EMERSON HERRERA TOLENTINO

FULL STACK AI ENGINEER / GENERATIVE AI

Porto Alegre, Brasil – Perú | +51 923 752 936 | +55 51 98128-9816
emersonheto@gmail.com | linkedin.com/in/emerson-herrera-tolentino

PERFIL PROFESIONAL

Ingeniero de Software con más de 8 años de experiencia, especializado en Machine Learning e Inteligencia Artificial Generativa. Combino una sólida base en arquitectura de software escalable con técnicas avanzadas de ML para construir soluciones inteligentes aplicadas a problemas reales. En los últimos dos años me he enfocado en diseñar y poner en producción sistemas de IA Generativa: chatbots empresariales, plataformas de Retrieval-Augmented Generation (RAG), ingesta multimodal de documentos, prompt engineering, integración de LLMs (OpenAI API), embeddings, bases de datos vectoriales (pgvector, Pinecone, Chroma) y Agentic AI. Recientemente construí una plataforma RAG multi-tenant y multi-sistema que atiende a varias aplicaciones bancarias desde una arquitectura compartida. Mi objetivo es construir soluciones de software e IA escalables, seguras, medibles y aplicables a problemas reales de negocio.

COMPETENCIAS TÉCNICAS

IA Generativa & Agentes: RAG (agéntico y adaptativo), LangChain, LangGraph, OpenAI API, Prompt Engineering, embeddings, bases de datos vectoriales (pgvector, Pinecone, Chroma, Qdrant), arquitecturas multiagente, MCP (Model Context Protocol), evaluación de LLMs y LLMOps

Machine Learning: Pandas, NumPy, modelos de visión-lenguaje (CLIP, SigLIP), fine-tuning y adaptación few-shot

Lenguajes & Frameworks: Python (FastAPI), Java (Spring Boot), C# (.NET), Node.js, PHP (Laravel, CodeIgniter), JavaScript/TypeScript

Frontend: Angular, React, Next.js, Vue.js, Tailwind CSS, Filament

Bases de Datos: PostgreSQL, Oracle, SQL Server, MySQL, MongoDB, Redis

Cloud & DevOps: AWS (EC2, S3, Lambda, RDS), Azure, Docker, CI/CD (GitHub Actions), Dokploy, Linux (Red Hat, Ubuntu)

Arquitectura & Integración: Microservicios, arquitectura hexagonal y clean, APIs REST/SOAP, JWT, sistemas distribuidos, automatización de procesos ETL

Calidad: TDD, JUnit, Mockito, PHPUnit, SonarQube/SonarCloud, Postman, Swagger · Scrum (SFPC™), Jira, Git

Idiomas: español nativo · portugués intermedio (buena comprensión) · inglés intermedio (técnico avanzado)

DESARROLLO ASISTIDO POR IA

- Flujo diario con agentes de código (Claude Code, Codex, Pi) sobre Gentle Shell y Gentle AI, con el flujo ODD (Organic Driven Development): exploración antes de cambiar, subagentes para implementar y verificar, TDD y revisión según el riesgo antes de cada commit.
- Memoria persistente con Engram: decisiones de arquitectura, hallazgos y estado de tareas entre sesiones.
- Entorno personalizado para proyectos reales: skills propias, servidores MCP (Oracle de solo lectura, Laravel Boost, logs de contenedores, GitLab, CodeGraph) y combinación de modelos según la tarea.
- Fábrica de agentes en VPS propio: issues ejecutados sin supervisión desde Telegram (n8n) y en tandas nocturnas, con PRs y CI.

EXPERIENCIA PROFESIONAL

Analista Programador II | Full Stack AI Engineer / GenAI & RAG Developer

Canvia · Perú (remoto) *Oct. 2025 — Actualidad*

- Cliente BCRP (Banco Central de Reserva del Perú): diseño, construcción y mejora de una plataforma RAG multi-tenant que atiende a 4 aplicaciones del banco (Portal BCRP, Portal de Investigación, NEXO 360 y Billetes), incluida la ingesta e indexación de miles de documentos y cientos de miles de fragmentos; su asistente está en producción en el Portal de Investigación (investigacion.bcrp.gob.pe).

- Gateway en Node.js que orquesta los sistemas cliente y backend en Python con herramientas compartidas; módulos de chat, búsqueda y prompts administrables por aplicación y por área.
- Pipelines RAG para artículos de investigación y documentos institucionales, con ingesta multimodal de PDFs e imágenes, búsqueda vectorial en Oracle 23ai, pgvector, Pinecone y Chroma, y LLMs vía OpenAI API y LangChain.
- Pipelines SQL para exportar los datasets de la Plataforma de Investigadores a RePEc, con modelado y optimización de consultas en Oracle; microservicios backend desplegados en AWS.
- Uso diario de agentes de IA con memoria persistente y skills propias del proyecto, con revisión humana y pruebas.
- Stack: Python, FastAPI, LangChain, Node.js, Java, React, Angular, PostgreSQL, Oracle, AWS, Azure, Docker.

Desarrollador Full Stack y de IA (freelance)

Certilab S.A.C. · Remoto *Sept. 2024 — Actualidad*

- Sistema de certificaciones en Laravel 11 y Filament, con más de 300 certificados emitidos: backoffice administrativo, intranet donde cada cliente consulta sus certificados y portal público de validación por QR con ID cifrado.
- Emisión de certificados, PDF firmado, generación de QR, roles y permisos, auditoría y almacenamiento en S3 (Docker).
- RAG y chatbot sobre certificados de calibración con ingesta multimodal y panel de consulta para la gerencia; versión agéntica con LangGraph, FastAPI y Qdrant (github.com/emersonheto/certilab-agentic-rag).
- Sitio web corporativo con Next.js 15 y React 19 y despliegue automatizado.

