Raiymbek Kokteubay

718-603-0712 | rk3230@columbia.edu | linkedin.com/in/raiymbek | github.com/raiymbek

Education

Columbia University August 2023 – May 2027
Bachelor of Arts in Data Science, Concentration in Economics *New York, NY*
Relevant Coursework: ML, AI, Linear Regression Models, Algorithms, Intro to Databases, Statistical Inference

Experience

Eulerion Technologies Inc. May 2025 – August 2025
Computer Vision Engineer *Cambridge, MA*

- Led a real-time, distance-adaptive 4K analytics pipeline (**Deep SORT** and **YOLO**) with hybrid two-stage multi-model fusion, image tiling, and parallelized async GPU inference; delivered **state-of-the-art** performance.
- Achieved **99.8% precision** and **99% accuracy** on benchmark test data for object identification tasks.
- Boosted inference throughput **30×** (1 FPS → 30 FPS) in live 4K streaming.
- Integrated **segmentation models (SAM(META), SAMURAI)**, attaining **90%** accuracy on complex scenes.
- Built applications with **FastAPI** and **PostgreSQL** integrating ML models for real-time gaming analytics.

Investbanq July 2024 – August 2024
Quantitative Research Intern *Singapore*

- Built ensemble regression models **(Random Forest, CatBoost)** to predict **Sharpe ratios**, achieving a **10%** improvement in predictive accuracy for systematic stock selection.
- Applied **Hierarchical Risk Parity (HRP)** decreasing portfolio volatility by **5%** relative to traditional methods.
- Backtested portfolio strategies under varying rebalancing schemes, achieving a **6%** cumulative return increase compared to standard single-time construction and periodic rebalancing.

Projects

CloudEx | *React, FastAPI, LangChain* July 2024 – Present

- Developing a **cloud-based AI powered IDE** to automate software development and deployment, with an interactive UI for file management and minimal bash scripting.
- Implementing **AI feature** to create and modify files, and execute bash scripts autonomously.

QPOML 2.0 | *AstroPy, PyTorch* June 2024 – Present

- Investigating the connection between energy and power density spectra for low-mass X-ray binaries using **neural networks, normalizing flow,** and **dimensionality reduction algorithms**.
- Designed a novel **SPENDER-inspired autoencoder** and tested it on thousands of space telescope observations of black holes, aiming to capture spectral-temporal patterns for astrophysical modeling.

Jane Street x GPU MODE Hackathon | *PyTorch, Triton, GPU Optimization* September 2025

- Placed **5th** in a live trading-style ML inference challenge optimizing accuracy, latency, and throughput.
- Implemented **dynamic batching**, applied **torch.compile**, and tested **Triton kernels** to accelerate models.
- Reduced average latency to **38.8 ms** (20× improvement) while sustaining **394 inferences/second** throughput.
- Explored kernel-level optimizations to reduce bottlenecks and achieve real-time inference speedups.

Columbia Data Science Datathon | *CatBoost, XGBoost, Random Forest* October 2023

- Placed **2nd** in the Patient Safety track for predicting drug-induced health risks with a four-person team.
- Led a team in cleaning and engineering features from medical datasets and trained **CatBoost, XGBoost,** and **Random Forest** models, reaching **86%** test accuracy.

Technical Skills

Languages: Python, MySQL, PostgreSQL, MongoDB, Neo4j, Java, JavaScript, C/C++
ML & Computer Vision: PyTorch, scikit-learn, YOLO, OpenCV, Deep SORT, ONNX, LangChain, WandB
Web Development: FastAPI, Flask, React.js, Django, Node.js
Data & Analytics: AstroPy, Pandas, NumPy, Matplotlib, Seaborn
Developer Tools & Cloud: Git, Docker, Google Cloud Platform