Sunny Son

(201) 496-4348 | sunnys2327@gmail.com | LinkedIn | GitHub | Website

Relevant Experience

Globalink AI Inc | Remote (Calabasas, CA)

Machine Learning Engineer — *Feb 2026 – Present*

- Designing an agentic BI decision engine combining ML models, LLMs, and Knowledge Graphs to generate actionable e-commerce merchant recommendations, reasoning across competing objectives including growth, margin optimization, and inventory risk
- Architecting inference pipeline integrating Knowledge Graph constraint modeling with LLM-driven reasoning to produce explainable, constraint-aware decisions on pricing, ad allocation, and SKU management
- Developing an evaluation framework measuring decision quality through merchant outcome metrics, A/B-tested recommendation lift, and consistency audits, leveraging few-shot learning and Bayesian updating for reliable decisions from limited operational data

Amazon | Los Angeles, CA

AI Engineer Intern — *Jun 2025 – Aug 2025*

- Engineered an **enterprise-scale RAG LLM** system serving 15,000 tables across 200 schemas, reducing data discovery time by 75% with automated DDL generation and intelligent table/column comments using **AWS S3**, **Knowledge Bases**, **Redshift**, and **Bedrock**
- Pioneered version control for data infrastructure by implementing **git-style versioning** with **md5 hash**-based difference tracking, enabling rollback capabilities and reducing schema conflicts by 60%
- Built an automated notification system with **Slack API integration**, cutting manual monitoring overhead by 90% and providing real-time updates on daily table schema modifications

NYU Langone Health | New York, NY

Data Scientist — *May 2023 – May 2025*

- Implemented **deep learning computer vision** models using U-Net **convolutional neural network** architectures with transfer learning, achieving 15% improvement in image segmentation accuracy for medical imaging analysis of over 2,000 microscopy samples
- Designed an end-to-end **machine learning pipeline** for biomedical data preprocessing, reducing processing time from 8 hours to 45 minutes while maintaining over 90% data quality through **feature engineering** and **data augmentation** techniques
- Conducted pathway analysis to identify key biological pathways involved in glucose and lipid uptake, transport, and oxidation, offering insights for targeted therapeutic interventions and contributing to advancements in cardiovascular research
- Streamlined bioinformatics pipelines by composing 2,000 lines of **R** code, facilitating comprehensive analysis of datasets with over 300,000 records, improving processing from 10 hours to 8.5 hours and ensuring reproducibility and reusability
- Implemented **dimensionality reduction** and **spacio-temporal** modeling to bridge theoretical gaps with trajectory inference, elucidating the impact of CXCL12's interaction with CXCR4, ACKR3 receptors in maintaining CD8+ T Cell presence within tumors

Education

New York University Center for Data Science — **GPA:** 3.63/4.0

Master of Science in Data Science — *May 2026*

New York University Leonard N. Stern School of Business — **GPA:** 3.73/4.0

Bachelor of Science in Business — *May 2023*

Programming Languages: Python, R, Java, SQL, JavaScript, TypeScript, HTML, CSS, Bash

AI/ML Frameworks: NumPy, Pandas, SciPy, OpenCV, Hugging Face Transformers, spaCy, PyTorch, Keras, Scikit-Learn, XGBoost, Weights & Biases, MLflow, Apache Airflow, Raytune, LoRA, Ollama

Cloud Platforms: Amazon Web Services (AWS), AWS S3, AWS Bedrock, AWS Redshift, NYU HPC (SLURM), Vercel, Supabase

MLOps Tools: Git, Singularity/Docker, Next.js, React, Three.js, Tailwind CSS

Specializations: Natural Language Processing, Computer Vision, Deep Learning, Machine Learning, Time Series Analysis

Certifications: Machine Learning Engineering for Production (MLOps) Specialization (Deeplearning.AI)

Projects

Extending Non-Vacuous Generalization Bounds for LLMs with Adaptive Per-Layer Weight

- Diagnosed and resolved an 857x performance bottleneck in a PyTorch training pipeline on A100 GPUs by implementing tensor caching across projection classes, reducing per-experiment added training time from ~7 days to ~12 minutes
- Extended the SubLoRA framework for PAC-Bayes generalization bounds on GPT-2 by developing adaptive per-layer subspace allocation and deploying training/evaluation pipeline on HPC, achieving competitive non-vacuous bounds on OpenWebText

Curiosity – AI Chat Platform with Conversation Branching, Dialogue Trees, and Multi-Provider Support

- Built a full-stack AI chat application (Next.js, Supabase, Vercel) with multi-provider LLM support (OpenAI, Anthropic, Gemini, local Ollama) featuring OAuth integration and a persistent vector-embedding memory system for retrieval-augmented context
- Designed an interactive text-selection interface where users highlight any passage to spawn contextual conversation branches — with inline AI summaries, custom action shortcuts, and a navigable tree visualization for exploring nonlinear dialogue paths

Skills and Interests

- **Skills:** Mandarin (Proficient), Spanish (Elementary), Public Speaking
- **Interests:** Running, Photography, Overwatch (Top 500), Smash Bros., Boarding, Weight Training, Calisthenics, Travel, Street Food, Archery, Counter Strike, Travel, Whistling (Taking Requests)