MUHAMMAD SHAHBAZ SARWAR AWAN

Full Stack & AI Engineer | Web Applications, LLM Systems & RAG Pipelines

Pakistan · Open to Remote | +92 302 0673674 | shahbazsarwar585@gmail.com | GitHub | LinkedIn

SUMMARY

Artificial Intelligence graduate from FAST (NUCES) who builds complete products, not just models. I work across the stack: React and Next.js on the front end, FastAPI and Node.js services on PostgreSQL behind it, and the LLM layer that makes the whole thing useful. Comfortable with RAG pipelines, agentic workflows, embeddings and vector search, grounded generation with source attribution, and containerized deployment. I have shipped AI automation for a US healthcare client, covering OCR and NLP clinical extraction along with ICD-10, CPT and HCPCS coding, and delivered production chatbots and workflow automation for international clients.

TECHNICAL SKILLS

Frontend: React, Next.js, JavaScript / TypeScript, HTML5, CSS, responsive UI, component architecture

Backend: FastAPI, Flask, Node.js, REST API design, microservices, authentication, async services

Generative AI & LLMs: RAG architecture, LLM orchestration, AI agents and agentic workflows, prompt engineering, embeddings and semantic search, retrieval tuning, model evaluation, token and latency optimization

Frameworks & Tools: LangChain, LangGraph, n8n, Git / GitHub, CI/CD, MLOps and model monitoring

Data & Retrieval: PostgreSQL, Supabase, SQL, vector databases (Pinecone, Chroma), OCR and NLP extraction

Cloud & DevOps: AWS, Docker, containerized deployment, environment configuration

Languages: Python, JavaScript / TypeScript, SQL, Java

EXPERIENCE

Paisol Technology | AI Engineer — Remote

- Owned the full build of **Chetti**, an internal platform that turns a single prompt into a deployable RAG chatbot and embeds it with one line of iframe code. Setup dropped from days of developer integration to minutes of self-serve configuration.
- Engineered the entire retrieval layer, covering ingestion, chunking strategy, embedding storage and grounded generation with source attribution, then measured it against a domain test set I built by hand.
- Chose open source LLMs over hosted APIs to keep per-deployment inference cost flat and client documents in-house, then tuned chunk size and retrieval depth to work within that constraint.
- Built a multi-turn booking assistant with explicit conversation state management that captures every required field in a single session, integrated with Google services for scheduling.

Aeromek AI | AI Engineer — Remote

- Replaced manual chart review with an automated clinical document pipeline that pulls demographics, diagnoses and procedures out of unstructured medical records using OCR and NLP.
- Implemented LLM based ICD-10, CPT and HCPCS code recommendation that returns ranked candidates with supporting evidence, moving human coders from manual lookup to quick validation.
- Built FastAPI microservices on PostgreSQL that check claims against payer requirements before submission, flagging non-compliant records ahead of rejection.

Independent Software & AI Engineer | Contract — Remote

- Delivered retrieval augmented assistants over client document sets as the sole engineer, handling ingestion, chunking strategy, retrieval tuning, front end integration, deployment and iteration.
- Standardized a reusable embeddable chatbot pattern across engagements, replacing bespoke per-client development with configuration driven deployment.
- Connected LLM reasoning to clients' CRM, inbox and scheduling systems through n8n workflows and REST integrations.

PROJECTS

HealPay, AI Medical Billing Automation (Final Year Project, graded A+): a full stack platform that ingests medical documents, extracts clinical data with OCR and NLP, recommends billing codes with LLMs and validates claims before submission. Built on FastAPI microservices and PostgreSQL, containerized with Docker.

Financial Recommender System: an LLM fine-tuned on financial data and paired with a retrieval layer so every recommendation can be traced back to its source.

EDUCATION

FAST, National University of Computer & Emerging Sciences — Graduated 2026

BS in Artificial Intelligence. Coursework in Agentic AI, NLP, MLOps, Machine Learning, Recommender Systems, Databases and Web Technologies.