

JUAN MONSALVE

SYSTEMS & COMPUTING ENGINEER | SOFTWARE ENGINEER — BACKEND, CLOUD & APPLIED AI

Bucaramanga, Colombia | +57 3175259997 | JuanMonsalve.23@hotmail.com | github.com/Moonsalve | linkedin.com/in/jmonsalve

PROFESSIONAL SUMMARY

Software Engineer with a strong foundation in Python, Java, and JavaScript, specializing in backend development, REST APIs, and applied AI systems. Proven ability to design scalable architectures, optimize databases, and implement containerized solutions using Docker and AWS. Recent work focuses on local-first AI products — LLM integration, embedding-based intent routing, and the latency, cost, and safety trade-offs that decide whether a model belongs on the critical path at all.

TECHNICAL SKILLS

Programming Languages

Python, Java, JavaScript/TypeScript, SQL, Bash

AI & Machine Learning

LLM integration (Ollama, local inference), sentence embeddings & semantic routing, speech-to-text (Whisper), structured tool calling with schema validation, ONNX Runtime / CUDA

Cloud & DevOps

AWS (EC2, S3, Lambda), Azure, Docker, Git/GitHub, CI/CD, Linux

Backend & APIs

Node.js, FastAPI, Express, RESTful APIs, Microservices architecture

Databases

PostgreSQL, MySQL, MongoDB, Redis

Practices

Test-driven development, typed contracts (Pydantic, TypeScript), technical documentation, Agile/Scrum

KEY SOFTWARE PROJECTS

APOLLO — Local-First Spanish Voice Assistant

Personal Project

Python 3.11 · Whisper · Sentence Embeddings · Ollama · Pydantic

2026

- Built a fully offline Spanish voice assistant that runs entirely on-device — **Silero VAD**, **faster-whisper large-v3-turbo** on CUDA, **Piper TTS**, and a local LLM served through **Ollama** — so no audio ever leaves the user's machine.
- Designed a **4-stage intent router** (literal → pattern → embedding similarity → LLM fallback) that resolves the majority of commands in **~2 ms** and keeps the 250–500 ms LLM call off the critical path; measured a **0.135 s end-to-end turn** (STT 0.121 s / routing 0.006 s / execution 0.008 s) on the target hardware.
- Diagnosed and fixed a semantic-routing failure where positive and negative cosine scores **overlapped** under any absolute threshold; solved it by mean-centering embeddings and introducing a negative `_fallback` class, converting the decision from absolute to relative — from unusable to **24/24 positives and 8/8 negatives**.
- Constrained LLM execution behind a **tool-call allowlist with strict Pydantic schemas** (`extra="forbid"`): the model can only emit validated tool calls, never shell commands or free-form paths. Covered by **307 unit tests**, with the router and skill layer testable off-device.

NovaTrack — Scalable Backend API for Task Management

Personal Project

- Designed and developed a high-performance RESTful API using **Python (FastAPI)** and **PostgreSQL**, implementing robust user authentication (JWT) and role-based access control.
- Containerized the microservices architecture utilizing **Docker** and orchestrated seamless deployment on **AWS EC2**, achieving 99.9% application uptime.
- Engineered automated CI/CD pipelines via **GitHub Actions** to streamline testing and deployment workflows, reducing code release cycles by 30%.
- Integrated comprehensive API documentation using Swagger UI, facilitating efficient frontend integration.

EM TORNEOS — League Management Platform (In Production)

Personal Project

Node · TypeScript · Prisma · PostgreSQL · Electron · Docker · Caddy

2026

- Built and **operate in production** a two-repository system — a typed **Node/Express API** and an **Electron** desktop client — used daily by an amateur football league; **22 domain models** covering tournaments, matches, debts, fines, credits and payroll.
- Deployed with **docker compose** on a self-hosted server: **Caddy** terminating TLS and renewing certificates unattended, **PostgreSQL** with no published port reachable only over the internal network, and migrations applied on container start.
- Implemented **daily backups plus a restore-drill script** that loads the latest dump into a separate database and verifies row counts — a backup that has never been restored is not a backup.
- Ship **self-updating desktop releases** from that same server, with the API address compiled into the installer so operators install and sign in with zero configuration; permissions follow the job, not the hierarchy.

PROFESSIONAL EXPERIENCE

Team Manager — Operations & Data Analysis

Nov 2025 – Aug 2026

WW Funcrafters JWA LLC

Remote

- Led a cross-functional team of 15+ employees, overseeing technical operations, quality assurance, and KPI achievement.
- Automated weekly performance reporting using **Python (Pandas)** and SQL to analyze productivity metrics, increasing reporting efficiency by 40%.
- Managed escalated technical cases, performing root-cause analysis to deliver effective resolutions and maintain 95%+ operational target achievements.

Customer Sales & Support

Feb 2024 – Nov 2025

WW Funcrafters JWA LLC

Remote

- Handled sales and customer support for US clients in English, in a remote, high-volume environment.
- Promoted to Team Manager in November 2025, taking over operations and reporting for the same team.

Customer Technical Support Engineer

Dec 2023 – Sept 2024

Concentrix Web Help

Remote

- Diagnosed and resolved complex application errors and functional issues through systematic root-cause analysis and log debugging.
- Collaborated closely with L2/L3 engineering teams by documenting reproducible steps and technical logs for software bugs, accelerating time-to-resolution.
- Consistently met quality and response-time KPIs in a high-volume environment while adapting to continuous software updates and architectural changes.

EDUCATION & CERTIFICATIONS

Systems & Computing Engineering (Bachelor's Degree)

2023 – 2026

Universidad Pontificia Bolivariana

Bucaramanga, Colombia

Certifications & Coursework: AWS Cloud Practitioner Fundamentals, Linux Fundamentals, Python Programming, SQL & Databases, Git & GitHub, Agile & Scrum.

Languages: Spanish (Native), English (Advanced, C1+).