Anshumaan Singh — AI Systems

+1 (934)-219-4495 anshumaanvsingh@gmail.com
linkedin.com/in/anshumaanvsingh github.com/KrossKinetic krosskinetic.github.io

Education

Stony Brook University **Aug 2023 – May 2027**
B.S. in Computer Science Honors — GPA: 3.88/4.0 *Stony Brook, NY*

- **Honors / Awards:** SUNY SOAR Fellow, URECA Fellow, University Scholars, Dean's List (2023 – 2025)
- **Coursework:** Systems Fundamentals I & II, Foundations of CS, Applied Linear Algebra, Theory of Computation Honors
- **Planned (Fall 2026):** Natural Language Processing, Compiler Design, Computer Networks

Technical Skills

ML Systems & Infra: vLLM, Slurm HPC, Docker, AWS (EC2/S3), Linux, Parallel Execution, Sandboxing, IPC
Languages: C (Systems), Python (PyTorch), Java, MATLAB, MIPS Assembly
Research & Optimization: LoRA/QLoRA, Model Distillation, RAG, MPC, CMA-ES, Transformer, SQLite3 Caching

Work Experience

Incoming AI & ML Engineering Intern **Jun 2026 – Aug 2026**
Bank of Montreal (BMO) *Berkeley Heights, NJ*

- Incoming software engineering internship focused on AI infrastructure and machine learning systems.

Undergraduate Research Assistant **Jan 2026 – Present**
Reliable Systems Lab, Stony Brook University *Stony Brook, NY*

- **Engineering a lightweight encoder-only attention architecture for multi-agent systems**, leveraging permutation equivariance to process jagged input arrays and eliminate data sorting overhead.
- **Improving UAV Agent target seeking efficiency by 34.6%** by architecting a Curriculum Weight Tuning API using CMA-ES through a grouped parameter strategy.
- **Enabling 100% collision-free horizontal scaling on the SeaWulf HPC cluster** by building a distributed isolation framework using dynamic sandboxing to eliminate MATLAB/MEX race conditions.

Undergraduate Research Assistant **Feb 2025 – Present**
LUNR Lab, Stony Brook University *Stony Brook, NY*

- **Improved Coding RAG accuracy by 5.4%** and **accelerated code rag bench runtimes by 73.5%** by fine-tuning CodeLlama-7B via LoRA and designing a parallelized vLLM routing system.
- **Reducing LLM inference costs by up to 100%** by architecting and integrating an SQLite3 caching layer directly into the model distillation pipeline.
- **Refactored the research implementation** of REPLUG to enable LM-Supervised Retrieval (LSR) fine-tuning for code generation tasks by architecting a training pipeline driven by a local vLLM server.

Software Developer Intern **Sep 2025 – Jan 2026**
Mailgator *Palo Alto, CA (Hybrid)*

- **Reduced validation latency by 99.9% (60m to <2s)** by architecting a scalable CI/CD pipeline on AWS EC2 with automated asynchronous integrity checks for a FastAPI backend.
- **Optimized system reliability** by refactoring LLM-driven parsing logic for deterministic data extraction, ensuring 100% accuracy across complex sender-recipient edge cases.

Projects

Automated Systems Fuzzer (C & Unix Systems) **Spring 2026**

- Engineered a high-performance C fuzzer utilizing **Unix signals and syscalls (fork, waitpid)** to stress-test executables via mutated input streams, achieving 100% process isolation and automated memory leak detection.

CMDFlow at HackPrinceton (Semantic Search & Local-First Systems) **November 2025**

- Engineered a local-first command-tracking system (FastAPI, MongoDB) that streams shell activity with **<1s latency**, performs automated PII scrubbing, and semantically indexes CLI commands for natural language search.