Purna Satya Karun Saride

ML Systems Engineer | LLM Inference & Deployment

USA | +1 (775) 300-9330 | purnasatyakarunsaride.24s@gmail.com | LinkedIn | GitHub

TECHNICAL SKILLS

Programming Languages: Python, C, C++, SQL, JavaScript (ES6+)

GPU & Systems Programming: CUDA programming, CUDA kernel development, GPU memory hierarchy, latency and throughput tradeoffs, profiling & debugging

ML and LLM Systems: LLM inference, inference latency and throughput, batching, tokenization, Hugging Face Transformers, embedding optimization, RAG pipelines, FAISS, chunking strategies, vector databases

Systems & Deployment: Docker, REST-based inference APIs, AWS (EC2, S3, Lambda), CI/CD (GitHub Actions), concurrency handling, profiling and debugging, RunPod

Data & Storage: PostgreSQL, MySQL, Supabase, schema design, structured and semi-structured data

Tools & Frameworks: FastAPI, LangChain, Git, GitHub, Linux, API testing (Postman)

PROFESSIONAL EXPERIENCE

Saatvik Advisors | AI Systems Engineer | USA **Aug 2025 - Dec 2025**

- Built production-oriented data pipelines to support LLM-based inference workflows over structured and semi-structured system logs, eliminating over 85% of manual preprocessing.
- Engineered deterministic complex keys across OSI layers (L3–L7) to enable reliable downstream ML inference and anomaly classification, reducing false-positive signals by 30%.
- Designed and optimized semantic retrieval paths using FAISS and transformer embeddings, achieving sub-second inference latency under realistic query loads.
- Applied tokenization-aware chunking and embedding optimization techniques to improve inference relevance and stability in retrieval-augmented ML systems.
- Collaborated with infrastructure and security stakeholders to align ML inference outputs with enterprise workflows, enabling traceable and explainable system-level decisions.

Saatvik Advisors | Network Data Engineer | USA **May 2025 - Aug 2025**

- Analyzed L3–L7 packet metadata to trace transport-layer behaviors impacting application-level system performance across 10K+ enterprise transactions.
- Classified TCP/IP behaviors (retransmissions, flow stalls, malformed segments) and mapped them to downstream service degradation observable by ML-driven monitoring systems.
- Designed structured metadata keys linking transport-layer signals to ServiceNow HRSD workflows, enabling ML assisted root-cause analysis and reducing troubleshooting time by 40%.
- Modeled normalized schemas to preserve end-to-end traceability between packet-level events and application outcomes, improving data integrity validation throughput by 2.5x.
- Partnered with cybersecurity and infrastructure teams to refine anomaly scoring logic used as input features for AI-assisted detection systems.

EDUCATION

University of Nevada, Reno **Dec 2025**

Master of Science in Computer Science GPA: **3.87 / 4.00**

Sasi Institute of Technology and Engineering, India **May 2023**

Bachelor of Technology in Computer Science and Engineering GPA: **3.20 / 4.00**

PROJECTS

Inference Runtime Optimization Pipeline (PyTorch, ONNXRuntime, TensorRT) *Independent Project*

- Benchmarked transformer inference across **PyTorch eager execution**, **ONNXRuntime CUDA**, and **TensorRT** to analyze latency, throughput, and batching behavior on GPU and CPU.
- Implemented CPU **INT8 dynamic quantization** and evaluated precision vs performance trade-offs, identifying batch-size-dependent overhead effects.
- Achieved up to **2.5–8× GPU throughput improvement** by eliminating Python runtime overhead via ONNX graph execution.
- Conducted hardware-aware analysis showing cases where **CUDA outperform TensorRT** due to GPU constraints, emphasizing empirical benchmarking over assumptions.

High-Throughput LLM Inference Serving Benchmarking (vLLM, GPU) *Independent Project*