

Rajasvi Vinayak Sharma

+1(858)319-9642 • rajasvi777@gmail.com • linkedin.com/in/rajasvi • rajasvi • rajasvi.github.io

EXPERIENCE

Moveworks (acquired by ServiceNow)

Jan 2025 - Present

Senior Machine Learning Engineer, Enterprise Search Relevance Team

Mountain View, CA

- **Post-trained** open-weight LLMs for listwise search reranking using **SFT and PEFT/LoRA** on multi-GPU FSDP, with a ListMLE + KL distillation objective and LLM-generated training labels over public and internal ranking datasets, lifting **NDCG from 0.65 to 0.70**.
- Shipped the fine-tuned reranker to production as the second-stage ranker, replacing a pointwise BGE cross-encoder and using compressed passage embeddings to make **listwise LLM reranking** feasible within the existing latency budget.
- Built the **first LLM evaluation framework** for search summarization serving **over 5,000 customers**, designing reference-less metrics for relevance, groundedness, citation validity, and multilingual fidelity, lifting summary relevance from **60% to 80%**.
- Established **ranking evaluation infrastructure** with labeled datasets for team-wide model and feature validation, and automated training data generation, reducing turnaround **10x** via **parallel LLM annotation of 500K query-document pairs**.
- Implemented a **listwise LLM annotator** for training and evaluation dataset creation, and designed a human-in-the-loop flow escalating ambiguous pairs to annotators to build **high-confidence ranking eval sets** with **60%** less human labeling effort.

Signos

Apr 2023 - Dec 2024

Machine Learning Engineer

Remote, CA

- Improved **post-meal glucose spike prediction accuracy from 80% to 87%** for **over 10,000 users** by running training-data ablations, correcting label noise, and engineering **continuous glucose monitoring (CGM)** derived and user behavior based features.
- Established **Latest Spike Time** as a predictor of weight loss through analysis across **500K weight logs and 200M CGM readings**, shipped as a user-facing feature, with the analysis contributing to research published in *Diabetes Technology and Therapeutics*.
- Increased **meal-logging engagement 12%** by shipping **real-time glucose event detection** that analyzes CGM curve shapes to prompt timely meal logging, feeding higher-quality training data into spike prediction models.
- Reduced **training data 10%** while retaining accuracy gains using cleanup and training techniques that scaled to a corpus of **over 1M free-text meal logs**; experimented with **meal embeddings** on the same corpus to enrich item representations.

Nvidia

Jun 2022 - Sep 2022

Data Scientist Intern

Santa Clara, CA

- Cut **alert response time from months to 1 week** by building an end-to-end **time-series anomaly detection tool** (Z-score thresholding, Kubeflow pipelines) to flag malicious activity across **over 1,000 categories and 10M gaming sessions**.
- Developed player **engagement metrics** via **regression analysis** to track progression and infer game completion, used as a feature in a **churn prediction model** and to target retention marketing.

Goldman Sachs

Jun 2018 - Aug 2021

Data Scientist, Analyst

Bengaluru, IN

- Fine-tuned **BERT, RoBERTa, and DistilBERT** models on **over 8M daily emails** to classify document zones (salutation, signature, disclaimer) and extract contact entities via a rule-based downstream pipeline, achieving **85.7% zone classification accuracy**.
- Built **XGBoost and LGBM** models to classify jurisdiction violations across **over 6M daily Bloomberg trader conversations** using semantic and temporal features, achieving **78% precision**.

EDUCATION

University of California - San Diego, California, US

Sep 2021 - Mar 2023

M.S. in Electrical and Computer Engineering (Major: Machine Learning and Data Science)

GPA: 3.91 / 4.0

Indian Institute of Technology (Banaras Hindu University) Varanasi, India

Jul 2014 - May 2018

B.Tech. in Electronics Engineering

GPA: 8.81 / 10

SKILLS

Languages: Python, Go, SQL

LLM Post-Training: Supervised Fine-Tuning (SFT), PEFT/LoRA, knowledge distillation, instruction tuning, synthetic data generation

LLM Evaluation: LLM-as-judge, reference-less metrics (groundedness, citation validity)

ML Systems: Information Retrieval (search, ranking, cross-encoders, bi-encoders), Recommendation Systems, NLP, Time-Series

Frameworks: PyTorch, Hugging Face (Transformers, PEFT), FSDP (Fully Sharded Data Parallel), scikit-learn, XGBoost

Infra / MLOps: Distributed ML training, Spark (PySpark, SparkSQL), Kafka, Kubeflow, MLflow, Weights & Biases, Databricks, AWS, GCP