

DAVID MOLIZANE

Software & AI Engineer

david.molizane@icloud.com | github.com/molizanee | linkedin.com/in/david-molizane

PROFESSIONAL SUMMARY

Software & AI Engineer with **2+ years** of experience building AI-powered and full-stack applications. Proven expertise architecting production **RAG (Retrieval-Augmented Generation) systems** with **vector databases (Qdrant, pgvector)**, **LLM integration (Langchain, Pydantic AI)**, **document ingestion, chunking, embeddings generation, and semantic retrieval**. Architected **multi-agent workflows** with state management via **LangGraph checkpoints** and deterministic routing through **semantic validation layers** achieving **40% conversion increase** and **60% support reduction**. Proficiency in building backend systems with **Python (FastAPI, SQLAlchemy)**, **Node.js & Bun.js (Express.js, Next.js, React.js, Vitest)**, **AWS (S3, EC2, Bedrock)**, **GCP (Bucket, Vertex AI, Cloud Run)**, and **Docker containers** and **CI/CD pipelines (GitHub Actions)**. Architected solutions that process huge amounts of data, increased company productivity, and reduced costs. Participated in the first AI Team at **Grupo Guanabara** (Brazil's largest bus transportation conglomerate with 20,000+ employees), successfully leading the implementation of a healthy AI culture, reducing AI costs by **30%** with LiteLLM proxy optimization, and eliminating **1+ day of manual work** through AI automation.

TECHNICAL SKILLS

AI & Data: RAG systems (40% retrieval accuracy improvement), Vector Databases (Qdrant, pgvector), LLM Integration (Langchain, Pydantic AI), Embeddings, Document Chunking, Semantic Search, Prompt Engineering, Synthetic Data, LiteLLM (30% cost reduction), LangFuse

Languages & Frameworks: Python, FastAPI, SQLAlchemy, JavaScript, TypeScript, Node.js, Bun.js, Express.js, GoLang, Go-chi, React, Next.js, Prisma ORM, Drizzle ORM

Data & Infrastructure: PostgreSQL, MongoDB, Redis, Microsoft SQL Server, MySQL, REST APIs, WebSockets

Cloud & DevOps: AWS (EC2, S3, Route 53, Bedrock), GCP (VM, Cloud Run, Bucket, Vertex AI), Docker, Kubernetes, Terraform, CI/CD (GitHub Actions), Observability (Grafana)

Testing & Practices: Jest, Vitest, React Testing Library, Agile/Scrum, Microservices

AI Safety & Guardrails: Content moderation (AWS Bedrock Guardrail), PII redaction (Microsoft Presidio), Deterministic response design, Hallucination mitigation

PROFESSIONAL EXPERIENCE

Molizane SYS (Freelancer)

February 2025 - Present

AI Engineer

- Developing **FestaPro AI Agent**, an agentic system built in **Python with FastAPI**. Features a **React Agent for WhatsApp Business API** that returns information about event venues and parties, answers contract questions, schedules venue visits, sends venue photos, and provides social links. Uses **Tortoise ORM** for schema management, **LangChain and LangGraph** for conversation state and flow routing, with checkpoints saved in **PostgreSQL**. Supports file uploads (PDF, TXT, DOCX) to **QdrantDB** for semantic context.

Grupo Guanabara

September 2025 - Present

AI Engineer

- Built **Commercial AI**, a Text-to-SQL AI Agent that receives business questions and returns answers using real company data. Architected with **n8n multi-workflow and multi-agent system**, **QdrantDB** for semantic search validation, **LiteLLM Proxy** for model consumption, **CubeJS semantic layer** for SQL payload generation with access to the data lake, and **Langfuse** for traces and evals. **Impact:** Executives gain quick business insights without relying on BI team, drastically reducing response time for business questions. Low cost achieved through strategic LLM selection: task-specific routing between Claude Sonnet 4.5, GPT-5, Nvidia Nemotron, Llama 4/3.1, and Qwen 3 based on cost-quality benchmarking.
- Designed **Guanabara Chat**, a complete AI stack for the entire company integrating **Open WebUI** for employee interface, **LiteLLM Proxy** for model and cost management per team, **AWS Bedrock** and **Google**

Vertex AI for models, **Langfuse** for traces and feedback, **Microsoft Presidio** for PII data masking, and **AWS Bedrock Guardrail** for content moderation. **Impact:** Secure AI access company-wide while keeping sensitive data within local infrastructure.

- Deployed **PCO**, a **WhatsApp Business Meta API** AI assistant helping bus drivers with incidents and support. Uses **QdrantDB** for incident embeddings ensuring deterministic responses and minimizing hallucinations in critical situations, **Redis DB** for conversation history management, and **Langfuse** for traces and evals. **Impact:** Drivers gain efficiency and accuracy on how to proceed in critical situations.

Grupo Guanabara

February 2025 - August 2025

Automation Developer

- Built **Punctuality**, a real-time vehicle delay alerting data pipeline. Consumes from a **WebSocket** bringing GPS data (service built in **Golang**), processes data in **n8n** with async architecture, executes queries on **AWS Athena**, and performs geolocation and time calculations for real-time delay alerts. **Processes over 864,000 vehicle data points per day** with low cost and high efficiency. **Impact:** Strategic business vision with massive generation of relevant data (routes, vehicles, and times with most delays), opening possibilities for ML to predict delays.
- Created **Orders Track**, a full-stack application (**React.js/Next.js** frontend, **Express.js/Bun.js** backend) providing an interactive real-time dashboard for delayed vehicle parts orders with automatic email alerts. Uses **Supabase auth** for user management and **PostgreSQL** for data storage. **Impact:** Logistics team **40% more efficient** due to optimized unified analysis of delayed order responsibilities.

Bravion

January 2024 - February 2025

Junior Full-Stack Engineer

- Designed and built the main microservice for **GenX**, a clinical genetics platform processing large volumes of patient genetic data and generating **PDF reports in under one minute**. Stack: **Node.js, TypeScript, Prisma ORM, PostgreSQL, react-pdf, Google Cloud Platform**.
- Developed core functionalities for **Cmorq**, an AI-powered cryptocurrency platform with **LGPD compliance**, building a secure **Next.js** application with **WalletConnect APIs** for user financial data protection.

Bravion

September 2023 - December 2023

Intern Full-Stack Engineer

- Designed a robust data migration process from legacy platform to **PostgreSQL**, including a **Python-based web crawler** capable of adapting to multiple legacy webpage layouts.
- Built and maintained the **Motorbooks** full-stack application with **FastAPI** backend and **React.js/Next.js** frontend, delivering a modern interface that improved usability and maintainability.

EDUCATION & LANGUAGES

Bachelor's Degree in Computer Engineering | Fundação Herminio Ometto | 2024 - 2030 (Expected, currently 3rd year)

Languages: English (Fluent - Professional Working Proficiency), Portuguese (Native)