EMERSON HERRERA TOLENTINO

FULL STACK AI ENGINEER / GENERATIVE AI

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

PROFESSIONAL SUMMARY

Software Engineer with 8+ years of experience, specialized in Machine Learning and Generative Artificial Intelligence. I combine a strong background in scalable software architecture with advanced ML techniques to build intelligent solutions applied to real-world problems. In the last two years I have focused on designing and shipping production-ready GenAI systems: enterprise chatbots, Retrieval-Augmented Generation (RAG) platforms, multimodal document ingestion, prompt engineering, LLM integration (OpenAI API), embeddings, vector databases (pgvector, Pinecone, Chroma) and Agentic AI. I recently built a multi-tenant, multi-system RAG platform serving several banking applications from a shared architecture. My goal is to build software and AI solutions that are scalable, secure, measurable and applicable to real business problems.

TECHNICAL SKILLS

Generative AI & Agents: RAG (agentic and adaptive), LangChain, LangGraph, OpenAI API, Prompt Engineering, embeddings, vector databases (pgvector, Pinecone, Chroma, Qdrant), multi-agent architectures, MCP (Model Context Protocol), LLM evaluation and LLMOps

Machine Learning: Pandas, NumPy, vision-language models (CLIP, SigLIP), fine-tuning and few-shot adaptation

Languages & 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

Databases: 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)

Architecture & Integration: Microservices, hexagonal and clean architecture, REST/SOAP APIs, JWT, distributed systems, ETL process automation

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

Languages: Spanish (native) · Portuguese (intermediate; good comprehension) · English (intermediate; advanced technical)

AI-ASSISTED DEVELOPMENT

- Daily work with coding agents (Claude Code, Codex, Pi) on Gentle Shell and Gentle AI, following ODD (Organic Driven Development): explore before changing, subagents to implement and verify, TDD and risk-based review before every commit.
- Persistent memory with Engram to keep architecture decisions, findings and task state across sessions.
- Environment tailored to real projects: custom skills, MCP servers (read-only Oracle, Laravel Boost, container logs, GitLab, CodeGraph) and model selection per task.
- Agent factory on my own VPS: unattended issue execution triggered from Telegram (n8n) and nightly batches, with PRs and CI as quality gates.

PROFESSIONAL EXPERIENCE

Analyst Programmer II | Full Stack AI Engineer / GenAI & RAG Developer

Canvia · Peru (remote) *Oct 2025 — Present*

- Client: Central Reserve Bank of Peru (BCRP). Designed, built and improved a multi-tenant RAG platform serving 4 of the bank's applications (BCRP Portal, Research Portal, NEXO 360 and Banknotes), including the ingestion and indexing of thousands of documents and hundreds of thousands of chunks; its assistant is live on the Research Portal (investigacion.bcrp.gob.pe).
- Node.js gateway orchestrating client systems and a Python backend exposing shared tools; chat, search and prompt modules configurable per application and per area.

- RAG pipelines for research articles and institutional documents, with multimodal ingestion of PDFs and images, vector search on Oracle 23ai, pgvector, Pinecone and Chroma, and LLMs through the OpenAI API and LangChain.
- SQL pipelines exporting the Researchers' Platform datasets to RePEc, with data modeling and Oracle query optimization; backend microservices deployed on AWS.
- Daily use of AI agents with persistent memory and project-specific skills, with human review and testing.
- Stack: Python, FastAPI, LangChain, Node.js, Java, React, Angular, PostgreSQL, Oracle, AWS, Azure, Docker.

Full Stack & AI Developer (Freelance)

Certilab S.A.C. · Remoto *Sep 2024 — Present*

- Certification system in Laravel 11 and Filament with 300+ certificates issued: admin back office, client intranet to look up their own certificates, and a public QR validation portal with encrypted IDs.
- Certificate issuance, signed PDF upload, QR generation, roles and permissions, activity auditing and S3 storage, running on Docker.
- RAG and chatbot over calibration certificates with multimodal ingestion and a query dashboard for management; agentic version with LangGraph, FastAPI and Qdrant (github.com/emersonheto/certilab-agentic-rag).
- Corporate website with Next.js 15 and React 19 and automated deployment.

