SRINIVAS GAMPASANI

Gen AI Engineer | ML Engineer | LLM Engineer | Data Engineer | Data Scientist | MLOps Engineer | NLP Engineer
USA | 3145483799 | srinivasgampasani7@gmail.com | linkedin.com/in/srinivasgampasani | srinivas-gampasani.github.io

PROFESSIONAL SUMMARY

Results-driven AI Engineer, Machine Learning Engineer, and Data Engineer with 3+ years of experience designing and deploying production-grade Generative AI systems, Machine Learning (ML) pipelines, and scalable data engineering solutions across healthcare and enterprise domains.

Expert in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), GraphRAG, Agentic AI, multi-agent orchestration (LangChain, LangGraph, LlamaIndex, CrewAI, AutoGen, MCP), LLM fine-tuning (LoRA, QLoRA, PEFT), RLHF, DPO, instruction tuning, hallucination mitigation, and LLM evaluation (RAGAS, TruLens).

Deep expertise in Natural Language Processing (NLP) including BERT, Transformers, HuggingFace, NER, and text classification, and Computer Vision including CNNs, Vision Transformers (ViT), CLIP, Diffusion Models, SAM, and OpenCV.

Proficient in Deep Learning frameworks (PyTorch, TensorFlow, JAX), classical Machine Learning (Scikit-learn, XGBoost, LightGBM), statistical modeling, time series, and anomaly detection. Skilled in building scalable ETL/ELT pipelines using Apache Spark, PySpark, Kafka, Airflow, and dbt, and data warehousing on Snowflake, Databricks, BigQuery, and AWS Redshift.

Experienced in end-to-end MLOps and LLMOps (MLflow, Docker, Kubernetes, CI/CD, ArgoCD, Kubeflow) and cloud-native deployment on AWS SageMaker, Amazon Bedrock, Azure ML Studio, Azure OpenAI, and GCP Vertex AI. Proven impact: 40% accuracy improvement, 60% compute cost reduction, 35% hallucination reduction, 30% faster deployment cycles, and 25% reduction in data processing time. Committed to Responsible AI, model explainability (SHAP, LIME), AI governance, fairness, HIPAA compliance, and data quality.

TECHNICAL SKILLS

Programming Languages: Python, SQL, PySpark, Java, Scala, TypeScript, JavaScript, R, Go, Bash

Generative AI and LLMs: Large Language Models (LLMs), GPT-4, GPT-4 Turbo, LLaMA, Gemini, Claude, Mixtral, Retrieval-Augmented Generation (RAG), GraphRAG, LLM Fine-Tuning (LoRA, QLoRA, PEFT), Instruction Tuning, RLHF, DPO, IPO, Prompt Engineering, Prompt Chaining, Chain-of-Thought Prompting, Tool Calling, Function Calling, OpenAI API, Azure OpenAI, AWS Bedrock, DSPy

Agentic AI and Agent Frameworks: LangChain, LangGraph, LlamaIndex, CrewAI, AutoGen, MCP (Model Context Protocol), A2A, Google ADK, Agent Space, Agent Builder, OpenDevin, Multi-Agent Orchestration, Supervisor-Worker Patterns, Human-in-the-Loop, Long-Term Memory, Autonomous Task Decomposition

LLM Evaluation and Responsible AI: RAGAS, TruLens, DeepEval, Guardrails AI, Nemo Guardrails, Hallucination Mitigation, Model Governance, Prompt Versioning, Preference Optimization, SHAP, LIME, Fairness and Bias Mitigation, AI Safety, Jailbreak Resistance, Responsible AI, AI Governance, HIPAA Compliance, Explainable AI

NLP and Computer Vision: Transformers, BERT, GPT, Sentence Transformers, HuggingFace Ecosystem, NER, Text Classification, Summarization, Topic Modeling, Semantic Search, Hybrid Search, BM25, Reranking, Embeddings, spaCy, OCR, CNNs, Vision Transformers (ViT), OpenCV, CLIP, Diffusion Models, SAM (Segment Anything), DINO, GroundingDINO, YOLO, Faster R-CNN, Object Detection, Image Classification, Image Captioning, Multimodal Models

Machine Learning and Deep Learning: PyTorch, TensorFlow, JAX, Scikit-learn, XGBoost, LightGBM, Random Forest, Neural Networks, LSTMs, Reinforcement Learning (RL), Statistical Modeling, Bayesian Methods, Time Series, Anomaly Detection, Graph ML, Feature Engineering, Hyperparameter Tuning, A/B Testing, Drift Detection, Ensemble Methods, Quantization, Distillation, Pruning, ONNX, vLLM, Weights and Biases

MLOps and LLMOps: MLflow, ClearML, Kubeflow, DVC, Docker, Kubernetes, Helm, ArgoCD, GitOps, CI/CD, Jenkins, GitHub Actions, FastAPI, Flask, RESTful APIs, gRPC, GraphQL, Model Monitoring, Model Versioning, IaC (Terraform, CDK), CloudFormation, OpenTelemetry, Prometheus, Grafana, Datadog, Agent Tracing, LLM Observability, Fault Tolerance, Logging and Alerting

