

# João Luiz Sotini

Data Scientist and AI Engineer

joaoluizxd@gmail.com | +55 (31) 991910177 | São José dos Campos, São Paulo | [linkedin.com/in/joao-luiz-sotini-da-silva](https://linkedin.com/in/joao-luiz-sotini-da-silva)  
| [github.com/SotiniJoao](https://github.com/SotiniJoao)

## PROFESSIONAL SUMMARY

AI Engineer and Data Scientist with more than 4 years of experience designing and delivering production-grade machine learning systems and scalable cloud infrastructure. Specialized in Generative AI, LLM applications, and RAG systems, with expertise in multi-agent orchestration, prompt engineering, and semantic search optimization using vector databases. Developed and implemented end-to-end ML pipelines on GCP, AWS, and Azure, using GPT-4o, Claude, Gemini, and frameworks such as LangChain and LangGraph to deliver measurable business outcomes. Hands-on experience fine-tuning models with LoRA and QLoRA, evaluating machine learning models including XGBoost and neural networks, and evaluating prompts with RAGAS and similar methodologies. Demonstrated technical proficiency in production-grade Python, developing guardrails for system safety, A/B testing, conversational systems architecture, and cross-functional collaboration, ensuring security, scalability, and performance in enterprise AI solutions.

## PROFESSIONAL EXPERIENCE

### Amaris Consulting

Abr 2026 – Presente

- Senior AI Engineer · São Paulo, SP (Remote)  
Worked as a Data Engineering and Machine Learning consultant assigned to Decathlon, leading the development and evolution of conversational and multi-agent AI systems to democratize access to the corporate data lake and automate the generation of metrics and insights.
  - Developed and evolved DataWiz, a multi-agent NL-to-SQL AI system used by hundreds of internal users, enabling natural-language queries of the corporate data lake with analysis-ready data and visualizations.
  - Implemented automated query scheduling in DataWiz.
  - Continuously improved the production system by refining prompts and validating outputs to enhance SQL query accuracy, response quality, performance, and robustness.
  - Orchestrated a stack with Python, SQL, Databricks, Chainlit, PydanticAI, Vertex AI, Gemini, Docker, and AWS, with observability through Datadog.
  - Developed a proof of concept for an AI agent to automatically generate and send promotional campaign analyses, OPECOs, combining Databricks data, deterministic indicator calculations, and Gemini interpretation through Vertex AI.
  - Implemented an end-to-end pipeline with validation, Markdown and JSON reports, and HTML email delivery through AWS SES, validating 10 reports for completed campaigns using production data.
  - Implemented anti-hallucination guardrails with a structured sheet containing 24 indicators, typed blocks, and automatic correction after rejection, in addition to developing a validator for numbers in pt-BR format.
  - Structured the analysis into deterministic and optional hypothesis stages, with supporting facts and validation paths, and implemented parallel batch execution, checkpointing, concurrency control, and separate exit codes.
  - Implemented idempotent email delivery with retry and backoff, unique Message-ID, and allowed domains, in addition to reconciling data with the business area's Tableau dashboard.
  - Applied TDD in vertical slices, developed an offline test suite with a simulated model, and documented architectural decisions in ADRs.

## • **Alana AI**

Fev 2026 – Mai 2026

Senior AI and Data Engineer · São Paulo, SP (Remote)

Built the search and data foundation for an AI shopping agent for fashion e-commerce, developing an LLM-based ingestion and enrichment pipeline and a multimodal search API used by the agent.

- Developed a search API with FastAPI and OpenSearch Serverless, combining vector, lexical, and hybrid search. Implemented Reciprocal Rank Fusion, increasing MRR by up to 87.7%, from 0.38 to 0.72, compared with the previous approach.
- Achieved MRR 1.0 in the evaluated scenarios with image search.
- Created an evaluation framework with LLM-as-judge, using NDCG, MRR, MAP, and Hit@k to compare search strategies, embedding models, and languages, guiding architecture decisions.
- Built idempotent ETL pipelines for an approximately 20,000-SKU VTEX catalog, with checkpointing, fault tolerance, and incremental delta sync, avoiding full catalog reprocessing.
- Performed enrichment with Gemini 2.5 Flash in three phases, with the occasion and style stages running in parallel, extracting approximately 30 attributes per product, including visual aspects, fit, occasion, and style, with prompts versioned in a database.
- Built an editorial looks pipeline with Gemini Vision, separating each look into individual pieces (616 pieces from 111 looks) and linking them to catalog products through multimodal search.
- Migrated text embeddings to a retrieval-specialized model (Google's text-multilingual-embedding-002), with 768 dimensions and 45% smaller vectors, correcting relevance issues in Portuguese and English search.
- Created a real-time size-based inventory filter through the VTEX API, applied to all search endpoints.

