SIVA NAGA RAJU PARIMISETTY

Senior AI-ML Engineer

parimisettysivanagaraju@gmail.com
+1-267-309-1872
Linkedin

Career Objective

Senior AI-ML Engineer with 9+ years of experience designing, building, and deploying scalable AI/ML solutions. Expertise in Generative AI, LLM pipelines, NLP, MLOps, and end-to-end model lifecycle management. Proven ability to deliver production-grade AI systems, optimize inference performance, and ensure enterprise-grade security and compliance across cloud platforms.

Professional Summary

- Senior **AI-ML Engineer** with 9+ years of experience designing, building, and deploying production-grade AI/ML solutions.
- Expertise in **Generative AI, Large Language Models (LLMs)**, and **end-to-end model lifecycle management** from data ingestion to deployment and monitoring.
- Proficient in **LLM engineering**, including **fine-tuning** (LLaMA2, Mistral), **prompt engineering**, and **attention mechanism optimization** for domain-specific applications.
- Skilled in building and optimizing Retrieval-Augmented Generation (RAG) pipelines using **LangChain, FAISS**, and **Pinecone** for high-precision knowledge retrieval.
- Extensive experience in **Natural Language Processing (NLP)** with **spaCy, NLTK**, and **LangChain** for tasks like sentiment analysis, NER, and topic modeling.
- Advanced knowledge of **MLOps tools** including **MLflow, DVC, Airflow**, and **Azure ML Pipelines** for reproducible, automated, and scalable workflows.
- Proven ability in deploying and scaling ML models as **microservices** using **FastAPI, TensorFlow Serving, TorchServe, KServe,** and **Triton Inference Server**.
- Strong background in **containerization** and **orchestration** with **Docker, Kubernetes, Helm**, and **AKS** for resilient, cloud-native deployments.
- Experience designing and managing secure, compliant AI infrastructure on **Azure (ML, OpenAI, AKS)** and **AWS (EC2, S3)** with enterprise governance.
- Adept at implementing **CI/CD pipelines**, **automated retraining**, and **model versioning** using **GitHub Actions, Azure DevOps**, and **Terraform**.
- Skilled in **performance monitoring, observability**, and **drift detection** using **Prometheus, Grafana**, and **Evidently AI** for production AI systems.
- Proficient in feature engineering, lifecycle management, and reusable pipeline development with **Pandas**, **NumPy**, and **Feast**.
- Experienced in **distributed training** and **inference optimization** using **Ray** for improved throughput, latency, and GPU utilization.
- Strong ability to translate model outputs into actionable business insights using **visualization tools** like **Tableau, Plotly**, and **Seaborn**.
- Committed to implementing **robust security protocols** including **RBAC, OAuth2**, and **Azure Key Vault** to ensure secure data handling and model deployment.

Technical Skills

- Programming & Scripting: Python, SQL, Pandas, NumPy, Statsmodels
- ML & Deep Learning Frameworks: PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM
- Large Language Models (LLMs) & Generative AI: Hugging Face Transformers, LLaMA2, Mistral, Prompt Engineering, Fine-tuning

- NLP & Text Analytics: spaCy, NLTK, LangChain, PromptFlow
- Model Deployment & Serving: FastAPI, TensorFlow Serving, TorchServe, Triton Inference Server, BentoML, KServe
- MLOps & Orchestration: MLflow, DVC, Apache Airflow, Argo Workflows, Azure ML Pipelines
- CI/CD & Automation: GitHub Actions, Azure DevOps, Terraform
- Cloud Platforms (Azure): Azure Machine Learning, Azure OpenAI, Azure Kubernetes Service (AKS), Azure Blob Storage, Azure Functions, Azure Key Vault, Azure Monitor
- Cloud Platforms (AWS): Amazon EC2, Amazon S3
- Containerization & Orchestration: Docker, Kubernetes, Helm
- Vector Databases & Search: FAISS, Pinecone
- Monitoring & Observability: Prometheus, Grafana, Evidently AI
- Data Visualization & BI: Tableau, Plotly, Seaborn
- Feature Store & Management: Feast
- Security & Governance: OAuth2, HashiCorp Vault, Role-Based Access Control (RBAC)

Experience

Senior AI-ML Engineer | Morgan Stanley, New York, NY

Jan, 2025 to Present

Enterprise LLM & RAG Platform for Financial Knowledge Automation
Designed and implemented a secure, scalable enterprise platform leveraging Large Language Models and Retrieval-Augmented Generation to automate financial knowledge extraction, insight generation, and context-aware querying for high-precision analytics and decision support.

