

Marcus Acosta

marcus.alan.acosta@gmail.com | LinkedIn | GitHub | marcusacosta.dev

EXPERIENCE

Bettorca

Co-Founder & Engineer

Remote

Sep 2025 - Present

- Owned **XGBoost** and **PyTorch** pipelines for prediction market scoring and calibration, orchestrating high-throughput ingestion, training, and continuous evaluation.
- Deployed **11 ONNX** models via **MLflow** with version-controlled artifacts, automated parity tests, and rollback support across **Python** and **Go**.
- Engineered real-time inference APIs that combined model probabilities with live inputs, achieving **4–31 ms** average response latency.
- Built production safeguards across **11.5M** model outputs, filtering **74%** of low-confidence or unfamiliar-input predictions before delivery.
- Instrumented ML telemetry that detected **500K–700K** daily input changes and triggered **40K–50K** stale-decision re-evaluations.
- Reduced external LLM calls by **70%** using **Qdrant RAG**, semantic caching, and request coalescing for repeated or similar requests.

OPEN SOURCE

mistral.rs

Contributor

GitHub

- Optimized PagedAttention configuration, preserving automatic fallback while making explicit requests fail clearly across loaders and scheduling.
- Added CPU-only regression tests for device, loader, scheduler, and SDK paths, preventing silent fallback from required PagedAttention configurations.
- Added pluggable external KV-cache connectors to live PagedAttention scheduling, with an in-memory reference implementation and runnable chat example.

torch tune

Contributor

GitHub

- Added native SimPO preference optimization with length-normalized scoring, preserving existing DPO/RSO behavior across LoRA and distributed training.
- Built a runnable training configuration and regression coverage for loss correctness, gradients, reference-free execution, and GPU integration.
- Added fail-fast LoRA target-module validation, preventing silently disabled adapter training; delivered CPU regression tests and a production-ready pull request.

SKILLS

Languages: Python, Go, Rust, SQL, TypeScript, JavaScript

ML/AI: PyTorch, XGBoost, scikit-learn, FLAML, MLflow, ONNX, Qdrant, LangGraph, LangSmith

Data Systems: PostgreSQL, DuckDB, Parquet, PyArrow, pandas

Frameworks: FastAPI, React, React Native

Infrastructure: Kubernetes, Docker, GitHub Actions, Git, Railway

ADDITIONAL EXPERIENCE

LifeTime Inc.

Facility Operations Supervisor

Walnut Creek, CA

Jan 2024 - June 2026

- Managed daily operations for a 5–7 person facilities team, coordinating workflows, inventory, vendors, safety inspections, and compliance documentation.

EDUCATION

Post-Baccalaureate Coursework in Computer Science

Oregon State University

- Data Structures, Discrete Mathematics, Databases, Software Engineering, Web Development

Bachelor of Arts, Sociology

UC Santa Cruz