Abhinav Avasarala

(984) 329 4539 — avasarala958@gmail.com — Linkedin: Abhinav Avasarala — github: Abhinav-Avasarala

Education

North Carolina State University (NCSU) **May 2027 (Expected)**

B.S. Computer Engineering *Raleigh, NC*

- **GPA:** 4.00/4.00
- **Courses**: Software Development Fundamentals, Data Structures & Algorithms, C & Software Tools, Statistics for Engineers, Multivariable Calculus, Operating Systems, Signals & Circuits, Machine Learning, Fundamentals of Logic Design

Technical Skills

Languages: Java, Python, JavaScript, C, C++, CUDA, SQL, Verilog, x86 Assembly
Cloud & DevOps: Goole Cloud Engineering Certificate, Google Cloud Platform (GKE, Cloud Run, VMs, Cloud SQL, Load Balancing, IAM, Functions), Docker, Terraform
Frameworks: ReactJS, FastAPI, Node.js + Express, Flask, CUDA, PyTorch
Databases & Tools: PostgreSQL, Pandas, NumPy, Git, Axios, REST APIs

Experience

iQuadra Information Services | *ReactJS, Axios, REST, Figma* **May 2025 – August 2025**

Software Engineer Intern *Remote*

- Developed a bug-tracking system using ReactJS front-end, Node + Express back-end, PostgreSQL database, and Axios integration
- Designed **12+ RESTful API endpoints** using Node + Express with PostgreSQL for persistence
- Improved bug resolution time by **20%** through priority-tagging and streamlined workflow
- Led a team of 3 front-end devs to build responsive UIs from Figma designs using React

North Carolina State University | *CUDA, cuBLAS, CUTLASS, PyTorch, C++* **Aug 2025 - Present**

Research Intern *Raleigh, NC*

- Working on accelerating a **16-bit diffusion** model by pushing inference to low-precision formats such as **INT8 and FP8 using NVIDIA CUDA, cuBLAS GEMMEx, and CUTLASS**, targeting lower memory footprint and latency without degrading output quality.
- Designed structured experiments to study precision bottlenecks, learning **GPU architecture and CUDA** from first principles and building prototype kernels and **PyTorch** extensions to evaluate multiple **quantization** paths.
- Implemented symmetric INT8 quantization for smaller networks, benchmarked FP8 variants, and proposed several scalable approaches for bringing these techniques to the full diffusion model.

Projects

Stock Sentiment Analyzer | *Python, FastAPI, FinBERT, HuggingFace*

- Built an **end-to-end ML pipeline** where users input portfolios and receive real-time news sentiment on stocks
- Combined relevant news from Reddit, Yahoo Finance, NewsAPI and applied **NLP sentiment classification**
- Compared FinBERT and Vader, selecting FinBERT for **92% test accuracy**

Mood-Based Music Recommendation App | *ReactJS, Node.js, Express, NLP Libraries, OpenAI API*

- Implemented a **full-stack web application** to recommend songs based on user input's sentiment analysis
- Integrated **Spotify API** to deliver 5 real-time personalized songs per query
- Utilized React for the frontend and Node.js with Express for the backend to handle **RESTful API integration**, and implemented secure user authentication and session management using **PostgreSQL** and **bcrypt hashing**
- Achieved **90%+ mood detection accuracy** using custom NLP logic

AI Wellness Copilot | *Chrome Extension (MV3), React, Express, Supabase, Pinecone, GPT-4.0-mini*

- Built a Chrome extension + Supabase + RAG backend in **36 hours**, shipping an AI wellness assistant that analyzes browsing activity and converts it into personalized focus insights.
- Indexed **5,000+ browsing events** using Pinecone embeddings to surface passive scrolling vs. deep-focus periods with **40% higher accuracy** than baseline heuristics.
- Led a team of three to build **3 interactive dashboards** highlighting focus patterns, break trends, and high-quality work sessions.

Leadership / Extracurricular

- **IBM-NCSU Pathfinder program:** Learned about various IBM & Red Hat technologies like Ansible & watsonx
- **180 Degrees Consulting:** Led a team of 4 student consultants to improve a non-profit client's marketing strategy