



Syed Ashar Raza

AI Engineer | Production AI | LLMs | RAG | AI Agents | Machine Learning | Generative AI

syedasharraza512725@gmail.com | +92 347-6371697 | Gujranwala, Punjab, Pakistan | linkedin.com/in/syed-ashar-raza

github.com/syed-ashar-raza

SUMMARY

AI Engineer focused on building reliable, production-oriented AI systems with Python. Experienced in LLMs, RAG, AI agents, machine learning, AI evaluation, security, inference infrastructure, observability, and MLOps. Built and validated end-to-end AI systems with automated testing, measurable evaluation, reliability controls, and reproducible engineering workflows.

PROFESSIONAL EXPERIENCE

Fargo Solve | AI Automation Lab, AI Automation Engineer | Voice AI & Workflows

01/2024 – 07/2024 | Remote

- Developed AI automation and workflow solutions using Python, LLMs, voice AI, APIs, and automation tools.
- Integrated AI-driven workflows with external services and business processes for conversational and automation use cases.
- Worked with LLM-based systems, speech technologies, APIs, and workflow orchestration.
- Focused on practical AI solutions connecting models, application components, and business workflows.

FAST NUCES | AI Research Lab, Applied AI Engineer - Final Year Project

06/2023 – 02/2024

- Worked on applied AI research involving large language models, model adaptation, and Urdu-language AI.
- Explored parameter-efficient model adaptation and LLM fine-tuning for domain-specific language tasks.
- Prepared datasets and conducted model experimentation and evaluation for language-model applications.
- Applied reproducible AI research and engineering workflows to develop and assess language-model solutions.

TechCraft Labs - AI internship Program, AI Engineer Intern | LLMs & RAG

01/2023 – 04/2023 | Gujranwala, Pakistan

- Developed an internal knowledge-based chatbot using LLMs, retrieval, and document-processing techniques.
- Prepared PDF documents for retrieval through cleaning, chunking, and source tracking.
- Integrated AI capabilities with application interfaces and backend components for knowledge-retrieval workflows.
- Contributed to testing, documentation, and demonstration of the AI application.

EDUCATION

Bachelor of Science in Artificial Intelligence (BSAI), IISAT University

11/2024 – Present | Gujranwala, Pakistan

Bachelor of Science in Artificial Intelligence (BSAI), focused on machine learning, deep learning, software engineering, data science, and applied AI systems.

SKILLS

Software Engineering & API Development — Proficient

Python • Software Architecture • Object-Oriented Programming • Design Patterns • REST APIs • FastAPI • API Design • Automated Testing • pytest • Git • GitHub • SQL • PostgreSQL • SQLite • Error Handling • Type Hints • CLI Development • Configuration Management

Generative AI, LLMs & AI Agents — Proficient

Large Language Models (LLMs) • Generative AI • LLM APIs • Prompt Engineering • Structured Outputs • JSON Schema • Tool Calling • Function Calling • AI Agents • Agent Architecture • Agent Workflows • Planning • Execution • Memory • Context Management • Model Routing • Provider Abstraction • Local LLM Inference

MLOps, Deployment & CI/CD — Proficient

MLOps • ML Lifecycle • Experiment Tracking • MLflow • Model Versioning • Model Serving • FastAPI Inference • Docker • Docker Compose • Kubernetes • Minikube • CI/CD • GitHub Actions • Health Checks • Readiness Checks • Model Validation • Reproducible Workflows • Deployment Automation

AI Security, Reliability & Distributed Systems — Proficient

AI Security • PII Detection • PII Protection • Security Evaluation • Input Validation • Security Testing • False-Positive Analysis • Reliability Engineering • Rate Limiting • Retry Policies • Bounded Retries • Circuit Breakers • Timeouts • Fault Isolation • Graceful Failure • Fail-Closed Design • Provider Failover • Distributed Systems

Machine Learning & Deep Learning — Proficient

Machine Learning • Deep Learning • PyTorch • scikit-learn • Model Training • Model Evaluation • Feature Engineering • Data Preprocessing • Dataset Preparation • Neural Networks • Classification • Embeddings • Hyperparameter Tuning • Model Validation • Model Serialization • NumPy • Pandas