- Engineered end-to-end **Retrieval-Augmented Generation (RAG)** pipelines using **LangChain** and **Azure OpenAI LLMs** to enable high-precision financial knowledge automation and context-aware query responses.
- Fine-tuned and optimized **LLaMA2** and **Mistral models** using **Hugging Face Transformers**, **incorporating domain-specific tokenization** and **attention mechanisms** to improve accuracy for financial NLP tasks.
- Architected and deployed scalable vector similarity search systems using **FAISS** and **Pinecone**, implementing index sharding and approximate nearest neighbor search for high-speed retrieval across large financial document sets.
- Deployed production-grade LLM microservices on **Azure Kubernetes Service (AKS)** using **KServe**, **configuring GPU scheduling**, **horizontal pod autoscaling**, and **resource affinity** to meet enterprise SLAs.
- Automated multi-stage LLM workflows with **Apache Airflow**, **orchestrating document ingestion**, **preprocessing**, **embedding generation**, **model fine-tuning**, and **secure deployment** across environments.
- Implemented experiment tracking, model versioning, and dataset management using **MLflow** and **DVC** to ensure reproducibility and structured comparison across **LLM** variants.
- Optimized distributed LLM inference using **Ray, implementing memory-efficient tensor storage, asynchronous batching**, and **parallel actor pools** to maximize throughput and minimize latency.
- Developed systematic prompt engineering and testing pipelines with **PromptFlow** to iteratively refine prompts and optimize generation quality across multiple **LLM models**.
- Standardized **model serving APIs** with **BentoML**, ensuring consistent endpoints, request validation, logging, and **metadata** exposure for secure enterprise integration.
- Secured model secrets, API keys, and credentials using **Azure Key Vault** and **RBAC policies**, enforcing enterprise compliance and access control across all ML services.
- Built **comprehensive monitoring** and **observability frameworks** with **Prometheus**, **Grafana**, and **Evidently AI** to track inference latency, GPU utilization, model drift, and embedding quality.
- Orchestrated secure document ingestion and vectorization pipelines from **Azure Blob Storage**, enabling efficient batch embedding generation and vector database indexing for **RAG workflows**.
- Designed fallback and ensemble strategies for multi-**LLM deployments**, implementing **dynamic query routing** and **versioned models** to ensure system reliability and contextual accuracy.
- Collaborated with cross-functional teams to integrate **LLM-driven insights** into secure financial dashboards, knowledge bases, and compliance systems, ensuring usability and operational alignment.

Environment: Python, LangChain, Azure OpenAI, Hugging Face Transformers, LLaMA2, Mistral, PromptFlow, FAISS, Pinecone, KServe, Ray, Airflow, MLflow, DVC, BentoML, Evidently AI, Prometheus, Grafana, Azure ML, AKS,

Azure Functions, Azure Blob Storage, Azure Key Vault, Azure Monitor, GitHub Actions, Terraform, Kubernetes, Helm

Senior ML Engineer | United Health Care, Edina, MN

Aug, 2023 to Dec, 2024

Enterprise Healthcare Predictive Analytics Platform

Designed and deployed an end-to-end predictive analytics platform for patient risk modeling and clinical decision support, leveraging machine learning on structured and unstructured healthcare data while ensuring HIPAA compliance and enterprise-grade security.

- Engineered predictive patient risk models using **PyTorch** and **TensorFlow**, integrating **multi-modal EHR data** and **unstructured clinical notes** to support proactive interventions and risk stratification.
- Orchestrated **end-to-end ML pipelines** using **Apache Airflow** and **Argo Workflows**, **automating data ingestion**, **feature engineering**, **model training**, **evaluation**, and **deployment** across secure environments.
- Implemented centralized feature management and serving using **Feast, ensuring consistent**, **high-fidelity feature inputs** for training and production inference while enabling feature reuse across models.
- Containerized **ML models** with **Docker** and deployed on **Kubernetes/AKS** using **Helm, enabling scalable, resilient**, and **GPU-optimized inference** in a cloud-native environment.
- Deployed low-latency inference APIs using **TensorFlow Serving**, **TorchServe**, and **Triton Inference Server** to provide real-time predictive capabilities for clinical applications.
- Secured patient data, model artifacts, and API credentials using **Azure Key Vault**, **HashiCorp Vault**, and **OAuth2**, ensuring compliance with **HIPAA** and enterprise security standards.
- Version-controlled datasets, model artifacts, and experiments using **DVC** and **MLflow** to ensure reproducibility, auditability, and traceability across all **ML workflows**.
- Designed and automated **CI/CD pipelines** using **GitHub Actions** and **Terraform**, streamlining model promotion from development to production while maintaining governance compliance.
- Optimized **GPU utilization**, **inference batching**, and **model throughput** on **AKS clusters** to ensure low-latency predictions and cost-efficient resource allocation at scale.
- Developed model interpretability dashboards using **SHAP** and **LIME**, enabling clinical teams to understand feature contributions and build trust in ML-driven healthcare decisions.
- Built and maintained data ingestion and preprocessing pipelines in **Python**, **processing structured EHR** and **unstructured clinical notes** from **Azure Blob Storage** for downstream modeling.
- Implemented automated retraining workflows with **Azure ML Pipelines** and **Azure Functions** to ensure models continuously adapt to evolving patient data and maintain predictive accuracy.
- Monitored production models with **Azure Monitor**, **Prometheus**, and **Grafana**, proactively detecting performance drift, anomalies, and infrastructure issues to ensure system reliability.
- Collaborated cross-functionally with data engineers, clinicians, and compliance teams to integrate **ML models** securely into enterprise healthcare applications while adhering to governance standards.