Data Engineering and Big Data: Apache Spark, Apache Kafka, Apache Airflow, Apache NiFi, Apache Hadoop, dbt (Data Build Tool), Delta Lake, Apache Iceberg, ETL/ELT Pipelines, Data Pipeline, Data Transformation, Feature Stores, Data Lakes, Data Modeling, Data Governance, Distributed Computing, Microservices, Trino, ClickHouse

Data Warehousing and Databases: Snowflake, Databricks, BigQuery, AWS Redshift, PostgreSQL, MongoDB, Neo4j, Redis, Elasticsearch, FAISS, Pinecone, Weaviate, Milvus, Chroma, Azure AI Search, MinIO, NoSQL

Cloud and AI Services: AWS (SageMaker, Amazon Bedrock, Glue, Athena, S3, Lambda, EKS, EC2, EMR, Step Functions, Redshift), Azure (ML Studio, OpenAI, AI Search, Databricks, Data Factory, Copilot Studio, Power Platform), GCP (Vertex AI, BigQuery, ADK, Agent Space, Cloud Composer, Cloud Storage), IAM, Security and Encryption, Cost Optimization

Business Intelligence and Visualization: Power BI, Tableau, SSRS, Plotly, Streamlit, Matplotlib, Dashboard Development, KPI Monitoring, Advanced Analytics, Data Visualization, Business Analytics, Regulatory Compliance

PROFESSIONAL EXPERIENCE

Generative AI Engineer and Data Scientist | Ascension Via Christi Health | Wichita, KS | 09/2024 - Present

- Architected production-grade RAG and GraphRAG pipelines using GPT-4, LLaMA, and Gemini via LangChain, LlamaIndex, Azure OpenAI, and Azure AI Search with hybrid search (dense plus sparse plus reranking), raising answer relevance by 40% and enabling real-time AI decision support for 500+ clinicians.
- Engineered multi-agent Agentic AI systems using LangGraph and Deep Agent Frameworks with chain-of-thought prompting, autonomous tool calling, MCP integration, long-term memory, Guardrails AI, and Nemo Guardrails output moderation, cutting LLM hallucinations by 35% and enforcing Responsible AI.
- Designed NLP pipelines for clinical summarization, NER, text classification, and structured extraction from 50,000+ patient records using BERT, Transformers, HuggingFace, spaCy, and Apache Spark, applying OCR and Knowledge Graphs (Neo4j) to reduce manual documentation effort by 30%.
- Applied LLM fine-tuning (LoRA, QLoRA, PEFT), instruction tuning, RLHF, DPO preference optimization, quantization, and distillation for clinical domain adaptation and evaluated fine-tuned models with RAGAS and TruLens, reducing compute cost by 60% via vLLM optimized inference.
- Engineered scalable ETL/ELT data pipelines processing 50,000+ clinical records using Apache Spark, PySpark, SQL, Snowflake, and Delta Lake, reducing data processing time by 30% through automated validation and monitoring.
- Implemented end-to-end MLOps and LLMOps pipelines with MLflow, Docker, Kubernetes (EKS), Helm, ArgoCD, CI/CD (Jenkins, GitHub Actions), and OpenTelemetry on AWS SageMaker and Azure ML Studio achieving zero-downtime deployments and 30% fewer release errors.
- Applied SHAP and LIME for model explainability and enforced HIPAA-compliant OAuth2 and RBAC data governance across all clinical data systems, and designed Power BI and Streamlit dashboards for KPI monitoring and executive stakeholder reporting.

Machine Learning Engineer and Data Engineer | Colgate-Palmolive | Topeka, KS | 12/2023 - 08/2024

- Built end-to-end Machine Learning pipelines for demand forecasting and predictive analytics using XGBoost, LightGBM, Random Forest, and Neural Networks (PyTorch, TensorFlow), achieving 20% prediction accuracy improvement through feature engineering, hyperparameter tuning, A/B testing, and ensemble modeling.
- Built NLP and text analytics models using Transformers, Sentence Transformers, and BERT for market sentiment analysis, semantic search, NER, topic modeling, and summarization to extract consumer insights from unstructured data.
- Designed and maintained scalable ETL/ELT pipelines using Apache Airflow, Apache Spark, PySpark, and dbt, integrated Snowflake data warehousing and Delta Lake, reducing data processing time by 25% and ensuring reliable data availability for downstream ML models.
- Deployed ML models as RESTful APIs via FastAPI and Docker with real-time Kafka and PySpark streaming inference pipelines, implemented model monitoring, drift detection, and automated retraining workflows using MLflow and CI/CD pipelines.
- Applied SHAP and LIME for model explainability, built Prometheus and Grafana observability dashboards cutting deployment time by 20%, and delivered Power BI and Tableau insights for 10+ global markets in Agile sprint cycles.

