Uday Lunawat

Senior Applied AI & Agent Systems Engineer | RAG, Multi-Agent Systems, LLMOps

(+91) 7020901969 | udaylunawat@gmail.com | udaylunawat.github.io | udaylunawat | uday-lunawat | udaylunawat

Summary

Senior Applied AI & Agent Systems Engineer with **7 years of experience** building **production-grade systems around pretrained models (LLMs & CV)**. Expertise in **Retrieval-Augmented Generation (RAG)**, prompt engineering, fine-tuning workflows, and **multi-agent orchestration** using LangGraph.

Proven track record of delivering **customer-facing AI solutions** by integrating models with tools, memory, and APIs. Led development of an enterprise AI platform supporting **400+ production agents and 17,000+ deployments**, with strong focus on **evaluation and reliability**.

Work Experience

Senior AI Platform Engineer (GenAI, LLMOps, CI/CD) – Fractal Analytics

Nov 2024 – Present

Tech: Python, GCP, LangGraph, FastAPI, Jenkins, OIDC, SQLModel, Docker, Alembic — *Pune*

- **Customer-Facing AI Systems:** Delivered agent-based AI solutions for enterprise stakeholders, integrating LLMs with tools, APIs, and business workflows to solve real-world telecom use cases.
- **Pretrained Model Systems:** Built production systems around LLMs (OpenAI, Claude, Gemini) using prompt engineering, structured outputs, and tool augmentation.
- **Multi-Agent Systems:** Built stateful orchestration workflows (**LangGraph**) using ReAct and hierarchical delegation for complex multi-step reasoning tasks.
- **Evaluation & Observability:** Built evaluation pipelines (Langfuse, semantic validation) with LLM-native metrics to guide optimization and system improvements.
- **Scale:** Platform supports **400+ production agents** and **17,000+ deployments**, reducing onboarding time to **<90 seconds**.
- **LLM FinOps:** Implemented cost-aware execution (token budgeting, adaptive context while maintaining accuracy SLAs.
- **Reliability Engineering:** Designed fault-tolerant systems (retries, fallbacks, partial recovery) improving success rate of multi-step workflows.
- **Security & Governance:** Enforced Zero-Trust architecture (OIDC, RBAC, secrets isolation) for multi-tenant deployments.
- **Leadership:** Led a team of 7 engineers; drove architecture, reviews, and enterprise adoption.

Senior Machine Learning Engineer – HCL

Apr 2024 – Jun 2024

Tech: Python, GCP, MLflow, Seldon Core, TensorFlow, Label Studio, Flask — *Bengaluru*

- Architected **human-in-the-loop (HITL)** evaluation frameworks integrating feedback loops into model validation.
- Developed scalable annotation pipelines using Label Studio and microservices.
- Deployed ML inference systems with Docker and Cloud Run ensuring reliability and observability.

Machine Learning Engineer – Chugani

Feb 2022 – May 2023

Tech: Python, AWS, MLflow, IaC, nbdev, Quantization — *Bengaluru*

- Built real-time liveness detection system achieving **150ms latency** via model optimization and quantization.
- Implemented MLflow for experiment tracking and modernized ML infrastructure using IaC.
- Refactored legacy ML pipelines into modular, documented Python packages using **nbdev**, improving maintainability, reuse, and collaboration.

Machine Learning Engineer – Yang

Apr 2019 – Jan 2022

Tech: Python, Airflow, AWS, ETL, Tableau — *Bengaluru*

- Built automated AI audit system processing real-time factory data streams for quality control.
- Delivered dashboards across 20+ vendor sites saving **200+ hours/week**.

Systems Engineer Intern –Infosys

Oct 2018 – Mar 2019

Analytics Intern – TCS

Jan 2018 – July 2018

Skills

Languages :	**Python**, SQL, Bash, Groovy
LLM & GenAI :	Hugging Face Transformers, RAG, Prompt Engineering, Fine-tuning (LoRA/PEFT), LangGraph, LangChain
Agent Systems :	Multi-Agent Systems, ReAct, Self-Reflection, Tool Augmentation, MCP, Workflow Orchestration
Cloud & Systems :	AWS, GCP, Docker, Kubernetes, FastAPI, Microservices, Terraform, CI/CD
MLOps & Serving :	MLflow, Seldon Core, Model Serving, Experiment Tracking, Feature Pipelines
Evaluation & FinOps :	LLM Metrics, Tracing (Langfuse), Cost Optimization
Data & ML :	Pandas, NumPy, TensorFlow, PyTorch
AI Tooling :	OpenAI, Gemini, Claude APIs, GitHub Copilot, Claude Code
Certifications :	**Anthropic Claude Certified Architect**, AppliedAI Course, Pyspark and Data Engineering

Education

B.Tech. in Information Technology RCOEM, Nagpur — *Apr 2014 – June 2018*

Diploma in Computer Engineering — *Apr 2011 – June 2014*