

Krish Sharma

ML SYSTEMS — GPU KERNELS — INFERENCE PERFORMANCE

krishs04@stanford.edu — (262) 402-0565 — [linkedin.com/in/krish-sharma-pop](https://www.linkedin.com/in/krish-sharma-pop) — github.com/krishs0404

EXPERIENCE

NVIDIA

Fall 2026

Deep Learning Performance Tooling Intern — Santa Clara, CA

- Developing agentic **SASS kernel generation** for **just-in-time (JIT) compilation** in TensorRT-LLM, supporting kernel optimization across the TensorRT inference team.
- Driving significant inference performance gains through generated **GEMV, mixture-of-experts (MoE), and scaled dot-product attention (SDPA)** kernels.

MatX

Summer 2026

Kernel Engineer Intern — Mountain View, CA

- Wrote **high-performance inference kernels** for an AI inference rack built on **MatX One**, targeting its custom accelerator architecture.

Hazy Research, Stanford University

2025 - Present

Research Engineer - GPU Systems / ML Kernels

- Developing custom **CUDA kernels** for **ThunderKittens**, a GPU kernel framework for deep learning workloads.
- Designed tiled kernels with explicit **shared-memory orchestration and warp-level parallelism** to maximize streaming multiprocessor utilization and throughput.
- Profiled instruction-level bottlenecks with **Nsight Compute**; analyzed memory coalescing and shared-memory bank conflicts to guide kernel optimization.

Tessera Labs

Jul 2025 - Present

Machine Learning Engineer Intern — Remote

- Shipped a production operations agent in one week; migrated the Python prototype into internal tooling, cutting build and integration time by **30%**.
- Built Workday and Salesforce agent integrations supporting **40+ workflows/day** and reducing manual operations touchpoints by **50%**.

Altemose Lab, Stanford Bioengineering

Feb 2025 - Sep 2025

Computational Biology Researcher — Stanford, CA

- Built high-throughput Snakemake pipelines for nanopore sequencing data, reducing analysis time by **70%**; trained a PyTorch basecaller with **92% accuracy on 1M+ long reads**.

SELECTED ML SYSTEMS PROJECT

AgentServe: Agent-Aware LLM Inference Engine

2026

Personal Project — github.com/krishs0404/agentserve

- Built an **OpenAI-compatible inference engine** implementing Llama 3.2 from scratch (RMSNorm, RoPE, GQA, SwiGLU), with safetensors weight loading and a **paged KV-cache allocator**.
- Designed **continuous batching** with agent-aware priority, overflow, and preemption policies to minimize DAG completion time; implemented prefix KV caching with LFU eviction for shared prompts.
- Built a benchmark harness measuring **throughput, time to first token, and latency** for A/B comparisons against FIFO scheduling and vLLM.

EDUCATION

Stanford University

Expected 2027

B.S. in Mathematics / Computer Science — Stanford, CA

Coursework: Machine Learning, Operating Systems, Statistical Inference, Linear Algebra, Reinforcement Learning

TECHNICAL SKILLS

Languages: C++, C, Python, CUDA, SASS, Bash, SQL

ML Systems: TensorRT-LLM, ThunderKittens, Nsight Compute, PyTorch, Transformers, NumPy

Infrastructure: Linux, Git, Docker, Kubernetes, Snakemake