

Rija Sohail

AI Engineer | Backend AI Systems | LLM Agents | RAG | Voice AI

Lahore, Pakistan | Open to Remote AI Engineering Roles

+92 334 815 5581

rijasohail3848@gmail.com | linkedin.com/in/rija-sohail-36962a275

github.com/rijasohail3848



Professional Summary

AI Engineer with hands-on experience building production-grade AI agents, RAG platforms, low-latency Voice AI systems, backend APIs, workflow automation tools, and full-stack AI applications. Strong background in Python, FastAPI, AsyncIO, LangGraph, LangChain, REST APIs, WebSockets, vector databases, retrieval pipelines, React, TypeScript fundamentals, and production debugging.

Experienced in taking ambiguous requirements from idea to shipped product: designing architecture, building backend services, integrating external APIs, managing stateful AI workflows, improving latency, debugging failures, documenting systems, and supporting production QA. Focused on building AI systems that are reliable, observable, fast, and usable beyond demos.

Core Skills

AI Agents and LLM Systems: AI Agents, Agentic Workflows, LangGraph, LangChain, Tool Calling, Function Calling, Multi-Turn State, Prompt Engineering, Structured Outputs, LLM Orchestration, Human Handoff

RAG and Search: Retrieval-Augmented Generation, FAISS, pgvector, Supabase, Embeddings, Hybrid Retrieval, Metadata Filtering, Query Rewriting, Document Grading, CrossEncoder Reranking, Source-Grounded Answers

Backend Engineering: Python, FastAPI, AsyncIO, REST APIs, Webhooks, Microservices, API Integrations, Queue-Based Workflows, Session Persistence, Authentication, Error Handling, Background Processing

Voice and Real-Time AI: Voice AI, Conversational AI, Speech-to-Text, Text-to-Speech, VAD, Barge-In, WebSocket Streaming, Low-Latency Pipelines, Real-Time User Interaction, Call Flow Automation

Frontend and Full-Stack: React.js, JavaScript, TypeScript Fundamentals, Vite, TailwindCSS, Axios, Streaming Interfaces, Multi-Session UI, Dashboards, Authentication, Frontend-Backend Integration

Reliability and Performance: Structured Logging, Latency Tracing, Retry Guards, Fallback Recovery, State Validation, Failure Classification, Conversation Trace Analysis, Prompt Testing, QA Debugging, Production Stabilization

Data and Infrastructure: PostgreSQL, MongoDB, DynamoDB-Style Persistence, SQL, Pandas, ETL Pipelines, Docker, Linux, Git, AWS Concepts, S3 Concepts, CI/CD Concepts

Professional Experience

AI Engineer / Software Engineer, AI Systems	Dec 2025 – Present HytGenX	Remote
- Built and shipped production-grade conversational AI agents for customer-facing workflows involving real-time interaction, API integrations, stateful sessions, retry handling, and human escalation. - Designed LangGraph and LangChain-based agent architectures for deterministic routing, tool execution, multi-turn state, retrieval, structured outputs, and controlled workflow transitions. - Developed scalable Python and FastAPI backend services involving REST APIs, async orchestration, webhook handling, authentication, session persistence, and production reliability controls. - Built real-time Voice AI workflows using WebSocket streaming, speech-to-text, text-to-speech, voice activity detection, interruption handling, and low-latency backend orchestration. - Implemented RAG systems using FAISS, pgvector, embeddings, metadata filtering, hybrid retrieval, reranking, query rewriting, document grading, and grounded answer generation. - Improved AI system reliability through retry guards, state validation, stale-input prevention, fallback recovery, structured logging, and escalation paths for unsafe or incomplete automation. - Optimized performance through streaming execution, caching, prompt tuning, retrieval optimization, provider configuration, and removal of unnecessary processing delays. - Debugged production issues across API calls, LLM outputs, state-machine transitions, speech pipelines, session stores, environment variables, logs, latency traces, and user workflow behavior. - Built cloud-ready session persistence patterns to support durable conversations, restart resilience, workflow recovery, and consistent multi-channel execution. - Worked directly with product, engineering, QA, operations, and business stakeholders to translate real user problems into stable AI workflow behavior. - Created implementation guides, debugging notes, architecture diagrams, QA reports, and reusable agent patterns to help teams ship faster while maintaining system quality. - Built React-based interfaces for AI products including multi-chat sessions, document upload, streaming responses, source display, authentication flows, and operational dashboards.		

