

Jitendra Yadav

+91 9610373417 | jy69518@gmail.com | linkedin.com/in/jityadav | github.com/Jitendra1707 | Hyderabad, India

AI Engineer | Agentic AI | Generative AI | MLOps | RAG | AIOps | GPU Orchestration | Cloud & Kubernetes

Education

- 2024
Indian Institute of Information Technology, Allahabad
B.Tech. in Information Technology, CGPA 8.45

Core Competencies

Agentic AI & LLM Engineering

Generative AI & VLM Solutions

AI/MLOps & Model Deployment

RAG & Knowledge Systems

AI Agent Orchestration

GPU & AI Infrastructure

Cloud & Kubernetes Engineering

AIOps & DevSecOps

AI Platform Architecture

Technical Skills

AI & Agentic Systems: Agentic AI, LLM/VLM Agents, LangChain, LangGraph, RAG, Qdrant, OpenAI APIs, MCP, Prompt Engineering, Agent Memory, Guardrails, Vector Databases

AI/ML & MLOps: vLLM, Hugging Face, MLflow, GPU Orchestration, Databricks, Model Deployment

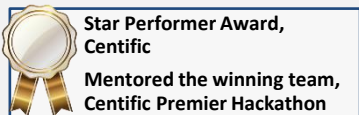
Cloud & DevOps: Azure, AWS, Kubernetes, Docker, Terraform, Azure DevOps, CI/CD, Production Deployment

Programming & Frameworks: Python, C/C++, FastAPI, SQL, PostgreSQL

Voice & Real-Time AI: Pipecat, STT/TTS, Plivo, Azure Speech Services, WebRTC, LiveKit

AI Infrastructure & Security: NVIDIA DGX Spark, DevSecOps, CloudOps, FinOps, SecOps

Achievements



Profile Summary

- AI Engineer & Technical Architect with **over 2 years** of experience in **Agentic AI, Generative AI, LLM/VLM Systems, RAG, AIOps, and MLOps**, building production-ready AI platforms, intelligent automation, and agentic solutions.
- Architected and delivered scalable AI/ML platforms across **Azure, AWS, Kubernetes, GPU Infrastructure, vLLM, LangGraph, FastAPI, Databricks**, with expertise in AI orchestration, agent pipelines, model deployment, and cloud optimization.
- Progressed from **AI/Cloud Engineering to Technical Architecture and Leadership**, leading teams and driving AI platform initiatives; **participated in 2 internal hackathons, including a winning team**, and delivered measurable impact through GPU optimization and cloud engineering, including **\$60K/month peak cloud savings and 70% faster inference**.

Work Experience

Centific Global Technologies | Hyderabad, India

AI Engineer (Technical Architect) — Digital Architecture Centre of Excellence (DAC) | Since Apr 2025

- Led AI/AIOps R&D for Centific's **AI Data Foundry (AIDF)**, onboarding 150+ users across cross-functional teams in the USA and India.
- Built an **AI-powered resume screening and talent assessment agent** for a company-wide hackathon, evaluating 6,000+ applicants and 2,000+ online assessments to shortlist the top 200 candidates.
- Developed a **PoC humanoid meeting agent** that autonomously joined live meetings, conducted natural conversations with human-like speech and turn-taking, leveraged multi-domain RAG, and delivered live web-based reports.
- Build **on-demand GPU compute offloading** across Azure, RunPod, Denvr, and on-premise clusters, reducing inference/training costs by 30–50% and peak cloud spend by ~70% (\$90K to \$30K).
- Re-architected the platform and developed a **CLI-based platform/runtime AI service installer**, reducing on-premise customer deployment timelines from 1 week to 1–2 days.

Associate Software Engineer (AI/MLOps) — Core AIDF Engineering Team | Jul 2024 – Apr 2025

- Build AIOps components and a DevSecOps-integrated **Manifest Engine**, standardizing secure AI service deployments and saving ~6 hours per release.
- Designed and developed a scalable **AI Orchestration Engine** supporting multi-step pipelines, dynamic payload exchange, and automated coordination across AI services.
- Automated **AI model deployments** from Hugging Face across AKS, Azure Databricks, RunPod, and on-premise environments; optimized GPU inference using vLLM, reducing inference time by 70%.
- Developed an in-house **Kubernetes GPU resource manager** to optimize resource utilization, workload scheduling, and operational control across on-premise GPU infrastructure.
- Took end-to-end ownership of a company-wide **AWS security incident** after costs escalated from ~\$20K to a projected \$150–200K/month; escalated the issue, rebuilt a sanitized AWS account, and strengthened security controls.

Cloud/DevOps Intern | Jan 2024 – Jul 2024

- Automated **Azure infrastructure provisioning** for VNets, App Services, and Storage using Terraform, reducing manual deployment effort by 70% and environment setup time by 60%.
- Built dynamic **CI/CD pipelines** using Azure DevOps REST APIs and runtime parameters, reducing release cycles by 50%.
- Developed a centralized **CloudOps/FinOps/SecOps dashboard** covering 100+ cloud resources, improving cloud visibility and reducing monthly cloud spend by 20–30%.

Projects

Agentic AI Counselor Platform | Multi-Channel LLM Agents

- Architected an **agentic AI platform** supporting PSTN voice, WhatsApp, email, and real-time video avatars with shared cross-channel memory using **FastAPI, LiveKit, Angular 19, and GPU services**.
- Built low-latency voice agents with **Pipecat, STT/LLM/TTS, Plivo/Azure Communication Services**, including turn detection, barge-in, and streaming responses.
- Implemented **hybrid RAG with Qdrant, reranking, episodic memory, and automated lead profiling/follow-ups**.
- Scaled diffusion-based avatar streaming to **10 concurrent 25 FPS streams per H100 GPU** with automated dialing, analysis, follow-ups, and human handoff.

Surveillance Agent | Multi-Camera Video Intelligence

- Built a **multi-camera video intelligence pipeline** using RTSP, **Cosmos-Reason2-8B, vLLM, and natural-language rules** for video transcription, alerting, and incident correlation.
- Reduced VLM inference load by ~80% using a **pixel-difference change gate**, increasing coverage from ~3 to 10–15 cameras per pod.
- Designed a tiered video memory architecture using **PostgreSQL and pgvector** and exposed camera intelligence through **REST, WebSocket, and MCP** for external AI agents.