

Lucas Gonçalves

AI/ML Engineer | LLM Agents & RAG · Computer Vision · NLP

Marília, SP, Brazil (GMT-3) · Remote-Ready

+55 14 99152-5951 · lucassg2015@gmail.com · linkedin.com/in/lucas-goncalvessz · github.com/LukeSantossz

Summary

AI/ML Engineer specializing in LLM agents, RAG systems, and applied machine learning across computer vision, NLP, and time-series forecasting, working end-to-end in Python, PyTorch, TensorFlow, and Rust. Sole developer of a FAPESP-funded production RAG system with semantic-entropy hallucination detection (Farquhar et al., Nature 2024) that emits continuous 0 to 1 confidence scores, shipped with 205 automated tests at ~83% coverage and strict-typed CI. Drove a LightGBM forecasting ensemble to 0.19°C RMSE (~75% improvement over a Prophet baseline) and engineered a Rust preprocessing CLI measured at ~42x Python throughput at 100K records. Seeking a remote AI/ML Engineer role on a distributed team building production-grade intelligent systems.

Key Achievements

- Shipped a FAPESP-funded production RAG system whose semantic-entropy detector replaced binary hallucination flags with a continuous 0 to 1 confidence score, hardened with 205 automated tests (~83% coverage), strict-typed CI (ruff + mypy --strict), and multi-stage Docker deployment, with zero dependency on paid inference APIs.
- Won 3rd place among 1,300+ entries at FETEPS 2025 with an on-device soil-texture classifier; full paper accepted for poster presentation at ICPA 2026 / 17th ConBAP.
- Drove a LightGBM ensemble to 0.19°C RMSE over 133K+ observations across 211 countries (a ~75% improvement over the Prophet baseline of 0.77°C), with SHAP-based feature attribution over the full pipeline.

Experience

Jacto | After-Sales Intern | Marília, SP, Brazil (Hybrid) | 11/2024 to Present

Global agricultural machinery manufacturer operating across 100+ countries, 3,000+ employees, industrial-scale after-sales operations.

- Automated the migration of Russian KLADR-classified fiscal address records into Salesforce during a dealership onboarding, handling extraction, Cyrillic-to-Latin transliteration, and geographic normalization across multiple regions, eliminating manual data entry from the process.
- Designed a computer-vision discard-auditing architecture (under internal review) combining photo-standardization checks, OCR cross-referencing of the captured slip against the system record, and part recognition, paired with an in-app camera module enforcing filename locks with automatic timestamping and geotagging for database-level evidence integrity.
- Partnered with engineering as the business-to-development bridge on an internal parts-return system, surfacing 10+ usability defects and validating business-rule changes across QA cycles that shifted the workflow from manual entry to majority-review operation.
- Audited and reconciled the field-parts return flow across 150+ master dealerships operating nationally and internationally, tracking millions of BRL in parts movement and processing write-offs through on-site and remote audits.

Projects

SmartB100 | Agricultural RAG Agent · FAPESP-Sponsored Research

github.com/LukeSantossz/sb100_agents

- Sole developer of a production RAG system (FastAPI + Qdrant + Ollama) implementing semantic-entropy hallucination detection (Farquhar et al., Nature 2024): generates N candidate answers, clusters them by semantic similarity, and computes Shannon entropy to output a continuous 0 to 1 confidence score instead of a binary flag.
- Integrated multi-provider verification (Groq / Ollama / OpenRouter) and expertise-adaptive responses (beginner / intermediate / expert); secured the /chat endpoint with bcrypt + JWT and per-IP rate limiting; runs fully offline with no paid API dependency.
- Hardened to a production bar: 205 automated tests at ~83% coverage, CI with ruff + mypy --strict, multi-stage Docker with healthchecks and log rotation, and a 5-step offline evaluation pipeline. Verification gate designed with a clean async-friendly interface as the natural independent-scaling extraction point.

VisioSoil | Geolocated Soil Texture Classification · 3rd Place, FETEPS 2025

github.com/LukeSantossz/visiosoil-app

- Built a cross-platform Flutter app (Riverpod, GoRouter, Drift/SQLite) running TFLite inference fully on-device in a background isolate across 5 soil-texture classes; the underlying MobileNetV2 classifier earned 3rd place among 1,300+ submissions at FETEPS 2025, with a full paper accepted at ICPA 2026 / 17th ConBAP.
- Engineered the full offline capture pipeline (camera, GPS geolocation, reverse geocoding, local persistence with schema versioning) behind a repository-pattern data layer decoupled from the persistence backend, with a reproducible two-phase MobileNetV2 transfer-learning training pipeline and CI running analyze, test, and APK build.

Global Temperature Forecasting Pipeline

github.com/LukeSantossz/weather-forecast

- Architected a temperature forecasting pipeline over 133K+ observations across 211 countries, applying IQR-based outlier treatment and lag/rolling feature engineering (1 to 21 days) to drive a LightGBM model to 0.19°C RMSE, a ~75% improvement over the Prophet baseline (0.77°C).
- Benchmarked 5 forecasting approaches (ARIMA, SARIMA, LightGBM, GradientBoosting, Prophet) and deployed an inverse-RMSE weighted ensemble (0.24°C RMSE) for risk diversification, plus dual-method anomaly detection (Z-score and Isolation Forest) and SHAP feature attribution; 37 passing tests, CI green, status complete.

Tweet Sentiment Analysis

github.com/LukeSantossz/tweet-sentiment-analysis

- Built an NLP fine-tuning pipeline (HuggingFace Trainer, early stopping) around a Twitter-specialized RoBERTa model on the TweetEval benchmark, establishing a zero-shot baseline of 0.71 macro F1 over a 12,284-sample test split as the target to beat, with a 6-function tweet-cleaning module (URLs, mentions, hashtags, emojis) and 21 unit tests.
- Engineered a high-throughput Rust preprocessing CLI (Rayon parallelism, Polars I/O) mirroring the Python reference implementation, measured at ~42x speedup at 100K tweets with output parity validated via grapheme-cluster emoji handling.

Technical Skills

Languages: Python, Rust, Dart, SQL

ML & Data: PyTorch, TensorFlow, TFLite, LightGBM, Scikit-learn, statsmodels, HuggingFace, SHAP, Pandas, Polars, NumPy

LLM & RAG: Qdrant, Ollama, RAG pipelines, semantic-entropy verification, FastAPI

Mobile & Backend: Flutter, Riverpod, Drift, GoRouter, Uvicorn, Gradio

Cloud & Tooling: GCP, AWS, Docker, GitHub Actions, Git, pytest, mypy, ruff, uv

Education

FATEC, São Paulo State Technology College | B.Tech, Big Data in Agribusiness | 01/2024 to 12/2026

Relevant coursework: Machine Learning, Artificial Intelligence, Data Structures, Databases (SQL & NoSQL), APIs & Microservices, Cloud Architecture, Statistics.

Languages: Portuguese (Native), English (B2, Upper-Intermediate)