MUHAMMAD SHAHARYAR SHABBIR BHATTI

Senior AI Engineer | Agentic AI & LLM Systems | Generative AI | Production ML & Cloud

+92 313 5532589 | shaharyarshabbir348@gmail.com | Rawalpindi, Pakistan | LinkedIn | GitHub | HuggingFace

PROFESSIONAL SUMMARY

Senior-minded AI Engineer with 3+ years designing, building, and deploying production-grade AI systems — agentic and multi-agent platforms, RAG architectures, generative-AI applications, and real-time voice agents — at scale. Proven record of shipping LLM-powered solutions that replace manual workflows and drive measurable impact, reducing operational overhead up to 90% and infrastructure costs up to 70%. Full-stack AI expertise across LLM fine-tuning, prompt engineering, RAG, vector databases, agent orchestration, model evaluation, and cloud deployment (AWS, Azure) with Docker/Kubernetes and CI/CD. Fluent with OpenAI, Anthropic Claude, and Gemini, and frameworks including LangChain, LangGraph, CrewAI, and LlamaIndex. Collaborates across product and engineering teams to turn AI concepts into reliable, scalable, cloud-native services with engineering rigor and responsible-AI practices.

CORE COMPETENCIES

AI & ML: Agentic AI, Multi-Agent Systems, LLMs, Generative AI, RAG Pipelines, LLM Fine-tuning, Prompt Engineering & Optimization, Embeddings, Semantic / Vector Search, Model Evaluation & Benchmarking, Deep Learning, Transformers, NLP, MLOps / AIOps, Model Monitoring

Frameworks & LLMs: LangChain, LangGraph, LlamaIndex, CrewAI, Hugging Face, OpenAI (GPT-4o), Anthropic Claude, Azure OpenAI, Azure AI Foundry, AWS Bedrock, PyTorch, TensorFlow, Scikit-learn, Gradio, Streamlit

Engineering & Deployment: Python, C++, SQL | FastAPI, REST APIs, WebSockets, Microservices, Docker, Kubernetes, CI/CD | AWS (EC2, Bedrock), Microsoft Azure AI, Vercel, Git / GitHub

Automation & Integrations: n8n, VAPI, Twilio, Firecrawl, Webhooks, HubSpot CRM, Microsoft Outlook API, LinkedIn API, Google Workspace, Amadeus, AviationStack, SerpAPI

Databases & Vector Stores: PostgreSQL, Supabase, SQLite, Cassandra | Pinecone, FAISS (vector databases)

Data & Visualization: Power BI, Pandas, NumPy, Matplotlib, Seaborn, EDA, React

PROFESSIONAL EXPERIENCE

AI Engineer (Mid-Senior) | Ikonic
Jun 2026 – Present

- Developed SafetyNexus, an AI-powered digital coaching and case-mapping platform for child welfare and domestic abuse caseworkers, built on the Safe & Together methodology.
- Engineered real-time streaming LLM coaching via Server-Sent Events (SSE), featuring safety/crisis screening interrupts, inline citations, document parsing (PDF/DOCX), and automated report generation.
- Built scalable multi-tenant services with vector-based knowledge retrieval (RAG), secure session tracking, and role-based access control (RBAC), powered by Anthropic Claude and AWS Bedrock.

AI Engineer | Aridian Technologies
Aug 2024 – Jun 2026

- Designed and deployed production-grade AI automation pipelines using n8n, FastAPI, and LLM agents, streamlining end-to-end business workflows and eliminating 90% of manual operations.
- Engineered a LangGraph-powered multi-agent financial analyst system — with role-based agents for research, analysis, and reporting — reducing manual analysis time by 60% through automated trend analysis, risk assessment, and report generation.
- Built PostgreSQL real-time data pipelines enabling automated querying and financial reporting, improving decision-making speed by 40%.
- Developed an AI voice assistant for flight booking and cancellations integrating STT/TTS, HubSpot CRM, webhooks, and n8n, achieving 95% accuracy and reducing customer handling time by 50%.

Associate AI Engineer | ZIYVO
May 2023 – Jul 2024

- Built Generative AI solutions including intelligent Q&A systems, document extraction pipelines, and AI-driven automation workflows for enterprise clients.
- Designed custom LLM pipelines and integrated REST APIs, reducing client operational costs by 30% and improving client satisfaction scores by 25%.

KEY PROJECTS

FinSight AI — Financial Intelligence Platform — *LangChain, CrewAI, Fine-tuned LLMs, Multimodal*

- Built an AI-powered platform for banks delivering document analysis, forecasting, and risk assessment, reducing analysis turnaround time by 70%.
- Fine-tuned DeepSeek on financial data using BERT as a judge model, achieving 86% domain accuracy and enhanced financial reasoning.

Multi-Agent Report Automation System — *LangGraph, Outlook API, FastAPI, PostgreSQL, LLMs*

- Designed a multi-agent pipeline with a Financial Agent and Transcription Agent that pulls email data from Microsoft Outlook and other sources, synthesises information, then auto-generates and delivers structured reports to clients.

- Implemented secure API integrations, structured handoffs between agents, and automated report dispatch, reducing manual reporting effort by 80%.

Financial P&L Data Pipeline & Live Dashboard — *Python, Pandas, SQLite, Gmail API, React*

- Designed and deployed an automated end-to-end financial reporting pipeline covering Gmail API ingestion, pandas-based preprocessing, and a SQLite ledger with SHA-256 deduplication, surfacing key KPIs via a live React dashboard — eliminating manual QuickBooks reporting for a US-based client.

Cost-Effective AI Calling Agent — *Twilio, OpenAI Realtime API, WebSockets, FastAPI*

- Replaced VAPI with a custom Twilio + OpenAI Realtime API voice agent, cutting voice infrastructure costs by 70% while achieving sub-second response times via bidirectional WebSocket audio streaming.

Enterprise Workflow Automation — *n8n, LLM-based routing, Webhooks, APIs*

- Designed end-to-end pipelines for lead intake, AI decision routing, and CRM updates, improving workflow efficiency by 60% and eliminating 90% of manual operations.

Website Chatbot — Voice + Chat RAG Agent — *RAG, LlamaIndex, LangChain*

- Built a 24/7 AI assistant with RAG pipelines over company knowledge bases, reducing manual query handling by 70% across voice and chat channels.

InsightAI — Self-Service ML Platform — *Python, Scikit-learn, Streamlit, Pandas*

- Built a no-code data science platform for non-technical users: CSV upload, preprocessing, multi-chart visualization, and training of Linear Regression, KNN, and Logistic Regression models; improved accessibility by 50%.

EDUCATION

B.Sc. in Artificial Intelligence | IQRA University, Islamabad *Sep 2021 – Jul 2025*
CGPA: 3.47 / 4.00

CA – AFC & CAF Level | SKANS School of Accountancy *Mar 2019 – Jun 2021*

CERTIFICATIONS

- Building AI Voice Agents for Production (Oct 2025) — Pipecat, WebRTC, real-time voice AI | DeepLearning.AI
- Long-Term Agentic Memory with LangGraph (2025) — Semantic, episodic & procedural memory-augmented AI agents
- Data Analytics & Power BI Modeling (May 2024) — Excel, Power BI | LinkedIn Learning

LANGUAGES

English (Full Professional) | Urdu (Native) | Potohari (Native)