Satyam Rai

United States | (+1) 240-476-1102 | satyamrai@outlook.com | linkedin.com/in/sr007 | github.com/satyamrai0511 | https://satyamrai0511.github.io

SUMMARY

Data Scientist with 3+ years of experience in building and optimizing data pipelines and predictive models for high-volume, structured and unstructured datasets. Skilled in Python, SQL and C/C++, with expertise in developing anomaly detection algorithms, domain-specific compilers and deep learning pipelines that improve data observability and reduce processing latency. Experienced in creating dashboards and reports to translate complex analyses into actionable insights for technical and non-technical stakeholders.

EDUCATION

University of Maryland, College Park | ***Master's, Data Science*** **(GPA: 3.7)** **Aug 2024 - Dec 2025**

Adamas University | ***Bachelor's, Computer Science & Engineering*** **(GPA: 3.9)** **Aug 2017 - Jul 2021**

TECHNICAL SKILLS

- **Evaluation & Benchmarking:** Automated Benchmarks, FlashInfer, Flash Attention, Regression Testing, Safety Metrics Validation
- **Languages:** C/C++ (Advanced), Python (Advanced), CUDA C/C++, SQL, Bash Scripting, Rust (Concepts)
- **ML & Research Stack:** PyTorch, JAX, TensorFlow, ONNX, vLLM, SGLang, MLC, HuggingFace, Scikit-Learn
- **Infrastructure & Systems:** GPU Kernel Development, JIT Compilers, Apache TVM, MLIR, Triton, cuTile, Distributed Inference
- **Cloud/DevOps:** GCP, AWS, Docker, Containerization, Orchestration, Cloud Computing, Telemetry Analysis
- **Domain:** Deep Learning Systems, High-Impact AI Workloads, LLM Inference Runtimes, Multi-modal Analysis, Anomaly Detection
- **Analytics & Reporting:** Machine Learning, Statistical Analysis, Data Visualization

EXPERIENCE

Google LLC | ***AI and Automation Engineer*** **May 2025 - Present**

- Innovated new AI systems technologies for efficient inference, leveraging C/C++ and Python to design extensible abstractions that significantly accelerated large language model (LLM) serving across robust production environments.
- Designed, implemented, and optimized custom GPU kernels using CUDA C/C++ and Triton for AI workloads, improving computational throughput and reducing latency for complex reasoning tasks across distributed clusters
- Collaborated closely with research engineers across deep learning frameworks to build efficient just-in-time (JIT) domain-specific compilers, hardening prototype LLM inference runtimes into reliable Google services
- Orchestrated distributed workloads on large GPU clusters, using containerization to streamline inference pipelines and validate attention kernel implementations against automated benchmarks, confirming improved performance
- Engineered automated pipelines for evaluating AI agent reasoning, integrating PyTorch and JAX to analyze system telemetry and refine safety guardrails for LLM serving engines

MSCI Inc. | ***Data Scientist*** **Nov 2023 - Aug 2024**

- Accelerated the transition of research prototypes into production by developing efficient inference system software, using Python and PyTorch to optimize core financial scoring engines, applying statistical analysis to refine models, and reducing latency by 28%
- Engineered high-throughput inference infrastructure using Apache TVM and ONNX, ensuring high-fidelity inputs for downstream predictive models and significantly reducing bottlenecks in complex risk valuation deep learning workloads.
- Designed and implemented extensible abstractions for high-performance computing pipelines, translating ambiguous risk signals into reliable, observable input streams that leveraged advanced GPU kernel optimization strategies, which improved the consistency of downstream risk model inputs
- Made the deep learning stack highly observable and reproducible by establishing automated documentation and data visualization dashboards, enabling cross-functional teams to reliably extend JIT compilers and predictive model architectures.
- Conducted deep performance analysis to identify runtime bottlenecks in complex risk models, collaborating with GPU arch teams to orchestrate systemic fixes utilizing C/C++ that substantially improved overall AI workload efficiency.

Trove Research Ltd. | ***Data Scientist*** **Oct 2021 - Oct 2023**

- Developed foundation models for forecasting market telemetry, created backtesting benchmarks, and used PyTorch, TensorFlow, and CUDA C++ to validate inference reliability against historical ground-truth data, confirming prediction accuracy within target thresholds
- Built domain-specific compilers and anomaly detection algorithms that automatically identified irregular patterns in high-volume telemetry logs, improving data observability and increasing downstream scoring accuracy by 15%, and consolidated data with SQL
- Optimized deep-learning ingestion pipelines and designed preprocessing steps for high-frequency market data using TensorFlow and CUDA acceleration, reducing processing latency by a measurable margin to enable real-time model inference at scale
- Implemented monitoring guardrails for LLM model drift and built interactive visualizations with Plotly to present findings to technical and non-technical stakeholders, supporting timely decision-making on model updates

SHA Infotech | ***Data Scientist*** **Sep 2020 - Sep 2021**

- Optimized algorithmic efficiency and redesigned database schema for core inference data retrieval using C/C++ and Python on GCP, cutting query latency by 20% and delivering reliable access for real-time dashboards
- Optimized algorithmic efficiency and database schema architecture for core inference data retrieval, leveraging C/C++ and Python to achieve a 20% reduction in query latency, and performed ad-hoc statistical analysis to validate data integrity.
- Implemented automated scraping agents in Python on GCP to collect and clean competitor data, producing accurate telemetry datasets that fed internal pricing algorithms and a machine-learning analytics pipeline, which improved pricing model accuracy by enabling timely data updates
- Implemented automated scraping agents to gather and clean competitor data, loading the datasets into a SQL-based machine-learning analytics platform and generating data visualizations that supported strategic decision-making, reducing analysis turnaround time