Environment: Python, PyTorch, TensorFlow, MLflow, DVC, Airflow, Feast, Kubernetes, Docker, Helm, Terraform, TensorFlow Serving, TorchServe, Triton Inference Server, Argo Workflows, Azure ML, AKS, Azure Blob Storage, Azure Functions, Azure Key Vault, Azure Monitor, SHAP, LIME, OAuth2, HashiCorp Vault, Git, GitHub Actions

ML Engineer | Databricks, San Francisco, CA

Nov, 2021 to Jul, 2023

Enterprise ML Platform & Predictive Analytics Pipelines

Developed and deployed scalable, end-to-end machine learning pipelines on the Databricks platform to support enterprise-wide predictive analytics. Focused on building reproducible, high-performance models, automating MLOps workflows, and enabling secure, real-time inference for business applications.

- Developed end-to-end ML pipelines using **PyTorch**, **TensorFlow**, and **XGBoost** to deliver high-accuracy predictive models for enterprise datasets, ensuring scalability and performance.
- Built and deployed **FastAPI microservices** on **Azure Kubernetes Service (AKS)** to provide scalable, high-availability endpoints for real-time model inference across enterprise workloads.
- Orchestrated **automated ML** workflows using **Apache Airflow**, **managing training, validation**, and **deployment pipelines** across multiple environments for operational consistency.

- Implemented **experiment tracking, model versioning**, and **dataset management** with **MLflow** and **DVC**, enabling reproducible ML lifecycle management and structured model comparisons.
- Containerized **ML models** using **Docker** and **managed deployments** with **Helm charts** on **Kubernetes/AKS** to ensure scalability, resiliency, and efficient resource utilization.
- Designed and automated **CI/CD pipelines** using **Azure DevOps** and **GitHub Actions** to streamline model build, testing, and deployment while enforcing governance and reproducibility.
- Integrated **Azure ML Pipelines** and **Azure Functions** to automate retraining workflows and batch inference processes for large-scale datasets, ensuring continuous model performance.
- Managed secure storage of datasets and model artifacts using **Azure Blob Storage** and **Azure Key Vault**, enforcing enterprise-grade security, access control, and compliance.
- Deployed production inference services using **TensorFlow Serving** and **TorchServe**, delivering low-latency, high-throughput API responses for enterprise ML applications.
- Implemented comprehensive monitoring dashboards using **Prometheus** and **Grafana** to track model performance, inference latency, GPU utilization, and detect model drift.
- Applied model interpretability tools **SHAP** and **LIME** to provide transparent, actionable insights to stakeholders and support regulatory and compliance requirements.
- Optimized **multi-model GPU workloads** and **container orchestration** strategies to improve inference throughput, latency, and cost-efficiency in production environments.
- Conducted **feature engineering** and **data preprocessing** using **Python**, **Pandas**, and **NumPy** to ensure data quality, consistency, and reliability across all ML experiments and pipelines.
- Collaborated with data engineers, product teams, and business stakeholders to translate business requirements into **technical ML solutions** and **integrate models** into operational workflows.

Environment: Python, PyTorch, TensorFlow, XGBoost, scikit-learn, FastAPI, Docker, Kubernetes, AKS, Helm, Terraform, Apache Airflow, MLflow, DVC, Azure ML, Azure ML Pipelines, Azure Functions, Azure Blob Storage, Azure Key Vault, Azure DevOps, TensorFlow Serving, TorchServe, Prometheus, Grafana, SHAP, LIME, GitHub Actions, JupyterLab

Data Scientist | Fannie Mae, USA

Jul, 2018 to Jun, 2021

Predictive Analytics & Business Forecasting Platform

Developed and deployed scalable machine learning models to forecast key business KPIs, optimize operational processes, and support data-driven decision-making. Focused on model interpretability, production deployment, and integrating predictive insights into business workflows.

