

Matthew Poncini

Full-Stack Software Engineer — AI Infrastructure, Data Systems, and Evaluation

San Francisco, CA

MJPon8896@gmail.com — LinkedIn — GitHub — monstra.bot

Professional Summary

Full-stack software engineer with an M.S. in Computer Science and B.S. in Applied Mathematics, focused on AI infrastructure, data-intensive applications, evaluation systems, and production workflows. Founder of Monstra, where I built and operate a full-stack platform spanning frontend interfaces, backend services, data pipelines, automated decision systems, external APIs, and production infrastructure. Experienced in Python, TypeScript, React/Next.js, PostgreSQL, ML evaluation, and designing systems that evolve quickly without sacrificing reliability, traceability, or maintainability.

Technical Skills

Languages: Python, TypeScript, JavaScript, SQL, C++, SAS

Full-Stack Engineering: React, Next.js, Node.js, REST APIs, PostgreSQL, Prisma, authentication, responsive interfaces, internal tools

AI / ML: PyTorch, Hugging Face Transformers, PEFT, TRL, LoRA, QLoRA, LLaMA-2, classification, feature engineering, clustering, model evaluation, prompt design

Data and Evaluation Systems: data pipelines, experiment tracking, versioned configurations, model outputs, benchmark comparison, failure analysis, statistical analysis, reproducible evaluation

Production Engineering: Docker, Linux, background workers, scheduled jobs, retries, idempotency, transactional state management, observability, recovery workflows, Vercel, Render

Experience

Founder & Full-Stack Software Engineer

San Francisco, CA

Monstra.bot

2026–Present

- Founded and built an end-to-end quantitative research and automation platform using Python, TypeScript, Next.js, PostgreSQL, REST APIs, and external integrations.
- Own the full product surface from database design and backend services through user-facing web and mobile interfaces, deployment, monitoring, and production debugging.
- Designed modular workflows for configuring algorithms, portfolios, execution settings, external integrations, and system state while abstracting complex backend behavior into understandable user experiences.
- Built data pipelines that transform raw market observations into normalized, clustered, and composite feature representations used by downstream research and decision systems.
- Developed evaluation infrastructure for comparing models, strategies, parameter configurations, and changing data universes using reproducible experiments and explicit information cutoffs.
- Built dashboards and product surfaces for monitoring portfolio state, algorithm behavior, execution history, system actions, positions, and performance.
- Engineered backend workflows that ingest data, evaluate model outputs, validate readiness, execute external API actions, and preserve attribution from input state through observed outcome.
- Designed reliability controls including retries, idempotency, transactional completion ordering, stale-output protection, exact-run matching, recovery jobs, and structured failure reporting.
- Migrated production validation from modeled historical execution to broker-executed paper trading, enabling evaluation through actual orders, fills, positions, and resulting portfolio state.
- Regularly redesign product and infrastructure components when real system behavior reveals weaknesses in earlier assumptions, balancing rapid iteration with long-term maintainability.
- Operate independently across frontend, backend, data, infrastructure, and product direction, identifying high-impact problems and driving them from ambiguous idea to deployed system.

Selected Machine Learning and Evaluation Projects

Fine-Tuning LLaMA-2-7B with QLoRA

2025

- Fine-tuned LLaMA-2-7B on a domain-specific financial question-answering dataset using PyTorch, Hugging Face Trans-

formers, PEFT, TRL, BitsAndBytes, and QLoRA.

- Built a complete training and evaluation pipeline with separate training and held-out test sets, covering tokenization, batching, GPU execution, generation, output extraction, and metric computation.
- Used 4-bit NF4 quantization and parameter-efficient LoRA adapters while keeping base-model weights frozen.
- Improved ROUGE-1 from 0.1205 to 0.2519 and ROUGE-Lsum from 0.0818 to 0.1526 after fine-tuning.
- Evaluated hallucination, truncation, verbosity, domain adaptation, and behavioral overfitting in addition to aggregate metrics.

LLM-Based Financial Signal Generation

2025

- Designed an end-to-end ML experiment testing whether economic-news language contained useful information for predicting subsequent foreign-exchange movement.
- Collected and labeled 1,160 historical news observations, aligned them with future EURUSD, EURJPY, and USDJPY price changes, and created directional classification targets.
- Fine-tuned three specialized LLaMA-2-7B QLoRA adapters and evaluated them against held-out examples.
- Identified model collapse toward the majority Neutral class despite improving aggregate accuracy, tracing the failure to class imbalance and insufficient discriminatory signal.
- Proposed revised sampling, weighting, feature, and objective strategies based on observed model behavior rather than treating benchmark improvement alone as success.

Open Shop Scheduling with Wisdom of Crowds

2025

- Built Python experiments using five independently evolving genetic-algorithm populations with different parameter configurations to solve open-shop scheduling problems.
- Designed fitness functions, selection and mutation workflows, experiment tracking, parameter comparison, and statistical evaluation.
- Aggregated independently optimized schedules through a Wisdom of Crowds approach and evaluated the effect on solution quality.

Additional Professional Experience

Staff Sergeant, 94H Test, Measurement, and Diagnostic Equipment Specialist

U.S. Army National Guard

May 2019–May 2026

- Led technical personnel supporting more than \$10 million in precision test and calibration equipment; diagnosed complex system failures through structured testing, validation, troubleshooting, and root-cause analysis.
- Managed technical workflows, documentation, accountability, and competing operational priorities in a high-reliability environment.

Mathematics and Computer Science Teacher

San Francisco, CA

ICA Cristo Rey

August 2021–June 2025

- Taught mathematics, statistics, programming, and quantitative reasoning while translating complex technical concepts for users with different levels of technical experience.
- Designed curricula, tools, assessments, and workflows around user needs and iterated based on direct feedback and observed outcomes.

Actuarial Intern

Remote

CSAA Insurance Group

January 2021–April 2021

- Supported pricing and retention research using SAS, statistical analysis, quantitative modeling, data validation, and business-impact analysis.

Education

University of Louisville

May 2026

Master of Science in Computer Science

Louisville, KY

Relevant study: generative AI, artificial intelligence, machine learning, optimization, algorithms, data mining, databases, software engineering, computer systems, applied statistics, and C++ programming.

Saint Mary's College of California

Spring 2018

Bachelor of Science in Applied Mathematics, Minor in Economics

Moraga, CA

Relevant study: probability, statistics, regression, numerical methods, mathematical modeling, optimization, economics, and game theory.