Likith Chilakalapadi Prabhu

AI/ML Engineer
(716) 750-8200 | likithchilakalapadiprabhu@gmail.com | Santa Clara, CA | linkedin.com/in/likith-chilakalapadiprabhu/

Summary

AI/ML Engineer with 5+ years of experience designing and deploying production-scale Generative AI, Agentic AI, LLM, and Machine Learning solutions for enterprise applications. Expertise in multi-agent systems, Retrieval-Augmented Generation (RAG), LLMOps, distributed GPU inference, and scalable AI platforms using Python, PyTorch, Kubernetes, and modern AI frameworks. Proven track record of building reliable, high-performance AI systems that improve response quality, optimize inference efficiency, reduce infrastructure costs, and deliver measurable business value through cross-functional collaboration.

Professional Experience

Nvidia **Mar 2025 – Present**
AI/ML Engineer Santa Clara, CA

- Developed Python-based multi-agent orchestration pipelines using PyTorch, NVIDIA NeMo, LangGraph, MCP, and MLflow, enabling intelligent enterprise workflows for over 35,000 employees while improving response accuracy by 19%.
- Built Retrieval-Augmented Generation pipelines using NeMo Retriever, hybrid search, embeddings, Cross-Encoder reranking, and vector databases, increasing enterprise knowledge retrieval accuracy by 21% across production AI applications.
- Optimized distributed GPU inference using TensorRT-LLM, Triton Inference Server, CUDA, FlashAttention, and dynamic batching, reducing inference latency by 34% while lowering infrastructure costs by 28%.
- • Designed autonomous agent workflows integrating planning, reasoning, memory, evaluation, and guardrails with LangGraph, MLflow Tracing, and prompt optimization for reliable AI task execution across complex business workflows.
- Developed scalable Python microservices using FastAPI, gRPC, Docker, Kubernetes, Redis, and PostgreSQL to power retrieval, orchestration, memory, and evaluation services for production AI platforms.
- Deployed cloud-native AI infrastructure on NVIDIA DGX Cloud using Kubernetes, Helm, and GPU Operator, automating model deployments, autoscaling GPU resources, and ensuring highly available inference services.
- Built distributed execution pipelines using Ray, Apache Kafka, asynchronous Python, and Kubernetes Jobs, enabling fault-tolerant agent orchestration with scalable scheduling, parallel processing, and automatic workload recovery.
- Integrated AI services with SAP, Salesforce, ServiceNow, SharePoint, and Confluence through secure tool calling, vector indexing, and governed access controls, enabling reliable knowledge retrieval across business platforms.
- Automated LLMOps and CI/CD workflows using GitHub Actions, MLflow, Docker, Terraform, Helm, and Kubernetes, streamlining model versioning, deployment automation, evaluation, and release management.
- Improved platform observability using Prometheus, Grafana, OpenTelemetry, MLflow, and PostgreSQL, implementing telemetry, audit logging, proactive monitoring, and agent evaluation for reliable AI system operations.

Accenture **Oct 2020 – Jul 2024**
Machine Learning Engineer India

- Developed Python-based AI pipelines using PyTorch, LangChain, Retrieval-Augmented Generation, prompt engineering, and MLflow, supporting intelligent business applications used by over 420,000 users.
- Built Retrieval-Augmented Generation solutions using LangChain, FAISS, Pinecone, hybrid search, Hugging Face embeddings, and Cross-Encoder reranking, improving knowledge retrieval accuracy by 21% across production AI applications.
- Optimized large language model inference using TensorRT-LLM, Triton Inference Server, CUDA, FlashAttention, and dynamic batching, reducing serving latency by 33% while lowering infrastructure costs by 29%.
- Designed intelligent AI workflows integrating memory, tool calling, hallucination detection, evaluation, and prompt versioning using LangGraph, LangSmith, DeepEval, MLflow Tracing, and Hugging Face Transformers.
- Developed scalable Python microservices using FastAPI, REST APIs, Docker, Kubernetes, Redis, PostgreSQL, and gRPC to deliver secure retrieval, orchestration, and enterprise AI services.
- Implemented AI infrastructure across AWS, Amazon Bedrock, Kubernetes, Terraform, GitHub Actions, IAM, and Secrets Manager, enabling secure, scalable, and highly available AI deployments.
- Built end-to-end MLOps and LLMOps pipelines using MLflow, Apache Airflow, GitHub Actions, AWS SageMaker, automated testing, and model registry for reliable AI model deployment workflows.
- Developed event-driven data pipelines using Apache Kafka, Apache Spark, Databricks, Airflow, Delta Lake, and AWS S3, enabling scalable document ingestion, embedding generation, and retrieval indexing.
- Integrated AI applications with Microsoft Teams, SharePoint, Salesforce, SAP, ServiceNow, OAuth 2.0, JWT, and RBAC, ensuring secure authentication and governed access across business platforms.
- Enhanced production observability using Prometheus, Grafana, OpenTelemetry, AWS CloudWatch, MLflow Tracing, and automated alerting, improving platform reliability and accelerating incident detection across AI services.

