Anshumaan Singh

☎ +1 (934)-219-4495 ✉ anshumaanvsingh@gmail.com

linkedin.com/in/anshumaanvsingh github.com/KrossKinetic krosskinetic.github.io

Stony Brook University **Aug 2023 – May 2027**

B.S. in Computer Science Honors — GPA: 3.93/4.0 *Stony Brook, NY*

- Honors / Awards: SUNY SOAR Fellow, URECA Fellow, University Scholars, Dean's List (2023 – 2025)
- Relevant Coursework: Algorithm, Systems, Data Structures, OOP, Software Development, Theory of Computation

Technical Skills

ML & AI Infra: LLM Fine-Tuning, RAG, Knowledge Distillation, PyTorch, TensorFlow, Multithreading,

Full-Stack & APIs: FastAPI, React.js, Node.js, JSX, Flask, PostgreSQL, MongoDB, MySQL, SQLite3, RESTful

Technologies: HPC/Slurm, Firebase, Google Developer Console, Docker, AWS (EC2, CI/CD), Linux/Unix

Languages: Python, C, Java, JavaScript, Kotlin, HTML/CSS

Professional Work Experience

Software Engineer (Generative AI & Backend) **Sep 2025 – Present**

Mailgator *Palo Alto, CA (Hybrid)*

- **Accelerating QA runtimes by 26×**, enabling **controlled testing across 700+ cases**, by engineering and deploying a RESTful mock server (FastAPI) with full CRUD support on AWS EC2.
- **Increasing data accuracy and system reliability** by resolving critical parsing bugs, enhancing OpenAI prompts for data extraction and ensuring comprehensive handling of sender-recipient edge cases.
- **Developing LLM systems for email analysis** with **a 5-person agile team**, building prompt QA and UX test infrastructure for ML backend (FastAPI, Node.js, React, PostgreSQL).

Undergraduate AI Research Assistant (Generative AI) **Jun 2025 – Present**

LUNR AI Lab, Stony Brook University *Stony Brook, NY*

- **Improving Coding RAG accuracy by +5.4% (MBPP Eval) and +1.6% (ODEX Eval)** by fine-tuning CodeLlama-7B on a custom 460K+ sample dataset in a 4-person team, targeting two ACL 2026 publication.
- **Achieving 73.5% faster benchmark runtimes** against existing baselines by designing a parallelized RAG benchmark system using vLLM and multiple commercial APIs.
- **Reduced LLM inference costs by up to 100%** by integrating SQLite3 caching system into the model distillation pipeline.

Software Engineer (Generative AI & Full Stack) **Feb 2024 – Oct 2024**

iGEM, Stony Brook University *Stony Brook, NY*

- **Achieved 90% retrieval accuracy** on embedded research documents by building a RAG Q&A chatbot using Transformers and LangChain, improving research wiki UX.
- **Helped secure over $50K in funding by leading a 3-person team** to develop a research wiki (Flask) that attracted 15+ stakeholders.

Projects

CMDFlow at HackPrinceton (Full Stack & Systems Engineering) **November 2025**

- Built a local-first, AI-powered command-tracking system (FastAPI, ReactJS, MongoDB) that streams shell activity ($< 1s$ latency), performs PII scrubbing, and semantically indexes commands for natural language search and automatic project-based organization.

NotiSentry: A Smart DND with LLMs (Generative AI and Android) **Jun 2025 – Present**

- Optimizing system performance to <**1% battery drain over 5 hours and achieving 1.3s worst-case latency** for an **intelligent notification filtering and summarization** app to minimize user distractions and improve focus using Firebase Gemini API and Jetpack Compose.

REPLUG LSR with vLLM (AI Research & Infrastructure) **June 2025**

- Refactored research implementation of REPLUG to enable LM-Supervised Retrieval (LSR) fine-tuning for code generation tasks by architecting a high-performance training pipeline with a local vLLM server.