

Muhammad Salman

AI Engineer | Agentic AI

Islamabad, Pakistan | +92 325-0427314

muhammad.salman1205@gmail.com | [linkedin.com/in/salman1205](https://www.linkedin.com/in/salman1205) | github.com/Salman1205

PROFILE

AI engineer building and operating production-grade Agentic AI and Generative AI applications. Sole engineer and technical owner of a platform in daily use by over 200 enterprise users that processes unstructured documents and audio into validated structured output. Hands-on across agent frameworks (LangGraph, LangChain, Anthropic Claude Agent SDK, Pydantic AI), LLM-powered workflows, prompt engineering, RAG, schema-validated structured outputs, Human-in-the-Loop review, AI safety guardrails, and model evaluation and monitoring frameworks, integrated with enterprise systems and third-party APIs.

TECHNICAL SKILLS

Agentic AI and orchestration: LangGraph, LangChain, Anthropic Claude Agent SDK, Pydantic AI, multi-agent workflows, orchestrator and specialist-agent patterns, AI orchestration, tool and function calling, agent state management, intelligent workflows, enterprise automation and AI automation solutions

LLM and Generative AI: production-grade LLM and Generative AI applications, prompt engineering and prompt optimisation, prompt governance and versioned prompt templates, Retrieval-Augmented Generation (RAG), embeddings, vector databases (ChromaDB, pgvector), semantic and hybrid search, prompt caching, context-window strategy, OpenAI, Anthropic Claude, Google Gemini, Groq

Structured output and document processing: schema validation with Pydantic and Pydantic AI, typed structured outputs consumed by downstream services, JSON schema-driven extraction, typed canonicalisation of monetary values, dates and names, deterministic verification of model claims against source data.

Intelligent Document Processing (IDP): extraction pipelines over unstructured source material, ingestion of files up to 2 GB, long-form transcription with speaker diarization, reconciliation against a system of record

HITL, safety and evaluation: Human-in-the-Loop (HITL) review workflows, AI safety guardrails, output validation before release, fallback and escalation paths, model evaluation and performance frameworks, offline evaluation harnesses using human-correction ground truth, performance monitoring, drift detection, structured decision logging and audit trails

Python and backend: Python (primary), FastAPI, Flask, asyncio, Node.js and Express, REST API design and integration, third-party and external tool integration, WebSockets, Server-Sent Events, webhooks, Redis queues, OAuth2, role-based access control, microservices

Cloud, data and DevOps: Microsoft Azure (Azure SQL Database), AWS (EC2, S3, Lambda), Docker, Kubernetes, Git, GitHub Actions CI/CD, pytest, Prometheus and Grafana, Airflow, MLflow, Kubeflow. PostgreSQL, MySQL, MongoDB, Redis, SQLite, schema design, ETL pipelines, Pandas, NumPy, PyTorch, TensorFlow, scikit-learn

PROFESSIONAL EXPERIENCE

AI Engineer - The Insolvency Group - United Kingdom (Remote)

December 2025 - Present

- Sole engineer and technical owner of a production Agentic AI platform in daily use by 200+ enterprise users: roughly 559 commits in five months across a Python and FastAPI service, a Node.js and Express API on MySQL, and a React front end. Built the agent layer as an orchestrator routing work to specialist agents across four execution modes, on the Anthropic Claude Agent SDK with Pydantic AI so every stage returns schema-validated structured output rather than free text downstream services must defensively parse.
- Developed and optimised LLM-powered workflows and extraction pipelines over unstructured input: ingestion up to 2 GB, long-form transcription with diarization, and roughly 131 configurable extraction questions per record fanned out concurrently under semaphore control. Built the RAG layer behind it with embeddings, vector search, hybrid retrieval and prompt-cache priming so a shared context prefix is not re-billed per question.
- Built the Intelligent Document Processing side: extracts structured fields from unstructured source material and reconciles them against an enterprise CRM record by record, with typed canonicalisation of monetary values, dates and names, plus gates preventing false discrepancy flags.
- Implemented Human-in-the-Loop workflows and AI safety guardrails: evidence-first prompting, a deterministic locator verifying each quoted claim exists in the source, output validation before content reaches a user, and a human reviewer retaining the final decision on anything consequential. Removed an earlier model-based safety gate after it produced a false pass in production, replacing it with a deterministic check that escalates rather than deciding.
- Developed the model evaluation and monitoring framework: an offline evaluation harness using 176 human reviewer corrections as ground truth, per-stage latency profiling, per-request cost tracking and structured decision logging. Prompts and rubric definitions are held as configuration rather than code, refined against reviewer feedback and measured against the harness before release. Reduced a manual two-hour process to about fifteen minutes per record.

AI Engineer - Anviro - Islamabad, Pakistan

June 2025 - August 2025

- One of the first three engineers. Owned architecture and infrastructure decisions and deployed customer-facing Generative AI applications to production on AWS at approximately 99% uptime, using RAG with embeddings and semantic retrieval to improve answer relevance and reduce hallucinated responses.

AI / Machine Learning Intern - AIM Lab - Islamabad, Pakistan

June 2024 - August 2024

- Led a team of five delivering a Generative AI system; deployed on AWS EC2 and Vercel, reducing hosting cost by 70%.

SELECTED PROJECTS

LaunchMind - Multi-agent orchestration with real tool use - Python, GitHub, Slack, SendGrid APIs

2026

- An orchestrator decomposes a goal and routes work to five specialist agents with a review and revision loop. The agents call real external tools rather than only generating text: opening GitHub branches and pull requests, dispatching Slack messages and email campaigns, posting inline review comments.

Voice AI Agent - Real-time conversational agent, structured output - Python, FastAPI, WebSockets, Twilio, Deepgram

2026

- Places a live phone call and holds a multi-turn conversation, returning a schema-validated JSON result. Every provider sits behind a typed interface so a model or vendor is swappable by configuration, with 60 automated tests covering state and failure modes including silence, interruption and early hangup.

MailAssist - Production Generative AI application, live at amanii.io - Next.js, TypeScript, PostgreSQL

2025 - 2026

- Built and deployed independently over seven months and 342 commits: RAG-grounded generation with output guardrails validating content before send, OAuth2 mailbox sync, role-based access control across three roles, REST APIs and third-party integrations.

Fraud detection pipeline with monitoring - Model evaluation and drift detection - XGBoost, LightGBM, Kubeflow, Grafana

2026

- Seven-stage pipeline with a conditional deployment gate: class weighting against SMOTE for severe imbalance, cost-sensitive learning, SHAP explainability, Prometheus and Grafana monitoring, drift simulation and alert-driven retraining.

EDUCATION

B.S. Data Science, FAST-NUCES, Islamabad, 2026. CGPA 3.3/4.00, Dean's List Fall 2025. ICPC Asia Regional: 134th of 2,000+ teams. Paper conditionally accepted at ICIT 2026.