AI Training and Cybersecurity Program Lead Jun 2025 – Aug 2025 **Catalytic Security** Remote

- Delivered workshops on AI security, secure LLM systems, prompt safety, cybersecurity fundamentals, and responsible AI adoption to 12+ teams.
- Mentored interns building AI-driven tools, automation workflows, and generative AI proof-of-concept products.
- Communicated technical risks, model behavior, implementation tradeoffs, and security recommendations to technical and non-technical audiences.

Teaching Assistant Jan 2025 – Jun 2025 **FAST-NUCES** Lahore, Pakistan

- Assisted 100+ students with Python, networking, socket programming, systems fundamentals, data structures, algorithms, and debugging.
- Provided hands-on technical support, lab guidance, and clear explanations of programming and system behavior.

AI/ML Research Intern Apr 2024 – Jun 2024 **Rhodium Tech** Remote

- Built transformer-based NLP pipelines for clinical entity extraction, information extraction, text preprocessing, classification, and structured retrieval.
- Supported experimentation with model evaluation, preprocessing strategies, domain-specific NLP workflows, and technical documentation.

Selected AI Systems and Product Work

Low-Latency Conversational Voice AI Platform

- Built production-ready AI voice workflows using real-time speech processing, backend orchestration, webhook routing, queue-based processing, API integrations, and stateful workflow management.
- Developed systems for transcription pipelines, call processing, retry handling, session state, structured form flows, and external API synchronization.
- Improved end-to-end response time by approximately 35
- Resolved race conditions and state corruption in streaming pipelines using per-session locking, state validation, fallback recovery, and controlled state transitions.
- Designed failure-safe handoff flows for missing input, invalid state transitions, repeated API errors, and cases requiring human support.

Enterprise RAG and Knowledge Assistant Platform

- Built a full-stack RAG product with 200K+ indexed document chunks, persistent conversations, source-grounded answers, document upload, and real-time streaming responses.
- Designed the backend using FastAPI, LangChain, LangGraph, FAISS, pgvector, Supabase, and REST APIs for retrieval, reranking, document grading, query rewriting, and answer generation.
- Implemented hybrid retrieval with cloud vector search and local FAISS fallback to improve reliability, availability, and continuity during service degradation.
- Added metadata filtering, smart routing, CrossEncoder reranking, document priority logic, and citation workflows to improve answer quality, traceability, and user trust.
- Developed a React-based frontend with authentication, multi-chat sessions, PDF/DOCX upload, source display, and streaming chat interactions.

Agentic Workflow Automation Platform

- Built a multi-interface AI agent platform for managing operational workflows through natural language chat, voice commands, tool execution, and API-based automation.
- Designed a LangGraph StateGraph architecture for deterministic tool execution, replacing manual agent loops with structured, testable, and maintainable state transitions.
- Integrated dynamic tool discovery and API interaction patterns to support extensible agents, customer-specific workflows, and scalable automation.
- Implemented caching, session persistence, and workflow optimization that reduced repeated API calls by 80
- Created reusable platform patterns for context management, tool validation, workflow routing, failure handling, structured outputs, and escalation.

Role-Based HRMS System

- Built a role-based HRMS platform with separated Admin, HR, and Employee workflows for access control, employee records, leave management, payroll-related flows, and dashboard routing.
- Implemented authentication, role-based navigation, backend business logic, database-backed workflows, and clean UI separation for different user permissions.
- Designed the system around real workplace process modeling, including user roles, restricted actions, data visibility, and maintainable full-stack structure.

Data and Automation Systems

- Developed Python/Pandas and SQL-style ETL pipelines to clean, map, validate, and reconstruct structured datasets with 10,000+ records and complex relational schemas.

- Built validation utilities for schema checking, deduplication, fuzzy matching, malformed record correction, and automated data quality improvement.
- Created automation scripts that reduced manual data errors and improved reliability of downstream reporting, search, and AI-enabled workflows.

Education

B.S. Data Science, FAST-NUCES Lahore

2021 – 2025 First Class

with Distinction Relevant Focus: Python, Data Structures, Algorithms, Databases, Machine Learning, NLP, Backend Systems, Data Engineering IELTS Academic: Band 8.0