## • **Zams**

Mar 2025 – Dez 2025

Machine Learning Engineer and Data Engineer · California, United States (Remote)

Developed conversational AI systems, semantic search, production Machine Learning pipelines, and document-processing solutions to optimize operations and support partnership discovery.

- Architected a conversational database interface with GPT-4o and Claude, using dynamic schema injection, few-shot learning, and validation guardrails for SQL generation, achieving approximately 3-second latency and reducing data analysis time by 80%.
- Developed an RAG- and Qdrant-based matching mechanism to identify alignments between Limited Partners and Emerging Managers, reducing partner discovery from weeks to hours through semantic search and ranking.
- Designed a production Machine Learning pipeline on GCP for postal routing optimization with HDBScan, reducing processing time by 65% and shipping operating costs by 30%.
- Built a document-processing system using AWS Textract OCR and LLMs to extract structured data from complex, multipage PDFs, reducing processing time from hours to minutes.

- **Zippia** Out 2022 – Mar 2025  
Data Scientist · San Mateo, CA (Remote)  
Developed recommendation systems, data pipelines at scale, production Machine Learning infrastructure, and generative AI solutions for a recruiting platform processing millions of daily events.
  - Generated a 15% revenue increase by developing a hybrid recommendation system with Neural Networks and XGBoost, trained on more than 10 million interactions, and implementing vector search with FAISS for semantic job matching.
  - Developed a data pipeline to process millions of daily behavioral events from MongoDB and Elasticsearch, creating low-latency feature engineering for real-time inference and personalization.
  - Built production ML infrastructure with Docker, CI/CD, and MLFlow for versioning, deployment, experiment tracking, and model management.
  - Developed a RAG system with GPT-4o, Mixtral, and LangChain for job-title standardization and normalization.
  - Developed an auto-fill solution with OpenCV and GPT-4o Vision to populate forms from document images.
  - Reduced MongoDB and Elasticsearch query latency by 40% through index optimization, query rewriting, and caching.
- **Harpia Health Solutions** Fev 2022 – Out 2022  
Data Scientist and ML Engineer · São José dos Campos, SP, Brazil  
Built medical imaging pipelines and developed deep learning models for automated disease classification and anomaly detection in diagnostic images.
  - Built a distributed medical imaging pipeline with PySpark to process hundreds of X-rays, CT scans, and MRI scans, improving efficiency by 8% through parallel computing.
  - Developed deep learning models with PyTorch and TensorFlow for disease classification and anomaly detection, reducing medical detection time by 42%.
  - Created interpretability tools with Grad-CAM for clinical validation and explainability of diagnostic recommendations.
  - Developed scalable training infrastructure with PySpark and MLFlow for experiment tracking, version control, and model lifecycle management.
- **Kyrally Digital Marketing** Jan 2021 – Jan 2022  
Data Engineer and Business Intelligence Analyst · São José dos Campos, SP, Brazil
  - Designed a marketing data pipeline integrating Google Ads, Google Analytics, and Facebook Ads APIs with internal databases to create a unified data warehouse for multichannel campaigns.
  - Reduced manual reporting time by 70% by automating data collection and transformation pipelines and enabling real-time dashboard updates.
  - Created interactive Business Intelligence dashboards and optimized SQL queries to extract performance indicators and support campaign optimization and budget allocation.

