EMERSON HERRERA TOLENTINO

FULL STACK AI ENGINEER / GENERATIVE AI

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

PERFIL PROFISSIONAL

Engenheiro de Software com mais de 8 anos de experiência, especializado em Machine Learning e IA Generativa. Combino uma base sólida em arquitetura de software escalável com técnicas avançadas de ML para construir soluções inteligentes aplicadas a problemas reais. Nos últimos dois anos, projetei e coloquei em produção sistemas de IA Generativa: chatbots corporativos, plataformas RAG, ingestão multimodal de documentos, prompt engineering, integração de LLMs (OpenAI API), embeddings, bancos vetoriais (pgvector, Pinecone, Chroma) e Agentic AI. Recentemente construí uma plataforma RAG multi-tenant que atende várias aplicações bancárias a partir de uma arquitetura compartilhada. Meu objetivo é construir soluções de software e IA escaláveis, seguras, mensuráveis e aplicáveis a problemas reais de negócio.

COMPETÊNCIAS TÉCNICAS

IA Generativa & Agentes: RAG (agêntico e adaptativo), LangChain, LangGraph, OpenAI API, Prompt Engineering, embeddings, bancos de dados vetoriais (pgvector, Pinecone, Chroma, Qdrant), arquiteturas multiagente, MCP (Model Context Protocol), avaliação de LLMs e LLMOps

Machine Learning: Pandas, NumPy, modelos de visão-linguagem (CLIP, SigLIP), fine-tuning e adaptação few-shot

Linguagens & 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

Bancos de Dados: 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)

Arquitetura & Integração: Microsserviços, arquitetura hexagonal e clean, APIs REST/SOAP, JWT, sistemas distribuídos, automação de processos ETL

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

Idiomas: espanhol nativo · português intermediário (boa compreensão) · inglês intermediário (técnico avançado)

DESENVOLVIMENTO ASSISTIDO POR IA

- Trabalho diário com agentes de código (Claude Code, Codex, Pi) sobre Gentle Shell e Gentle AI, com o fluxo ODD (Organic Driven Development): exploração antes de alterar, subagentes para implementar e verificar, TDD e revisão por risco antes de cada commit.
- Memória persistente com Engram: decisões de arquitetura, descobertas e estado das tarefas entre sessões.
- Ambiente personalizado para projetos reais: skills próprias, servidores MCP (Oracle somente leitura, Laravel Boost, logs de contêineres, GitLab, CodeGraph) e combinação de modelos conforme a tarefa.
- Fábrica de agentes em VPS próprio: execução não assistida de issues acionada pelo Telegram (n8n) e lotes noturnos, com PRs e CI como controle de qualidade.

EXPERIÊNCIA PROFISSIONAL

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

Canvia · Peru (remoto) *Out. 2025 — Atual*

- Cliente BCRP (Banco Central de Reserva do Peru): projeto, construção e melhoria de uma plataforma RAG multi-tenant que atende 4 aplicações do banco (Portal BCRP, Portal de Pesquisa, NEXO 360 e Cédulas), incluindo a ingestão e indexação de milhares de documentos e centenas de milhares de fragmentos; seu assistente está em produção no Portal de Pesquisa (investigacion.bcrp.gob.pe).
- Gateway em Node.js que orquestra os sistemas clientes e backend em Python com ferramentas compartilhadas; módulos de chat, busca e prompts configuráveis por aplicação e por área.

- Pipelines RAG para artigos de pesquisa e documentos institucionais, com ingestão multimodal de PDFs e imagens, busca vetorial em Oracle 23ai, pgvector, Pinecone e Chroma e LLMs via OpenAI API e LangChain.
- Pipelines SQL para exportar os datasets da Plataforma de Pesquisadores para o RePEc, com modelagem e otimização de consultas em Oracle; microsserviços backend implantados na AWS.
- Uso diário de agentes de IA com memória persistente e skills próprias do projeto, com revisão humana e testes.
- Stack: Python, FastAPI, LangChain, Node.js, Java, React, Angular, PostgreSQL, Oracle, AWS, Azure, Docker.

Desenvolvedor Full Stack e de IA (freelance)

Certilab S.A.C. · Remoto *Set. 2024 — Atual*

- Sistema de certificações em Laravel 11 e Filament, com mais de 300 certificados emitidos: backoffice administrativo, intranet onde cada cliente consulta seus certificados e portal público de validação por QR com ID criptografado.
- Emissão de certificados, upload do PDF assinado, geração de QR, papéis e permissões, auditoria e armazenamento em S3 (Docker).
- RAG e chatbot sobre certificados de calibração com ingestão multimodal e painel de consulta para a gerência; versão agêntica com LangGraph, FastAPI e Qdrant (github.com/emersonheto/certilab-agentic-rag).
- Site corporativo com Next.js 15 e React 19 e implantação automatizada.