Technical Skills

Programming Languages: Python, SQL, Bash
Generative AI & LLMs: Generative AI, Large Language Models (LLMs), Agentic AI, AI Agents, Multi-Agent Systems, Retrieval-Augmented Generation (RAG), Prompt Engineering, Tool Calling, Model Context Protocol (MCP), Prompt Optimization, LLM Evaluation, Agent Evaluation, LLMOps, Reinforcement Learning from Human Feedback (RLHF), Hallucination Detection
Machine Learning & NLP: Machine Learning, Deep Learning, Natural Language Processing (NLP), PyTorch, Hugging Face Transformers, NVIDIA NeMo, Transformer Models, Embeddings, Semantic Search, Hybrid Search, Cross-Encoder Reranking, Vector Databases
Inference & AI Serving: TensorRT-LLM, Triton Inference Server, NVIDIA NIM, vLLM, CUDA, FlashAttention, Dynamic Batching, GPU Inference, Distributed Inference, Inference Optimization, Model Serving, Model Quantization
AI Frameworks & MLOps: LangGraph, LangChain, LangSmith, MLflow, MLflow Tracing, DeepEval, Ray, Apache Airflow, GitHub Actions, Model Registry, CI/CD, Automated Testing, Model Versioning, Continuous Deployment
Backend & Distributed Systems: FastAPI, REST APIs, gRPC, Microservices, Asynchronous Python, Apache Kafka, Apache Spark, Databricks, Kubernetes Jobs, Docker, Redis, PostgreSQL
Cloud & Infrastructure: AWS, Amazon Bedrock, AWS SageMaker, AWS S3, AWS IAM, AWS Secrets Manager, NVIDIA DGX Cloud, Kubernetes, Helm, NVIDIA GPU Operator, Terraform, API Gateway
Data Engineering & Enterprise Platforms: FAISS, Pinecone, Delta Lake, Vector Indexing, Enterprise Search, SAP, Salesforce, ServiceNow, Workday, SharePoint, Confluence, Microsoft Teams, OAuth 2.0, JWT, RBAC
Observability & Monitoring: Prometheus, Grafana, OpenTelemetry, MLflow Tracing, AWS CloudWatch, Application Insights
Soft Skills: System Design, AI Solution Architecture, Cross-Functional Collaboration, Stakeholder Management, Technical Leadership, Technical Communication, Agile Development, Mentoring

Project

EvalForge: LLM Evaluation and Regression Testing Platform
Technologies: Python, FastAPI, PyTorch, Hugging Face, MLflow, PostgreSQL, Redis, Docker, AWS

- Engineered an automated LLM evaluation platform that benchmarked more than 20 model, prompt, and retrieval configurations across over 2,000 test cases using correctness, groundedness, citation accuracy, hallucination rate, latency, and cost metrics.
- Implemented deterministic and LLM-as-a-judge evaluation pipelines using semantic similarity, context precision, answer relevance, and safety scoring, reducing manual evaluation effort by 72%.
- Designed configurable regression-testing gates that detected 93% of intentionally introduced quality regressions and automatically blocked releases when groundedness, critical-test coverage, latency, or cost thresholds were violated.
- Integrated MLflow experiment tracking and PostgreSQL dataset versioning, improving experiment reproducibility and reducing model and prompt comparison time by 60%.
- Built asynchronous FastAPI services with Redis-based result caching and Docker deployment, reducing repeated evaluation runtime by 48%, achieving less than 250-millisecond P95 API latency, and supporting more than 50 concurrent evaluation jobs.

Certifications

NVIDIA Certified Associate – Generative AI LLMs
AWS Certified Machine Learning Engineer – Associate
Databricks Certified Generative AI Engineer Associate
Certified Kubernetes Administrator (CKA)

Education

Master of Science in Computer Science
University at Buffalo, New York

Bachelor of Technology in Software Engineering (Integrated)
Vellore Institute of Technology, Vellore