Zijun Liu

San Diego, CA · liuzijun6688@gmail.com · 530-220-8681 · Github · Personal Portfolio Website · huggingface

Summary

UCSD Data Science & Astrophysics student specializing in **ML engineering, LLM fine-tuning, and multi-agent systems**. Proven track record of implementing reinforcement learning frameworks (RLAIF), engineering **RAG pipelines for production LLM grounding**, building end-to-end data pipelines, and deploying optimized models in resource-constrained environments.

Education

University of California, San Diego - B.S. Astrophysics, Data Science minor, *Sep 2025 ~ June 2027*, **GPA: 3.98 (Institutional) | 3.88 (Cumulative)**

University of California, Davis - B.S. Physics (Astrophysics Emphasis), Data Science minor, *Sep. 2023 ~ June 2025*, **GPA 3.84**

Skills/Tools

- **Languages: Python, C++**, SQL, Bash/Shell, R
- **Machine Learning & AI:** PyTorch, **PyTorch Lightning**, **RLAIF, GRPO**, Hugging Face (TRL, Transformers), Edge Inference (llama.cpp, GGUF), Model Quantization (4-bit), PEFT/LoRA, **Variational Autoencoder**, XGBoost, LightGBM, Temporal Convolutional Network, Scikit-learn, Optuna, CrewAI (Multi-Agent Systems), **RAG**
- **Data Engineering & Systems: REST APIs (C++ httplib)**, **CMake Build System**, PostgreSQL, DuckDB, Polars, IPC, BigQuery, ETL Architecture, Chroma (Vector Store)
- **Cloud & MLOps:** AWS (Lambda, SageMaker, S3), Docker, Git/GitHub CI/CD, Linux Environment, Parallel Computing

Experiences

XView LLC — AI/ML Engineer Intern *(Apr 2026 – present)*

- **Architected a 5-agent CrewAI content pipeline** (Trend Analyst → Content Creator → Platform Adapter / SEO Strategist → Analytics Reviewer) with Claude 3.5 Sonnet and Chroma RAG over Markdown product docs; generates **35 platform-tailored drafts per run** across 5 social platforms, cutting content production time by **~90%**.
- **Drove a ~60% reduction in brand factual errors** by engineering and evaluating a RAG integration; established phased evaluation protocol (prompt A/B testing, chunking strategy comparison) to iteratively improve output quality, achieving **~30% human revision rate** pre-publish.

Rothenberg Wealth Strategies — AI/ML Quantitative Research Intern *(Jan 2026 – Apr 2026)*

- **Conducted systematic falsification of 40 LLM-generated alpha factors** sourced from academic literature; implemented out-of-sample vectorized backtesting with transaction cost modeling and blocked deployment of all 40 spurious signals by proving negative post-cost returns.
- **Architected an iterative feature selection pipeline** using **LightGBM** to evaluate 200+ engineered features; applied **Mutual Information** scoring to prune highly correlated pairs and eliminate **OOM errors** during **backtesting**.
- Built an 8-layer **Temporal Convolutional Network (TCN)** for probabilistic quantile forecasting (P10/P50/P90); stabilized training via **gradient norm clipping** and **dropout**; maximized GPU throughput by resolving I/O bottlenecks with async **PyTorch DataLoaders** and **memory pinning**; achieved **~15% Sharpe improvement** over LightGBM.

- Engineered a three-stage signal pipeline over ~10M rows (~1,000 US equities): a self-supervised **Conv1D-VAE** compresses market windows into **512-dim latent representations**; an **LGBM Scout** + **meta-learning Gatekeeper** filter low-confidence signals via **CPCV**; achieved **~7% win rate improvement** over baseline LightGBM.

Himiway Intelligent Technology USA — Data Science intern *(Sep 2025 – Dec 2025)*

- **Architected a scalable SQL/Python ETL pipeline** for e-bike sales forecasting; refactored Jupyter notebooks into a production-ready codebase (argparse, Makefile orchestration) to streamline continuous data ingestion.
- **Developed and deployed an end-to-end XGBoost model** to optimize regional store locations and inventory allocation, successfully accelerating the executive decision-making cycle by **75% (from 4 weeks to 1 week)**.
- **Designed and evaluated a months-long strategic A/B test** on in-store customer experiences (fat-tire vs. thin-tire); leveraged **Tableau** dashboards to analyze key metrics including Conversion Rate and Revenue per Visitor, driving data-informed retail operations.

Projects

KataGo × LLM - Explainable Go AI | *Project Lead - 2025 ~ Present* | demo video

- **Spearheaded an explainable Go AI** by fine-tuning Qwen3-8B via GRPO, elevating bot strength from a 10k baseline to ~7k in real-world testing while delivering natural-language move rationales for university Go club players.
- **Resolved severe reward hacking** across a 113k-row dataset by engineering a RLAIF pipeline; replaced static metrics with a -0.5 format-validation penalty gate and dynamically scaled policy/score-lead weights for lopsided positions.
- **Architected a zero-network-overhead C++ proxy** bridging the Lizzie GUI with local LLMs; achieved ~42.5 tok/sec inference throughput on resource-constrained 8GB VRAM edge devices via 4-bit GGUF model quantization.

Breaking Barriers Hackathon (AWS × Deloitte × AT&T) — *Finalist (Top 8/32) - 2025* - demo (refresh if can't see moving dots)

- Architected a real-time anomaly detection system for simulated tail-risk events, implementing a **0.5–1s micro-batching pipeline** to efficiently process LLM-generated synthetic data streams.
- Engineered **lagged spatio-temporal features** via client-side sliding windows for stateless backend processing; trained an **XGBoost classifier**, tuning thresholds via **PR AUC** to minimize False Negatives on imbalanced data.
- Deployed ultra-low latency model inference via **AWS Lambda (<300ms)**, streaming anomaly triggers to an S3-hosted **Amazon Location Service** interactive dashboard for real-time visualization.

ASTR199 - Stellar Parameter Prediction via Deep Learning (with Prof. Theissen) | *ML Researcher - 2026* - final report

- **Engineered an ETL pipeline** processing ~1M rows of stellar photometry across 7 catalogs; utilized DuckDB for high-throughput cross-matching and NA handling, integrated with Pandas for sigma clipping.
- **Overcame multi-task collapse** in joint-loss training (via Homoscedastic Uncertainty Loss) by architecting a two-stage sequential deep learning pipeline, bypassing the need for computationally expensive sub-models to predict fundamental stellar parameters.
- **Formulated a physics-informed feature space** in Pandas by engineering 171 color indices and 19 absolute magnitudes; injected Stage-1 T_eff predictions as Hertzsprung-Russell diagram proxies to break color-log g degeneracy, driving **log g R^2 from 0.519 to 0.833**.