KASHIF ALI

AI Engineer – Agent Evaluation & Governed Systems

Karachi, Pakistan - kashif.ali.arain.it@gmail.com - GitHub - LinkedIn - Portfolio

PROFESSIONAL SUMMARY

AI Engineer specializing in production-grade LLM agent evaluation, governed agent runtimes, and deterministic guardrail enforcement. Experienced in building multi-step tool-using agents, evaluation frameworks, human-in-the-loop (HITL) approval workflows, and model-agnostic backend microservices using Python, FastAPI, and Qdrant. Skilled in converting complex business rules into code-enforced validation gates, establishing audit logging for data provenance, and running regression checks across prompt/model iterations to ensure safe, scalable deployment.

TECHNICAL SKILLS

Agent Governance & Eval: Agent Evaluation Frameworks, Regression Testing, Human-in-the-Loop (HITL) Queues, Deterministic Guardrails, Data Provenance & Audit Logging, Statistical Quality Sampling
AI Platforms & Architecture: Multi-Model Integration, Tool Calling, Structured Outputs, RAG Architectures, Vector Search (Qdrant), Agentic Workflows, Voice AI (AssemblyAI)
Backend Software Engineering: Python, Async FastAPI, TypeScript, Pydantic, REST APIs, SQLite, PostgreSQL, Docker, Git/GitHub, Background Workers, Automated Testing

EXPERIENCE

Backend AI Engineer – Intern **July 2026 – Sep 2026**
FlyRank AI *Remote*

- Engineered production-oriented backend microservices and REST APIs using Python, FastAPI, and Pydantic, enforcing strict schema validation and error-handling guardrails.
- Integrated model-agnostic LLM providers, retrieval pipelines, and persistence layers with complete request/response audit tracking to log data provenance across execution runs.
- Architected queue-oriented background workers to isolate asynchronous data processing tasks and execute validation gates before committing persistent state changes.
- Authored clear technical design documentation, unit tests, and staged implementation pipelines adhering to strict code review standards.

SELECTED AI & GOVERNANCE PROJECTS

Enterprise AgentOS – Production Multi-Agent Automation Platform *Python, FastAPI, LangGraph, PostgreSQL, Redis, MCP*

- Building a production-oriented AI agent platform for executing multi-step business workflows through controlled tools, multi-agent orchestration, and human-in-the-loop approvals.
- Designing a modular backend architecture for API-driven execution, persistent workflow state, role-based access control, audit logging, and background task processing.
- Focusing on reliable agent execution, structured outputs, tool permissions, and observable workflows to demonstrate production-oriented AI backend engineering.

Tailgate Quote – Governed Voice-First AI Agent System *Python, FastAPI, AssemblyAI, React, TypeScript, SQLite*

- Architected a tool-using voice AI agent system with deterministic validation rules enforced in backend code rather than relying on non-deterministic LLM output.
- Engineered hard deterministic catalog search and pricing calculation engines that explicitly prevent LLM price hallucinations and enforce auditability.
- Implemented a Human-in-the-Loop (HITL) "Proof of Hearing" validation queue, requiring verification of calculated quote evidence against raw transcript statements before authorization.
- Built structured FastAPI evaluation and review endpoints with state persistence in SQLite to track execution decisions, quotes created, and approval states.

PatientMemory – Governed Patient Record Retrieval System *Python, RAG, Qdrant, LLM APIs*

- Built a context-aware medical record retrieval pipeline implementing parent-child document chunking and vector search in Qdrant for natural language queries.
- Enforced context-isolation rules to verify retrieval accuracy and data provenance before feeding context to model context windows.

Real-Time Sign Language Translation Engine *Python, PyTorch, DETR, FastAPI, OpenCV, React*

- Trained a 53-class DETR object detection model on 15,600+ samples, establishing rigorous mAP@0.50 testing thresholds (achieving 94.6% mAP) to evaluate inference regressions.

EDUCATION

Bachelor of Science in Information Technology **2022 – 2026**
Quaid-e-Awam University of Engineering, Science & Technology (QUEST) *Nawabshah, Pakistan*
Relevant Area Highlights: Software Engineering, Database Systems, Artificial Intelligence, Machine Learning, Data Structures & Algorithms, API Architecture