Rayaan Attari

rta2125@columbia.edu • (321) 890-7155 • github/attarira • linkedin/in/attarira

EDUCATION

Columbia University, Fu Foundation School of Engineering and Applied Science — **New York, NY**
M.S. in Computer Science, Machine Learning Track [GPA: 3.78] — Dec 2025
Coursework: Applied ML, NLP, Unsupervised Learning, Deep Learning, Big Data Analytics, Distributed Systems

Grinnell College — **Grinnell, IA**
B.A. in Computer Science and Mathematics [GPA: 3.94] — May 2022

SKILLS

Core Languages & Systems: Python, Java, SQL, TypeScript
ML Stack: PyTorch, TensorFlow, Scikit-learn, Transformers, Statistical Modeling
Cloud & DBs: Google Cloud Platform (Certified), AWS, Docker, Kubernetes, PostgreSQL

PROFESSIONAL EXPERIENCE

Routerr Health (Columbia BuildLab) — **New York, NY**
Founding ML Engineer — Aug 2025 – Present

- Accelerated Hospital-at-Home scheduling by **2.5x** using combinatorial optimization and predictive modeling systems for the Clinical Route Optimization and Scheduling (CROS) platform, ensuring timely deployment of clinical staff to patient homes
- Built probabilistic time-series models (XGBoost, PyTorch) to generate **95%** prediction intervals for intraday patient demand and clinician availability, enabling uncertainty-aware staffing decisions and reducing scheduling overhead by **40%**
- Fine-tuned a domain-specific language model on scheduling context and external signals such as availability and traffic to drive real-time decisions like dynamic clinician rerouting, with hard guardrails to enforce clinical and operational constraints
- Directed cross-functional product initiatives, collaborating with design and engineering to experiment, A/B test, and launch data-driven features, improving clinician resource utilization by **28%** across pilot deployments in New York City

Avati Consulting Solutions — **Mumbai, IN**
ML Engineering Intern — May 2025 – Aug 2025

- Engineered an AI-driven Early Warning System within a risk management and compliance software suite to forecast credit default and portfolio risk **3-6** months in advance using financial, behavioral, and macroeconomic signals across **15+** banks
- Architected and deployed predictive risk models using statistical and ensemble methods (logistic regression, random forests, gradient boosting), achieving **0.78** AUC and enabling proactive detection of at-risk accounts for credit monitoring
- Optimized inference with feature pruning and batch scoring, reducing end-to-end latency by **4x** for real-time risk evaluation
- Collaborated with business stakeholders to translate model outputs into actionable risk insights by building an LLM-based reporting pipeline using automated data labeling and QLoRA fine-tuning, reducing manual report generation by **60%**

Perficient — **Dallas, TX**
Software Engineer — Jun 2022 – Mar 2024

- Pioneered the end-to-end modernization of large-scale customer analytics platform by architecting an event-driven microservices solution on Google Cloud that increased modularity and reduced average deployment time by **40%**
- Migrated large-scale on-prem databases to Cloud SQL, designing ETL workflows to cleanse and transform data; streamlined data retrieval and reduced query response time by **3x**, improving SLA compliance for customer-facing applications
- Engineered and deployed high-performance, low-latency RESTful APIs and ML inference pipelines for real-time predictive analytics using Vertex AI; integrated Cloud Firestore and Pub/Sub to achieve **sub-50ms** data processing times for queries
- Implemented automated model training, evaluation, and deployment pipelines with Bayesian hyperparameter optimization for sales reward models, reducing retraining time by **30%** while serving **100k+** concurrent users

RESEARCH & PROJECTS

Media Watcher: AI-powered News Intelligence | *Python, Gemini, NLP, Sentiment Analysis* — Jan 2026
Utilized search grounding and fine-tuned sentiment models to produce structured, citation-backed risk intelligence; implemented schemas to extract entities and risk scores in real-time, reducing manual compliance review time by **~70%**

Spike Sorting: Efficient Dimensionality Reduction for Neural Recordings | *Python, PyTorch, Scikit-Learn* — Dec 2025
Benchmarked **20+** dimensionality reduction and clustering methods (UMAP, t-SNE, VAEs) on large-scale neural embeddings, achieving **10x** compression with **85%** information retention; results under review for publication

FinSearch: Q&A for Regulatory Finance | *Python, LangChain, RAG, Pinecone* — Oct 2025
Built an agentic RAG system for financial document Q&A using LangChain and Pinecone, incorporating function calling and self-correction loops to iteratively refine responses, improving answer accuracy by **50%** over baseline