Arshitha Ippagunta

AI/ML Engineer
(716) 907-9757 | ippaguntaarshitha@gmail.com | San Francisco, CA | linkedin.com/in/arshitha-ippagunta

Summary

AI/ML Engineer with 5+ years of experience designing, developing, and deploying production-scale Machine Learning, Generative AI, and Agentic AI solutions. Expertise in Python, PyTorch, LLMs, RAG, MLflow, MLOps, distributed inference, Kubernetes, AWS, and cloud-native AI infrastructure for building scalable, low-latency intelligent systems. Proven ability to improve model accuracy, optimize inference performance and cost, and automate the end-to-end ML lifecycle from data engineering to production deployment and monitoring.

Technical Skills

Programming Languages: Python, SQL, Java, Bash
Machine Learning & AI: PyTorch, Scikit-learn, Spark ML, Feature Engineering, Model Training, Model Evaluation, Hyperparameter Optimization, Deep Learning, Transfer Learning, Machine Learning, Artificial Intelligence
Generative AI & LLMs: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Multi-Agent Systems, Agentic AI, Prompt Engineering, Embedding Models, Hybrid Retrieval, Semantic Search, Cross-Encoder Reranking, Vector Search, AI Safety, Guardrails, LLM Fine-Tuning, Model Routing
MLOps & Model Serving: MLflow 3.0, MLflow Model Registry, Model Deployment, vLLM, Ray, Ray Serve, GPU Optimization, Experiment Tracking, Workflow Management, Automation, Scalability
Data Engineering: PySpark, Apache Spark, Apache Airflow, Apache Kafka, Delta Lake, Databricks Workflows, ETL Pipelines, Data Ingestion Pipelines, Feature Store, Unity Catalog
Cloud & Infrastructure: AWS, Amazon EKS, Amazon EC2, Amazon S3, IAM, Docker, Kubernetes, Helm, Terraform, Linux, Cloud-Native Architecture
Backend & Distributed Systems: FastAPI, REST APIs, gRPC, Application Programming Interfaces (APIs), Microservices, Asynchronous Processing, Event-Driven Architecture, Redis, PostgreSQL, Distributed Workflow Orchestration
DevOps, CI/CD & Observability: GitHub Actions, CI/CD, Prometheus, Grafana, Model Monitoring, DevOps
Software Engineering: Software Engineering, Software Development, Systems Engineering, Scripting
Soft Skills: Stakeholder Management, Technical Leadership, System Design, Communication, Agile Development, Leadership, Troubleshooting, Planning, Innovation, Continuous Improvement, Decision Making, Research

Professional Experience

AI/ML Engineer
Databricks | San Francisco, CA — Jun 2025 – Present

- Architected scalable multi-agent AI workflows using Python, PyTorch, MLflow 3.0, and Ray, enabling reliable enterprise task automation supporting over 8,000 enterprise customers across production AI applications.
- Improved response accuracy by 29% through advanced RAG optimization, hybrid retrieval, cross-encoder reranking, prompt engineering, and continuous evaluation pipelines using Python, Vector Search, and MLflow.
- Reduced inference cost by 31% by optimizing enterprise-scale GPU inference with vLLM, Ray Serve, intelligent model routing, dynamic batching, and Kubernetes, while maintaining low-latency production performance.
- Developed enterprise Retrieval-Augmented Generation pipelines using Python, Databricks Vector Search, embedding models, hybrid retrieval, and semantic ranking, delivering grounded responses across large-scale knowledge repositories.
- Engineered scalable LLM serving infrastructure using PyTorch, vLLM, and MLflow 3.0, enabling reliable high-performance GPU inference, automated deployment, and continuous evaluation for enterprise AI applications.
- Engineered model evaluation pipelines using Python and MLflow 3.0, monitoring response quality, hallucination detection, and safety guardrails to improve enterprise AI reliability.
- Architected cloud-native AI backend microservices using FastAPI, REST APIs, Docker, Kubernetes, Helm, AWS EKS, and Amazon S3, enabling secure, highly available, and scalable enterprise inference services.
- Implemented event-driven orchestration pipelines with Ray, Kubernetes Jobs, Redis, and PostgreSQL, enabling reliable asynchronous workflow execution and improving scalability across enterprise AI production systems.
- Implemented secure AI platform infrastructure leveraging Unity Catalog, Delta Lake, Feature Store, Kubernetes, and AWS, enabling governed model deployment and enterprise-scale machine learning operations.
- Operationalized end-to-end MLOps workflows using Python, MLflow 3.0, GitHub Actions, Docker, and Kubernetes, accelerating reliable model deployment, continuous monitoring, and enterprise AI release cycles.

