NIRMAN TATERH

Brooklyn, NY | +1-929-290-6179 | nt2613@nyu.edu
linkedin.com/in/nirman-taterh | nirmant.vercel.app

SUMMARY

ML/AI Engineer with 4+ years shipping production systems spanning NLP, RAG, and algorithmic finance. Expertise in deep learning (Transformers, LLM fine-tuning), reinforcement learning, statistical analysis, and systematic trading systems. Delivered 18–40% accuracy/latency gains and +50% CTR improvements across deployments. Published researcher in RL-based research (Spring 2021); ACL-tier NLP work pending. GPA: 3.7/4.0.

EDUCATION

New York University - M.S. Data Science | Sep 2023 – Present
GPA: 3.7/4.0 | Coursework: Robo Advisors & Systematic Trading (ongoing), NLP, Data Science & AI for Business, Statistical Modeling
K. J. Somaiya College of Engineering - B.Tech. Electronics & Telecom | Jun 2019 – Jun 2023
GPA: 3.68/4.0

EXPERIENCE

Johnson & Johnson
Data Science Intern - LLM / NLP | Mar 2025 – Aug 2025 | New York, NY · Remote
- **+18% accuracy (biomedical decision-support)** by identifying evaluation gaps and iterating transformer-based NLP pipelines (Python, HuggingFace) vs prior baseline models, improving clinical prediction reliability
- **45s to 31s preprocessing latency (-14s, -30%)** by redesigning ETL pipelines (Python, Spark, AWS SageMaker), eliminating bottlenecks across 3 biomedical R&D workflows
- **4 to 5 experiments/week (+1, +25%)** by introducing AI-assisted workflows (Cursor, Copilot) and building shared evaluation tooling, accelerating model iteration cycles

Miko
Machine Learning Engineer / Conversational AI Developer | Nov 2024 – Jan 2025 | California, USA · Remote
- 24% to 19% hallucination rate (-5 pts, -20%) across 5K+ interactions by redesigning RAG pipelines and tool-calling systems (Python, LangChain, FAISS), improving response grounding.
- **80 to 100 req/s throughput (+20, +25%)**, **3.2% to 2.7% error rate (-0.5 pts, -15%)** by building automated evaluation pipelines and optimization loops (HuggingFace, MLflow), stabilizing production inference.
- 2.**1 to 2.3 features/week (+0.2, +10%)** by mentoring 3 engineers on AI-assisted debugging workflows (Cursor, Copilot), reducing iteration and code review time

Miko
NLP Intern | Jun 2022 – Sep 2022 | Mumbai, India
- 0.694 to 0.867 F1 (+0.173, +25%) on multilingual classification (5 languages) by fine-tuning SIEBERT transformers with domain adaptation
- **0.62 to 0.84 BLEU (+0.22, +35%)** on response generation quality by optimizing transformer-based multilingual pipelines

Paramount Global (Last.fm)
Music Recommendation Systems Engineer | Sep 2020 – Feb 2021 | London, UK · Remote
- **2.8% to 4.2% Click Through Rate (+1.4 pts, +50%, p-value < 0.05)** across 1.9M users by building hybrid recommender systems (collaborative filtering + NLP metadata features) for 93K artists
- **240 ms to 168 ms inference latency (-72 ms, -30%, <200 ms SLA)** by applying model quantization, distributed training (4 GPUs), and pipeline optimization under constrained infrastructure

PROJECTS & PUBLICATIONS

NLP Research - Implicit Prosodic Encoding in Language Models *(Pending Publication)*
- **Probed latent prosodic encoding in transformer LMs (GPT-2, BERT)** by designing PyTorch-based probing classifiers, demonstrating implicit structure capture without explicit supervision
- **Built end-to-end transformer encoder probing pipeline** for controlled linguistic evaluation, targeting submission to a top-tier ML venue

Large-Scale Review Trust Modeling
- **97.7% recall, 96.1% accuracy** on 26.7M reviews by building fraud detection pipeline (logistic regression, random forests, BiLSTM-attention) using Python and Spark vs traditional rule-based baselines
- **1.47 to 0.46 MAE (-1.01, -69%)** in sentiment scoring by fine-tuning DeBERTa and designing trust-weighted composite scoring for product reliability

Bitcoin Investing Assistant ***(ICIVC 2021, Proceedings in Adaptation, Learning and Optimization, Vol. 15)***
- **+10% return vs benchmark strategies** by developing FinRL-based reinforcement learning agent for portfolio optimization
- **+14% accuracy (financial NLP)** by integrating FinBERT sentiment analysis into trading signal pipeline vs baseline models

TECHNICAL SKILLS

Languages: Python, SQL, C++, TypeScript, Scala, R
Machine Learning & AI: Deep Learning, Transformers (BERT, T5), LLM Fine-Tuning, Reinforcement Learning (RL), Supervised & Unsupervised Learning, Bayesian Methods, Model Evaluation
NLP & LLM Systems: Natural Language Processing (NLP), Retrieval-Augmented Generation (RAG), Prompt Engineering, HuggingFace, LangChain, LlamaIndex, FAISS, Vector Databases, Large Language Models (LLMs)
Systems & Data Engineering: Spark, ETL Pipelines, Distributed Systems, SQL Optimization
MLOps & Cloud: Docker, Kubernetes, MLflow, AWS (SageMaker), Google Cloud Platform (GCP), Git, Model Serving / Deployment
Tools: Cursor, GitHub Copilot, Claude