

Gilson Vieira de Souza

Machine Learning Engineer | GenAI & AI Engineering | AI Systems Architecture

São Paulo, Brazil | [linkedin.com/in/gilson](https://www.linkedin.com/in/gilson) | github.com/gil-son | gilson.dev.br
gilson.inspire@gmail.com | +55 11 94053-4609

Machine Learning Engineer with 5+ years in technology, including nearly 2 years focused on GenAI, AI Engineering, MLOps and FMOps, built on a strong Software Engineering foundation. Designs and scales enterprise AI systems running in production for millions of users in the financial sector. Works across the full lifecycle: LLM orchestration (LangChain, LangGraph, Google ADK), RAG and vector databases, agentic workflows and tool calling, and model deployment on AWS SageMaker, EKS and Kubernetes. Owns AI systems architecture end to end, bringing software engineering rigor to AI through Spec-Driven Development (SDD), Harness Engineering, Design Docs, evaluation, guardrails and observability (LangFuse, MLflow). AWS certified, English C1, experienced in global teams.

TECHNICAL SKILLS

- **AI & GenAI:** Generative AI, LLMs, RAG, AI Agents, NLP/NLU/NLG, Model Context Protocol (MCP), Guardrails.
- **Frameworks & Orchestration:** Python, LangChain, LangGraph, Google ADK, FastAPI, PyTorch, Hugging Face (Transformers, Fine-tuning), Ollama.
- **AI Systems Architecture & Context Engineering:** AI systems architecture for LLM-based products, Spec-Driven Development (SDD), Harness Engineering, Design Docs, Context Management & Engineering, Prompt Engineering, Evaluation.
- **Data Science & ML Foundations:** Supervised and unsupervised modeling, applied statistics, optimization, feature engineering, model evaluation metrics.
- **MLOps, FMOps & Cloud:** AWS (SageMaker, EKS, CloudWatch, Lambda, S3, Glue), Azure, Docker, Kubernetes, Terraform (IaC), CI/CD, GitHub Actions, MLflow, LangFuse, Vector Databases (FAISS, Weaviate, pgvector).
- **Software Architecture & Engineering:** Clean Code, SOLID, Design Patterns, Microservices, Java (Spring Boot, Spring Batch), Go, Apache Kafka, Apache Camel.
- **Observability & AI-Assisted Engineering:** LangFuse, MLflow, Datadog, Grafana, Splunk, CloudWatch; Claude Code, ChatGPT, Gemini.
- **Languages:** Portuguese (native), English (C1 – fluent professional communication).

EXPERIENCE

IBM – São Paulo, Brazil

Machine Learning Engineer / AI Engineer (04/2025 – Present)

- Architect and scale enterprise GenAI applications, AI Agents and RAG systems using LangChain, LangGraph and vector databases (FAISS, Weaviate), delivered to production for millions of banking customers.
- Design and implement end-to-end MLOps/FMOps pipelines and model deployments on AWS SageMaker and Amazon EKS using Terraform (IaC) and CI/CD automation.
- Integrate multi-provider LLMs (OpenAI, Anthropic) through Model Context Protocol (MCP) and custom tool-calling workflows in Python.
- Own AI systems architecture from design doc to production: component boundaries, API contracts, schema validation, context management and failure modes across LLM, retrieval and orchestration layers.
- Apply Spec-Driven Development (SDD) and Harness Engineering to build the scaffolding around models — prompt and context harnesses, tool interfaces and evaluation loops — so AI behavior stays testable and reproducible across releases.
- Implement production observability, experiment tracking and AI guardrails with LangFuse and MLflow for trace monitoring, logging and evaluation of model quality.
- Define evaluation criteria and regression checks for LLM features, keeping releases measurable rather than anecdotal.

Mid-level Backend Developer (08/2022 – 04/2025)

- Engineered high-throughput, fault-tolerant backend services with Spring Boot, Spring Batch, Apache Camel and Apache Kafka for core financial infrastructure.
- Built and managed cloud infrastructure with Terraform (IaC) across AWS environments, with continuous integration and deployment pipelines.

- Wrote unit and integration test suites following Clean Code and SOLID principles to keep the codebase reliable under frequent change.

Software Developer – Junior (06/2021 – 07/2022)

- Automated card services for over 600k users at Santander using a microservices architecture.
- Enhanced backend systems and built message queues for real-time data processing.
- Gained hands-on experience with containerized environments and cloud deployments.

UNICID – São Paulo, Brazil

Data Analyst – Junior (07/2019 – 05/2020)

- Built ETL processes to extract insights from enrollment and retention data.
- Supported marketing and operational strategy with data dashboards, helping grow the student base.

SELECTED PROJECTS & OPEN SOURCE

- **Local-first AI Automation Platform** — Solo-built platform using FastAPI, LangChain/LangGraph, pgvector and Ollama, with design docs and spec-driven workflow driving the architecture.
- **Language & AI Engineering Lab** — Open 16-module structured curriculum on LLM and AI engineering, from fundamentals to agentic systems and evaluation.
- **Machine Learning Lab** — Open educational repository of ML material organized by learning paradigm, including a consolidated reference on model evaluation.
- **System Design for AI/ML** — Four-phase study and portfolio repository covering ML and GenAI system design patterns.

EDUCATION & CERTIFICATIONS

- **MBA in Software Engineering with AI** – In progress
- **B.S. in Computer Science** – UNICID, São Paulo (01/2017 – 12/2020)
- **AWS Certified Machine Learning Engineer – Associate**
- **AWS Certified Cloud Practitioner**