Machine Learning Engineer
Accenture | India Aug 2020 – Jul 2024

- Engineered scalable feature engineering and Spark-based model training pipelines using Python, PySpark, Spark ML, MLflow, and Delta Lake, enabling consistent machine learning workflows across enterprise datasets.
- Improved batch inference throughput by 34% through Spark-based processing, optimized feature engineering, model caching, and Spark execution tuning using Python, PySpark, MLflow, and Databricks.
- Reduced end-to-end model deployment time by 40% through implemented MLflow pipelines, Docker packaging, Kubernetes orchestration, Helm deployments, and GitHub Actions CI/CD workflows.
- Increased production model prediction accuracy by 12% through advanced feature engineering, automated hyperparameter optimization, experiment tracking, and continuous model evaluation using Python, Scikit-learn, and MLflow.
- Developed production-ready machine learning backend microservices using Python, FastAPI, REST APIs, gRPC, Docker, Kubernetes, and Redis, enabling scalable real-time batch and online inference services.
- Engineered scalable data ingestion and feature pipelines using Python, Apache Airflow, Kafka, PySpark, Delta Lake, and Databricks Workflows, ensuring reliable processing of large-scale enterprise datasets.
- Designed cloud-native machine learning infrastructure on AWS using Amazon EKS, S3, EC2, Docker, and Kubernetes, supporting scalable model training and reliable enterprise inference workloads.
- Implemented model deployment pipelines on AWS using Amazon EKS, MLflow Model Registry, GitHub Actions, Terraform, and Kubernetes, improving deployment consistency across enterprise machine learning environments.
- Established secure enterprise machine learning platforms on AWS integrating Amazon S3, IAM, PostgreSQL, Redis, Docker, and Kubernetes, ensuring centralized governance, monitoring, and reliable production operations.
- Implemented comprehensive observability and infrastructure monitoring using Prometheus, Grafana, Fluent Bit, MLflow metrics, Kubernetes, Docker, PostgreSQL, and Redis, enabling proactive performance optimization and reliable production operations.

Projects

Enterprise RAG Knowledge Assistant
Tech Stack: Python, FastAPI, LangChain, LlamaIndex, FAISS, OpenAI/Llama 3, Sentence Transformers, Docker, Kubernetes, MLflow, AWS S3, PostgreSQL

- Architected and developed an enterprise Retrieval-Augmented Generation (RAG) platform using Python, FastAPI, LangChain, LlamaIndex, and FAISS, enabling citation-backed question answering across 50K+ enterprise documents through scalable semantic retrieval and optimized LLM inference.
- Built automated document ingestion, embedding generation, vector indexing, and evaluation pipelines with Sentence Transformers, MLflow, AWS S3, and PostgreSQL, improving retrieval precision by 35%, reducing hallucinations by 42%, and increasing system throughput by 40%.

AI Document Intelligence Platform
Tech Stack: Python, FastAPI, Azure Document Intelligence, Haystack, LangChain, OpenAI, FAISS, PostgreSQL, Redis, Docker, Kubernetes, AWS

- Engineered an AI-powered Document Intelligence platform using OCR, LangChain, Haystack, FAISS, and Large Language Models to extract, classify, index, summarize, and answer questions across 100K+ enterprise documents with over 96% extraction accuracy.
- Developed scalable FastAPI microservices with Redis caching, asynchronous processing, Docker, Kubernetes, and AWS, optimizing hybrid retrieval, document search, and LLM evaluation to reduce document review time by 65%, improve search relevance by 38%, and lower processing latency by 45%.

Certifications

- Databricks Certified Generative AI Engineer Associate
- AWS Certified Machine Learning Engineer – Associate
- AWS Certified Solutions Architect – Associate
- Certified Kubernetes Application Developer (CKAD)

Education

Master of Science in Computer Science & Engineering
University at Buffalo

Bachelor of Technology in Computer Science & Engineering
Sree Vidyanikethan Engineering College