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

Way Tech (projetos WIN Empresas / Universidad Continental) · Peru (remoto) *Nov. 2024 — Out. 2025*

- Desenvolvimento de plataformas acadêmicas e institucionais para a Universidad Continental (mais de 24.000 estudantes em Huancayo, Arequipa, Cusco e Lima) com Angular e microsserviços em C# (.NET) e Java.
- Consumo e documentação de APIs RESTful com Swagger e melhoria de desempenho em sistemas acadêmicos de alto tráfego.
- Automação de processos financeiros e de negócio com Node.js, garantindo a integração entre sistemas.
- Gestão e otimização de bancos SQL (MySQL, SQL Server, Oracle, PostgreSQL) e NoSQL (MongoDB) de alto tráfego.
- Diagnóstico e resolução de incidentes em sistemas Java, PHP, C# e AWS, com documentação técnica de cada solução.

Systems Engineer

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

- Cliente Telefónica del Perú: sistemas distribuídos e microsserviços de telecomunicações, com processos diários de altas e baixas de clientes e dados financeiros de faturamento, reclamações e comissões.
- Processamento em tempo real da telemetria de modems residenciais de todo o Peru, por zona: microsserviço ETL em Python que integra o estado da rede de várias APIs, gera alertas e alimenta a gestão de clientes e os dashboards de monitoramento.
- Aplicações web com Java Spring Boot, PHP (Laravel, CodeIgniter), .NET, Python, PowerBuilder, Angular, React e Vue.js.
- Integração de serviços REST e SOAP com segurança JWT e serviços da AWS e do Azure.
- Administração de MySQL, Oracle, SQL Server, MongoDB e servidores Linux Red Hat; automação ETL com Shell.
- Testes unitários com JUnit, Mockito e Postman e análise de código com Sonar, aplicando metodologias ágeis.

Analista de Sistemas de TI

Multilab · Pueblo Libre, Peru *Fev. 2020 — Mar. 2021*

- Sistemas clínicos transacionais com Laravel e CodeIgniter: novas funcionalidades, relatórios e papéis e permissões.
- Modelagem e otimização de bancos de dados; serviços na AWS (S3, EC2) e controle de versões com GitLab.
- Suporte técnico especializado, remoto e presencial, para a continuidade operacional das unidades.

Chefe da Área de Desenvolvimento de Software

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

- Liderança de uma equipe de 2 desenvolvedores, com entregas no prazo e otimização de sistemas críticos para postos dos grupos Gasocentro e Irasa (Lima e Ica).
- Planejamento e execução de projetos de software sob medida, alinhados aos objetivos do cliente.

Software Developer

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

- Soluções sob medida e sistemas transacionais com C#, ASP.NET, PHP (Laravel), JavaScript, TypeScript e Vue.js; modelagem e otimização de bancos de dados SQL Server e MySQL.

PROJETOS EM DESTAQUE

Assistente de reuniões com transcrição local

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

- Bot que entra nas reuniões, biblioteca de gravações com transcrição local (WhisperKit), separação de falantes e resumos com LLM; implantado em VPS com Dokploy, CI e SonarCloud, construído com agentes de IA em PRs revisados.

PUBLICAÇÕES / PESQUISA

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

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

- Benchmark de CLIP, SigLIP, SigLIP2 e FuseLIP para detecção de danos em desastres naturais a partir de imagens e texto. Trabalho liderado por Mariana Risco Cosavalente, com o Prof. Claudio Jung (UFRGS).

FORMAÇÃO ACADÊMICA

Aluno Especial — Programa de Pós-Graduação em Computação (PPGC)

UFRGS, Brasil · CMP263 Aprendizado de Máquina · candidato ao mestrado *2026/2 — Em andamento*

Programa de Especialização em AI Engineer — Diploma de Excelência (19/20)

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

- RAG, agentes, MCP, aplicações multimodais, avaliação de LLMs, guardrails e LLMOps; projeto final: sistema multiagente aplicado a processos empresariais.

Bacharelado em Engenharia de Software

UTP — Universidad Tecnológica del Perú *2021 — 2024*

Programa de formação em Microsserviços com Java e Spring Boot

Lite Thinking *Jun. 2024 — Set. 2025*

Profissional Técnico em Desenvolvimento de Software — curso técnico superior de 3 anos

SENATI *Mar. 2016 — Jul. 2019*

CERTIFICAÇÕES

- **DSRP (2026), Diplomas de Excelência:** Avaliação, Observabilidade e LLMOps (19) · Agentes Multimodais (19) · Agentes de IA e MCPs, RAGs Agênticos (18) · Prompt Engineering e RAGs (18) · Fundamentos de Engenharia de IA (18) · Processamento de Linguagem Natural.
- **Tecylab:** Especialista em DevOps Multicloud (2025) · Especialista em Angular (2025) · Sênior em Microsserviços NodeJS com Cloud (2025) · Sênior em Programação com Spring Boot em Multicloud (2024).
- Google Cloud Platform — DSRP (2024) · Scrum Foundation Professional Certification (SFPC™) — CertiProf (2024)