Sri Harsha Mudumba

(515)-916-3011 — srim@iastate.edu — LinkedIn — GitHub — Portfolio — Ames, Iowa (IA)

Education

Master's in Computer Engineering, Iowa State University, Iowa, IA — Aug 2023 – Aug 2025
GPA: **3.77/4.0**

Technical Skills

AI Frameworks: PyTorch, TensorFlow, ONNX, IREE, MLIR, scikit-learn
Programming: Python, C++, C, SQL, Shell scripting
Model Optimization: Pruning, Quantization (1-bit, 8-bit), Knowledge Distillation, Early-Exit Inference, TinyML
Other: LLMs, RAG Systems, Linux, AWS, Docker, Git, CI/CD, Oracle Databases (11g/12c/19c)
High Performance computing tools: OpenMP, MPI, Slurm, Parallel Computing
Hardware/Simulation: CIMLOOP, Heterogeneous CiM (SRAM, RRAM, MRAM), Performance Benchmarking

Research

DURACIM — Durable Compute-in-Memory Early-Exit Deep Learning Framework GitHub
Jan 2024 – Present
Technologies: Python, PyTorch, ResNet-50, Reinforcement Learning, CIMLOOP Simulator

- Enabled **on-device deep learning** by designing an early-exit **ResNet-50** that adapts to hardware constraints while maintaining accuracy.
- Delivered **21.7% lower energy use** and **10.5% faster inference latency** on resource-limited devices, extending system lifetime from **92 to 165 days**.
- Integrated **CIMLOOP cycle-accurate simulation** to co-design algorithms with device-level constraints, proving RL as a scalable approach to optimize energy–accuracy trade-offs.
- Demonstrated that reinforcement learning discovers **superior energy–accuracy frontiers**, enabling **privacy-preserving on-device deployment** of models.
- Research publication in progress.

Key Projects

TorchWeave-LLM: Continuous Batching Inference Server (present) GitHub
Technologies: Python, PyTorch, Hugging Face Transformers, FastAPI, Docker

- Designed and deployed a custom **LLM serving system** to support **multi-user, low-latency inference** at scale.
- Implemented **continuous batching with async scheduling**, reducing **time-to-first-token (TTFT)** by 35% and boosting **2–5**× the throughput.
- Built per-request **KV-cache management** and token streaming via **Server-Sent Events (SSE)** to ensure smooth real-time user interaction.
- Containerized optimizer and server as independent services with shared artifact storage, enabling reproducible builds and scaling to **ECS/Kubernetes**.

ARGUS: Multimodal Retrieval-Augmented Generation System GitHub
Technologies: Python, LangChain, LangGraph, OpenAI, FAISS, AWS, Docker

- Built a hybrid Retrieval-Augmented Generation for **semantic search** and **re-ranking** of text documents using **FAISS similarity**.
- Integrated **LangGraph** for **modular orchestration** of retrieval and response-generation agents, enabling flexible pipeline.
- Designed and deployed the pipeline in a **containerized AWS environment** using **Docker**, and implemented **CI/CD practices** for continuous integration and deployment.
- Achieved **sub-second query response** on **100+ documents** with scalability tested for **100,000+ embeddings**.

LEXA: Lightweight Local Retrieval-Augmented Generation GitHub
Technologies: Python, FastAPI, TinyLlama, HuggingFace, SentenceTransformers, NVIDIA RTX 4060

- Architected a fully-local **RAG** system with **TinyLlama 1.1B** optimized for GPU inference on **NVIDIA RTX 4060**, implementing custom CUDA memory management.
- Developed GPU-accelerated early exit logic with adaptive confidence thresholding, reducing mean response time by **40%** and GPU power consumption by **30%**.

Professional Experience

Associate Software Engineer, Cognizant Technology Solutions, Bengaluru, India — Aug 2020 – Jul 2023
`(Database Administration Team)`

- Designed and maintained **scalable data and compute pipelines**, improving overall reliability and efficiency across enterprise-scale systems.
- Automated **Linux-based workflows** for replication, patching, and monitoring, **reducing downtime by 50%** and ensuring consistent high availability in production environments.
- Optimized performance across large-scale deployments, reducing **latency, I/O bottlenecks, and memory overhead**, leading to significant gains in efficiency and resource utilization.
- Collaborated with cross-functional teams to deliver **secure, high-performance infrastructure**, supporting mission-critical workloads under strict reliability requirements.