Harshith Kantamneni

kantamneniharshith@gmail.com | +1 (414) 916-5799 | Milwaukee, WI
linkedin.com/in/hk4231 | github.com/harshithkantamneni | harshithkantamneni.github.io

SUMMARY

Master's graduate in Electrical and Computer Engineering (UW-Madison, Dec 2025) with hands-on experience in CUDA and Triton kernel optimization, LLM inference profiling, and multi-agent and retrieval system design. Strong foundation in C++, Python, parallel programming, and GPU architecture, skilled at translating low-level GPU profiling into measurable latency and throughput gains. Fluent across the full stack, from hand-tuned GPU kernels up to complete AI systems. Curious and collaborative, I like to understand systems down to the metal before I build on top of them.

TECHNICAL SKILLS

GPU, Architecture & Performance: CUDA, Triton (kernel DSL), PyTorch, kernel optimization, kernel fusion, Tensor Cores, mixed precision (fp16), GPU memory hierarchy, roofline and memory-bandwidth analysis, NVIDIA Nsight Systems, Nsight Compute, quantization, ROCm/HIP, gem5, McPAT

AI Systems: LLM inference profiling, FlashAttention/SDPA, KV-cache decode, RAG (retrieval-augmented generation), multi-agent orchestration, Model Context Protocol (MCP) server design, Claude Code / Agent SDK, agent evaluation

Retrieval & Memory: Hybrid search, reciprocal-rank fusion (RRF), cross-encoder reranking, sqlite-vec, SQLite FTS5

Languages & Infra: Python, C++, C, Linux, Git, GitHub, Docker, CI/CD, Slurm, Supabase, Railway

PROJECTS

Triton vs cuBLAS LLM Kernel Benchmarking (GitHub) Jan 2026 - Present
Triton, CUDA, PyTorch, NVIDIA Nsight

- Fused linear, bias, and **GELU** into one **Triton** kernel, cutting small-batch FFN latency **up to 1.73x** by eliminating two **HBM** round-trips and two kernel launches.
- Benchmarked **76 LLM-shaped** GEMM shapes on an **A100**, with the autotuned Triton GEMM peaking at **213 TFLOP/s** (**68%** of fp16 tensor-core peak).

Edge LLM GPU Profiling on Jetson Orin Nano (GitHub) Jan 2026 - Feb 2026
CUDA, NVIDIA Nsight Systems, PyTorch, FlashAttention

- Profiled attention and **KV-cache decode** on a 15 W **Jetson Orin Nano** with **Nsight Systems**, isolating the per-step kernel-launch overhead characteristic of small-batch edge decode.
- Measured a **5.5x** p50 latency gap between fused **SDPA** (FlashAttention backend) and a hand-rolled materialized-attention baseline at sequence length 1024.

bert: Long-Context Memory for AI Coding Assistants (MCP) (GitHub) May 2026 - Present
Python, MCP, sqlite-vec, sentence-transformers

- Built **bert**, a local **Model Context Protocol (MCP)** server giving AI coding assistants (Claude Code, Cursor, Codex) searchable memory of a whole codebase via reranked **hybrid retrieval**.
- Outscored naive context-window truncation **0.85 to 0.10** on a held-out QA eval while sending **4.6x fewer** input tokens.
- Scored **0.745 nDCG@10** on **BEIR** scifact (beating published BM25 by **0.08**) and held **0.75** on a **3M-token** corpus where truncation scores **0.00**.

The Obsidian Archive: Autonomous Documentary Pipeline (GitHub) Feb 2026 - Apr 2026
Python, Claude API, MCP, Supabase, Railway

- Built a **13-stage** autonomous pipeline that researches, scripts, renders, and uploads documentaries to a live YouTube channel, deployed on **Railway**.
- Gated every upload behind per-stage quality checks and a script-scoring pass across **8 dimensions** with automatic rewrites of weak sections.

Autonomous ML Research Lab (Multi-Agent System) (GitHub) Mar 2026 - May 2026
Python, C, Apple Metal, Claude API

- Built an autonomous system of **31 role-specialized agents** that run an end-to-end **ML research loop**, with all training and experiments running on-device and a human only setting direction.
- Directed the agents to build a **from-scratch C17 training engine** (hand-written autodiff, a **Mixture-of-Experts** transformer, 4-bit **QAT**) with no PyTorch or TensorFlow.
- Built the integrity layer (pre-registration, unanimous phase gates, anti-forgery sign-off verification) that stopped an agent forging **5 approvals**, with no recurrence in subsequent runs.

HIVE: Autonomous Multi-Agent Software Org (GitHub) Mar 2026 - May 2026
Rust, libp2p, Python, Claude Code

- Built and operated **HIVE**, an autonomous software org of **45 role-specialized agents** across **203 plan-build-evaluate cycles**, under ADR-driven governance with anti-forgery co-signing.
- Directed the agents to build a **62K-line Rust** peer-to-peer reasoning engine gated by **1,300+ tests**, including a forward-chaining **Rete-II** engine and a local differential-privacy layer.

ML-Guided CUDA Kernel Configuration (GitHub) Jan 2025 - May 2025
UW-Madison course project | Python, PyTorch, CUDA, Slurm

- Trained a PyTorch MLP surrogate that predicts CUDA kernel runtime at $\mathbf{R^2 = 0.96}$, replacing exhaustive autotuning with argmin config selection.
- Validated predicted block and grid configs to **within 5%** (median) of the measured optimum, with CUDA event-timing automated on **Slurm**.

MI300X Memory-System Characterization & gem5 Modeling Jan 2025 - May 2025
UW-Madison course project | HIP, gem5, McPAT, Python

- Characterized the **AMD MI300X** memory hierarchy with **HIP** microbenchmarks (**4.6-4.8 TB/s** write bandwidth), then tuned full-system **gem5** from an **89%** latency error down to **within 5%** of hardware measurements.
- Modeled in-order and out-of-order CPUs in **gem5** across three AXPY kernels, extracting CPI, latency, and power with **McPAT**.

EDUCATION

University of Wisconsin-Madison *Sep 2024 - Dec 2025*
M.S. Electrical and Computer Engineering USA

- Relevant Coursework: High Performance Computing, Advanced Computer Architecture, Machine Learning, Fault-Tolerant Computing

Vellore Institute of Technology *2020 - 2024*
B.Tech in Electronics and Communication Engineering India