RAKSHITH MATHAD

(213) 272-7811 | rakshith1262k@gmail.com | LinkedIn | GitHub | New York City

EDUCATION

University of Southern California, Los Angeles, California - Master of Science in Applied Data Science GPA: 3.7/4 AUG 2022-MAY 2024

Coursework Highlights: Machine Learning for Data Science and AI, Applications of Data Mining, Predictive Analytics, Fairness, Security and Privacy in AI

KLE Technological University, India - Bachelor of Engineering in Computer Science GPA: 3.95/4 AUG 2018-JUN 2022

Coursework Highlights: Data Structures and Algorithms, Data Mining, Machine Learning, Distributed & High-Performance Computing, Cloud Computing

SKILLS

Languages Python, SQL, C, C++, CUDA (Intermediate)

Tools & Technologies Analytics, Deep Learning, Machine Learning, A/B Testing, GCP Vertex AI, AWS, Docker, Git, Informatica ETL, Web Scraping/Automation, AI Research, Parallel Programming, Numpy, Pandas, Scikit learn, LangChain, Nvidia DGX Server, Nsight Profiling, Nvidia NeMo, LLMOps, RAG, VectorDBs, LLM Training and Inference techniques, HuggingFace, FastAPI, RESTful API, Responsible AI, Linux, MongoDB, Selenium, Hadoop HDFS, PySpark, PowerBI/Tableau, LLMs, Generative AI, MapReduce, Azure, Google Cloud Platform, BigQuery, HiveQL, Horovod, ArcGIS, MLOps, Jenkins CI/CD, NLP, Forecasting, Unsupervised ML, Generative Models

Certifications NVIDIA CUDA Computing ,Juniper Networks Certified Associate, TensorFlow/PyTorch for Artificial Intelligence, Agile, TCP/IP

EXPERIENCE

CVS Health - Software Engineer - Generative AI, New York City, NY JUN 2024- PRESENT

- As a **generalist AI Engineer,** I am working in the **Conversational AI** customer service team to build a large-scale complex **RAG and Rule based** FastAPI chat application on **GCP Google AI Platform**. **My contributions to the project has an impact of ~ $6 MM/yr and I received a company-wide award for high and vibrancy impact in the company**. Worked with **OpenAI GPT PTUs and Gemini LLMs,** employed **Feature Store** and Matching Engine to build **async/multithreaded/parallel** semantic search based information retrieval and answer generation modules, resulting in improved latency and caching strategy. Helped in profiling and **scaling this production AI system** to **~20,000 req/hr impacting around 300k customers everyday** leveraging **Kubernetes GKE** and **Vertex AI**.
- Performed complex data ingestion and preprocessing on GCP **BigQuery** and **BigTable**, including chunking and embedding strategies. Took part in design reviews, building chat **Agents**, improving semantic search by **fine-tuning** the embedding model service with bi-encoders etc.
- Also worked with Airflow, **K8s**, Jenkins CI/CD**, LLM Evaluation and regression framework**, scraping/automation, feature flags, On call rotations with prod deployments, advanced prompt engineering, load testing, AI safety and guardrails etc. **I constantly produced multiple impactful PRs every week, exceeding expectations.**

AEG Entertainment Group - Data Engineer Intern, Los Angeles, CA FEB 2024-MAY 2024

- Built complex **Azure** Data Factory pipelines for large data processing, used **PySpark** in **Databricks** for batch API processing and Dynamics CRM data migration via OData REST APIs. Handled Parquet/Avro for streaming and financial data reporting.

CVS Health - Analytics Engineering Intern, New York City, NY MAY 2023- AUG 2023

- Optimized Hadoop HDFS big data pipelines, processing **100M+** rows with HiveQL, cutting latency by 40%. Designed scalable ETL workflows and **boosted customer campaign growth** strategy leveraging **Tableau**, advanced OLAP, data blending, and predictive modeling.

AlphaICs Corporation - AI Software Engineer Intern, India JAN 2022-JUL 2022

- Developed deep learning applications for Object Detection, Recurrent Models for Visual Attention, Recommendation Systems (Matrix Factorization, Collaborative Filtering) etc for faster **AI Inference** on a custom **AI processor**.

Samsung R&D Institute -AI Research Intern, India NOV 2020-JUN 2021

- Led a team of 3, and worked on the SOTA **Generative AI PyTorch** research Implementations for the task **of Semantic Image Editing using Conditional GAN** with spatially adaptive normalization, for Image manipulation with controlled generative-fill. Performed Semantic Segmentation with DeepLab-V2 to curate a 20k-image dataset and trained generative models on an **Nvidia DGX** cluster.

PROJECTS

STUDY AND IMPLEMENTATIONS OF PARALLELISM FOR LLM INFERENCE ON GPUs

- Deployed and served **BERT Transformer** on an RTX GPU in **PyTorch**, **ONNX**, **NVIDIA TensorRT** runtimes and profiled it to study performance bottlenecks using **NVIDIA Nsight**. Building a small CUDA Kernel to infuse into Pytorch for potential optimized inference on a custom NN.
- Used **Ray Serve** to deploy a async **TinyLlama** on a local GPU using FastAPI. Working on FSDP/DDP training simulations using Ray/DeepSpeed.

BETTER LLM (*LLAMA 2*) WITH RETRIEVAL AUGMENTED GENERATION USING GUARDRAILS

- Utilized **LangChain** and **Llama 2** to build a RAG-based chatbot with a custom database integration, powered by **Pinecone** VectorDB and HuggingFace Sentence Transformer. Integrated **NVIDIA NeMo Guardrails** and LoRA LLM finetuning techniques. Exploring an e2e pipeline to LoRa finetune, quantization training, and improve inference using NVIDIA Triton Inference Server.

RESEARCH EXPERIENCE

- **Published** and authored paper, "*Performance Analysis of Distributed Deep Learning using Horovod on Image Classification*", at **IEEE ICICCS**
- **Machine Learning Engineer (Information Sciences Institue (ISI) – Los Angeles, CA) –** Developed algorithms to leverage ML privacy-preserving techniques to detect anomalies infused by adversary client nodes in a **Federated Learning** system for **Homomorphic Encryption.**