(+91)9836316925
davechirav@gmail.com

CHIRAV DAVE

Pune, India

SUMMARY

Lead AI/ML engineer and system architect focused on designing and deploying scalable LLM solutions for real-world impact. Skilled in translating research into production-ready systems with high performance and reliability. Passionate about making AI open, safe, and widely accessible through responsible innovation. I've created a local-first, open-source LLM platform with full backend/frontend integration.

EDUCATION

MS, CS, Arizona State University | GPA 3.92 *August 2017 - December 2019*

BTech, CS, Maulana Abul Kalam Azad University of Technology | GPA 8.67/10 *July 2012 - July 2016*

TECHNICAL SKILLS & COURSEWORK

Languages:	Python, GOLANG, C++, SQL, Bootstrap
ML Technologies:	Pytorch, Huggingface, vLLM, Numpy, TensorFlow, OpenCV, Pandas, Scikit-learn, Matplotlib
Cloud	DevOps: Docker, Kubernetes, Helm, AWS
Databases:	MySQL, DynamoDB, OCI NoSQL, PostgreSQL

WORK EXPERIENCE

AI/ML Lead, Full-Stack Backend Developer | A10Networks - Incedo | Pune *June 2023*

- Leading the R&D team at A10Networks to drive the company's AI vision by creating innovative solutions for building and deploying LLM guardrails to safeguard applications against Cyber attacks.
- Designing a containerized solution for seamless deployment of our guardrails on any on-prem Kubernetes cluster using Helm.
- Designed and implemented a highly scalable system capable of retrieving up to 10K attacked IP addresses from a data warehouse containing 65M entries in as little as 13 secs. The system allows users to apply filters based on attack categories such as botnets, reflectors, and command and control (C2) servers, allowing for the targeted retrieval of relevant data.
- Implemented an enhanced rate-limiting solution in Golang to efficiently manage incoming requests. By processing a certain number of requests and queuing the remaining ones, the solution resulted in significant improvements over the previous Flask implementation, including a 45% reduction in CPU usage, a 38% reduction in memory usage, and a 70% improvement in response time for searching attack entries based on a specific IP address.

Technology Stack: Python, GOLANG, Flask, FastAPI, PyTorch, vLLM, DynamoDB, Docker, Helm, Kubernetes

Research Engineer, Lead | DiDi Labs | California *Nov 2021 - April 2023*

- Architected & Implemented a Graph-Based Neural Network to forecast the heading direction of pedestrians within a scene to aid our car's decision-making process.
- Boosted the efficiency of the model from 65% to 72%.
- Optimize and deploy models using techniques such as model compression, pruning, and quantization to improve performance and reduce computational cost.
- Constructed different metrics to measure the accuracy of our algorithms.

Technology Stack: C++, Python, PyTorch

Backend Software Developer | AmazonGO | Washington *Feb 2020 - Sept 2021*

- Designed Java APIs based on SQS to receive requests from clients and utilize vision-based algorithms for predicting the ultimate shopping events.
- Collaborate with data scientists and software engineers to take machine learning models from research and development to production.

Technology Stack: Python, Java, Numpy, Scikit-learn, OpenCV, SQS, Guice

OPEN SOURCE PROJECTS

GPTs from Scratch | Github

- Built a GPT-style model in PyTorch from scratch inspired by Llama and Mistral.
- Codebase designed for modularity and extensibility to support experimentation and learning principles.

LLMs for Everyone – Fully Offline | Github

- Developed an open-source platform that enables easy offline use of Hugging Face models with integrated back-end and front-end.
- Simplified running LLMs locally without the internet - no cloud or API dependency.
- Delivered a seamless LLM chat and completions playground.