

MANAN JAIN

+91 9897236924 · jainmanan1804@gmail.com · manan.info · Bangalore, India



SUMMARY

Applied AI & Platform Engineer who builds the infrastructure that makes LLM agents **observable, safe, and auditable** in regulated enterprise environments. Currently leading an **agent observability and guardrails platform** that detects **PII/PHI and secrets**, supports **red teaming**, and flags unacceptable agent behavior in one integrated view. Previously architected a Kubernetes-based **LLM inference platform on 500+ GPUs** (~80% utilization, ~60% cost reduction) and operated **150+ EKS clusters** serving 40+ engineering teams. Comfortable working directly with stakeholders to turn ambiguous needs into production systems.

SKILLS

AI & Agent Platforms: LLM inference, agent observability, guardrails, PII/PHI & secrets detection, red teaming, AI safety & auditability

Programming & Automation: Python, Shell Scripting

Cloud Platforms: AWS – EKS, ECS, EMR, EC2, S3, IAM, VPC

Container & Platform Engineering: Kubernetes, Docker, Helm, Cilium, Karpenter, Flyte, GPU scheduling, Cluster Operations

Infrastructure as Code: Terraform, Infrastructure Automation

CI/CD & GitOps: Argo CD, GitOps, TeamCity, AWS CodePipeline, CodeBuild, CodeDeploy, Bitbucket, Jenkins

Observability & Monitoring: Prometheus, Amazon Managed Prometheus, Grafana, Datadog, Splunk, Metrics & Alerting

Data & Platform Components: Kafka, Redis, MinIO

Security & Compliance: Wiz, Sysdig, Dependency-Track, Cloud Security, Vulnerability Scanning

WORK EXPERIENCE

ZS ASSOCIATES

Engineering Specialist (Consultant)

India

06/2026 – Present

- Lead the design and delivery of an **integrated agent observability platform** that gives teams a single place to monitor LLM agents in production and catch behavior that is not acceptable for enterprise use.
- Built detection into the agent harness for **PII/PHI exposure and leaked secrets** across prompts, tool calls, and responses, so sensitive data is caught before it leaves the agent.
- Integrated **red teaming** into the platform to probe agents for unsafe or policy-violating behavior before and after release, turning ad hoc testing into a repeatable safety check.
- Unified traces, safety findings, and policy violations into one view so engineering and governance teams can **audit agent actions end-to-end** and act on issues quickly.

Senior Cloud Engineer

07/2024 – 06/2026

- Architected a **Kubernetes-based LLM inference platform** leveraging Spot-backed GPUs. Optimized scheduling policies, fault tolerance, and retry strategies to maintain **~80% utilization across 500+ GPUs**, driving a **~60% reduction in compute cost** while sustaining large-scale batch processing.
- Implemented **Amazon Managed Prometheus and Grafana** for observability across **150+ EKS clusters**, and ran an internal uptime monitoring platform covering **2,500+ endpoints**, evolving basic monitoring into reliable alerting across control plane, ingress, infrastructure, and application metrics while keeping ingestion costs low.
- Designed a **centralized GitOps platform using Argo CD** to manage Kubernetes add-ons at scale, supporting **120+ applications across 150+ clusters**, with bulk deployments and Argo CD Notifications across 100+ projects.
- Researched and deployed **50+ Kubernetes add-ons**, including MinIO, Wiz, Kafka, Redis, and Dgraph, plus core integrations such as S3/EFS CSI drivers, Step Functions, and **DCGM Exporter** for GPU telemetry.
- Prototyped **EMR on EKS for Spark workloads** with Karpenter for dynamic scaling, and built reusable **Terraform modules** that provision **80+ AWS resources through a single automated pipeline** for multi-tenant, multi-environment deployments.

Cloud Engineer

09/2022 – 07/2024

- Deployed and operated **100+ applications on Amazon ECS** and delivered **250+ application deployments on Amazon EKS** using TeamCity, Argo CD, Helm, and GitOps, stabilizing pipelines that broke under parallel deployments and scale.
- Owned **20+ COTS platforms**, including Dependency-Track, Sysdig, and Argo CD, ensuring they met **security, scalability, and compliance** requirements before onboarding application teams.
- Built bulk-upgrade automation to patch **500+ CI/CD pipelines in parallel**, and led migrations (**EC2 → ECS/EKS, ECS → EKS**) that moved teams to container-native patterns.
- Led the **Sysdig POC and enterprise rollout**, analyzing application runtime behavior and deploying across **500+ products in roughly one month** using automated rollout tooling.

Cloud Engineer Intern

11/2021 – 09/2022

- Owned day-to-day **AWS operations** and served as an **escalation point for complex production issues**; built server bootstrapping automation and wrote **cloud deployment standards and governance guidelines** to prevent compliance drift.

CERTIFICATIONS

Claude Certified Architect - Foundations — Anthropic

Verify ↗

NVIDIA-Certified Associate: AI Infrastructure & Operations (NCA-AIIO) — NVIDIA

Verify ↗

Platform Engineering Certification — PlatformEngineering.org

Verify ↗

EDUCATION

B.Tech, Computer Science & Engineering — MIT-ADT University, Pune

08/2022 · GPA: 7.9