DHRUVIL BHATT

Seattle, WA • bhattdb@uci.edu • (949) 231-9789 • LinkedIn • www.dhruvilbhatt.com

EDUCATION

University of California, Irvine (Irvine, CA) September 2022 – December 2023

Master of Computer Science **GPA: 3.95/4.0**

Coursework: Advanced Algorithms, Artificial Intelligence, Parallel and Distributed Computing, Computer Security

DA-IICT (Gandhinagar, India) August 2018 – May 2022

Bachelor of Technology in Information and Communication Technology **GPA: 3.8/4.0**

Coursework: Data Structures, Databases, Networking, Machine Learning, HPC, Computer Architecture, Data Analysis

TECHNICAL SKILLS

AWS (SageMaker, Bedrock, EKS, ECS, EC2, S3, IAM), Azure, **vLLM**, TensorFlow, Hugging Face, **Distributed computing/HPC** (MPI, OpenMP, Ray), **Inference optimization**, Python, C++, Java, JavaScript, SQL/NoSQL, Docker

EXPERIENCE

Amazon Web Services - Bedrock/Sagemaker (USA) October 2024 – Present

Software Engineer (Machine Learning)

- Designed a **version-aware extension framework for vLLM** that builds on its plugin architecture, enabling **runtime-selective**, version-guarded patches across 15+ LLM families on AWS Bedrock (Llama, DeepSeek, etc). This work was published on vLLM’s official engineering blog and featured on their LinkedIn page.
- Architected a **unified container build and validation pipeline for Bedrock and SageMaker** LMI, using EKS-based ephemeral environments to spin up and tear down endpoints for integration, chaos, and regression tests—**reducing container count by 30%** for AWS Bedrock and simplifying maintenance.
- Extended the system with SageMaker-specific compatibility testing, automating Inference Component provisioning and validating LMI containers against runtime features (**sticky routing, Multi-LoRA, orchestration**) for full parity between EKS and SageMaker environments.
- Built **cross-region LLM benchmarking infrastructure** that reduced latency bias by 30% and compute cost by 40% through ECS-based resource co-location and adaptive on-demand provisioning.

QBurst (USA) February 2024 – October 2024

Software Developer Intern

- Developed a scalable, **cloud-based grocery store portal** using React, TypeScript, and Redux.
- Improved UI refinement through optimized API calls, enhancing interactions with MongoDB and .NET backend for **30% performance improvement**.
- Managed code deployment through **Azure DevOps**, utilizing pipelines for automation and SonarCloud for quality checks to streamline development and ensure code integrity.

Synaptics Incorporated (USA) June 2023 – September 2023

Software Developer Intern

- Engineered an **Azure-based** calling application, integrating Microsoft Graph API v1.0, MSAL (Microsoft Authentication Library), Graph APIs, and MS Teams with Synaptics’ headsets for audio testing.
- **Bridged the gap between business and technology**, and integrated speech recognition (web services) feature into the Azure calling application for generating speech-to-text transcripts for headset debugging.
- Created a software application on host side to parse UART (universal asynchronous receiver transmitter) data formatted in binary HCI format. Implemented extraction of debug messages from received data.

DA-IICT Research Lab (India) January 2022 – June 2022

Machine Learning Intern (Researcher)

- Led a team of 3 researchers in curating the **largest open-source Corporate Credit Rating dataset using ETL data pipeline**, consisting of 7805 data points, 4 times larger than the previous largest dataset (Dataset Link).
- Devised a set of time-independent strategies (using **Explainable AI/ML techniques**) based on financial ratios to **help corporate firms attain investment grade rating** with a mean precision value of **95%.**
- Employed the GraphViz package in Python to visualize the Decision Tree model (research paper listed on **SSRN's Top 10 download list**).

Institute for Plasma Research (India) October 2020 – August 2021

Distributed and Parallel Computing Intern (Researcher)

- Designed an efficient serial algorithm code in C++ for generating synthetic images of plasma **using ray tracing data analysis techniques**, and linear programming (for solving quartic equations).
- Collaborated with a fellow researcher to **model data for a 3D pinhole camera setup** and visualize it using **data labeling/annotation techniques.**
- Parallelized the serial algorithm using OpenMP, MPI, and High Performance/Distributed Computing (HPC) techniques, achieving a **2100%** speedup, reducing the synthetic image creation time to **less than 0.65 seconds**

PROJECTS

Mac Terminal-Embedded Portfolio Website | *React.js, Node.js, JavaScript, CSS* | (Link to portfolio)

- Crafted a **fully interactive and responsive Mac terminal-embedded website**, enabling users to explore my entire professional journey using the shell.
- Garnered traction from over **10,000 developers** globally for showcasing the portfolio.