Akash Dasarraju

(469) 428-0652 | dasarrajuakash@gmail.com | sky1219.github.io
linkedin.com/in/akash-dasarraju-143520240 | github.com/Sky1219

Low-Level GPU Kernels & ML Systems | LLM Inference & Performance Engineering

Experience

Researcher, CARAML Lab (Prof. Rishabh Iyer) Mar. 2026 – Present
Submodular ML lab; I implement and run experiments for 2 in-prep papers, with PhD mentors *Richardson, TX*

- Beat C2LIP, a published state-of-the-art fine-tuned CLIP model, by 3% on fine-grained object–attribute reasoning (SugarCrepe, SugarCrepe++) with only 81k COCO image–caption sets, by implementing a submodular set loss that learns from several positives and hard negatives at once (targeting CVPR)
- Developing a submodular method that picks which past conversation turns an LLM agent keeps under a token budget, compared against top-k and MMR, self-hosting Qwen2.5-32B with vLLM on an A6000 (targeting ICML)

AI Engineer (sole developer), SoyMedical *(medical supplies distributor, $3M annual revenue)* Summer 2026

- Tripled the products a buyer can evaluate while halving the time, with 30% fewer underperforming products ordered, via a RAG-based agent pipeline now in the ordering workflow (since run on thousands of products)
- Built and deployed it solo as a scheduled production job: Pinecone vector search, Keepa API for Amazon prices, and Redis caching, returning validated JSON with automatic retries, and a second LLM ranking products

Personal Projects (Solo)

hipinfer | *C++, HIP/CUDA* | github.com/Sky1219/hipinfer | sky1219.github.io/hipinfer | Colab

- Built an LLM inference engine in HIP with every kernel from scratch (no cuBLAS or vendor libraries) supporting dense and MoE models, multi-GPU tensor parallelism, int8 weights, speculative decoding, and KV-cache eviction; tested on NVIDIA, compiles for AMD
- Verified against Hugging Face: WikiText-2 perplexity within 0.01% on every model, 99.4–99.9% top-1 token agreement, a greedy token-match gate, and kernel tests against PyTorch
- Beat llama.cpp at single-request decode on every model tested on 3 GPUs (5–16% on 1–3B models, 2080 Ti, at 81–87% of memory bandwidth); 23–31% faster than vLLM on an A100; GEMM at 92–95% of cuBLAS on Turing
- 1.5–1.7x faster decode on 2 RTX 2080 Tis via tensor parallelism, with a custom all-reduce 1.4x faster than NCCL on small messages; runs 7–8B models (Llama-3.1-8B, Mistral-7B) too large for one card
- 1.6–1.8x faster decode than fp16 with int8 weights (16–26% faster than llama.cpp Q8_0, +0.04–0.23% perplexity)
- Up to 1.39x faster decode with speculative decoding (1B draft, 3B target), output token-for-token identical to plain greedy decoding in all 105 test runs on each of 2 GPUs

SubCache | *Python, PyTorch* | github.com/Sky1219/SubCache | sky1219.github.io/SubCache

- 2.8x higher throughput from keeping 25% of the KV cache on long coding-agent sessions (3x more requests per GPU, 5% perplexity increase), measured in my open-source compression benchmark
- Benchmarked 9 eviction policies vs. H2O and R-KV on agent sessions, LongBench, retrieval, and MATH-500
- 1.1–2.5% better than H2O on repetitive agent sessions with my submodular eviction method; profiling showed selection took 71% of decode time, so I wrote GPU kernels for it in hipinfer (6–25x faster than PyTorch)

Leadership

HackUTD, Organizer (Leadership Team), UT Dallas Jan. 2026 – Present

- Organize a 1,300-person hackathon: built a RAG/MCP Discord bot (near-100% answer success) that cut pre-event organizer questions 80% vs. prior years, and a 3D campus map (three.js) used by hundreds; led 5 AI workshops, tripling turnout (30 to 100+); leading 10+ volunteers

Technical Skills

Languages & Tools: Python, C++, Java; CUDA, HIP/ROCm, NCCL, PyTorch, vLLM, Hugging Face
GPUs: NVIDIA Turing & Ampere (2080 Ti, A6000, A100)

Education

The University of Texas at Dallas Richardson, TX
B.S. in Computer Science, CS Squared Honors Program; GPA: 3.62 *Aug. 2025 – May 2029 (expected)*

- Coursework: Data Structures, Prob. & Stats, Comp. Arch. (honors); Linear Algebra; Math of Computing (senior)
- National Merit Finalist; ACT 36