SHREYAS KULKARNI

+1-669-977-9227 | skulkarni.sk.18@gmail.com | LinkedIn | Github | Publication | Certifications

Education

San Jose State University Aug. 2023 – May 2025
Master of Science in Computer Engineering *San Jose, CA*
Coursework: Network Security, Machine Learning, Data Mining, Deep Learning, Recommendation Systems

Technical Skills

Languages: Python, C/C++, SQL (Postgres), Shell Scripting, MATLAB
Frameworks: Pandas, Numpy, Scikit-learn, PyTorch, Scipy, Flask, FastAPI, PySpark, Kafka
ML & GenAI: AutoML, AWS SageMaker, MLFlow, AWS SageMaker, CodePipelines, LangChain, OpenAI, Llama 3.1, HuggingFace, Milvus, FAISS, Linear Algebra, Probability, Foundation Models
DevOps Tools: Git, GitLab, Docker, Kubernetes, Prometheus, Jenkins, Github Actions

Experience

Deep Learning Project Assistant Aug. 2024 – Present
San Jose State University *San Jose, CA*

- Utilizing Semantic Textual Similarity (STS), zero-shot (all-MiniLM-L6-v2) and pre-trained **generative models** (BERT, Llama 3.1) to build AI-driven recruitment platform for candidate-job matching
- Optimizing low-latency inference models deployed on AWS SageMaker, integrating Milvus vector database with FastAPI and AWS **Elastic Kubernetes Services**, achieving sub-100ms response time
- Integrating real-time data streams using **Kafka** and EventHub to enable asynchronous processing and efficient data ingestion and extracted structured data using LangChain and OpenAI

MLOps Engineer Aug. 2021 – Jul. 2023
PricewaterhouseCoopers *Bengaluru, India*

- Built a high-performance FastAPI backend for real-time financial document summarization using **LangChain** and GPT-3, delivering a ROUGE-2 score of 0.80 and processing over 1,000 documents simultaneously with 99% uptime
- Collaborated on end-to-end **MLOps pipeline** employing MLflow for model tracking and versioning, trained models (XGBoost, LSTM) for time series forecasting, enhancing prediction accuracy by 20% with 2x retraining speed
- Managed distributed data workflows with **PySpark** and MLflow retraining pipelines, accelerating real-time decision-making by 15% for global sales forecasts, maintaining scalability and fault-tolerant architecture
- Developed a centralized platform using Linux Server and PostgreSQL to unify data from Jira and Azure DevOps, reducing data hunting time by 50% and implemented CI/CD pipeline for automated deployment
- Containerized a multi-client product on Azure using **Docker** to streamline deployment, and monitor KPI, by automating REST API data ingestion, and orchestrating workflows with Apache **Airflow** DAGs
- Implemented application monitoring using **Prometheus** and **Grafana**, creating dashboards to track health and performance metrics and enabling proactive issue identification

Projects

Brain MRI classifier | *Link* Jan. 2025 – Feb. 2025

- Developed an end-to-end deep learning pipeline for multi-class brain MRI image classification, achieving an accuracy of **97%** using VGGNet and transfer learning along with MLOps tools like DVC and **GitHub Actions**
- Deployed the model on AWS using Docker and CI/CD pipelines, enabling scalable real-time image classification with seamless data access from AWS S3, processing over 7,000 MRI images efficiently.

Waymo 3D semantic segmentation | *Link* Aug. 2024 - Dec 2024

- Achieved a final mIoU of **0.70** and mAcc of **81%** on 1TB Waymo Open Dataset by optimizing the PointceptV3 semantic segmentation **transformer** for large-scale LiDAR data, enhancing performance across 23 class labels
- Improved mIoU by **22%** on ScanNetPP through fine-tuning, combining ScanNetPP training with Waymo dataset and utilized the high-performance computing units and managed GPU nodes for optimized performance

Nvidia ODSC hackathon Oct. 2024

- Built a data pre-processing pipeline with Nvidia's NeMo Curator to clean and organize legal Q&A data from the Law-Stack Exchange dataset, enhanced data quality and reached 87% accuracy in fine-tuned label generation
- Fine-tuned a large language model (BERT) with **LoRA** adapters leveraging Nvidia's NeMo framework and optimized hyperparameters to enhance legal tag prediction accuracy by 40% compared to baseline model