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

Way Tech (WIN Empresas / Universidad Continental projects) · Peru (remote) *Nov 2024 — Oct 2025*

- Built academic and institutional platforms for Universidad Continental (24,000+ students across Huancayo, Arequipa, Cusco and Lima) with Angular and microservices in C# (.NET) and Java.
- Consumed and documented RESTful APIs with Swagger and improved performance of high-traffic academic systems.
- Automated financial and business processes with Node.js, ensuring integration across systems.
- Managed and optimized SQL (MySQL, SQL Server, Oracle, PostgreSQL) and NoSQL (MongoDB) databases for high-traffic transactional workloads.
- Diagnosed and resolved incidents in Java, PHP, C# and AWS systems, documenting each solution.

Systems Engineer

Indra · San Isidro, Peru (remote) *Mar 2021 — Aug 2024*

- Client Telefónica del Perú: distributed systems and microservices for telecommunications, with daily customer activation and deactivation processes and financial data on billing, claims and commissions.
- Real-time processing of home modem telemetry across Peru by zone: Python ETL microservice integrating network status from several APIs, raising alerts and feeding customer management and monitoring dashboards.
- Built web applications with Java Spring Boot, PHP (Laravel, CodeIgniter), .NET, Python, PowerBuilder, Angular, React and Vue.js.
- Integrated REST and SOAP services with JWT security and AWS and Azure services.
- Administered MySQL, Oracle, SQL Server and MongoDB databases and Red Hat Linux servers; ETL automation with Shell.
- Unit testing with JUnit, Mockito and Postman and code analysis with Sonar, following agile practices.

IT Systems Analyst

Multilab · Pueblo Libre, Peru *Feb 2020 — Mar 2021*

- Transactional clinical systems with Laravel and CodeIgniter: new features, reporting, and roles and permissions.
- Database modeling and optimization; AWS services (S3, EC2) and version control with GitLab.
- Specialized remote and on-site technical support to keep all sites operating.

Software Development Area Manager

MControl Energy Technology SAC · Lima, Peru *Jul 2019 — Jan 2020*

- Led a team of 2 developers, ensuring timely delivery and optimizing critical systems for service stations of the Gasocentro and Irasa groups (Lima and Ica).
- Planned and executed custom software projects aligned with client objectives.

Software Developer

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

- Custom solutions and transactional systems with C#, ASP.NET, PHP (Laravel), JavaScript, TypeScript and Vue.js; SQL Server and MySQL database modeling and optimization.

SELECTED PROJECTS

Meeting Assistant with On-Device Transcription

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

- Bot that joins meetings, recording library with on-device transcription (WhisperKit), speaker separation and LLM summaries; deployed on a VPS with Dokploy, CI and SonarCloud, built with AI agents through reviewed PRs.

PUBLICATIONS / RESEARCH

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

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

- Benchmark of CLIP, SigLIP, SigLIP2 and FuseLIP for natural-disaster damage detection from images and text. Work led by Mariana Risco Cosavalente, with Prof. Claudio Jung (UFRGS).

EDUCATION

Special Student — Graduate Program in Computer Science (PPGC)

UFRGS, Brazil · CMP263 Machine Learning · applying to the Master's program *2026/2 — Ongoing*

AI Engineer Specialization Program — Diploma of Excellence (19/20)

Data Science Research Perú · 92 academic hours *Mar 2026 — Aug 2026*

- RAG, agents, MCP, multimodal applications, LLM evaluation, guardrails and LLMOps; final project: a multi-agent system applied to enterprise processes.

Bachelor's Degree in Software Engineering

UTP — Universidad Tecnológica del Perú *2021 — 2024*

Training Program in Microservices with Java and Spring Boot

Lite Thinking *Jun 2024 — Sep 2025*

Professional Technician in Software Development — 3-year technical degree

SENATI *Mar 2016 — Jul 2019*

CERTIFICATIONS

- **Data Science Research Perú (2026), Diplomas of Excellence:** Evaluation, Observability and LLMOps (19) · Multimodal Agents (19) · AI Agents and MCPs, Agentic RAG (18) · Prompt Engineering and RAG (18) · AI Engineering Fundamentals (18) · Natural Language Processing.
- **Tecylab:** DevOps Multicloud Specialist (2025) · Angular Specialist (2025) · Senior in NodeJS Microservices with Cloud (2025) · Senior in Spring Boot Programming on Multicloud (2024).
- Google Cloud Platform — DSRP (2024) · Scrum Foundation Professional Certification (SFPC™) — CertiProf (2024)