Sri Harsha Mudumba

sri19harsha98@gmail.com | +1 (515)916-3011 | IA, USA | LinkedIn | GitHub | Portfolio

SUMMARY

Machine Learning and Systems Engineer with 3+ years of experience building scalable data and ML systems. Strong foundation in performance profiling, reliability, and production pipelines, with hands-on work in transformers, embeddings, and evaluation-driven iteration. Interested in personalization and recommendation systems where sequential signals, disciplined experimentation, and shipping impact matter.

TECHNICAL SKILLS

Programming & Data Processing: Python, C++ (modern standards), SQL, Bash, Data Structures and Algorithms, Linux Systems.
ML Inference & Optimization: Inference optimization, latency and throughput profiling, memory efficiency, early-exit architectures, quantization (post-training and inference-aware), sparsity-aware execution, model benchmarking and performance validation.
Machine Learning Frameworks: PyTorch, TensorFlow, Hugging Face, ONNX Runtime, IREE.
Natural Language Processing & GenAI: Text Preprocessing, Tokenization, Named Entity Recognition (NER), Topic Modeling, Sentiment Analysis, Transformers, Hugging Face, LangChain Retrieval-Augmented Generation (RAG), Vector Databases (FAISS, Pinecone)
Deployment & ML Systems: FastAPI, REST-based model serving, batch and real-time inference pipelines, experiment tracking (MLflow), model versioning, CI/CD for ML systems, Docker, Kubernetes (basic)
Cloud & Data Engineering: AWS (S3, EC2, Lambda), Azure Machine Learning, ETL Pipelines, Batch & Real-Time Inference Optimization.

PROFESSIONAL EXPERIENCE

AI Engineer Intern, *HCLTech* *Jul 2025 – Present | Remote, USA*

- Assisted senior engineers in problem framing, exploratory data analysis, and feature preparation for machine learning use cases, contributing to the timely delivery of multiple internal and client-facing proof-of-concept initiatives.
- Built transformer driven NLP and embedding workflows for enterprise text, boosting extraction quality by 20 percent based on internal evaluation sets.
- Prototyped semantic search and reranking using embeddings and vector retrieval, improving top k relevance by 25 percent.
- Used MLflow to track experiments and model versions, making it easy to reproduce runs and compare feature and training changes.
- Helped ship batch and real-time inference endpoints with FastAPI and Docker, adding latency checks and reliability-focused logging for day-to-day debugging.
- Worked closely with data engineers and stakeholders to turn requirements into measurable goals and deliver iteratively.

Software Engineer, *Cognizant Technology Solutions* *Aug 2020 – Jul 2023 |KAR, INDIA*

- Operated and optimized production-grade data systems that served latency-sensitive analytics and machine learning applications, building a strong foundation in performance tuning, reliability, and failure handling under strict SLAs.
- Performed deep performance profiling across query execution, memory utilization, and I/O paths, identifying system bottlenecks and improving end-to-end data throughput for compute-intensive workloads.
- Supported large-scale data ingestion and feature-serving workflows by ensuring data correctness, consistency, and timely availability, providing a stable data foundation for downstream analytics and ML inference pipelines.
- Collaborated cross-functionally with application, infrastructure, and security teams to deliver highly available systems, strengthening debugging skills across distributed systems and performance-critical environments.

RESEARCH EXPERIENCE

DURACIM – RL-Guided Early-Exit Co-Design for Compute-in-Memory Systems | Reinforcement Learning, Python, PyTorch

- Designed sequential decision policies using PPO and Q learning to optimize accuracy latency and energy tradeoffs under deployment constraints.
- Built an experiment pipeline that integrates PyTorch model execution with CIMLOOP based cost modeling to estimate per layer energy and device endurance trends.
- Achieved **86 to 88 percent relative inference energy savings** on ResNet 50 compared to full depth execution while keeping accuracy within about one percentage point.
- Evaluated endurance behavior under multiple mapping configurations and policy choices, showing the effective RRAM lifetime **to 1.02 × 10^7 years** under the endurance model, even in all-RRAM configurations.
- Publication in progress

FEATURED PROJECTS

TorchWeave-LLM – Continuous Batching Inference Server | Python, PyTorch, FastAPI, Docker **(Link)**
Developed a continuous batching inference server for LLMs using asynchronous scheduling, achieving 4× throughput improvement and 35% lower latency across multi-user concurrent workloads.

LEXA – Lightweight Local RAG with TinyLlama 1.1B | CUDA, PyTorch, FastAPI, Sentence Transformers **(Link)**
Built a GPU-optimized local RAG system integrating TinyLlama with adaptive early-exit logic, reducing average inference latency by 40% and GPU power consumption by 30% during multi-query workloads.

EDUCATION

Master of Science, *Iowa State University* Aug 2023 – Aug 2025 | IA, USA
Computer Engineering (Computing and Networking Systems)
Bachelor of Engineering, *Amrita Vishwa Vidyapeetham* *Aug 2023 – Aug 2025 | KAR, INDIA*