Rajeshwari Sah

rajeshwari.sah@gmail.com **(858)-729-4048** **rajeshwari-sah** **Sunnyvale, CA**

Education

- **University Of California, San Diego** — San Diego
  Masters of Science in Computer Science — *2021 - 2023*
- **R.V College of Engineering, Bengaluru** — India
  Bachelors of Engineering in Computer Science — *2012 - 2016*

Experience

- **Lead Applied Scientist, Lifio.ai (a subsidiary of Fresh Gravity)** June 2023 – Present
  AI Agents for Clinical Trial Protocol Document
  - Orchestrated the product development of AI Agent system for **clinical trial protocol design**, balancing speed of trial execution, revenue impact, and regulatory compliance.
  - Fine-tuned LLMs with protocol and CTGov data, integrating RAG to generate protocol sections from summaries.
  - Designed a **retrieval and ranking model**, leveraging multi-task learning to optimize **criteria approval rates, compliance rejection rates, and protocol acceptance rates**.
  - Implemented ML model deployment workflows in **GCP Workflows and Vertex AI**, enabling seamless model training, deployment, and monitoring in production environments.
  - Achieved **$1M in revenue impact**, **60%+ adoption within three months**, and **15% reduction in protocol amendments**.

  Extraction of Entities from PubMed Publications and Clinical Documents
  - Built a team of 3 engineers and data scientists for designing, implementing, and optimizing **LLM architectures**, resulting in a **15% increase in model performance**.
  - Developed novel techniques for fine-tuning LLMs(**medAlpaca & LLama**) on domain-specific tasks improving task-specific accuracy by 25% .
  - Designed automated CI/CD pipelines for ML models using **Kubernetes and Airflow**, improving deployment efficiency.

  Advanced Clinical Document Summarization using Domain-Adapted Transformers
  - Spearheaded a multidisciplinary team of NLP experts, medical professionals, and data engineers to create a highly specialized clinical document summarization system.
  - Designed an architecture leveraging BERT, **ClinicalBERT**, and **OCR** pipelines to process and summarize scanned clinical documents.
  - Fine-tuned on a mixture of medical literature and electronic health records, achieved an average summarization **accuracy improvement of 20%**.

  Tax Collector and Assessor Document Processing Using RAG and LLaMA Models
  - **Led a team of 4 engineers** to design and architect a pipeline for inferring column structural parsing of unstructured documents, improving entity mapping accuracy for database.
  - Fine-tuned LLaMA models on domain-specific tasks, improving SQL-like query generation accuracy for handling varied document formats.
  - Leveraged Retrieval-Augmented Generation (**RAG**) with LLaMA models to enhance document processing, **boosting retrieval efficiency by 20%.**
  - Implemented **MLflow** for experiment tracking, **hyperparameter tuning**, and **model versioning, streamlining model development and deployment.**
- **Senior Software Engineer, Visa** July 2016 - Feb 2021
  Fraud Detection on Transactions
  - Led efforts to enhance Visa's fraud detection system using time series analysis and graph analytics.
  - Implemented anomaly detection algorithms and improved feature engineering, resulting in a 8% reduction in false positives and enhanced customer experience.

Data localisation

- Led data India's largest data migration for localising payment transactional data.
- Coordinated with 10+ teams to identify data sources, pipelines and tables to be migrated.
- Created a robust, fault tolerant and low latency pipeline for migration of data capable of handling streaming data of 5TB per day.

- **Applied Science Intern, Amazon** Aug 2022 - Dec 2022

Hierarchical Entity Matching with Topic Classification for Amazon Catalog

- Designed an unified product representation for existing products on Amazon using their product description, images, videos.
- Proposed a novel hierarchical structure using Amazon's product catalog taxonomy. Generated content and concept embeddings using using **cross modal contrastive learning**.
- Re-categorized 500M products into existing taxonomy, improving classification precision by **12%**.
- Utilized **PySpark on AWS EMR** for distributed data processing and feature engineering, improving efficiency in handling large-scale datasets.
- Deployed models on **AWS SageMaker** with real-time inference endpoints, integrating with Redshift for large-scale data storage and querying.

- **ML Researcher, UCSD-Teradata** July 2021 - Feb 2022

A Query-Aware Database Tuning System with Deep Reinforcement Learning

- Architected a module for SQL queries vectorization considering query type, table, operations and cost information.
- Used the the **actor-critic networks** to find optimal DB configs.
- Worked with RL based **Deep deterministic policy gradient (DDPG) model** for index selection, reducing the query run time by **15%**.

Skills

- **Languages:** Python, Go, NodeJs, Java, C, C++, Scala, Bash
- **Databases:** BigTable, Spanner, ScyllaDb, DynamoDb, Cassandra, Redis, Memcache, PostGres, MySQL, ElasticSearch, Vector Db, ChromaDB, Pine
- **Cloud:** Amazon Web Services (AWS),Google Cloud Platform (GCP) and Databricks
- **Tools:** Google Cloud Dataflow (Apache Beam), Apache Flink, Google PubSub, Kafka, Airflow, Kubernetes (GKE), Docker, Jenkins, Spinnaker, Prometheus, Grafana, GCS, S3, SQS, Kinesis
- **ML Related Tools:** Tensorflow, Pytorch, Keras, Distributed Training, Vertex AI Platform, MLFlow, Kubeflow, LangChain, LLamaIndex, ChainLit , Tfserving, various data processing libraries for analysis like numpy, pandas, seaborn, matplotlib etcs
- **Modeling strategies:** LLMs, AI Agents, LLAMA, Alpaca, Mistal, MedAlpaca, ClinicalGPT, RoBerTa, RAGs, FFM, Wide and deep ranking, learning to Rank, Multi-task learning, Two Tower models, Multi-Model Mixture of Experts, Transformers, Natural Language Processing (NLP), OCR Models, Deep Learning, Scikitlearn, SQL, Apache Spark, Pyspark

Research Projects

- *DNN, DCN and DeepFM techniques for Movie & Recipe rating prediction* *Feb 2022*
  - Implemented these approach on MovieLens and Food.com data.
  - Improved the RMSE by 10% over baselines with best achieved by DCN.
  - Selected in **top 5 in Kaggle competition**.
- *News Articles Recommendation using Neural Contextual Bandits (NeuralUCB)* *Nov 2021*
  - Implemented **NeuralUCB** to predict reward and upper confidence bounds computed from the network to guide exploration.
  - Extended the algorithm to include piecewise stationary reward functions to capture time variance of the news articles.
  - Improved **NDCG@5 by 4.8%** over the Linear UCB algorithm.

Publications

- Surveillance video based robust detection and notification of real time suspicious activities in indoor scenarios *CCSEIT, May 2016.*
- Approach and Analysis of Machine Learning Techniques for Crime Classification and Prediction *Workshop Proceedings of ICDM, July 2017.*