Owned environment health and deploys (Reeve deploy runbook, Freya uptime-monitor hardening, AWS EC2 and Railway operations) and hardened Reeve's multi-channel agent gateway (auth gate, idempotency, loop detection).

Backend Developer (concurrent) — AccelOne (US client), USA - Remote Aug 2023 - Jan 2024

- Built and operated the backend for a video platform in Django / PostgreSQL on AWS, owning the storage, upload and update pipelines for very large media files with asynchronous, queue-based processing on AWS SQS.
- Modeled video metadata for filtering and search and implemented search with Elasticsearch.
- Integrated Stripe for payments and content management on the platform.

Backend Developer & AI Engineer — Wordbox, Mexico / United Kingdom - Remote Jan 2022 - Nov 2024

- Integrated Google and OpenAI LLMs into the content pipeline for automated transcription, translation, and adaptive learning, serving 500,000+ users across 10+ markets and lifting user satisfaction 35%.
- Designed asynchronous, event-driven processing with Redis and Google Cloud Pub/Sub to decouple services across the content platform.
- Refactored core services from Python to Go, reducing API latency 40% and supporting a 50% increase in concurrent users.

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

- Built APIs serving predictive ML models, primarily in Python with PHP (Laravel) services, achieving 95% test coverage; improved CI/CD, accelerating deployment cycles 40% (Indra, Peru).
- Built backend systems and APIs primarily in Go, also delivering PHP (Laravel) services, for gaming, finance, and DNA-sequencing data processing (Pretty Technical, Netherlands; client: Elimbio, USA).

Backend Developer — Red Cientifica Peruana, Lima, Peru Jul 2020 - Jan 2022

- Internationalized Yachay (yachay.lat), a web hosting and domain platform serving 11,000+ businesses, driving 25% revenue growth in new markets.

Software Engineer / IT Support Analyst L3 — Codrise (Vilbra Group), Lima, Peru Dec 2017 - Jul 2020

- Progressed from L3 IT Support to Software Engineer, building web applications and REST APIs in PHP (Laravel, CodeIgniter and pure PHP); integrated satellite applications with transactional databases and reporting services.

EDUCATION

Bachelor of Systems Engineering — National Technological University of South Lima, Lima, Peru 2019

SKILLS

Languages	Python, Go, Java, TypeScript / Node.js, PHP, SQL, Bash
Backend & APIs	FastAPI, Django, Flask, Spring Boot, Laravel, CodeIgniter, REST, GraphQL, gRPC, WebSocket, SQLAlchemy, Alembic
Async & Messaging	AWS SQS, Google Cloud Pub/Sub, Celery, Redis, event-driven workflows, n8n orchestration
Cloud & Infrastructure	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)
Databases & Search	PostgreSQL, Redis, Elasticsearch, Qdrant, pgvector, Typesense, Neo4j
AI / LLM Engineering	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
Monitoring & DevOps	Grafana, Sentry, PostHog, GitHub Actions, CI/CD
Frontend	React, Next.js (App Router), TypeScript, Tailwind / shadcn/ui, Yjs / CRDT

LANGUAGES

Spanish — Native (C2) · English — Professional working proficiency (C1) · Italian — Elementary (A2)