---

|           |                                                                                                                                                                      |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EDUCATION | <b>International University of Applied Sciences</b> <span>Fev 2024 – Feb 2027</span><br>Master of Science em Data Science · Berlin, Germany (Online)                 |
|           | <b>Federal University of São Paulo (UNIFESP)</b> <span>Jan 2018 – Jan 2023</span><br>Bachelor of Science em Science and Technology · São José dos Campos, SP, Brazil |

---

**Técnicas:** LLM Development, Prompt Engineering, RAG Systems, Vector Databases, Multi-Agent Systems, Machine Learning, Deep Learning, Neural Networks, Computer Vision, Natural Language Processing (NLP), Transformers, Recommender Systems, MLOps, Data Engineering, ETL/ELT Pipelines, A/B Testing, Data Visualization, Few-Shot Learning, Chain-of-Thought Prompting, Function Calling, Document Chunking, Context Optimization, Retrieval Optimization, NL-to-SQL, Lakehouse Architecture, Semantic Search, Vector Indexing, Feature Engineering, Clustering, OCR, Document Processing, Model Interpretability, Business Intelligence, Multichannel Analysis, NL-to-SQL Conversational Interfaces, Semantic Search, Vector Embeddings, Schema Injection, SQL Query Generation, Medical Imaging AI, Postal Route Optimization, Document Extraction, Data Warehouse Architecture, Cloud-Native Architecture, Structured Data Integration, Real-Time Query Processing, Schema Optimization, Distributed Data Processing, Medical Image Analysis, Disease Classification, Image Anomaly Detection, Apache Airflow, Reinforcement Learning, HDBSCAN, Grad-CAM, Data Warehouse, Observability, AI Guardrails, Search Ranking, Pipeline Checkpointing, Data Modeling, API Data Extraction, Index Optimization, Data Validation, Test-Driven Development (TDD), Automated Testing, Event-Driven Architecture, Parallel Processing, Concurrency Control, Idempotency, Retry and Backoff, Data Reconciliation, Model Regression, XGBoost, Pydantic, PydanticAI, AWS SES, Databricks SQL, Databricks SDK, OAuth 2.0, AWS OpenSearch Serverless, Cloud IAM, VTEX APIs, Markdown, JSON, YAML, Experiment Tracking, Model Registry, OCR with AWS Textract, AWS, GCP, Databricks, Vertex AI, Gemini, Chainlit, Apache Spark, PyTorch, TensorFlow, OpenCV, FAISS, Qdrant, OpenSearch, Elasticsearch, MongoDB, SQL Server, BigQuery, MySQL, Google Ads API, Google Analytics API, Facebook Ads API, Data Warehousing, Hybrid Recommendation Systems, Machine Learning in Production, LLM Vision, Schema Validation, Cloud Computing, Distributed Architecture, Reciprocal Rank Fusion, LLM-as-Judge, NDCG, MRR (Mean Reciprocal Rank), MAP (Mean Average Precision), Hit@k, Hybrid Search, Lexical Search, Multimodal Search, Image Search, Multilingual Embedding Models, LLM-Based Data Enrichment, Medical Image Processing, Anomaly Detection, AWS Aurora, AWS Secrets Manager, AWS Textract, VTEX, FastAPI, OpenSearch Serverless, Docker, MLflow, CI/CD, Git/GitHub, Datadog, Pandas, NumPy, Boto3, Tableau, Google Cloud Storage, Machine Learning Engineering, Search and Information Retrieval Systems, Vector Search, Search System Evaluation, Multimodal Data Processing, Catalog Enrichment, Incremental Data Synchronization, Idempotent ETL, REST API, Medical Imaging Pipelines, Image Classification, Low-Latency Feature Engineering, Query Optimization, Caching, Google Analytics, Facebook Ads, Multichannel Data Integration, AI Agent Architecture, Agent Orchestration, Google Gemini 2.5 Flash, Claude

