Dillan Khurana

703-608-5688 | dillan.khurana@gmail.com | linkedin.com/in/dillan-khurana | dillank.dev

Education

University of Virginia Charlottesville, VA
B.S. Computer Science, B.A. Mathematics *Aug. 2022 – May 2026*
Relevant Coursework: Natural Language Processing (NLP), Machine Learning, Advanced Data Structures and Algorithms, Reinforcement Learning, Computer Vision, Advanced Software Development, Probability

Experience

Amazon Web Services (AWS) May 2025 – Aug. 2025
Software Development Engineer Intern *Arlington, VA*
- Developed an MCP server using FastMCP, Docker, and PostgreSQL to deliver automatic code review feedback, reducing average review cycles by 43%.
- Built agentic tools within the MCP server to access the file system, integrate with code review APIs, and persist issue watchlists in PostgreSQL.
- Created a full-stack web application (React + Python REST) to manage regional build notifications for 22,000+ SNS subscribers, eliminating 600+ hours annually spent on notification template redeployments.
- Engineered backend infrastructure with AWS CDK (DynamoDB, S3, API Gateway, Lambda, IAM) to support auditing, template storage, and APIs, establishing a scalable foundation for future notification features.

Alarm.com June 2024 – Aug. 2024
Software Engineer Intern *Tysons, VA*
- Engineered an automated testing platform to benchmark Multi-modal Large Language Models (MLLMs), replacing manual evaluation methods and doubling testing speed.
- Designed an algorithm to quantify MLLM image recognition performance using k-means clustering, OpenAI vector embeddings, and cosine similarity.
- Constructed a web interface and pipeline for collecting, pre-labeling, validating, and storing image snapshots using Azure Blob Storage, ASP.NET, and C# to reduce dataset procurement time by 83%.

10Pearls June 2023 – Aug. 2023
Software Engineer Intern *Tysons, VA*
- Developed an LLM-powered sentiment analysis application to flag high-risk posts, saving pharmaceutical clients $16K+ per incident. Increased accuracy by 34% via PyTorch + HuggingFace PEFT/QLoRA fine-tuning.
- Built an admin dashboard with FastAPI, React, and MySQL to generate and visualize sentiment analysis reports, eliminating command-line use. Containerized backend services with Docker.
- Built a CI/CD pipeline with GitHub Actions, pytest, and Selenium, improving application reliability and security.

Research & Projects

UVA ML Researcher | *Transformers, PyTorch, Reinforcement Learning* Aug. 2025 – Present
- Conducting research under Prof. Cong Shen on Transformer architectures across NLP, audio, and control systems.
- Trained and evaluated large (∼100 M-parameter) language and audio models on H100 GPUs with PyTorch and NumPy to study cross-modal architecture transferability. Implemented grouped query attention and KV-caching.
- Exploring Looped Transformers and RL post-training to increase reasoning performance with fewer parameters.

Agentic Developer Framework | *Python, llama.cpp, PyTorch, Numpy, LangChain* June 2024 – Present
- Built an open-source framework (DillAgent) for agentic LLM applications, offering abstractions for LLMs, tools, agents, and executors as a modular alternative to LangChain.
- Designed asynchronous execution graphs supporting agentic workflows such as Retrieval Augmented Generation (RAG), multi-agent supervision, and tool execution.
- Fine-tuned and quantized local LLMs with llama.cpp and PyTorch to support OpenAI function-calling, enabling 4-bit 8B+ model integration with the framework.

Technical Skills

Languages: Java, Python, C/C++, SQL (Postgres), JavaScript, Typescript, HTML/CSS, C#, Bash, Linux/Unix
Frameworks: React, Node.js, FastAPI, Flask, Django, LangChain, Express.js, PyTest, JUnit, Svelte, FastMCP
Developer Tools: Git, Docker, Google Cloud Platform, AWS, Azure, VS Code, Visual Studio, IntelliJ, Vercel
Libraries: PyTorch, NumPy, pandas, TensorFlow, Transformers, Matplotlib, Tailwind, scikit-learn, RTK Query