Harshith Kantamneni

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

SUMMARY

Agentic AI engineer (M.S. ECE, UW-Madison, Dec 2025) who designs, runs and audits autonomous **multi-agent systems** and an **MCP server**, directing coding agents and owning architecture, orchestration and **evaluation**. Also merged Triton kernels in flash-linear-attention and tests in NVIDIA cuDNN frontend.

PROJECTS

Autonomous ML Research Lab (GitHub) Apr 2026 - Sep 2026
Python, Anthropic Claude API, Claude Code, tmux, stream-JSON

- Designed and ran an ML research lab of 31 **Claude Code** agent roles on a from-scratch C17 training engine.
- Diagnosed a director agent forging **5 PI sign-offs** and shipped a signature verifier, a runner gate and an evaluator check.
- Audited that fix against the lab's runner logs and found all three parts defective, with the gate executing on **0 of 270** eligible sessions. Shipped an improved detector with regression tests.
- Corrected my own savings estimate from 6-9x to **2-3x** after moving sub-dispatches to cheaper model tiers. Added a rate-limit gate to stop no-op sessions during limit waits.

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

- Built a local **MCP server** giving Claude Code and other MCP hosts searchable project memory with **hybrid retrieval**.
- Benchmarked it on a 1.26M-token synthetic corpus: it held **0.90** accuracy, level with agentic grep, where 200K-window full-context fell to 0.08. On code, its multi-turn search reached 0.86 against 0.66 for one-shot retrieval. Published the results, including where grep scored higher.

The Obsidian Archive: Autonomous Content Pipeline (GitHub) Feb 2026 - Sep 2026
Python, Claude API, ElevenLabs, fal.ai, MCP, Supabase, Railway

- Built a 13-stage **autonomous pipeline** that researches, writes, renders and uploads a documentary from one topic.
- Added atomic checkpoints and **resume**, automatic retries, an optional **budget cap** on Claude spend, and Telegram and Discord failure alerts.

OPEN-SOURCE EXPERIENCE

GPU Kernel Contributor Mar 2026 - Present

- Merged **Triton** attention kernels and KV-cache decoding into **flash-linear-attention** (#1319, #1412) and tests into **NVIDIA cuDNN frontend** (#1287). NVFP4 kernel in review at NVIDIA TransformerEngine (#3591).

SKILLS

Languages: Python, C++, CUDA, Triton
Agent systems: Agentic AI, multi-agent orchestration, MCP, tool use, sub-agents, agent memory
Evaluation and reliability: Agent evaluation, evaluator agents, cost optimization, budget caps, retries, checkpoint and resume
LLM APIs and retrieval: Claude API, Claude Code, OpenAI API, RAG, hybrid retrieval, reranking
Data and infrastructure: Vector search, sqlite-vec, SQLite FTS5, Supabase, Railway, Linux, Git, Docker, CI/CD

EDUCATION

University of Wisconsin-Madison, M.S. Electrical and Computer Engineering Sep 2024 - Dec 2025
Coursework: High Performance Computing, Advanced Computer Architecture, Machine Learning, Fault-Tolerant Computing

Vellore Institute of Technology, B.Tech Electronics and Communication Engineering 2020 - 2024