**Linguagens de Programação:** Python, SQL

**Ferramentas:** GPT-4o, GPT-3.5, Claude Opus, Claude Sonnet, Claude Haiku, LLAMA, Mixtral, LangChain, LlamaIndex, LangGraph, Qdrant, FAISS, Pinecone, Weaviate, PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, Apache Spark, PySpark, BigQuery, MongoDB, Elasticsearch, MySQL, PostgreSQL, SQL Server, Redis, Docker, MLFlow, Git, FastAPI, Flask, Streamlit, Pandas, NumPy, OpenCV, Seaborn, AWS Textract, AWS S3, AWS Secrets Manager, AWS Aurora, GCP Cloud Storage, GCP Vertex AI, GCP IAM, Azure Blob Storage, OpenSearch, Databricks, Chainlit, Datadog, GCP OpenSearch, VTEX, HDBSCAN, Grad-CAM, Google Ads, Google Analytics, Facebook Ads, Gemini, Aurora, IAM, AWS, GCP, Google Ads API, Google Analytics API, Facebook Ads API, OpenSearch Serverless, Cloud Storage, Vertex AI, CI/CD, AWS SES, boto3, pytest, GitHub, AWS OpenSearch Serverless, Tableau, Reciprocal Rank Fusion, Git/GitHub, Google Cloud Storage, Python SDKs, Databricks OAuth M2M, ADRs (Architecture Decision Records)

**Comportamentais:** Problem Solving, Cross-Functional Collaboration, Technical Communication, Analytical Thinking, Project Management, Adaptability, Strategic Thinking, Systems Architecture, Systems Thinking, Process Optimization, Continuous Improvement, Fault Tolerance, Architectural Decision Documentation, Cross-Functional Stakeholder Validation, Technical Leadership, Stakeholder Validation

|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| PROJECTS | <div><div>CValore</div><div>Out 2025 – Presente</div><div>Co-founder and AI Engineer (with Luiz Raveduti Zafiro)</div><div>Co-founded a pre-launch SaaS that imports résumés, evaluates their quality and alignment with a specific job description, and rewrites only fields that do not pass review. Developed the stateless agentic backend in Python/FastAPI and the web application in Next.js/TypeScript. - Developed a multi-provider LLM agent using Claude, GPT, and Gemini through LiteLLM, tool calling, and task-specific models: a larger model selects the tool, while a smaller model drafts the response. - Implemented model fallback for rate limits and timeouts, as well as SSE streaming with token usage and cost updates on each turn. - Designed a deterministic pipeline before the LLM: a pure-Python audit identifies issues, and an LLM-as-judge reviews only ambiguous fields. - Classified job-description keywords as substantiated, merely stated, or missing, preventing the agent from inventing experience. Also developed a benchmarking environment to compare models by cost and quality. - Implemented privacy by design: replaced personal data with tokens before LLM calls and prevented résumé text from being stored in logs. Evaluated masking company names, which cost 6.8% in quality, and kept this feature disabled.</div></div> |
|          | <div><div>Wedding Website: Ateliê Digital de Casamentos</div><div>Out 2025 – Presente</div><div>Co-founder (with Luiz Raveduti Zafiro) · Next.js 16, TypeScript, PostgreSQL, Supabase, Prisma, QStash, Gemini 2.5, Claude Haiku</div><div>Led the product repositioning from a standalone wedding website to a full-service digital studio offering each couple's visual identity, monogram, invitations, Instagram posts, WhatsApp communication, and published website. - Implemented the first version of the creative assets offering, generating materials based on each couple's visual identity; the team later evolved this foundation into the current product.</div></div>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|          | <div><div>Mobile Application for Assessing Treatment Response in Hemifacial Spasm Using Machine Learning — UNIFESP</div><div>Fev 2022 – Jan 2024</div><div>First author · Machine Learning, Computer Vision, dlib, HOG, SVM, iBUG 300-W, Eye Aspect Ratio, Google Sheets</div><div>Published by Springer (IFMBE Proceedings, CLAIB &amp; CBEB 2022). In partnership with the Department of Ophthalmology at Escola Paulista de Medicina, developed a mobile application that assesses response to botulinum toxin treatment from videos recorded on a mobile phone, replacing manual blink counting and infrared equipment. Delivered a chart to the physician within minutes. Fine-tuned a dlib facial-landmark model (HOG + SVM) on the iBUG 300-W dataset, with more than 11,500 images, reducing the model from 68 to 12 eye points. Implemented an automatic, individualized blink-detection threshold based on the Eye Aspect Ratio (Soukupová &amp; Čech), adapted to different eye sizes without manual calibration. Validated the model against counts from two ophthalmologists, achieving a normalized RMSE of 0.0995 and approximately 95% of blinks detected. Recorded each patient's history in Google Sheets for before-and-after treatment comparison.</div></div>                                                                                                 |