SRAVAN PUSULURI

Generative AI Engineer · LLM Inference & MLOps · Open to Remote

New York, NY · (518) 888-0259 · sravanpusuluri03@gmail.com · LinkedIn· GitHub

PROFESSIONAL SUMMARY

Results-driven Generative AI Engineer with 4+ years building LLM inference pipelines, RAG architectures, agentic AI systems, and GPU-accelerated ML solutions at Fortune 500 enterprise scale. At MetLife, delivered a 42% inference latency reduction and 35% GPU memory savings on A100 clusters using vLLM, TensorRT-LLM, and INT8 quantization. Expert in conversational AI, function calling, and multi-agent orchestration (LangChain, LangGraph, LlamaIndex, ReAct). Deep experience with vector databases (Pinecone, ChromaDB, OpenSearch), Amazon Bedrock, and open-source LLMs (Llama 3, Mistral). Committed to responsible AI practices and production-grade guardrails in regulated finance and healthcare environments. AWS Certified Generative AI Developer – Professional (in progress).

PROFESSIONAL EXPERIENCE

MetLife (Fortune 50) | *Generative AI Engineer – LLM Optimization & Inference* Sep 2024 – Present · NY

- Optimized LLM inference stack using **vLLM** and **TensorRT-LLM**, cutting p95 response latency by **42%** while preserving output consistency across 3B-parameter model variants in production.
- Deployed quantized **INT8 GPTQ** pipelines with KV-cache streaming and head pruning, reducing GPU memory footprint by **35%** on NVIDIA A100 clusters — enabling 2× model concurrency at the same hardware cost and deferring an estimated **$500K in additional GPU hardware procurement**.
- Integrated **DeepSpeed MII** for multi-GPU orchestration via NCCL and CUDA Graphs, achieving near-linear scalability across 8-GPU nodes.
- Engineered dynamic batching in **Triton Inference Server**, increasing GPU utilization from ~58% to **83%** under peak load — equivalent to extracting the compute value of ~3 additional A100s from existing hardware, avoiding an estimated **$120K+ in annual cloud GPU spend**, with no degradation to p95 latency SLAs.
- Converted fine-tuned models to **ONNX Runtime** for edge deployments, achieving **1.7× token throughput** vs. baseline PyTorch — enabling real-time inference at the edge while reducing per-inference compute cost by an estimated **~40%**, eliminating the need for cloud GPU calls on latency-sensitive workloads.
- Architected **RAG pipelines** using LangChain and vector databases (Pinecone, ChromaDB), grounding LLM responses in enterprise knowledge bases with measurable hallucination reduction.
- Built production-grade inference monitoring with **Prometheus + CloudWatch**, enabling real-time latency alerting and automated horizontal scaling triggers.
- Automated CI/CD via **Docker + GitHub Actions**, compressing release cycles from weekly to daily and improving deployment reliability by **50%**.
- Led GPU performance investigations using **Nsight Systems**, diagnosing memory fragmentation bottlenecks and reducing inference variance by 27%.

Sage Softtech | *Machine Learning Engineer* Mar 2021 – Jul 2023 · INDIA

- Engineered production ML system on **50M+ EHR records** — one of the largest healthcare datasets in the project's history — designing a patient readmission risk engine (Random Forest + XGBoost) that improved high-risk detection accuracy by **18%**, directly enabling earlier clinical intervention for at-risk patients.
- Built **clinical NLP pipeline** using spaCy NER to extract comorbidities, medication patterns, and diagnosis codes from unstructured physician notes — boosting predictive AUC by **0.09** (a gain equivalent to labeling 3M+ additional training records), demonstrating the same text-understanding fundamentals now core to modern LLM applications.
- Owned end-to-end ML lifecycle: designed **Airflow-orchestrated pipelines** for automated ingestion, feature engineering, and model retraining, compressing retraining cycles from **4 hours** → **45 minutes** — a 5× speedup that enabled weekly model refreshes without engineering intervention.
- Deployed **production inference APIs (FastAPI + Docker on AWS EC2)** serving real-time readmission risk scores to **200+ physicians daily** — a live, high-stakes clinical decision support system operating under strict latency and availability requirements.

- Implemented **MLflow + Prometheus** for full model lifecycle governance: experiment tracking, performance versioning, and automated data drift detection — ensuring regulatory-grade reproducibility in a compliant environment.
- Operated within a **regulated healthcare environment** (data governance, audit trails) — the same risk and compliance mindset now applied to financial AI systems at MetLife.