RAG & Information Retrieval — Proficient

Retrieval-Augmented Generation (RAG) • Embeddings • Semantic Search • Similarity Search • Retrieval Pipelines • Document Processing • PDF Processing • Text Chunking • Context Construction • Grounded Generation • Source Tracking • Retrieval Evaluation • Answer Evidence Evaluation • Abstention • Deduplication • Vector Retrieval

AI Evaluation, Testing & Reproducibility — Proficient

AI Evaluation • Regression Testing • Continuous Evaluation • Evaluation Gates • Retrieval Evaluation • Answer Evaluation • Abstention Testing • Deterministic Evaluation • Test Automation • Unit Testing • Integration Testing • Smoke Testing • Benchmarking • Reproducibility • Quality Gates • Failure Testing • Performance Testing

AI Infrastructure, Observability, PEFT & Multimodal ML

— Proficient

AI Inference Infrastructure • Model Serving • vLLM • Inference Benchmarking • Streaming Inference • Provider Routing • Request Tracing • Prometheus • Metrics • Observability • Health Monitoring • Latency Analysis • p50/p95 Benchmarking • Throughput Measurement • Performance Engineering • PEFT • LoRA • Multimodal ML

LANGUAGES

- English— Working Proficiency
- Urdu — Bilingual Proficiency
- Punjabi — Native Proficiency

PROJECTS

Nexora , <i>Production AI Inference Infrastructure Platform</i>	09/2026 – Present
Built a production-oriented AI inference infrastructure gateway with provider routing, bounded retries, circuit breaking, rate limiting, authentication, streaming, health/readiness checks, and Prometheus observability. Validated local Kubernetes deployment and benchmarked vLLM Qwen3-0.6B inference with reproducible p50/p95 latency measurements.	
Verity , <i>AI Evaluation & Regression Platform</i>	09/2026 – 10/2026
Engineered a production-oriented AI evaluation platform for reproducible system testing, deterministic metrics, per-case thresholds, latency budgets, baseline regression gates, and persisted evaluation reports. Implemented CLI and FastAPI interfaces with Docker and CI integration, achieving 5/5 evaluation checks with a 100% pass rate and a 142 ms mean latency on the example evaluation.	
Orion , <i>MCP & Modern Agent Architecture</i>	05/2026 – 08/2026
Engineered a modern agent architecture using Model Context Protocol (MCP) concepts for structured tool integration and extensible AI workflows. Designed controlled tool interactions, modular agent components, and standardized interfaces to improve interoperability and maintainability across AI agent systems.	
Syntrix , <i>Multimodal ML Engineering Platform</i>	05/2026 – 07/2026
Engineered a production-oriented ML system focused on reproducible training, model evaluation, inference serving, benchmarking, observability, and multimodal experimentation. Implemented PEFT/LoRA experimentation, automated testing, CLI tooling, deterministic workflows, and reproducible inference performance measurements.	
RAGForge , <i>Production Retrieval-Augmented Generation System</i>	02/2026 – 04/2026
Engineered a production-oriented RAG system for document ingestion, semantic retrieval, and grounded LLM question answering. Implemented embeddings, chunking, source tracking, deduplication, retrieval evaluation, answer-evidence checks, deterministic abstention, and continuous evaluation gates.	
Sentinel , <i>AI Security Evaluation Platform</i>	12/2025 – 02/2026
Engineered an AI security evaluation platform focused on PII protection, security controls, deterministic security testing, and automated evaluation. Implemented security detectors and validation workflows with 15 automated tests and a 10/10 deterministic security evaluation on a local test dataset.	
AgentForge , <i>Production AI Agent Engineering Platform</i>	11/2025 – 02/2026
Engineered an AI agent platform with LLM-driven planning, structured tool selection, bounded multi-step execution, memory, and controlled failure handling. Implemented schema-validated decisions, execution limits, deterministic testing, and reliability controls for agent workflows.	
ModelOpsForge , <i>Production ML Lifecycle & MLOps Platform</i>	05/2025 – 09/2025
Developed an end-to-end ML lifecycle platform covering reproducible training, evaluation, model versioning, API serving, and operational monitoring. Achieved 89.1% accuracy, 83.71% F1, and 96.10% ROC-AUC on the evaluated classification model, with automated testing and CI validation.	