ZOHAIB AHMED

AI SOLUTIONS ENGINEER

APPLIED AI SYSTEMS | AGENT RELIABILITY | LLM EVALUATION | PRODUCTION DEPLOYMENTS

AI Solutions Engineer specializing in production deployment of LLM-powered applications and multi-agent systems. I build agent reliability platforms with evaluation pipelines, replay validation, observability, and governance workflows that drive measurable impact. Passionate about shipping trustworthy AI systems that are scalable, reliable, and enterprise-ready.

Production deployment validation using Cloud Run, Prometheus metrics, and automated regression testing.

87%	18% → 6%	43	p50 57ms p95 270ms	1,000+	99%+
Evaluation Pass Rate Across 145+ Tests	Hallucination Rate Reduction through eval-driven optimization	Req/Sec Throughput Sustained under load testing	End-to-End Latency	Automated Evaluations (JSONL artifacts)	Workflow Success Rate in production tests

EXPERIENCE

Agent Reliability, Governance & Evaluation Platform

Mar 2025 – Present

Python • FastAPI • LangGraph • Docker • Cloud Run • Prometheus • JSONL

- Designed and deployed a multi-tenant agent reliability platform supporting workflow orchestration, replayable execution artifacts, experiment tracking, and leader board-driven optimization.
- Implemented governance metrics including task success, citation coverage, political reliability, unsafe-action blocking, tenant scoring, compliance scoring, and workflow latency measurement.
- Built approval-gated enterprise AI workflows with tenant isolation, replay validation, governance policy enforcement, workflow lifecycle management, and operational dashboards.
- Implemented incident governance workflows with approval timelines, audit visibility, role-scoped actions, violation tracking, and incident resolution workflows.
- Instrumented Prometheus observability, JSONL audit artifacts, rate limiting, compliance metrics, and automated testing integrated into a repo-wide suite with 145+ passing tests.
- Achieved 43 req/sec throughput at ~99% success rate while reducing hallucination rates from 18% to 6% through evaluation-driven retrieval optimization and governance enforcement.

Autonomous AI Research Agent (Multi-Agent System)

Mar 2025 – Present

LangGraph • OpenAI • FastAPI • Qdrant • Redis • Docker • LangSmith

- Designed Planner, Researcher, Retriever, Writer, and Reviewer agent architecture using LangGraph with tool-calling workflows with human-in-the-loop checkpoints.
- Integrated retrieval, web search, memory management, and citation generation with evaluation-driven reflexion loops reducing hallucinations by ~35%.
- Built LangSmith observability and tracing with prompt traces, tool traces, latency analytics, and quality scoring for each agent step.
- Deployed as production FastAPI services on Cloud Run with persistent memory and streaming responses.

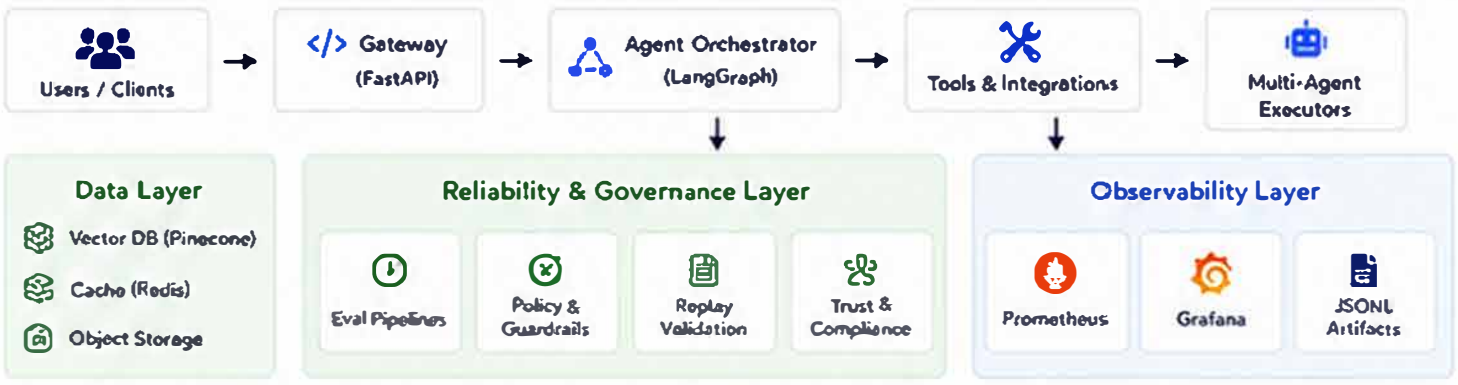
RAG Evaluation & Observability Pipelines

Jan 2025 – Mar 2025

Python • FastAPI • LangChain • Pinecone • Prometheus • Grafana

- Built modular evaluation pipelines for RAG systems with faithfulness, relevancy, context precision, citation accuracy, and answer correctness scoring.
- Implemented regression testing and replay validation using JSONL datasets and automated metrics.
- Built dashboards with Grafana to monitor latency, token usage, error rates, and evaluation trends.
- Reduced retrieval latency by 37% and improved answer relevancy by 22% through iterative tuning.

ARCHITECTURE OVERVIEW



KEY IMPACT

87%	18% → 6%	43	p50 57ms p95 270ms	1,000+	99%+
Evaluation Pass Rate	Hallucination Reduction	Req/Sec Throughput	Latency	Automated Evaluations	Workflow Success Rate

San Jose, CA

408-648-8785

azohaib.0150@gmail.com

linkedin.com/in/zohaib-a-1a0017174

github.com/Electricpaper77/ai-rag-eval-platform

CORE COMPETENCIES

- Agentic AI & Multi-Agent Systems
- LLM Evaluation & Observability
- RAG / Vector Search
- Workflow Orchestration
- Incident Governance
- Trust & Compliance

TECHNICAL SKILLS

Languages: Python, SQL

Frameworks: FastAPI, LangGraph, LangChain

AI/ML: OpenAI API, RAG, Embeddings, Agents

Databases: Pinecone, Qdrant, Redis, PostgreSQL

Infrastructure: Docker, Kubernetes, GCP (Cloud Run)

Observability: Prometheus, Grafana, LangSmith

Tools: Git, Github Actions, JSONL, PyTest

ADDITIONAL EXPERIENCE

PayPal – Cloud Support Engineer
Jan 2023 – Feb 2024
Resolved Auto Failures, reduced MTTR 25–30%.

Tesla – Systems Operations Associate
Jan 2020 – Dec 2020
Maintained uptime in high-throughput systems

Home Depot – Technical Support Associate
Jul 2017 – Dec 2019
Resolved 40–60 issues daily in high-volume env.

EDUCATION

San Jose State University, San Jose, CA — Dec 2025
B.S. Management Information Systems (MIS)
Relevant Coursework: Systems Analysis, Database Management, Data Structures, Statistics, Business Intelligence, Information Systems

CERTIFICATIONS

- AWS Cloud Practitioner
- AWS Cloud Solutions Learning
- Cisco Cybersecurity Fundamentals
- Google Cloud Associate Cloud Engineer
- DeepLearning.AI – Generative AI with LLMs
- LangChain Academy – Agents AI

ADDITIONAL

Clearance: Eligible to work in the U.S.

Strengths: Problem Solving, Communication, Leadership, Ownership, Adaptability

Interests: Building reliable AI systems, automation, open-source, and product innovation