Varun Venna

973-517-0645 | vvenna@g.ucla.edu | linkedin.com/in/varun-venna | varunv22.github.io/

Education

University of California, Los Angeles Los Angeles, CA
Bachelor of Science in Computer Science *Sep. 2022 – June 2026*

- Relevant Coursework: Data Structures and Algorithms, Software Construction, Computer Organization, Computer Systems Security, Computer Systems Architecture, Design of Digital Systems, Operating Systems

Experience

Software Engineer/Co-Founder June 2024 – Present
Swyft *Los Angeles, CA*

- Developed a mobile app for UCLA students to order groceries using Expo, React Native, and WooCommerce API
- Integrated user accounts with WooCommerce for seamless order tracking and real-time inventory updates
- Implemented payment gateways and optimized API calls with local caching for faster load times
- Implemented push notifications to alert users about order status updates, ensuring timely communication throughout the delivery process

Software Engineering Intern March 2024 – July 2024
Brev.dev (Acquired by NVIDIA) *San Francisco, CA*

- Tested and fixed over 30 Jupyter Notebooks for LLMs and multi-modal models including Mistral 7B, NVIDIA's NeMo Framework, TensorRT-LLM, LLaVa, and the StreamingLLM framework
- Upgraded the CLI by doing a large overhaul and strengthening/removing commands as necessary
- Integrated a cloud service named Crusoe into the console through multiple functions that allow users to create, delete, stop, start, and retrieve information about an instance
- Created a Golang script capable of creating and terminating Windows/Linux instances via Microsoft Azure

Software Engineering Intern Oct. 2023 – Feb. 2024
Avolta *Toronto, ON*

- Established and maintained routing framework with Flask and SQL for a mobile vehicle monitoring app while enhancing security with a bcrypt hashing function and parametrized queries
- Collaborated in a team to develop functionality and security for password change/reset, data accesses, login, register user, etc.

Projects

AI Coding Assistant | *LangChain, OpenAI, Python, AstraDB, GitHub API*

- Built a custom AI agent for answering detailed queries about issues in Github repositories using Retrieval-Augmented Generation (RAG) and LangChain
- Integrated AstraDB for vector store management, leveraging OpenAI embeddings to retrieve GitHub issues
- Implemented a tool to fetch and process GitHub issues, enhancing coding productivity with AI-powered insights
- Implemented note-taking functionality within the AI agent to track and store user queries and responses for improved task management and follow-up

ShoulderMe | *Express, HTML/CSS, React Native, Expo, MongoDB, Node, Postman*

- Implemented an AI therapist using the Gemini API, alongside a calendar feature that enables users to monitor their moods over the month, functioning as a diary with both text and audio entries
- Developed a mental health mobile/web app using the MERN stack, Expo, and React Native
- Designed a matching algorithm using clustering to match users with peers who are compatible with their interests

Technical Skills

Languages: Java, Python, C/C++, SQL , JavaScript, HTML/CSS, GoLang, Lisp
Technologies: React, Node.js, Flask, Azure, AWS, Crusoe, VastAI, Postman, Swagger, Docker, MongoDB, Redux, Shell, Git, OpenCV, Triton, Expo, CUDA
Libraries: Pandas, NumPy, Matplotlib, TensorFlow, TensorRT, Ollama, OpenAI, LLaVa