DILLAN KHURANA

linkedin.com/in/dillan-khurana

EDUCATION

University of Virginia *Expected May 2026*
Bachelor of Science, Computer Science *3.7/4.0 GPA*
Bachelors of Arts, Mathematics

- Dean's List - School of Engineering and Applied Sciences
- Coursework: Machine Learning, Computer Vision, NLP, Reinforcement Learning, Advanced Software Dev, Advanced Algorithms (DSA 3), Computer Systems/Org 2, Cyber Security, Survey of Algebra, Partial Differential Equations, Probability

EXPERIENCE

Amazon Web Services (AWS) Arlington, VA
Software Development Engineer Intern May 2025 - Aug 2025

- Developed an **MCP server** for **Amazon Q CLI** using **FastMCP**, **Docker**, and **PostgreSQL** to deliver developer-specific code review feedback derived from past review history and comments, reducing average team review cycles by **43%**.
- Created **agentic tools**, **resources**, and **prompts** within the MCP server to access the local file system, integrate with internal code review APIs, and persist common issue watchlists in PostgreSQL.
- Designed and implemented an admin dashboard with **React** and **Python REST APIs** to manage regional build notifications for **22,000+** SNS subscribers, saving **600+ hours** annually by eliminating manual code changes and redeployments for notifications.
- Engineered and deployed backend infrastructure using **AWS CDK (IaC)**, including **DynamoDB**, **S3**, **API Gateway**, **Lambda**, and **IAM** to support auditing, template storage, and API functionality, establishing a foundation to add future notification and auditing features.

Alarm.com Tysons, VA
Software Engineer Intern Jun 2024 - Aug 2024

- Engineered an automated testing platform to benchmark Multimodal **Large Language Models** (LLMs), replacing manual evaluation and doubling testing speed.
- Devised a data-driven approach to validating LLM outputs using **k-means clustering**, OpenAI embeddings, and cosine similarity that quantified model performance on image recognition tasks.
- Utilized **Azure Blob Storage**, **ASP.NET**, and **C#** to build an admin dashboard and pipeline for collecting, pre-labeling, validating, and storing image snapshots, reducing dataset procurement time by **83%.**
- Refactored the OpenAI integration service to allow for custom model selection, message configurations, and role selection while maintaining data privacy requirements.

10Pearls Tysons, VA
Software Engineer Intern Jun 2023 - Aug 2023

- Developed an LLM-powered sentiment analysis application to flag high-risk social media posts, saving pharmaceutical clients **$100,000**+ per incident. Fine-tuned LLM using **PyTorch, HuggingFace Transformers/PEFT,** and **QLoRA** to increase accuracy by **34%**.
- Implemented concurrent request handling in the sentiment analysis pipeline, leveraging parallel processing to eliminate inference bottlenecks and achieve a **5x** increase in throughput.
- Built an admin dashboard using **FastAPI**, **React.js**, and **MySQL** to enable clients to generate and visualize sentiment analysis reports, eliminating the need for command-line interactions.
- Constructed a **CI/CD** pipeline with end-to-end (E2E) testing using **GitHub Actions**, **pytest**, **Docker**, and **Selenium**, ensuring robust application reliability and security for clients.

PROJECTS

DillAgent | *https://github.com/buzzoo123/dillagent* *Jan 2024 - Present*

- Built DillAgent, an open-source framework for LLM-powered applications that provides abstractions including LLMs, tools, agents, and agent executors to enable flexible AI workflows and custom logic integration. Custom alternative to LangChain.
- Implemented modular **agentic workflows** including **Retrieval Augmented Generation (RAG)**, agent planning, tool execution, and evaluation using asynchronous dependency-driven **execution graphs**.
- Fine-tuned and quantized local LLMs using **Llama CPP** and **PyTorch** to adhere to Open AI function calling conventions, allowing 4-bit quantized 8b+ parameter LLMs to integrate with the framework.

Python, Llama CPP, PyTorch, HuggingFace, Vector Stores, Milvus, Numpy, Pandas, Sklearn, LangChain, LangGraph, GenAI

Demon Time | *https://www.demontime.top/* *Dec 2024 - Present*

- Developed a full-stack open-source platform for technical interview preparation that enables users to follow community-created roadmaps and utilize spaced-repetition techniques.
- Designed the platform using a **React TypeScript** and **Tailwind CSS** frontend, a **FastAPI GraphQL** backend with browser cookie authentication, and **PostgreSQL** with **SQLAlchemy** for database management. Containerized backend services using **Docker.**
- Integrated the **Anthropic** API and **one-shot prompting** techniques to facilitate real-time problem-solving conversations, helping users master topics through interactive sessions.

Typescript, Python, React, RTKQuery, Tailwind, FastAPI, Postgres, SQLAlchemy, Docker, AWS, Nginx, CloudFront, GraphQL

TECHNICAL SKILLS

Programming Languages: Java, Python, C, C++, C#, JavaScript, Typescript, HTML/CSS, SQL, Bash
Technologies: AWS, Azure, Docker, React, Postgres, Git, Numpy, Sci-kit Learn, LangChain, FastAPI, FastMCP, Next.js, Kubernetes, Node.js