TECHNICAL SKILLS

GenAI & LLMs	vLLM, SGLang, TensorRT-LLM, DeepSpeed MII, ONNX Runtime, Triton Server, Amazon Bedrock, OpenAI API, Llama 3, Mistral, Mixtral, LangChain, LangGraph, LlamaIndex, RAG, Prompt Engineering, Function Calling, Structured Outputs, LoRA/QLoRA Fine-tuning, RLHF, PEFT, Conversational AI
Agentic AI	LangGraph, AutoGen, ReAct, ReWoo, Multi-Agent Orchestration, Tool Use, LangSmith, AI Guardrails, Responsible AI, Context Window Optimization
Vector & Search	Pinecone, ChromaDB, Weaviate, OpenSearch, pgvector, Embedding Pipelines, Semantic Search, Hybrid Search
ML & DL	PyTorch, TensorFlow, XGBoost, Random Forest, LSTM, ResNet, Anomaly Detection, Time Series, NLP, spaCy, Hugging Face Transformers
GPU & Perf	CUDA, NCCL, Nsight Systems, INT8 Quantization, KV-Cache Optimization, Dynamic Batching, GPU Profiling, A100/H100
MLOps & Cloud	MLflow, Airflow, Docker, Kubernetes (K8s), GitHub Actions, AWS (EC2, SageMaker, Bedrock, CloudWatch), Azure AI, GCP, FastAPI, Prometheus, Weights & Biases, RAGAS, DeepEval
Languages	Python, C, C++, Rust, SQL, Scala, Spark, R, Pandas, Keras

SUNY Albany | MS *Data Analytics — Full-Time Study & Independent AI Research* Jan 2024 – Dec 2025 · Albany, NY

Completed Master of Science in Data Analytics full-time. Conducted independent research on LLM inference optimization and RAG pipeline design, building foundational expertise directly applied at MetLife.

Coursework: Deep Learning, Distributed Systems, Statistical Computing, Machine Learning.

PROJECTS

Enterprise RAG Chatbot with Agentic Workflows | LangChain · *LangGraph · LlamaIndex · Amazon Bedrock · ChromaDB · LangSmith · AWS Lambda*

- Architected multi-agent conversational AI system using LangGraph (ReAct pattern) and Amazon Bedrock (Llama 3 / Claude 3), grounding responses in a 100K+ document enterprise corpus via ChromaDB with semantic + hybrid search chunking strategy.
- Implemented **AI Guardrails and responsible AI controls** using Amazon Bedrock Guardrails for content filtering, PII redaction, and hallucination mitigation — meeting enterprise compliance requirements.
- Integrated **LangSmith observability** for full prompt/response tracing, latency profiling, and automated regression testing of retrieval quality; measured with RAGAS achieving 91% faithfulness score.
- Deployed as serverless API on AWS Lambda + API Gateway with streaming response and function calling/tool use support, handling 500+ concurrent requests at <800ms p95 latency.

Fraud Detection System — Real-Time Anomaly Scoring | Python · *Isolation Forest · Autoencoders · XGBoost · AWS SageMaker · CloudWatch*

- Built fraud detection pipeline on highly imbalanced financial transaction data using Isolation Forest and Autoencoder-based anomaly detection, achieving **94% recall** on fraudulent transactions.
- Applied SMOTE and threshold optimization for class imbalance; deployed as real-time scoring API on SageMaker with CloudWatch alerting for sub-second high-risk transaction flagging.

Multi-Model Time Series Demand Forecasting | Python · *LSTM · Prophet · ARIMA · Apache Airflow · AWS*

- Developed ensemble forecasting system comparing ARIMA, Prophet, and LSTM, achieving **12% lower MAPE** vs. baseline; enabled 30-day ahead supply chain forecasts with automated seasonality detection.

EDUCATION & CERTIFICATIONS

State University of New York, Albany Graduated 2025

Master of Science in Data Analytics · Coursework: Machine Learning, Deep Learning, Distributed Systems, Statistical Computing

Certifications (Completed / In Progress)

- AWS Certified Machine Learning Engineer – Associate (MLA-C01) (In Progress)
- AWS Certified AI Practitioner (AIF-C01)
- NVIDIA Deep Learning Institute – Accelerating LLM Inference (In Progress)
- NVIDIA Deep Learning Institute – AI for All: From Basics to GenAI Practice