Henrique Pimenta Souza

AI Systems Architect | Senior AI Engineer | Enterprise GenAI Specialist

São Paulo, Brazil | +55 11 98110-4828 | pimenttasouzza02@gmail.com | LinkedIn

Professional Summary

AI Systems Engineer specialized in designing, deploying, and scaling Generative AI systems in production, particularly within regulated environments. Proven experience architecting large-scale multi-agent ecosystems (+5,000 active users, +60,000 AI conversations/month), structuring deterministic, auditable, and compliance-oriented layers on top of probabilistic LLM models.

Strong expertise in Agentic Workflows, Enterprise RAG architectures, AI Governance, and Explainable AI (XAI), combined with a solid foundation in backend architecture, high-concurrency microservices, and *context engineering* for business-critical applications.

Technical Skills

Generative AI & Intelligent Systems: Multi-Agent Systems, Agentic Workflows, LangGraph, LangChain, RAG (Enterprise & Production), LLM Validation Strategies, Deterministic Guardrails, Explainable AI (XAI), AI Governance, OpenAI Whisper, OCR + LLM Systems, Hugging Face Embeddings, Botpress

Backend & Architecture: Python (FastAPI, Async I/O, Pydantic v2, aiohttp), .NET Core (C#), Java (Spring Boot), Microservices, High-Concurrency APIs, Deterministic Validation Layers, Domain-Driven Design (DDD), Clean Architecture, RabbitMQ

Data & Infrastructure: PostgreSQL, MySQL, MongoDB, Pandas, PyArrow, Vector Databases (FAISS), Snowflake (Data Lake), Docker, AWS, CI/CD, Grafana (Observability), Streamlit

Integration & Governance: Zendesk API, Human-in-the-Loop (HITL), Make (Integromat), Traceability and Audit Logging, Asynchronous Operations Security

Professional Experience

AI Systems & Automation Engineer — Oct. 2024 – Present
Omni Saúde — *São Paulo, Brazil*

- **Production multi-agent architecture:** Designed and orchestrated an LLM-based support ecosystem integrated with Zendesk using a Human-in-the-Loop (HITL) architecture, ensuring safe escalation to human agents.
- **Operational scale:** +5,000 active users and +60,000 AI-driven conversations per month in a regulated healthcare environment.
- **Measurable impact:** Increased autonomous resolution rate from 55% to 88% within two months and reduced support ticket volume by 40%.
- **Backend engineering:** Developed high-concurrency microservices using FastAPI and Async I/O, implemented type-safe validation with Pydantic v2, and created deterministic control layers over LLM reasoning.
- **Governance and traceability:** Implemented guardrails, interaction auditing, context control, and performance optimization using Pandas and PyArrow.
- **Workflow integration:** Built end-to-end pipelines connecting Zendesk and Typeform to AI-driven workflows via Make.

CEO & Founder — Bridge (MVP) — Feb. 2025 – Feb. 2026
AI-Based Recruitment Platform (SaaS B2B2C) — *São Paulo, Brazil*

- Led the design and execution of an AI-native platform focused on Explainable AI (XAI), bias mitigation, and privacy-by-design (LGPD-ready compliance).
- **Product stage:** MVP with 25 active users; NPS 10/10 in UX, functionality, and perceived impact.
- **Market validation:** Pilot conducted with The Growth Hub (CEO: Anderson).

- **AI architecture:** Implemented RAG pipelines for contextual candidate-job matching, explainable justification generation, and cultural fit logic.
- **Strategy and product:** Conducted TAM/SAM/SOM, PESTEL, Porter's Five Forces, and VRIO analyses; defined roadmap by strategic horizons and risk × impact prioritization.

Full Stack Developer Aug. 2024 – Sept. 2025
Stefanini *São Paulo, Brazil*

- **Enterprise AI solutions:** Developed RAG-based agents with Context Engineering over a 10GB+ document base, including extraction, normalization, and standardization in Python.
- **Software architecture:** Applied DDD principles in .NET Core APIs and produced technical documentation (C4 Model, sequence and class diagrams).
- **Frontend:** Built modular interfaces in Angular and Vue.js integrated with backend services via Axios.

Highlighted Projects

SAI App (Stefanini) — LLM
AI-powered application for intelligent automation and productivity enhancement.

- Developed frontend components using Angular and Vue.js (TypeScript).
- Built scalable APIs using .NET / ASP.NET Web API applying DDD principles.

Sales Insights API — RAG NL→SQL (Study Project)
Independent project focused on designing a RAG pipeline for validated natural language to SQL conversion.

Implemented schema-aware contextual retrieval, query validation, and humanized response generation. Technologies: Python, FastAPI, PostgreSQL, SQLAlchemy, Docker, LangChain.

OCR + LLM API — Intelligent Document Processing
Extraction pipeline using PaddleOCR and Transformers for resume and tax document interpretation, including summarization and Q&A with MongoDB logging.

Jira Automation & Speech-to-Text Pipeline
Data extraction via Jira API with Grafana monitoring; integrated OpenAI Whisper with GPU acceleration (CUDA) for large-scale transcription.

Education

Systems Analysis and Development (Associate Degree) — São Paulo Tech School (SPTech) Expected Graduation: Jun. 2026
Database Technician Certification — Etec Ferraz de Vasconcelos Dec. 2021

Certifications & Languages

Languages: English (Professional Working Proficiency)
Google Certifications: Foundations of Project Management; Data Analytics
Additional Skills: SOLID Principles, Clean Code, Scrum, Kanban