- Developed predictive models using **XGBoost**, **LightGBM**, and **scikit-learn** to forecast critical business KPIs, enhancing operational efficiency and strategic planning.
- Built and deployed Flask-based **REST APIs** to serve production-ready **ML models** on **AWS EC2 instances**, enabling real-time inference and integration with business applications.
- Designed and maintained reusable feature engineering pipelines using **Pandas** and **NumPy** to ensure consistent, scalable data transformations across multiple projects.
- Implemented **NLP preprocessing** and **text analytics pipelines** using **spaCy** and **NLTK** for tasks such as document classification, sentiment analysis, and entity recognition.
- Containerized machine learning applications using **Docker** to standardize deployment environments across development, staging, and production.
- Managed end-to-end data workflows, including ingestion, storage, and retrieval from **AWS S3** and **relational databases (MySQL, PostgreSQL)** to support training and inference pipelines.
- Conducted statistical modeling and time series analysis using **Statsmodels** to identify trends, support business forecasting, and inform decision-making.
- Developed **interactive data visualizations** and **dashboards** using **Seaborn** and **Plotly** to communicate model insights, performance metrics, and business impacts to stakeholders.
- Applied **hyperparameter tuning** and **cross-validation strategies** to optimize gradient boosting models, improving prediction accuracy, robustness, and generalization.
- Implemented version control and collaborative development workflows using **Git** and **GitHub** to track experiments, manage codebases, and ensure reproducibility.
- Validated **model performance** across diverse datasets and production environments to ensure consistency, reliability, and alignment with business objectives.

- Collaborated with cross-functional teams including business analysts, engineers, and product managers to translate requirements into technical **ML** solutions.
- Monitored model **performance post-deployment** and conducted **periodic retraining** to maintain accuracy and adapt to evolving data patterns.
- Ensured compliance with **data governance** and **security standards** throughout the **model development** and **deployment lifecycle**.

Environment: Python, scikit-learn, XGBoost, LightGBM, Statsmodels, Pandas, NumPy, Flask, Docker, Seaborn, Plotly, AWS S3, AWS EC2, Git, GitHub, spaCy, NLTK, MySQL, PostgreSQL, Jupyter

Python Developer | Allerin Tech, Mumbai, India

Mar, 2016 to Jun, 2018

Enterprise ETL & Business Intelligence Automation Platform

Developed and maintained Python-based ETL pipelines, predictive models, and interactive dashboards to streamline data processing, enhance reporting accuracy, and deliver actionable business insights across multiple enterprise domains.

- Developed robust **ETL pipelines** using **Python**, **Pandas**, and **NumPy** to preprocess, clean, and transform multi-source enterprise datasets for **BI reporting** and **analytics**.
- Built predictive models using **scikit-learn** to generate business insights and forecasting outputs, integrated into **interactive dashboard visualizations** for stakeholder review.
- Optimized complex **SQL queries** in **MySQL** and **PostgreSQL** to enhance data retrieval performance and efficiency for large-scale datasets and analytical workflows.
- Automated data aggregation and transformation processes using **Python scripting**, reducing manual reporting efforts and improving data delivery timelines by over 30%.
- Designed and implemented data validation checks and consistency rules within **ETL pipelines** to ensure high-quality, reliable datasets for **downstream BI applications**.
- Developed interactive **Tableau dashboards** using **processed datasets** to provide stakeholders with real-time, actionable business insights and performance metrics.
- Created reusable **Python modules** for feature engineering and data preprocessing, standardizing workflows across multiple projects and improving development efficiency.
- Implemented **version control** and **collaborative coding practices** using **Git**, maintaining reproducibility and organization across codebases and analytical scripts.
- Conducted exploratory data analysis using **NumPy** and **Pandas**, producing statistical summaries and trend reports to support management decision-making.
- Integrated structured database outputs seamlessly into **end-to-end BI** and reporting workflows, enabling automated, scalable analytics delivery.
- Collaborated with cross-functional teams including data analysts, business stakeholders, and IT to gather requirements and deliver tailored data solutions.
- Documented **ETL processes**, **data models**, and **dashboard functionalities** to ensure knowledge transfer and operational continuity.
- Supported the **deployment** and **maintenance** of **BI tools** and **data pipelines** in staging and production environments.
- Contributed to continuous improvement initiatives by identifying bottlenecks in data workflows and implementing performance optimizations.

Environment: Python, Pandas, NumPy, scikit-learn, MySQL, PostgreSQL, Tableau, Git, Jupyter Notebook, Linux, REST APIs

Education

Bachelor of Computer Science & Technology

Gudlavalleru Engineering college, AP, India — Jul, 2012 to May, 2016