Full Stack Developer | Angular, Java, .NET & Process Automation

Way Tech (proyectos WIN Empresas / Universidad Continental) · Perú (remoto) *Nov. 2024 — Oct. 2025*

- Desarrollo de plataformas académicas e institucionales para la Universidad Continental (más de 24.000 estudiantes en Huancayo, Arequipa, Cusco y Lima) con Angular y microservicios en C# (.NET) y Java.
- APIs RESTful documentadas con Swagger y mejora de rendimiento en sistemas académicos de alto tráfico.
- Automatización de procesos financieros y de negocio con Node.js, asegurando la integración entre sistemas.
- Gestión y optimización de bases SQL (MySQL, SQL Server, Oracle, PostgreSQL) y NoSQL (MongoDB) de alto tráfico.
- Diagnóstico y resolución de incidentes en sistemas Java, PHP, C# y AWS, con documentación técnica de cada solución.

Systems Engineer

Indra · San Isidro, Perú (remoto) *Mar. 2021 — Ago. 2024*

- Cliente Telefónica del Perú: sistemas distribuidos y microservicios de telecomunicaciones, con procesos diarios de altas y bajas de clientes y datos financieros de facturación, reclamos y comisiones.
- Procesamiento en tiempo real de la telemetría de módems de hogares de todo el Perú, por zonas: ETL en Python que integra el estado de red de varias APIs, genera alertas y alimenta la gestión de clientes y el monitoreo.
- Aplicaciones web con Java Spring Boot, PHP (Laravel, CodeIgniter), .NET, Python, PowerBuilder, Angular, React y Vue.js.
- Integración de servicios REST y SOAP con seguridad JWT y servicios de AWS y Azure.
- Administración de MySQL, Oracle, SQL Server, MongoDB y servidores Linux Red Hat; automatización ETL con Shell.
- Pruebas unitarias con JUnit, Mockito y Postman, y análisis de código con Sonar, aplicando metodologías ágiles.

Analista de Sistemas de TI

Multilab · Pueblo Libre, Perú *Feb. 2020 — Mar. 2021*

- Sistemas clínicos transaccionales con Laravel y CodeIgniter: nuevas funcionalidades, reportes y roles y permisos.
- Modelado y optimización de bases de datos; servicios en AWS (S3, EC2) y control de versiones con GitLab.
- Soporte técnico especializado, remoto e in situ, para la continuidad operativa de las sedes.

Jefe de Área de Desarrollo de Software

MControl Energy Technology SAC · Lima, Perú *Jul. 2019 — Ene. 2020*

- Dirección de un equipo de 2 desarrolladores, con entregas oportunas y optimización de sistemas críticos para estaciones de servicio de los grupos Gasocentro e Irasa (Lima e Ica).
- Planificación y ejecución de proyectos de software a medida, alineados a los objetivos del cliente.

Software Developer

MControl Energy Technology SAC · Lima, Perú *Nov. 2017 — Jul. 2019*

- Soluciones a medida y sistemas transaccionales con C#, ASP.NET, PHP (Laravel), JavaScript, TypeScript y Vue.js; modelado y optimización de bases de datos SQL Server y MySQL.

PROYECTOS DESTACADOS

Asistente de reuniones con transcripción local

Python, Docker, Dokploy, WhisperKit, PostgreSQL *2026*

- Bot que se une a las reuniones, biblioteca de grabaciones con transcripción local (WhisperKit), separación de hablantes y resúmenes con LLM; desplegado en VPS con Dokploy, CI y SonarCloud, construido con agentes de IA en PRs revisados.

PUBLICACIONES / INVESTIGACIÓN

Coautor — "Benchmark of Multimodal Vision-Language Models for Damage Detection"

LatinX in CV (LXCV) Workshop, CVPR 2026 *Jun. 2026*

- Benchmark de CLIP, SigLIP, SigLIP2 y FuseLIP para detección de daños en desastres naturales a partir de imágenes y texto. Trabajo liderado por Mariana Risco Cosavalente, con el Prof. Claudio Jung (UFRGS).

FORMACIÓN ACADÉMICA

Alumno Especial — Programa de Posgrado en Computación (PPGC)

UFRGS, Brasil · CMP263 Aprendizaje de Máquina · postulando a la maestría *2026/2 — En curso*

Programa de Especialización en AI Engineer — Diploma de Excelencia (19/20)

Data Science Research Perú · 92 horas académicas *Mar. 2026 — Ago. 2026*

- RAG, agentes, MCP, aplicaciones multimodales, evaluación de LLMs, guardrails y LLMOps; proyecto final: sistema multiagente aplicado a procesos empresariales.

Bachiller en Ingeniería de Software

UTP — Universidad Tecnológica del Perú *2021 — 2024*

Programa de formación en Microservicios con Java y Spring Boot

Lite Thinking *Jun. 2024 — Sept. 2025*

Profesional Técnico en Desarrollo de Software — carrera técnica de 3 años

SENATI *Mar. 2016 — Jul. 2019*

CERTIFICACIONES

- **Data Science Research Perú (2026), Diplomas de Excelencia:** Evaluación, Observabilidad y LLMOps (19) · Agentes Multimodales (19) · Agentes de IA y MCPs, RAGs Agénticos (18) · Prompt Engineering y RAGs (18) · Fundamentos de Ingeniería IA (18) · Procesamiento del Lenguaje Natural.
- **Tecylab:** Especialista en DevOps Multicloud (2025) · Especialista en Angular (2025) · Senior en Microservicios NodeJS con Cloud (2025) · Senior en Programación con Spring Boot en Multicloud (2024).
- Google Cloud Platform — DSRP (2024) · Scrum Foundation Professional Certification (SFPC™) — CertiProf (2024)