Aneri Rana

anerirana2000@gmail.com | (718) 808-5622 | linkedin.com/in/aneri-rana | github.com/anerirana

EDUCATION

University of Massachusetts Amherst, USA **Sept 2022 - May 2024**
Master of Science in Computer Science with Data Science concentration 3.85/4 GPA
Courses: Advanced Machine Learning Theory, Natural Language Processing, Applied Statistics

Savitribai Phule Pune University, India **June 2014 – May 2018**
Bachelor of Engineering, Computer Science

TECHNICAL SKILLS

- Programming Languages: Python, R, SQL, MATLAB, C++, Java
- Data Science: PyTorch, TensorFlow, HuggingFace, Keras, Numpy, Pandas, scikit-learn, Apache Spark
- Tools and Technologies: Google Cloud Platform, Kubernetes, Docker, Tableau, Apache Airflow, Terraform, Ansible

PROFESSIONAL EXPERIENCE

Upstream Tech, Remote, United States
Machine Learning Engineer **Jun 2024 – Present**

- Deployed **production-ready ML models** trained on gauge data from 41 watersheds, supporting 15 organizations after evaluating 300+ experiments.
- Engineered a more robust short-term **LSTM** forecasting model by introducing hyperparameters for multi-source weather sampling, boosting model generalization and improving average **KGE** from **0.8 to 0.86**.
- Built an internal **Streamlit dashboard** to benchmark Upstream's hydro forecasting **performance metrics** against competitors, empowering sales and marketing teams.
- **Lead customer onboarding** by configuring production setup and delivering tailored hydro forecasting models, successfully onboarding 7 new organizations to date.

Goldman Sachs, Remote, United States
Graduate Student Researcher **Feb 2024 – May 2024**

- Built a **Retrieval-Augmented Generation** (RAG) system for legal credit agreement QA by integrating ColBERT-based retrieval with **Large Language Models** (LLMs), advancing **GenAI** capabilities for enterprise document intelligence.
- Pioneered a novel Self-Rewarding RAG framework that enabled **self-supervised training**, reducing costly data annotations by 70%.
- Improved model alignment efficiency by 30% by training models using **Direct Preference Optimization** (DPO), eliminating the need for separate reward models and reducing training complexity compared to RLHF.
- Researched 30+ **task-specific prompts** for question paraphrasing, reward generation, and document-grounded QA, resulting in more reliable and context-aware LLM outputs.

C3.ai, Redwood City, California
Data Science Intern **Jun 2023 – Aug 2023**

- Developed a robust logs error analysis framework for C3's core platform, reducing service interruptions by 15% and enhancing overall user experience.
- Identified and prioritized 30 critical bug fixes in the platform codebase through efficient analysis of error log summaries, utilizing regular expressions and clustering algorithms.
- Established an efficient pipeline integrating server metadata into error logs summary model for root cause analysis of server shutdowns/restarts, reducing bug discovery and debugging time from 1 day to 1 hour.

HSBC Technologies, Pune, India
Senior Software Engineer **Mar 2020 – Aug 2022**
Software Engineer **Aug 2018 – Feb 2020**

- Implemented a cloud-based batch training and inference pipeline using **Docker** containerization with **Apache Airflow** on **Google Kubernetes Engine** (GKE), enabling scalable and reliable ML deployment.
- Reduced model processing time by 50% by engineering a **distributed computing framework** with **Apache Spark** on **Hadoop YARN**, accelerating large-scale data processing and training tasks.

- Orchestrated the development of a plug-and-play **API infrastructure** on GKE with integrated documentation for production use, facilitating deployment of the first two ML models.
- Developed a multi-task **deep learning** model with **0.93 precision** and **0.96 recall** to auto-repair transaction fields, reducing 200K+ manual corrections monthly in the UK.
- Trained a **BiLSTM** model (**F1 score: 0.94**) for payment address classification, automating HSBC audit reporting via a Data Studio dashboard with probability-based confidence scoring.

PROJECTS

Common Sense Reasoning through Winograd Schema Challenge **Jan 2023**

University of Massachusetts Amherst, Natural Language Processing Project

- Researched state-of-the-art large language models for commonsense reasoning, with a focus on the Winograd Schema Challenge (WSC).
- Modeled WSC as a **text-generation** task and conducted zero/few-shot experiments using **Chain of Thought** (CoT) and **parameter efficient prefix tuning** on GPT-3.5 and Flan-T5 (small to XXL).
- Strategized few-shot and CoT prompting using cosine similarity with random sampling to select optimal in-context examples and achieved **91.22% accuracy** matching current SOTA (91.28%) on the Allen Institute Leaderboard.

Mars Spectrometry | Detect Evidence for Past Habitability **Jun 2022**

ML Challenge, NASA

- Used evolved gas analysis (EGA) mass spectrometry data from Curiosity Rover missions to predict chemical compositions that indicate past livable conditions on Mars.
- Ensembled various combinations of metric learning algorithms like LMNN, NCA, ITML, and classifiers such as Extra Trees and LightGBM to predict the composition of chemical compounds in the geological samples.
- **Ranked 18th globally** out of 713 participants with a log loss of 0.14 on the test data set.

Game AI **Nov 2023**

University of Massachusetts Amherst, Reinforcement Learning Project

- Applied Actor-Critic and Reinforce with Baseline Reinforcement Learning (RL) algorithms to train AI agents proficient in playing OpenAI Gymnasium's cart-pole, mountain car, and cliff walking games.
- Developed custom gaming environments, including 687-GridWorld and Snake game, for specialized RL training of AI agents.

COURSES AND CERTIFICATIONS

- **Computer Science Research Mentorship Program Scholar** [CSRMP 2023a] - Accepted to a mentorship program by **Google** in support of my pursuit of machine learning research pathways. (Feb 2023 to Jun 2023)
- Google Cloud Certified “Associate Cloud Engineer”. (Aug 2019)

VOLUNTEERING

- Developed new features and resolved user-reported bugs for **Keras**, **Snorkel**, and Avalanche **Open Source** Projects. (May 2021 – Mar 2022)
- Led website management and feature development using Python and Django as Super Volunteer and Web Maintainer for Women in Machine Learning (WiML), advancing its mission to support and amplify the impact of women in ML. (Oct 2021 – Aug 2022)
- Facilitated a workshop on “Leveraging Open-Source Tools for NLP” at the **International Conference for Machine Learning** (ICML, a top-tier ML conference) with research associates from the Georgia Institute of Technology and the University of Toronto. (Jul 2021)
- **Emceed** talk and keynote sessions for an entire day at the **Open Data Science Conference**. Introduced speakers, handled live broadcasting, and managed issues during the session. (Sept 2021)