Machine Learning Engineer | Maxton Technology Pvt. Ltd., | Bangalore, India | 08/2022 - 08/2023

- Developed supervised and unsupervised Machine Learning models (Scikit-learn, XGBoost, CNNs, LSTMs) for demand forecasting, anomaly detection, and clustering, improving model accuracy by 15% through iterative feature engineering, statistical modeling, Bayesian methods, and hyperparameter optimization.
- Applied Computer Vision techniques using OpenCV, CNNs, Vision Transformers (ViT), CLIP, SAM, and OCR for image classification, object detection, document intelligence, and automated data extraction workflows.

- Built NLP models for text classification, NER, topic modeling, and sentiment analysis, and deployed production models via Flask, Docker, and RESTful APIs on GCP Vertex AI with statistical cross-validation (AUC, RMSE, F1).
- Designed scalable ETL/ELT pipelines using Python, PySpark, and Apache Spark on GCP (Vertex AI, BigQuery, Cloud Storage), reducing pipeline latency by 25%, and ensured data integrity through automated validation, data quality checks, and version-controlled codebases in Agile environments.

FEATURED PROJECTS

Enterprise AI Knowledge Retrieval System (RAG + Agentic AI)

- Architected production RAG and GraphRAG system for 100,000+ enterprise documents with hybrid semantic and sparse retrieval, cross-encoder reranking, chain-of-thought reasoning, and Knowledge Graphs (Neo4j), achieving sub-second latency, 35% accuracy improvement, and 99.9% uptime.
- Built LangGraph multi-agent Agentic AI with MCP tool integration, supervisor-worker patterns, autonomous task decomposition, human-in-the-loop controls, and Guardrails AI, deployed via FastAPI, Kubernetes (EKS), ArgoCD, GitOps, OpenTelemetry, and Terraform IaC on AWS and Azure.
- **Tech Stack:** LangChain, LangGraph, FAISS, Pinecone, Weaviate, Azure OpenAI, GraphRAG, Neo4j, FastAPI, Docker, Kubernetes, Terraform, OpenTelemetry

LLM Fine-Tuning and Multi-Agent Orchestration Pipeline

- Built end-to-end LLM fine-tuning pipeline with LoRA, QLoRA, PEFT, RLHF, DPO preference optimization, quantization, and distillation, reducing compute cost by 60%, evaluated with RAGAS and TruLens, and tracked via MLflow LLMOps with drift detection and automated retraining triggers.
- Designed multi-agent orchestration framework using LangGraph, CrewAI, and AutoGen with tool calling, session-aware memory, fault tolerance, and human-in-the-loop, deployed via vLLM optimized inference, Datadog observability, CI/CD pipelines, and Docker and Kubernetes production services.
- **Tech Stack:** LoRA, QLoRA, PEFT, RLHF, DPO, PyTorch, HuggingFace, LangGraph, CrewAI, AutoGen, RAGAS, TruLens, MLflow, vLLM, Datadog

Demand Forecasting ML System and Full MLOps Pipeline

- Built scalable demand forecasting engine using XGBoost, Neural Networks (PyTorch), and ensemble methods with automated feature engineering, A/B testing, and drift-triggered retraining, achieving 20% prediction accuracy improvement deployed on AWS SageMaker as RESTful APIs.
- Implemented full MLOps pipeline with MLflow, Apache Airflow, dbt, Docker, Kubernetes, CI/CD (GitHub Actions), real-time Kafka streaming for feature generation, Snowflake feature store architecture, and Power BI dashboards for SLA and pipeline monitoring.
- **Tech Stack:** XGBoost, PyTorch, Apache Airflow, MLflow, Snowflake, Kafka, AWS SageMaker, Docker, Kubernetes, Power BI

Real-Time Anomaly Detection and Streaming ML System

- Engineered streaming anomaly detection pipeline using Apache Kafka and Isolation Forest for real-time enterprise data monitoring achieving 95%+ precision, integrated feature engineering, automated drift detection, and MLflow experiment tracking for continuous retraining.
- Deployed scalable inference APIs with FastAPI and Docker on AWS, orchestrated data workflows using Apache Airflow, implemented model versioning and A/B testing, and built Prometheus and Grafana observability dashboards enabling fully automated ML lifecycle management with 99.9% uptime.
- **Tech Stack:** Kafka, Isolation Forest, Scikit-learn, PyTorch, FastAPI, Docker, MLflow, Apache Airflow, Grafana, AWS

EDUCATION

Master of Science in Artificial Intelligence | Saint Louis University, St. Louis, MO | 08/2023 - 12/2025

Bachelor of Technology in Computer Science | Godavari Global University, India | 08/2019 - 05/2023

CERTIFICATIONS

- AWS Certified Data Analytics
- Future Skills Prime Certified Artificial Intelligence, Data Science and Natural Language Processing (NLP)
- Coursera Certified Machine Learning