

SAICHARAN RAMINENI

904-615-0799 | csramineni@gmail.com | [linkedin.com/in/saicharanramineni](https://www.linkedin.com/in/saicharanramineni) | github.com/GodlyDonuts | saicharanramineni.com

EDUCATION

University of Central Florida

Bachelor of Science in Computer Science, Burnett Honors College; **GPA: 4.0**

Orlando, FL

Expected May 2029

EXPERIENCE

Undergraduate Research Assistant

University of Central Florida - Dr. Shyam Kattel Laboratory

Orlando, FL

Mar 2026 - Present

- Curated and quality-controlled **terabytes of computational-catalysis data**, transforming noisy simulation outputs into structured, training-ready datasets for large-scale forward/inverse modeling and catalysis-specific ML research workflows.
- Built **forward and inverse catalyst-design models** mapping structures to behavior and target properties back to candidates; trained **NVIDIA Nemotron Ultra** with SFT/GRPO and GLM 5.2 teacher supervision for catalysis reasoning and scientific workflows.
- Scaled scientific-model training across **32+ NVIDIA H100 GPUs** with gradient checkpointing, DeepSpeed ZeRO-3, and InfiniBand all-reduce for long-running scientific workloads, cutting full experiment turnaround from **weeks to under 72 hours**.

PROJECTS

Sai: Vision-Based Voice-Native macOS Agent | *Amazon Nova, FastAPI, WebSockets, PyAutoGUI, Computer Vision*

- Won Best Agentic AI (\$8,000)** at the Amazon Nova Hackathon (**13,000+ participants**) for a voice-native macOS agent that reasons over screenshots and executes tasks across any app and display resolution from pixels alone, without DOM or accessibility APIs.
- Engineered a tiered Nova 2 Lite-to-Nova Pro pipeline with streaming speech recognition, intent repair, and **~200ms simple-command routing**, escalating only multimodal workflows to the full vision agent through FastAPI and WebSockets.
- Built a **25-step Plan-Act-Verify loop** with fresh screenshots, click-critic correction, coordinate validation, **1-6-action cycle detection**, and resolution-independent Retina-to-**[0,1000]^2 mapping** with LANCZOS resampling and PyAutoGUI execution.

Psi: From-Scratch Autograd & GPU ML Stack | *C++17, CUDA, Metal, Autograd, Transformers*

- Built Psi entirely in **C++17 with no ML framework**: reverse-mode tensor autograd, BPE, GQA/RoPE/SwiGLU/RMSNorm transformer, AdamW, and checkpoints; **all 12 tensor operations matched analytic gradients to ~1e-12 in double precision**.
- Trained four **115K-574K** TinyStories models on **~200M tokens**; the **354K-parameter** model matched TinyStories-1M's **8/6/6/5** median across **36 held-out completions** at one-third the parameters under identical evaluation settings and scoring rubric.
- Wrote a bit-exact Metal matrix-multiplication kernel reaching **883 GFLOP/s** on M1 (**34% of peak, 6x naive**); built a separate 2-bit ternary kernel with **16x smaller weights** while sustaining full fp32-matmul throughput on the same Metal backend.
- Ported the full autograd/model/training path to CUDA/cuBLAS; preserved **bit-identical CPU/V100/H100 loss** and cut 100 steps from **150.7s to 14.1s (10.7x)**, while custom H100 GEMM reached **2.47 TFLOP/s** on a 2048^3 matrix shape.

Crisis Averted: Voice-Orchestrated UN Data Platform | *Next.js, React, Three.js, FastAPI, Databricks*

- Won 3rd Place Overall** at Hacklytics 2026 (Georgia Tech, **1,200+ participants**) for a sovereign, voice-orchestrated platform analyzing **5 GB of UN data** through distributed agents with real-time ingestion, retrieval, and conversational access.
- Architected a distributed multi-agent system and voice-orchestrated Databricks pipeline using Kafka Streams for **sub-millisecond command dispatch** across services; connected ingestion, transformation, embedding, retrieval, and voice orchestration.
- Self-hosted an Actian Vector DB on Vultr for vector intelligence, semantic retrieval, and low-latency query serving, executing **sub-100ms searches across 18,000+ embeddings** for real-time responses over the UN dataset.

OPEN SOURCE

- Merged **21 TensorFlow contributions (199k stars)**: fixed SiLU/ELU/*erfinv* precision, TensorList scatter/set crashes, TFLite FFT leaks and high-rank stack overflows, and concurrent-GPU *boolean_mask* corruption; added focused regression coverage.
- Landed a **PyTorch core** contribution (**102k stars**) in TorchInductor, enabling static Triton launchers to allocate device-side TMA global scratch on serialized FX graph-cache hits and validating repeated non-default-stream and persistent-GEMM launches.
- Merged **3 PRs into Unsloth (72k stars)**: restored vLLM 0.22+ fast inference with a meta-path-finder shim, made cross-platform *uv* installer paths space-safe on macOS/Linux, and hardened sandboxed Python execution for UTF-8-safe output.

TECHNICAL SKILLS

Languages: Python, C++17, Rust, Java, TypeScript, Dart, JavaScript, C, SQL, Bash

Application Development: React, Next.js, Node.js, Express, Spring Boot, FastAPI, REST APIs, GraphQL, WebSockets, OAuth/JWT

Data & ML: PostgreSQL, Redis, Kafka, Spark, Databricks, PyTorch, TensorFlow, CUDA, Triton

Cloud & Developer Tools: Docker, Kubernetes, AWS, GCP, Linux, Git, CMake, Nginx, GitHub Actions